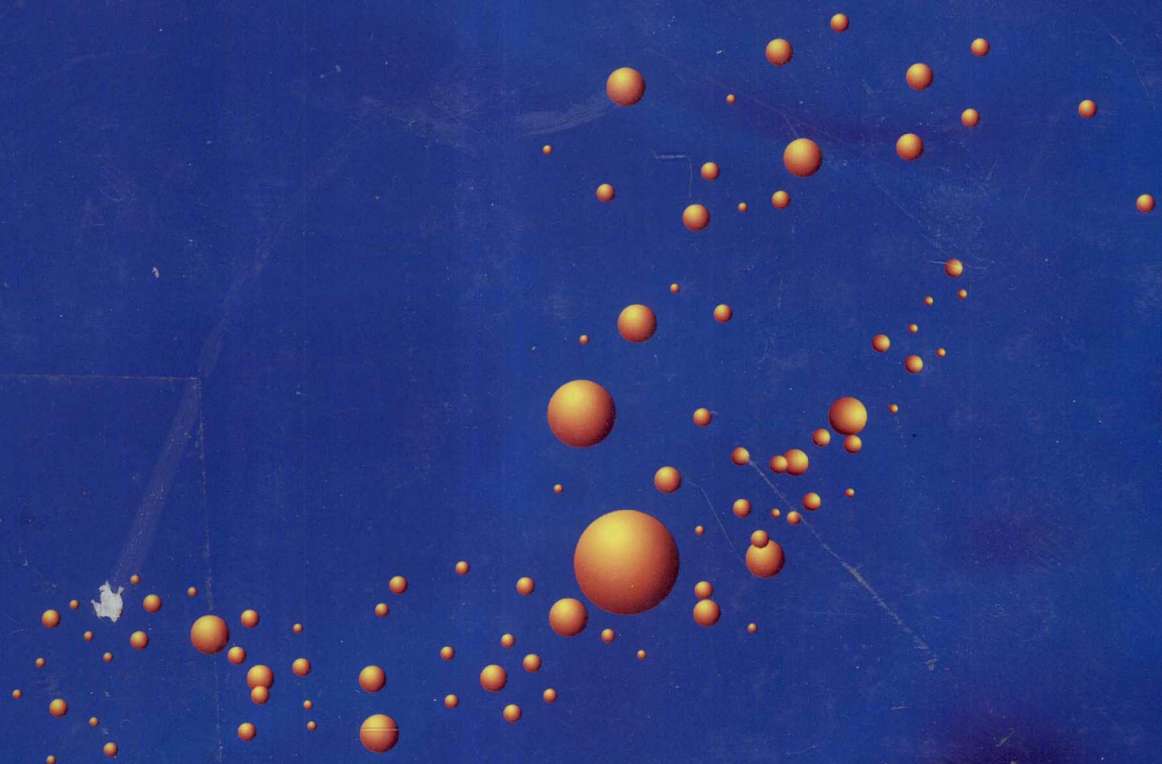
生物医学语义技术

SEMANTIC TECHNOLOGY FOR BIOMEDICAL SCIENCE

[主编] 李劲松　黄智生



Zhejiang University Press
浙江大学出版社

生物医学语义技术

Semantic Technology for Biomedical Science

主　编　李劲松　黄智生
副主编　俞思伟　包含飞　郑　锴

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

生物医学语义技术 / 李劲松，黄智生主编. —杭州：浙江大学出版社，2012.7

ISBN 978-7-308-10129-5

Ⅰ.①生… Ⅱ.①李… ②黄… Ⅲ.①生物工程—医学工程—语义学—研究 Ⅳ.①R318②H030

中国版本图书馆 CIP 数据核字（2012）第 137194 号

生物医学语义技术

李劲松　黄智生　主编

责任编辑	张凌静(zlj@zju.edu.cn)
封面设计	王波红
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	杭州中大图文设计有限公司
印　　刷	浙江印刷集团有限公司
开　　本	787mm×1092mm　1/16
印　　张	27.5
字　　数	670 千
版 印 次	2012 年 7 月第 1 版　2012 年 7 月第 1 次印刷
书　　号	ISBN 978-7-308-10129-5
定　　价	58.00 元

