中文领域本体自动构建理论与应用研究

Theory and Application of
Chinese Domain Ontology Automatic Construction

◎ 刘柏嵩 著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

本书得到国家社会科学基金项目《领域本体的自动构建和应用研究》(No. 08CTQ014)的资助

中文领域本体自动构建理论与应用研究

刘柏嵩 著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目(CIP)数据

中文领域本体自动构建理论与应用研究 / 刘柏嵩著.
—杭州：浙江大学出版社，2014.6
ISBN 978-7-308-13257-2

Ⅰ.①中… Ⅱ.①刘… Ⅲ.①计算机网络—情报检索
—研究 Ⅳ.①G354.4

中国版本图书馆 CIP 数据核字(2014)第 098445 号

中文领域本体自动构建理论与应用研究

刘柏嵩 著

责任编辑 石国华
封面设计 刘依群
出版发行 浙江大学出版社
（杭州市天目山路 148 号 邮政编码 310007）
（网址：http://www.zjupress.com）
排 版 杭州星云光电图文制作有限公司
印 刷 杭州日报报业集团盛元印务有限公司
开 本 710mm×1000mm 1/16
印 张 11.75
字 数 220 千
版 印 次 2014 年 6 月第 1 版 2014 年 6 月第 1 次印刷
书 号 ISBN 978-7-308-13257-2
定 价 38.00 元

前 言

目前，互联网不仅已经成为人们进行信息交流和知识获取的工具，而且成为人类生存和发展的重要工具。现有的互联网技术通过超文本系统实现了对因特网上信息资源的迅速定位。然而，这些互联网技术并没有对信息的含义进行描述，导致互联网上难以实现信息内容处理的自动化和智能化。为此，互联网研究者提出了语义 Web 技术来解决该问题。语义 Web 的目标是使计算机能够理解和自动处理 Web 上的内容，为实现智能化的 Web 应用提供必要的基础。而领域本体获取正是构建语义 Web 的关键。由于领域本体描述了一种形式化的、共享的概念化模型，因此能够为 Web 内容提供机器可读的显式的语义信息。

同时，领域本体也被认为是解决“信息和知识孤岛”问题的最有前途的方法。知识内在的异质性和分布性已经严重地阻碍了知识在多主体和应用系统之间的共享和重用，而领域本体为该领域的知识提供了显式的概念化的规范说明，因此人类和计算机能够以一种结构化的方式共享和重用领域知识。在近年来提出的 Web 2.0 和 Library 2.0 中，领域本体的作用更加突出，本体是提供各种 2.0 服务标准接口的关键所在。

领域本体的构建和学习已成为制约信息检索、数字图书馆、信息抽取和机器翻译等许多智能信息处理任务的瓶颈。这是因为这些研究领域要取得突破性的进展，并不能完全依赖于计算机的运算速度和空间大小，更离不开对问题领域相关知识的需求。

众所周知，领域本体的人工构建是一项耗时耗力和繁杂的任务。如何以半自动或自动的方式获取领域本体已成为语义 Web、知识管理、智能信息处理等多个领域的重要研究课题。领域本体自动构建的目标就是以自动或半自动的方式从各类数据源中学习领域本体。领域本体快速而高效地获取是构建未来语义 Web 的基础；是实现信息和知识共享的基本方法；也是开发智能信息处理系统和实现新一代数字图书馆的关键。

目前本体自动构建（或学习）系统处理和分析的语料大多是基于西文，这方面研究国内起步较晚，在已开发的可支持中文语料处理的本体构建系统中缺乏对中文文本的深入分析。由于中文表达缺乏形态上的变化，词与词间在书写上并没有界标，中文分词的歧义问题因此尤为突出，从而使得中文概念及其关系的自动抽取更为困难。本书就是在这种背景下，为了能够有效地克服传统本体构建方法所存

在的费时费力、难以处理中文语料等方面的问题，提出综合运用多种自然语言处理和机器学习方法，采用分层技术原理，提出了一种新的分层本体学习方法体系，来实现中文领域本体的自动构建。并开发了一种基于 Web 的多策略本体自动构建平台 GOLF，然后讨论了本体演化和本体评价问，对本体构建平台 GOLF 进行了实验和评价。本书的主要研究内容包括：

