语义网的主要功能及其在数字图书馆中的应用*

□ 欧石燕 胡珊 / 南京大学信息管理学院 南京 210093

摘要：语义网自诞生以来，其发展过程一直处于不断的调整变化中，新的语义网标准规范不断推出，其功能和应用也不断向深度和广度扩展。文章首先对语义网的诞生与发展过程进行了回顾与分析，然后通过对语义网应用的调研归纳总结出语义网的主要功能，最后对语义网功能在数字图书馆中的应用进行了分析与阐述。

关键词：语义网，关联数据，数字图书馆
DOI：10.3772/j.issn.1673—2286.2014.03.001

1 语义网的诞生与发展

自1991年万维网（简称Web）诞生以来，Web已经发展成为一个拥有亿级页面的巨大分布式信息空间，为用户提供海量的信息服务。二十多年来，Web虽然经历了一系列变化与发展，但是无论是1.0时代的只读静态网页，还是2.0时代的交互式信息平台，基于超文本格式的非结构化文档之网（web of documents）的特性一直都没有改变，这使得当前Web还远远不能满足人们对信息共享和处理的需要，主要表现在[1]：（1）当前Web只能供人阅读和进行信息共享，计算机并不能“理解”Web的内容，并在“理解”的前提下处理和利用这些信息；（2）即使目前有大量网页的内容是由来自底层数据库的结构化数据自动生成，但是网页一经生成，信息反而失去了在数据库中的结构化特征，而这一特征对于机器理解和处理信息是非常有用的；（3）人们虽然能在HTML网页中建立超链接关系，但是却无法在生成这些网页的底层数据间建立关联关系，导致Web底层的海量数据孤立而分散地存在着，无法进行集成和互操作，形成了一个个信息孤岛。

正因如此，人类对当前Web的利用无法得到软件工具的很好支持。一直以来，伴随着Web诞生的搜索引擎是人们搜索和使用Web信息的几乎唯一的、不可或缺的工具。 但是以关键词匹配为核心的Web搜索引擎同样面临着一些严重问题[1]，如（1）高匹配但低精度；（2）低匹配或者无匹配；（3）查询结果对查询词高度敏感；（4）用户必须自行在检索结果中浏览和定位所需文档并从中抽取有用信息进行集成。虽然研究者们试图采用各种手段提高搜索引擎的智能性和精度，但是上述问题仍无法从根本上得以解决，这同样归因于当前Web非结构化、非语义化的信息表示形式，使得搜索引擎只能通过简单的关键词匹配而非语义匹配来搜索相关信息，并且关键词之间只能通过简单的布尔关系而非准确的概念关系来描述。

面对着当前Web在信息表达、组织、检索中存在的严重缺陷与不足，有两种可能的途径来解决上述难题[1]：一种考虑是，采用人工智能和自然语言处理技术开发出更为复杂的程序来对网页上的非结构化信息进行处理。但是很遗憾，人工智能技术本身有着很大的局限性，再精妙的机器（计算机程序）也无法真正像人一样进行理解和思考，这一途径无疑遇到了无法突破的瓶颈。另一种考虑是，能否采用一种适于机器理解和访问的新方式来表达Web上的内容，从而方便机器的处理？这就是语义网想法的最初由来。1998年，万维网的发明人伯纳斯-李在他的Web设计笔记里首次提出了对语义网的设想，即“一个在某种程度上类似于全局数据库的数据之网（web of data）”[2]。2001年5月，伯纳斯及其合作者在*Scientific American*杂志上发表了题为“The Semantic Web”的论文[3]，系统论述了他对下一代万维网架构语义网的蓝图，这篇论文同时也被认为是语义网诞生的标志。

* 本文系国家社科基金重点项目“语义网应用技术体系和发展战略研究”（编号：11AZD121）的研究成果之一。

语义网的目标是通过给万维网上的文档添加能够被计算机所理解的语义（Meta data），让计算机能够“理解”分布在网上的信息和知识，并在“理解”的前提下更好地处理、利用这些信息和知识，从而使整个Web成为一个支持全球化知识共享的智能信息服务平台。由此看出，语义网相对于现有万维网的最大优势是“机器可理解”，它对Web的扩展可以使得Web具有知识理解及一定的推理和自动处理能力，它的出现给Web带来了革命性的变化，使人和机器协同工作、理解并处理Web上的信息成为可能。因为语义网的建立极大地涉及了人工智能领域的部分，与Web 3.0智能网络的理念不谋而合，因此语义网也被看作是Web 3.0的重要特征之一。