版权所有　翻印必究　　印装差错　负责调换

浙江大学出版社发行部邮购电话　(0571)88925591

内容简介

本书系统地介绍了语义技术及其在医学和生命科学上的应用。内容包括语义技术的基础理论、技术与方法，医学与生命科学语义数据应用，医学语义检索技术，生命科学语义技术，中医语义技术，以及医学信息系统中的语义技术等。最后展望了生物医学语义技术的应用前景。

本书主要作为生物医学信息学相关专业的研究生教材，也可作为医学与生命科学、计算机科学与工程、信息科学与技术等领域研究人员、工程技术人员、教师、学生的参考书。

编 委 会

主　编　李劲松　黄智生

副主编　俞思伟　包含飞　郑　锴

编　者　（按姓氏笔画排序）：

丁宝芬	南京中医药大学
王华琼	浙江大学
包含飞	上海中医药大学
李亚子	中国医学科学院
李劲松	浙江大学
罗志伟	日本神户大学
周　强	上海中医药大学
周天舒	浙江大学
郑　锴	美国密歇根大学
胡　臻	美国明尼苏达大学
俞思伟	武汉大学
姜　赢	北京师范大学珠海分校
钱　庆	中国医学科学院
黄智生	荷兰阿姆斯特丹自由大学
董建成	南通大学

统　稿　周天舒　王华琼

此为试读，需要完整PDF请访问：www.ertongbook.com

Editor Committee

Chief Editors: LI Jingsong HUANG Zhisheng

Vice-Chief Editors: YU Siwei BAO Hanfei ZHENG Kai

Committee Members:

BAO Hanfei	Shanghai University of Traditional Chinese Medicine, China
DING Baofen	Nanjing University of Chinese Medicine, China
DONG Jiancheng	Nantong University, China
HU Zhen	University of Minnesota, U. S. A.
HUANG Zhisheng	Vrije University Amsterdam, The Netherlands
JIANG Ying	Beijing Normal University Zhuhai Campus, China
LI Jingsong	Zhejiang University, China
LI Yazi	Chinese Academy of Medical Sciences, China
LUO Zhiwei	Kobe University, Japan
QIAN Qing	Chinese Academy of Medical Sciences, China
WANG Huaqiong	Zhejiang University, China
YU Siwei	Wuhan University, China
ZHENG Kai	University of Michigan, U. S. A.
ZHOU Qiang	Shanghai University of Traditional Chinese Medicine, China
ZHOU Tianshu	Zhejiang University, China

前　言

从20世纪末到21世纪初，信息科学与计算机技术领域最重大的科学进展之一就是万维网的诞生与普及。它深刻地影响了整个人类社会的发展进程，同时也给各个领域的科学研究带来了一个全新的信息和知识环境。语义网及语义技术的发展为处理万维网上浩瀚的信息和知识资源提供了一种全新的方法，已经在诸多领域得到了广泛的应用。医学及其生命科学的研究成为了语义技术发展与应用最为活跃也最富有成效的领域之一。近十多年来，在计算机科学家与医学、生命科学领域专家的共同努力下，它已经取得了丰硕的研究成果，极大地丰富了现代医学与生命科学研究的信息资源整合与知识分析管理的手段，促进了现代医学与生命科学的研究发展。

在这样一个以万维网为主要信息与知识资源的研究环境下，系统地介绍语义技术及其在生物医学领域的发展与应用，具有特别重要的意义。为此，我们特意组织国内外医学与生命科学领域语义技术研究的著名专家学者撰写本书。本书主要的设计目标之一是作为生物医学信息学相关专业的研究生教材。此外，由于本书涉及的内容极其广泛，不仅涉及语义技术的基础理论，而且还覆盖了语义技术在生物医学应用中的方方面面，故本书对于医学及生命科学、计算机科学与工程、信息科学与技术等领域的学生、教师和研究人员都将具有一定的参考价值。

全书共分十二章。各章节是由各位专家学者相对独立完成的，每一章都反映了语义技术以及生物医学某一方面的最新成果与发展趋势。在本书的撰写过程中，各章编写者在百忙之中精心组织所负责的有关章节的内容，并对部分其他章节的内容进行认真的审阅，付出了大量的心血。在此，我们对全体参编者的努力工作深表谢意。感谢解放军301医院的叶玲博士为审阅本书第十章关于语义技术及其在生命科学中的应用所提出的许多宝贵意见和建议。感谢各位参编专家所在单位相关科研人员、研究生的大力支持和无私帮助。