一是在综合现有本体构建方法和技术的基础上，提出了一种领域本体自动构建分层体系，从而可逐步实现中文领域本体的自动构建，其中包括术语自动抽取、概念抽取、实例学习和分类关系学习和非分类关系抽取等多种关键技术。在对现有本体构建学习方法做大量改进的基础上，完全实现了本体自动构建全过程的无缝集成。

二是开发了多语种领域本体自动构建平台 GOLF，采用 Web 文档作为本体学习源，进行了多个领域、多个语种的实验。同时，实现本体自动构建过程中的本体演化管理，并对结果本体进行评价和反馈。

三是在领域本体自动构建中引入多策略学习方法，以提高本体抽取的效率和质量。各种学习算法的组合框架采用概率组合分布，能根据不同的语料特征为每个算法设定权值，由此增强了平台对多种领域语料的适应性。课题实验对比分析表明，在结果本体的准确率和召回率两个指标方面，GOLF 系统比著名的 Text2Onto 系统要好一些。

四是探讨了本体自动构建方法 GOLF 在数字图书馆等实际领域的应用，通过该方法可以有效地对网络信息进行组织，实现海量信息资源的高效检索、元数据的自动生成及个性化服务等内容。

本书研究成果的主要学术价值在于，建构了相对完善的领域本体自动构建的学术框架体系，并采用一种全新视角对中文领域本体自动构建的核心元素进行了探讨，以汉语自由文本为研究对象，围绕领域本体构建的对象：领域概念、领域概念间关系、领域概念实例知识，研究相应的自动获取的方法模型、核心算法及其应用。通过本书的研究，探索了领域本体自动构建的基础理论，并提出了一种分层构建的新方法，完善了中文本体构建工具的开发，实现跨领域中文本体的自动构建。在当前互联网的大数据时代背景下，对情报信息处理、知识组织等相关学科发展铺设了一定的理论基础，起到了很好的推动作用。

本书相比于传统的本体自动构建理论与方法研究，具有以下三方面创新之处：

一是解决了快速自动化构建领域本体的关键问题。自从把本体概念应用到信息科学、人工智能以及图书馆情报等很多领域以来，国内外的许多学者对本体的研究就主要集中在理论方面，而涉及具体的领域的本体构建的研究和实践却相对较少。主要原因是构建领域本体是一项巨大的系统工程，如果按照传统的人工构建叙词表以及结构化词表的方法去构建领域本体已不太现实，因为那样需要大量的人力、物力和时间。如何才能迅速地构建领域本体已经成为了本体研究人员的难

题。而本书就是研究如何迅速地构建领域本体中概念之间的关系，去解决本体构建的关键问题。通过本书的研究，提出了一种迅速地构建本体中概念之间关系的方法体系，进而达到节省构建本体的人力、物力和时间，加快本体构建的进度，最终实现中文领域本体的快速构建。

二是提出了一种新的本体自动构建模块化方法以及结果本体的评价标准。在本书提出本体构建分层体系的方法指导下，开发一个本体学习平台 GOLF。该平台采用模块化设计理念，能灵活地对不同的功能模块进行组合，从而进一步提高系统的可扩展性。本体工程师可以针对不同的领域特征和语种配置不同模块，能集成不同的语言学模式和各类语义词典。本书提出一种新的基于贝叶斯决策的本体评价模型 RiMOE，并采用 RiMOE 模型对 GOLF 的学习过程和结果本体进行评价。GOLF 平台基本实现了跨领域、跨语种的全自动无监督本体自动构建功能，经语料测试和本体评价，其性能良好。

三是在本体自动构建的多语种自适应性方面进行了有益的尝试。本书所提出的方法和工具能够很好地处理多语种语料，与同类系统相比，对中文的处理能力明显加强，特别针对中文的语言学特征，引入 HowNet 和 WordNet 等多个语义词典，并可开放地年度计划其他语料词典，添加了对应于中文文本的语言学模式和停用词表，并自动生成新的生词表。与目前主要的本体构建系统相比，多语种自动适应的性能有明显改善。

该研究方向目前整体还处于探索阶段，未来仍有大量的工作要做。由于时间的限制，本书没有讨论所构建本体的编码形式，即以哪种表达形式来表示结果本体，也缺少对本体的自动扩充问题进行研究。另外，由于人类语言普遍存在歧义性，以及 Web 海量数据的稀疏性特征，从海量文档集中快速进行精确的领域概念识别以及抽取概念之间的语义关系仍存在较大挑战。本书目前主要考虑了本体的层次分类结构，兼顾考虑概念之间的其他语义关系（包括多元关系、多元属性和非分类关系），对本体实例的扩充涉及不多。