在语义网出现的最初十年，语义网的各项语言规范得到制定和完善，如RDF、RDFS、OWL、SKOS、SPARQL等，各种语义网实现工具也相继被开发出来，譬如，RDF三元组存储器3Store、Virtuoso和AllegroGraph，RDF数据转换工具RDFizers，语义网开发工具包Jena，本体编辑器Protégé和OntoEdit，本体推理机Pellet、RacerPro和FaCT++，从而使语义网技术有了在实践中进行应用的可能。但是，相比Web当初的发展，语义网的发展速度还显得比较迟缓，对语义网的研究主要集中在高校和研究机构，缺乏企业界的普遍参与，除了在少数专业领域（如医学和生物），几乎没有对广大Web用户有价值的语义网数据集的存在。原因主要在于：在语义网研究的早期，过分强调推理的必要性，大量依靠本体进行知识建模和语义标注，使得语义网的应用只能局限在特定领域的狭小范围，而无法扩展至Web级的海量数据，因此严重制约了语义网在整个Web上的推广与应用。

2006年，语义网的发明者伯纳斯-李进一步提出了关联数据的概念[4]。关联数据是指在网络上发布、共享、连接各类数据、信息和知识的一种方式，是推荐的语义网最佳实践[4]。关联数据从技术上来说虽然很简单，然而却正在使Web发生深刻的改变，它摒弃了语义网早期过度依赖本体进行知识建模和数据集成的做法，强调RDF数据的互联与Web访问，促进了数据之网（Web of Data）的创建，为语义网的大规模应用奠定了基础。严格说来，数据之网并不能算是真正的语义网，因为它主要强调数据结构化和关联，还远未达到伯纳斯所设想的语义与智能的程度，因此可将其看作是语义网的一个子集或初级阶段。近年来，许多机构纷纷采用关联数据作为发布结构化数据的一种途径，从而构成了一个全球的数据空间。该数据空间的出现源自于语义网研究社区的努力，特别是得益于万维网联盟（W3C）的“Semantic Web Education & Outreach”工作组支持的“关联开放数据（Linking Open Data，简称LOD）”项目。截至2011年9月，在LOD云中已有310亿个RDF三元组，被5.04亿个RDF链接所连接，所关联的数据集已达到295个[5]。整个LOD云以DBPedia①为核心，囊括了地理、政府、媒体、生命科学、图书馆、用户生成内容等领域的数据以及一部分跨领域数据，其中图书馆及其相关领域（如教育、出版）的关联数据集有87个，约占整个LOD云的9.33%[5]。目前，LOD云中的数据几乎以每3年2个数量级的速度在增长，以致到了几乎无法计量的地步。

在注重将现有结构化数据以关联数据方式在Web上进行发布的同时，如何实现海量的传统网页向语义网的自然过渡更是值得考虑的问题。通过基于本体的语义标注将非结构化的HTML网页完全转换为结构化的RDF数据，意味着完全放弃传统Web及背后的成熟技术，这种做法不仅复杂，而且事实证明也不被广大Web用户所接受。将RDF数据以RDF/XML格式直接嵌入到XHTML网页中也是不可行的，因为这是目的和结构完全不同的两种表示格式，互不兼容。万维网联盟（W3C）和Web民间社区殊途同归地分别提出了三种功能相同的Web语义标注格式：来自W3C的RDFa格式[6]、来自民间的微格式[7]和来自HTML5规范的微数据[8]，这三种格式虽然具体的规则和表示不同，但是本质上都是通过在原有HTML/XHTML网页中嵌入语义标签，从而将传统网页在人类可读的基础上提高到机器可读、可理解的状态。2011年6月，微软、谷歌和雅虎三大引擎联手发起了schema.org网站[9]，以帮助网站管理员在网页里使用结构化数据标记来帮助搜索引擎更好地理解网页里的内容，那些符合schema.org标注格式的网站，不仅能提高用户检索效率，也同时能增加网页被检索到的可能，这一措施无疑进一步促进了语义网技术的普及与应用。