本书所介绍的部分研究工作以及本书的出版得到了国家自然科学基金项目

(No. 61173127)、国家“863 计划”项目(No. 2009AA045300)、国家科技支撑项目(No. 2011BAH15B08)、中央高校基本科研业务费专项资金,以及浙江大学"海外一流学科伙伴计划"专项资金等的资助,在此一并感谢。

由于时间仓促,编者水平有限,加之编者分散在世界各地,很难汇聚一堂就编著内容展开充分研讨等原因,本书在内容安排与描述上出现纰漏与瑕疵在所难免,敬请各位同行专家及广大读者不吝赐教指正。

李劲松　黄智生

2012 年 6 月于杭州、阿姆斯特丹

目　录

此为试读,需要完整PDF请访问: www.ertongbook.com

第一章　语义技术导论

第一节　语义技术概述

以计算机技术为基础的信息技术已经成为现代科学和现代社会的重要基石，已经是现代人类社会密不可分的组成部分。近几十年来，计算机网络技术的发展，促进了信息技术的网络化和分布式处理方式的新技术的发展。特别是万维网的诞生，使得信息技术面临了一个全新的信息处理环境。现在，已经很难想象，一个现代化的信息处理系统能够脱离网络环境而存在。由于互联网上的信息资源主要是以万维网的形式提供的，所以，研究面向万维网的信息处理技术成为现代信息处理技术的核心主题，这也是本书所涉及的核心主题之一。

这个面向万维网的信息处理环境具有以下特征：

分布性(distribution)：我们所要处理的信息资源不再是孤立地存储于单一的计算机系统之中，而是分布性地存在于万维网系统之中。

异构性(heterogeneity)：我们所要处理的信息资源的表达格式可能是千差万别的。

海量性(scalability)：我们信息处理系统所涉及的信息资源，从数量上看，可以是极其巨大的；从某种意义上(即从现有的万维网上的所有信息资源)看，甚至可以说是浩瀚无穷的。

不一致性(inconsistency)：万维网上的部分信息资源汇集在一起，从内容上看有可能是相互矛盾的。

动态性(dynamics)：网络信息资源总是处于实时和动态变化之中的，即不存在一个绝对封闭的信息或知识范围。

由于面向万维网的信息处理环境具有以上特征，因此完全靠人来完成这些信息的分析处理以获取有用的知识越来越不现实。这促使科学家们寻找和开发新的信息处理技术来解决或者是缓解这个问题，促使万维网之父 Tim Berners-Lee 提出了语义万维网(简称语义网)(The Semantic Web)[①]的思想，希望对现有网络信息资源作语义标注，使得人们能够更方便、更快捷地找到网络信息。

语义网的主要思想是采用逻辑语言作为描述工具来刻画各种信息内容(即它们的语义)。我们把这些基于逻辑语言描述的信息称为知识，把这种面向语义网的信息和知识处理技术统称为语义技术(semantic technology)。

语义技术的主要技术特点是：

①形式表达：采用某种形式化的语言来描述网络信息资源。

---

① http://en.wikipedia.org/wiki/Semantic_Web.

②推理支持：采用某种逻辑推理工具来分析数据，并能获得数据表达背后的间接内容。这个从数据中获得蕴含的间接内容的过程，就是推理(reasoning)。

近年来，语义技术作为新一代信息技术的基础受到了普遍的关注，并在许多领域(如生命科学领域、信息检索领域、交通信息管理领域等)得到了广泛应用，而且其所展现的优越性逐渐被业界所认识，为信息系统的智能化分析和决策支持提供了重要的技术支持。

语义技术采用国际统一标准的基于语义的数据表达语言，如资源描述框架(resource description framework，RDF)数据和网络本体语言(Web ontology language，OWL)的表达方式，使得数据独立于特定的系统。引入了基于知识表示和推理的方式，使得基于语义的信息系统能够方便地在现有基础上实现许多基于知识分析的功能扩展，这大大地缩短了新系统开发的周期，而且提供许多通过现有数据检索不能获得的而只能通过推理才能获得的信息。