今后的研究工作，将围绕互联网这一海量信息源，利用云计算技术和大数据分析技术，进一步增加本体自动构建实验的广度和深度，尤其是 2.0 类型应用包括如微博和维基等的数据处理，对本体实例扩充问题进行深入研究。我们将着重对自动获取的结果本体进行综合量化评价（Evaluation），包括检查本体一致性、定量评价分析等。在今后研究中需进一步加强对语料的深层次语义分析和不确定性知识的处理。

目 录

九层之台，起于累土。

——《道德经》

第1章 绪 论

1.1 研究的背景及意义

领域本体(Domain Ontology)又称领域实用分类系统，是指对该领域里的知识进行表述的词和术语[①]，编制者根据该领域的知识结构将这些词和术语组成等级类目，采用面向对象的方法按需要将一些类目进行更细致的定义(包括特性、限制、推理规则等)。本体与传统知识分类工具的一个显著区别，就是系统中的概念、实例、特性等内容都是计算机可读懂的，因此本体中的知识可以方便地被其他系统再利用。

本体已经在数字图书馆 DL、知识工程 KE、自然语言处理 NLP、信息检索、网络异构信息的集成、软件复用和语义网站等领域中广泛应。主要应用场景有：(1)基于本体的信息抽取和数据集成；(2)基于本体的信息查询，特别是在数字图书馆和智能搜索引擎中；(3)领域本体的应用。如在生物信息学中目前广泛应用的 Gene 本体，该本体虽然只包括了 Part-Of 等简单的语义关系，但是对生物信息学界已经产生巨大影响；(4)各类语义 Web 服务；(5)海量资源的在线元数据管理和各类自动信息发布与推送。

目前，互联网不仅已经成为人们进行信息交流和知识获取的工具，而且成为人类生存和发展的重要工具[②]。现有的互联网技术通过超文本系统实现了对因特网上信息资源的迅速定位。然而，这些互联网技术并没有对信息的含义进行描述，导致互联网上难以实现信息内容处理的自动化和智能化。为此，互联网研究者提出

---

① 秦健. 实用分类系统与语义网：发展现状和研究课题. 现代图书情报技术，2004. 1：16－23.

② Berners-Lee T.，Hendler J. and Lassila O. The Semantic Web. Scientific American，2001，284(5)：34－43.

了语义 Web 技术来解决该问题[①]。语义 Web 的目标是使计算机能够理解和自动处理 Web 上的内容，为实现智能化的 Web 应用提供必要的基础。而领域本体获取正是构建语义 Web 的关键。

同时，领域本体也被认为是解决“信息和知识孤岛”问题的最有前途的方法。知识内在的异质性和分布性已经严重地阻碍了知识在多主体和软件实体之间的共享和重用，领域本体为该领域的知识提供了显式的概念化的规范说明，

因此人类和计算机能够以一种结构化的方式共享和重用领域知识。在近年来提出的 Web 2.0、Library 2.0 和关联数据中，领域本体的作用更加突出，本体是提供各种 2.0 服务标准接口的关键所在。领域本体的构建和学习已成为制约信息检索、数字图书馆、信息抽取和机器翻译等许多智能信息处理任务的瓶颈。

例如，现有的 Web 搜索引擎（包括 Google、Yahoo 和百度等）采用的是基于关键词的全文检索技术，难以查询同一概念的不同词汇表示形式，难以查询同一主题的不同内容表示形式，更难以查询具有上下位关系、部分整体关系、主题关联关系、因果关联关系等语义关系的信息[②③④⑤⑥]。一个根本的原因是基于关键词的信息检索技术缺乏知识处理能力和理解能力。把信息检索从基于关键词层面跨越到基于语义或知识层面（即本体层），才是解决问题的根本和关键。

然而，领域本体的人工构建是一项耗时耗力和繁杂的任务。早期的本体构建方法主要诞生在具体的开发项目，为具体的项目实践服务。主要的工作有 Protégé[③]、KAON[④] 为代表的本体建模理论与工具设计方法研究，其主要优点是提供了特定的建模工具，不足在于对用户的依赖程度非常高，从而影响了本体建模的可扩展性。如英国曼彻斯特大学的 OpenGalen 项目利用分类表《疾病国际分类表》为基础，经过领域专家和知识工程师的合作，历经十年努力而建成。目前 OpenGalen 这个大型的医学本体，包括 2 万多个手术程序，1 万多个解剖学概念，1 万多种药物及其相关的概念。美国 Syracuse 大学的 Qin Jian 教授探索了将 GEM（教育资料网关）中的受控词表转化为本体的原理和原则框架[⑥]。

另一类研究是以 OntoEdit，Text2Onto，TANGO 和 OntoLearn 等为代表的半

---

① Berners-Lee T., et al. A Framework for Web Science. Foundations and Trends in Web Science, 2006, 1(1).

② Amal Zouaq, Dragan Gasevic, Marek Hatala. Towards Open Ontology Learning and Filtering. Information Systems, 2011, 36: 1064－1081.

③ Berners-Lee T., Hendler J. and Lassila O. The Semantic Web. Scientific American, 2001, 284(5): 34－43.

④ Berners-Lee T., et al. A Framework for Web Science. Foundations and Trends in Web Science, 2006, 1(1).

⑤ Albert Weichselbraun, Gerhard Wohlgenannt, Arno Scharl. Refining Non-taxonomic Relation Labels With External Structured Data To Support Ontology Learning. Data & Knowledge Engineering, 2010, 69: 763－778.

⑥ Qin J., Hernández N. Building interoperable vocabulary and structures for learning objects. Journal of the American Society for Information Science and Technology, 2006, 57(2): 280－292.

自动本体建模工作[①②③]，其主要特点是能够在用户辅助下，通过对初始领域本体以及相关领域文档的处理，构建领域本体。这些研究主要集中在概念和关系的获取。目前还没有一个相对成熟的领域概念获取的方法，也无法自动地为获取到的非分类关系赋予语义。

同时，在对中文语料的处理和中文本体构建等方面缺乏必要的支持。由于英语和汉语分别属于屈折语和孤立语，汉语具有如下特点：(1)形态的变化较少，没有格、性、数的变化标志；(2)汉语的词序严格，如果词序不同，意义也将会不同；(3)主要的语法手段之一是虚词；(4)汉语书写系统以采用词标的形式为主，词与词之间缺乏明显的形态界限。汉语的这些特征，就决定了针对英语等其他语言文本的本体构建方法，并不能完全适用于汉语文本的领域知识获取[④]。

本体的构建过程非常费时费力，需要完整的系统化、工程化的方法来支持，特定的领域本体尚需要专家的参与与交互[⑤]。目前通用的大规模本体比较少，大多数本体是针对某个具体应用领域或应用而构建的。在实际应用中，不同本体之间常常需要进行扩充、映射与合并处理，并可能根据特定的需要从一个大的本体中抽取满足需求的小的本体等操作。此外，当实际的知识体系发生了改变时，之前构建的本体必须作出相应的演化，以保持本体与现实的一致性，这都是当前本体构建所需考虑的问题。

综合相关研究分析，当前领域本体的构建主要存在以下问题[⑥⑦]：(1)大多采用手工方式，一旦遇到复杂的领域就费时费力，也很容易出错；同时无法进行大规模扩展，要让大量的用户和领域专家专门构建本体存在相当的困难。(2)建设过程无规范。在建立各自的领域本体时均采用不同的标准、建模方法，所以构建的本体不通用；缺乏统一的本体构建体系结构概念和方法，不用实现本体重用。(3)成果没有评价标准。本体的评价方法没有统一的标准，更没有标准的测试集。不能对本体的建设成果进行合理评价，必然影响到本体的应用和进化过程。(4)缺乏对中文语料的处理。当前本体开发系统处理和分析的语料多是基于西文。

本体构建过程中如何集成现有的不同本体？怎么才能有效地大规模地构造本体？如何维护领域本体及其进化过程？这一系列的问题需要专门的方法论作为指

---

① Marta Sabou. Learning Web Service Ontologies: An Automatic Extraction Method and Its Evaluation. ISWC2005.

② Bozsak E., et al. KAON—Towards a Large Scale Semantic Web. In Proc. of the 3rd ICEWT. Heidelberg, 2002: 304－313.

③ Van Damme, et al. FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies. SemNet, 2007: 57－70.

④ 杜小勇，李曼，王珊. 本体学习研究综述. 软件学报，2006，17(9)：1837－1847.

⑤ 刘萍，高慧琴，胡月红. 基于形式概念分析的情报学领域本体构建. 图书情报知识，2012，(3)：20－26.