与传统的信息处理技术相比，语义技术的优越性主要表现在以下几个方面：

①数据共享。现有的许多领域的数据均有其对应的语义数据，如地理语义数据集(GeoNames)，维基百科数据集(DBPedia)等，可以非常方便地被融合到其他语义数据中，实现最大程度的数据共享。

②知识表达。由于采用基于知识表示及其本体表达的技术，语义系统能够方便地在知识层面上进行分析、模拟和管理，代替现有的大量人工干预的枯燥工作。

③决策支持。引入知识处理，提高了处理问题的精度和效率，提供了知识管理与推理，对宏观把握信息系统提供决策支持。

语义技术的上述优势使得它的应用领域变得非常广泛。它可以应用于各种不同的情景中，如应答机、自动内容标志、基于概念的检索、内容注解、动态用户界面等。在生物医学领域，随着各个生物系统(biological systems)产生越来越多的数据，以及多学科交叉产生各种各样的新数据，传统的生物医学研究已经逐渐走向了转换医学和个性化医学阶段，而这些新研究必定会产生重要影响。这些研究能否顺利展开，很大程度上依赖于多学科的结合以及对于数据的分析。然而，大量异构、分立、具有复杂语义特征的数据使得生物医学信息学面临着极大挑战。生物医学现象本身复杂的特性，使得生物医学数据也具有多维特性。语义技术逐渐走向成熟，从而越来越多地被应用于生物医学信息学中。

第二节 语义万维网

语义网是指能够表达语义的信息而能被计算机自动信息处理的万维网，它最早是由万维网之父 Tim Berners-Lee 提出的，其经典文献是发表于 2001 年《科学美国人》杂志，题为“The Semantic Web”的论文。语义网作为现有万维网的改进和补充，便于解决现有网络信息爆炸性增长所带来的巨大的处理和搜索问题。语义网的核心问题就是如何有效地表达语义信息，使之能够被计算机有效处理。

从语义网的发展起源来看，语义网是人工智能领域和网络技术相结合的产物。人工智能领域中的知识工程研究从孤立的知识库系统逐渐发展到基于内部网(专用网)、外联网的信息系统集成，最后扩展到整个因特网。在这个研究过程中，逐渐加深了对知识表示和推理的认识，并总结出一些新的描述和推理方法。另一方面，万维网经过十几年的发展，积累的

海量数据需要一种新的、机器可以自动完成的方式来处理和管理。因此，当这两个领域的积累都比较成熟，而且有了需求时，就必然会走向结合。

语义网建立的基础，是知识的概念化和形式化以及相应的推理，并且它和人工智能有着深厚的渊源关系。因此，许多分析都需要从人工智能领域的角度来考察。但是由于应用环境不同，两者间还存在着一些差异。例如，从人工智能的逻辑学派和认知学派的观点来看，知识和概念化是人工智能的核心。传统的人工智能系统，虽然采用一些共同遵守的公共概念或者是一致的定义，但是一般都有它们各自狭义的、特有的用于信息推理的规则集合。尽管数据能够从一个系统转换到另一个系统，但由于系统间的推理规则通常以完全不同的形式存在，而致使一个系统的规则不能用于其他系统。从这一点上看，传统的人工智能系统是一种集中、孤立(专有)的系统。同样，语义网也是以知识的概念化表示为基础展开的。语义网中的知识，就是一系列对资源的建模及描述。资源是一个非常广泛的概念，它可以是网站、网页，甚至网页的某一个部分的内容。该描述采用某种形式的符号和表达式对网络上的与该资源相关的其他资源，以及这些资源之间的关系进行刻画。但是，和传统人工智能系统不同，语义网的知识表示的特殊性在于，它本身要符合网络的分散性和通用性。知识的表示本身可能是由众多的独立团体或个人，以各种各样的方式来提供，而这些知识却又要求能够被各种各样的应用实现共同理解，并且在一定的逻辑规则的指导下进行推理，所以语义网上的知识具有创建上的分散性，同时又具有应用上的通用性。综合上述差异，语义网更侧重大数据量、低智能化的应用环境，重在使用推理获得间接数据；而人工智能侧重小数据量、高智能化的应用环境。这是语义网和传统的人工智能系统的一个非常重要的区别。