⑥ 陈晓美，毕强. 面向文本的领域本体学习方法与应用研究综述. 图书情报工作，2011，55(23)：27－31.

⑦ Amal Zouaq, Dragan Gasevic, Marek Hatala. Towards Open Ontology Learning and Filtering. Information Systems, 2011, 36: 1064－1081.

导，目前该方向的研究尚处于探索阶段，没有形成成熟的研究体系与方法论。此外，本体构建不仅需要理论上的探讨和研究，还需要从实际语料和素材中方便地构造出本体。如何能利用成熟的软件系统平台辅助用户构建本体？此类软件系统能在哪些方面发挥自动化或者半自动化的作用？诸如此类已成为该方向急待解决的问题。

单个本体的手工构建都很费时费力，而每一个领域需要多个本体，现实世界存在无数个需要本体的领域。因此，如何以自动的方式获取领域本体已成为语义Web、知识管理、智能信息处理等多个领域的重要研究课题[①]。领域本体快速而高效地获取是构建未来语义 Web 的基础；是实现信息和知识共享的基本方法；也是开发智能信息处理系统和实现新一代数字图书馆的关键[②]。

1.2 本书的主要内容、基本思路和方法

1.2.1 主要内容

面向中文语义 Web，以领域本体自动构建为主要目标，研究这个构建过程中相应的基本理论和核心方法。建立领域概念、领域概念关系，以及领域概念实例知识的获取和评估的模型和方法。在此基础上，开发一个面向汉语文本的本体自动构建工具，并探索将其应用于数字图书馆和电子政务系统中。主要内容如下：

1. 领域本体自动构建的基础理论和核心方法

包括领域概念、概念间语义关系，以及领域概念实例知识的获取和评估的模型和方法。研究从文本中如何获取三种领域概念及其定义，即领域实体概念、领域属性概念和领域关系概念。首先从本体论和领域应用的角度，分析三种领域概念的判断准则；然后建立领域概念的获取和评估的模型和方法。

领域概念间的语义关系获取研究，从词汇和概念、词法和句法、领域概念属性的物理性和可变性，以及它们所能承担的语义角色等方面，构建领域概念之间的系统完整的多维语义关系分类体系，并研究各类语义关系的元性质。在此分类模型的指导下，进一步研究领域概念关系的学习和评估的模型和实现算法，建立由各种语义关系关联的领域概念空间[③]。

建立领域概念属性的语义分类体系（taxonomy）；研究概念实例的物理属性和社会属性知识、显式属性和隐式属性知识的学习和评估的模型和方法。

2. 本体构建工具的开发，实现跨领域中文本体的自动构建

目前的本体自动构建工具的功能都非常有限，它们都仅能处理某几种类型的数据源，获取部分领域本体对象，尤其是不能处理中文语料。本书将在现有的开发

① 刘萍，高慧琴，胡月红. 基于形式概念分析的情报学领域本体构建. 图书情报知识，2012，(3)：20－26.

② 章成志，王惠临. 面向数字图书馆应用的多语言领域本体学习研究. 图书情报工作，2011，55(2)：11－16.

③ 刘柏嵩. 基于本体的知识管理关键技术研究. 情报学报，2005，(1)：75－81.

工具基础上，拟开发出了一个稳定的、整合的、能够完成多种学习任务且能处理中文源的本体自动构建工具。并探讨其在电子政务和数字图书馆中的应用。

3. 对本体开发过程和结果的多策略评价

构建好的领域本体需要确认与评价，检验是否满足了所提需求，是否满足本体的建立准则，本体中的概念及其关系是否完整等。本体自动构建作为一种无监督的学习技术，对其进行评价较为困难，尤其是标准测试数据集的建立和标准结果的制定。通过本书的研究将计算将建立结果本体的评价标准，以及对本体开发过程的评价模型。

1.2.2　重点和难点

面向开放领域文档的中文文档术语自动抽取、概念学习、实例学习、概念间分类关系和非分类关系学习是本研究的关键，同时本体构建过程中的本体演化和本体评价方法也是难点。拟解决的关键问题：

关键问题 1：提出一种基于语义层面的混合式的通用领域概念获取方法。

关键问题 2：挖掘中文文本中以显式或隐式方式存在的领域概念之间语义关系，尤其是非分类关系的获取。

关键问题 3：自动快速构建领域本体骨架及对结果本体的评价。

1.2.3　基本思路和方法

由于本体自动构建是当前信息管理领域的热点与前沿问题之一，研究方法上紧紧抓住研究内容中的关键问题，注重吸收与创新相结合、理论研究与应用实践相结合。研究结果的创新性与可行性需通过实际应用进行检验，拟采用“理论研究—系统架构—关键技术—应用、测试评价与完善”的技术路线（如图 1-1 所示）。

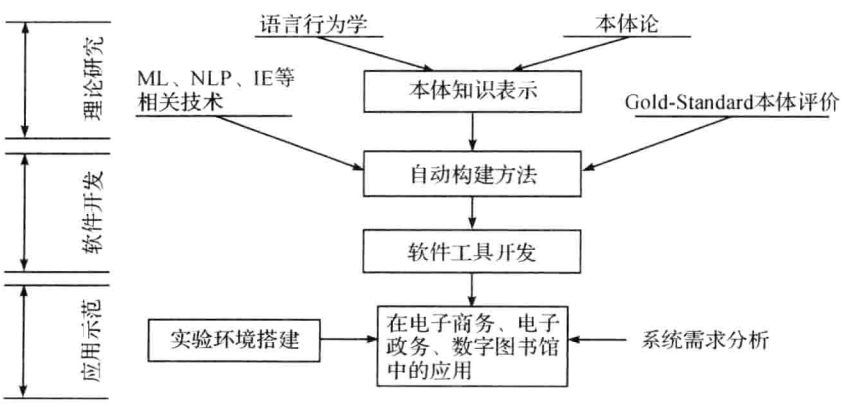

图 1-1　本研究的技术路线

1.2.4 创新之处

研制语义角色驱动的通用领域概念获取方法与工具。目前的领域概念获取方法主要是停留在句法层面，我们从句法层面跨越到语义层面，通过识别语句的主动词和语义角色来获取领域概念。在此基础上，研制一种基于语义层面的集成式的通用领域概念抽取模型。同时，也把各类 Web 2.0 网站上的标签、大众分类(Folksonomy)作为领域概念候选，以扩大本体概念范围。

采用多策略的机器学习的本体关系获取。基于统计方法的概念关系获取主要依赖人工或同义词词典等资源来识别语义关系的类型，基于模式方法又存在召回率低的问题。本书设计了利用多策略的机器学习方法来提取概念关系，包括并列模式引导的学习方法和基于语义聚类的学习方法。

实现跨领域多语种的本体构建，各个知识领域能自适应。与同类系统(大多只能处理西文文本)相比，可较好处理多语种，在对中文的处理能力上明显加强，能够对中文文本进行分词和词性标注。特别是针对中文的语言学特征，添加了对应于中文文本的语言学模式和停用词表，从而可实现汉语领域本体的自动构建。

不以规矩，不能成方圆。

——《孟子》

第2章　领域本体构建综述

2.1　传统领域本体构建方法及存在的问题

随着互联网的迅速普及，海量新信息快速产生。如何科学地管理、组织和维护这些海量信息以便为用户提供有效的服务成为一项重要而迫切的难题。本体是共享概念模型的形式化规范说明，能够从语义和知识层次上对信息进行描述。自被提出以来就引起了国内外众多科研人员的关注，并在许多领域得到了广泛的应用，如知识工程、数字图书馆、信息检索、异构信息的处理和语义网等。本体应用的基础是构建本体，传统的手工构建方式虽然可以保证质量，但费时费力。

Gruber 提出构建本体的五条原则[①]：

(1)明确性和客观性原则：本体应能对所定义术语的内涵有效地说明；定义应该是客观的，与领域背景相对独立；定义应该尽量完整，所有的定义应该采用自然语言进行说明。

(2)一致性：要求由本体所推导出来的概念含义与本体中的概念含义本身保持一致。本体定义的公理，以及用自然语言说明的文档都应该一致。

(3)可扩展性：在本体中进行新概念添加时，不能对原有的内容修改。

(4)最小编码偏差：本体不能依赖于某一特定的符号表示方法。

(5)最小本体承诺：只要能满足特定的知识共享需求，本体的承诺应尽量做到最小，对所建模对象给出尽可能少的约束条件。

此后，其他研究人员还陆续补充其他几条原则：

---

① Farquhar A., Fikes R., Rice J. The Ontolingua Server: A Tool for Collaborative Ontology Construction. Int. Journal of Human-Computer Studies, 1997, 46(6): 707－727.