网络本体语言已于 2004 年由国际万维网组织(W3C)通过，成为国际标准的网络本体描述语言，它构成了语义网技术的重要基础之一。OWL 是资源描述语言 RDF/RDFS 的进一步扩充，主要增加了本体描述的许多特征，如逻辑运算(特别是逻辑否定等)的描述能力。我们将在第五章具体介绍网络本体语言。图 1-1 是著名的语义网技术层次图，从中可以看出，唯一资源标识(uniform resource identifier，URI)、国际资源标识(internationalized resource identifier，IRI)，以及 Unicode 技术构成了语义网的最底层基础。因为 URI 实现了网络信息资源的唯一标识，IRI 作为对 URI 的补充，允许使用 Unicode 来标识网络资源，Unicode 实现了不同语言的符号集的单一表示，从而实现了在信息的数字化表达层面的最基础的技术统一。

可扩充标记语言(extensible markup language，XML)的引入使得我们能够通过标签

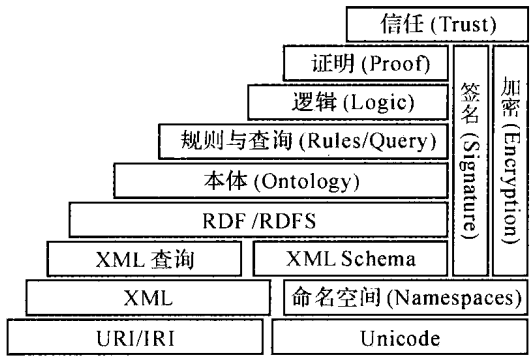


图 1-1　语义网技术层次图

(tag)来表达半结构化的数据。名字空间(Namespaces)的使用使得概念的 URI 前缀表达有了更方便的表达方式。在此基础上引入 XML 框架模式(XML Schema)及其 XML 查询语言,使得我们可以用它来表达许多不同领域的元数据。RDF/RDFS 所提供的一系列技术使我们能够方便地描述和询问网络数据资源。

使用标准的网络本体语言来描述信息仅仅完成了本体的最基本的性质描述,更多的本体数据性质可能需要更强的本体描述能力,也就是通过规则(rule)来描述。除此之外,还需要用本体查询(query)与管理语言对本体数据进行询问和基本推理处理。这就是图 1-1 中所展现的规则与查询层(Rules/Query)所要求的内容。

对于特定的应用系统,我们还需要在这基础之上建立一个特定的逻辑推理层,也就是图 1-1 中所展显的逻辑(Logic)层,使之能够满足具体应用系统的基本要求,而基于应用逻辑层所建立的证明层(Proof)才完成一个具体应用系统的要求,使之能够成为一个相对完整的应用系统。考虑到信息安全问题,我们还需要考虑其系统的可信度问题,以及个人签字系统(Signature)和加密(Encryption)等信息安全问题,这就是信任层(Trust)的系统要求了。

现有的语义网技术开发与研究主要集中在逻辑层以下。特定领域的本体开发经过近十年的发展,特别是与网络本体语言的结合,已经日趋成熟。

当前语义网研究开发涵盖了许多万维网的全球规范与前沿技术,例如,互联数据、群体语义网、语义网搜索、智能数据集成、语义网信息挖掘,以及海量万维网计算。经过多年来的基础研究以及规范化工作,初具规模的语义网已经为万维网上语义技术的研究提供了众多研究方向与广阔应用前景,下文将着重讨论该技术在智能信息检索、企业间数据交换及知识管理、万维网服务这三个方面的应用。

面对海量信息,智能信息检索一直是科研人员的重要课题。但是,正如本章第一节所言,万维网上传统的信息表示方法使信息检索面临了种种窘境。因此改进信息检索的重要方法之一就是整理和重新规范万维网上的信息。如今,万维网在保持高速发展期间产生了大量的普通超文本标记语言(hypertext markup language,HTML)页面,整理这些信息的实质性问题就是如何从 HTML 页面中提取语义信息,构建能够描述这些页面的本体(ontology)。手工实现这一过程需要耗费大量的人力、物力。因此,可行的办法是采用本体学习系统实现本体的自动或半自动提取,不仅对文本信息可以采用语义网的方法来加强智能检索,而且还可以对多媒体信息结合模式识别和对象提取技术实现基于内容的检索,国外已有这方面的文献报道。前人对传统万维网信息内容模型、信息检索和信息提取、计算语言学、机器学习等方面展开了大量的研究,并取得了很多成果,为网络信息的整理打下了很好的基础。

企业间的数据交换和知识管理一直是基于万维网的电子商务和企业资源规划(enterprise resource planning,ERP)系统的重要组成部分,现有很多项目都围绕着企业万维网知识管理而展开。这些项目潜在的假设就是,企业提供的万维网信息结构可以转化成为一个巨大的知识库。这种转化的重要基础就是利用基于本体的元数据来对企业发布的信息或企业的内部文档进行标注。围绕这一假设,需要开发一系列的相关技术和工具,如企业知识的建模、标注工具、本体的提取工具、本体推理工具等。

当前万维网正在从一个文本、图片、音频、视频的信息提供者向服务的提供者转变,这种转变体现了"网络就是计算机,软件就是服务"的思想。产业界目前推动的网络服务(Web

Service)通过万维网向外界提供了如何调用自身功能/服务的说明。由于在网络环境下的分布式计算涉及平台的异构性,因此它的核心技术包括 XML 作为数据传输和交换的标准格式,以简单对象访问协议(simple object access protocol,SOAP)作为发送和接收 XML 数据的基本消息协议。底层的传输则采用超文本传输协议(HTTP)、文件传输协议(FTP)、简单邮件传输协议(SMTP)、互联网协议第四版(IPv4)、互联网协议第六版(IPv6)等因特网协议。服务的描述、查找和发布则采用了 Web Service 描述语言(WSDL)、通用描述、发现与集成服务(UDDI)等协议。当前,服务并没有以本体为基础,基本上还是采用标准化分类的方式来描述服务的功能、提供者以及如何访问服务并与之交互。因此它们对服务的描述能力非常有限,而且缺少灵活性。学术界在语义网研究中提出了基于本体的一些服务描述语言如 DAML-S 等,这些语言为语义网和 Web Service 的结合提供了一个良好的契机。通过创建语义网的语义描述,使得 Web Service 能够被机器理解、对用户透明,同时这种描述能够被代理商自动处理,实现 Web Service 之间的交互性。

Cycorp 是目前世界上最大的知识库和推理引擎,它利用其在知识表示、机械推理、自然语言处理、语义数据集成、信息管理和搜索最前沿的创新,提供一个语义中间件,在以知识为基础的应用阵列开发能力方面提供了大量的语义支持技术;它依靠结合了无与伦比的常识与一个强大的本体推理引擎和自然语言接口,形成对新的知识密集型应用的开发。图灵奖获得者 Edward Feigenbaum 赞美 Cycorp 拥有世界上最大的知识库。

基于语义网的搜索引擎 Swoogle,是一个针对互联网上的语义网文档、术语以及数据的搜索引擎。Swoogle 利用一种搜索器系统来发现 RDF 文档以及内置有 RDF 内容的 HTML 文档,然后会针对这些文档及其组成部分进行推理,并在其数据库之中记录和索引具有实际意义的元数据。

第三节 元数据与本体

在语义网思想发展的初期,人们主要期待的是,希望对现有网络信息资源作语义标注,使得人们能够更方便、更快捷地找到网络信息。由于描述网络数据的需要,科学家们开发了一系列元数据描述语言,如 RDF/RDFS 等。出于对语义分析进一步细化的需要,科学家规定了本体描述语言(如 OWL),并开发了种种特定领域的本体。所谓本体,可以简单地将它理解为特定知识领域中满足共同约定的常识部分,这对于特定领域信息分类是必要的一步。

本体是语义网的核心。按照 Gruber 的定义,本体是感兴趣领域共享的概念化的显式规约。作为一个规约,本体需要通过某种语言表达,通常被看成知识库中满足共同约定的常识部分。2004 年 2 月 10 日,与 RDF、RDFS 和 OWL 语言有关的 12 个技术规范正式发布,标志着语义网的本体语言及理论基础已经奠定。同年 2 月 25 日,W3C 成立了"Semantic Web Best Practices and Deployment"工作组,宣告"Semantic Web Activity"进入第二阶段。2007 年 11 月,RDF 查询语言技术规范 SPARQL 正式发布。与此同时,为鼓励发展语义网应用,国际语义网大会自 2003 年以来每年都举办语义网挑战赛。目前,OWL 已成为基于语义网本体描述语言的国际标准,也成为各类语义知识库常用的描述语言,如著名的英文 Wordnet 已经有对应的 OWL 格式表达,Cyc 大型知识库系统也有其对应的 OWL 文件。OWL 语言根据其不同的应用需求可分为不同的子语言,如 OWL Lite、OWL DL 和 OWL Full 等。其

中最常用的是 OWL DL 语言，它建立在描述逻辑(DL)的逻辑体系之上，是对应着一阶谓词逻辑中可判定的一个子语言。OWL 2 是 OWL 基本版本的进一步扩充，其逻辑语义对应着描述逻辑中的 SROIQ 语言。

用 RDF/RDFS 表达的数据被统称为元数据(meta data)。规则语言如 SWRL 是对本体语言 OWL 的进一步扩展。元数据、本体与规则描述构成了网络结构化的数据，通常被称为网络知识，简称知识。我们使用网络结构化数据来对现有万维网上的非结构化的数据进行语义标注(annotation)，构成了语义网的数据基础。结构化的网络数据连同被语义标注的数据被称为语义数据(semantic data)，它们连同其他网络数据构成了网络信息资源的全集。网络信息资源结构图如图 1-2 所示。RDF 和 RDFS 的语义模型是基于三元组结构的，OWL 是 RDF/RDFS 语言的进一步扩充，所以在语义网与本体技术研究领域，语义数据的规模通常是以三元组(triple)的数量来度量的。

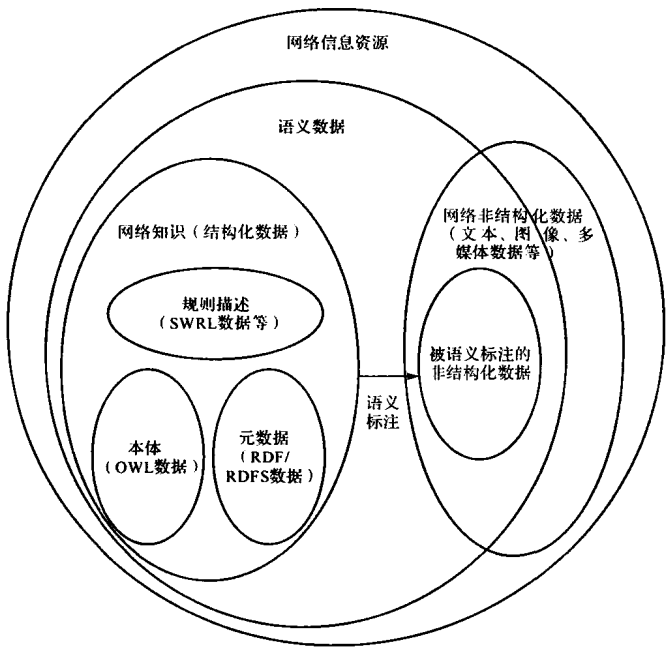


图 1-2 网络信息资源结构图

语义技术已经深入人类知识领域的方方面面，现在已经很难找到一个领域可以宣称与语义技术无关。且不说生命科学领域，食品与农业领域已有许多研究人员在做与语义技术相关的工作，就是在冷僻的领域，如石油勘探与开采、红学研究、政治学分析等，都有人在开发本体产品。

语义技术能带来经济效益的应用不胜枚举，其巨大的技术潜力之一是它能够代替大量的人工干预和分析数据的枯燥工作，如对海关的大量进出口数据进行预处理和筛选，再如应用于价格比较网站和信息推荐网站等。这些应用都具有以下特征：它需要人工低智能化的干预，而不能完全被自动化处理，否则效果会较差。如音乐下载推荐网站，目前使用的技术，要么由简单的字符串匹配来决定推荐信息，要么是分析以往用户的下载习惯来决定，要么是人工预先安排推荐的信息。语义技术在这些方面能够取得最佳的效果。

此为试读，需要完整PDF请访问：www.ertongbook.com