2 语义网的主要功能

通过对2008至2013年“语义网挑战竞赛（Semantic

① DBpedia是从Wikipedia词条中抽取出结构化数据并将其以关联数据形式在网络上发布的项目，见http://wiki.dbpedia.org/Datasets。

Web Challenge）[②]”中的84个语义网应用进行调研，以及对其它零星语义网应用案例和部分文献进行梳理，笔者对语义网的主要功能进行了归纳总结。在本文中，笔者定义：语义网功能是指计算机程序或系统依靠语义网技术能够完成的一项特定任务或实现的一个特定目的，多个功能的有机结合构成一个完整的面向终端用户的语义网应用，如语义知识管理系统、语义门户、语义推荐系统、语义数字图书馆等。

（1）语义标注（Semantic Annotation）

语义标注是指给传统的HTML网页添加语义信息，将其升级为机器可理解的语义网的过程[10]。早期的Web语义标注主要依赖于本体进行，首先要采用本体语言形式化地定义标注词汇（即本体的类和属性）及其语义，然后利用这些词汇作为语义标签对非结构化信息进行标注，将其转化为与使用的本体相兼容的RDF数据（即本体的实例），这种语义标注方式也被称为本体填充（Ontology Population）。基于本体的语义标注实施起来比较复杂，不仅需要构建本体，而且还需要学习不同于传统HTML/XHTML格式的全新数据格式，不利于普通用户掌握。为此，W3C和语义网民间社区分别推出了RDFa、微格式和微数据三种Web语义标注语言，只采用简单的语法就能够充实已有网页的语义，避免使用重量级的语义网知识。目前，包括Google在内的几家搜索引擎巨头大力提倡采用微格式对传统网页进行标注[9]，这一举措无疑将促使大量人类和机器同时可读可理解的语义网页的出现，推动语义网的发展。语义标注可以通过手动、半自动和自动的方式进行，已经有大量的基于本体的标注工具存在，如SHOE Knowledge Annotator[11]、AeroDAML[12]等，支持手动和半自动标注，但是构建大规模语义网应用关键还是要实现对大批量HTML网页的自动标注，这往往需要采用自然语言处理、文本挖掘和机器学习技术来进行，譬如进行命名实体识别、关系抽取、语义相似度计算等。

（2）结构化数据的RDF化转换（RDF Conversion）

结构化数据的RDF化转换是指将特定应用的结构化或半结构化数据，如关系型数据库、Excel数据表、MARC格式的书目数据、BibTex格式的书目数据等， 转换为RDF数据。目前已经存在着大量RDF自动化转换工具，譬如SIMILE项目中开发的RDFizers工具组件可实现近20种结构化数据的转换[13]。自关联数据出现以来，出现了一些工具能够将关系型数据库直接以关联数据的形式在网络上发布，譬如D2R Server允许用户使用D2RQ映射语言自定义映射关系，将关系型数据库中的数据转换为RDF数据并将使之可通过Web访问[14]；内容管理系统Drupal能够借助导入的本体将结构化数据进行RDF化转换并发布为关联数据[15]。

（3）数据集成（Data Integration）

语义网一个很大的优点就是能够将来自多个数据源的异质数据进行集成。通过共同的领域本体，可以将同一领域的不同来源、不同格式的数据进行整合。譬如，基于同一书目本体，能够将MARC格式、DC格式、BibTex格式的书目数据转换为统一的RDF格式进行整合[16]。对于具有相关性的不同领域的RDF数据，可以通过建立RDF语义链接构成关联数据，譬如，将语义书目数据、基于FOAF本体的表示个人/组织机构的数据、SKOS的受控词表相互关联，形成一个更大范围的图书馆书目数据的语义视图[16]。

（4）语义推理（Semantic Inference or Semantic Reasoning）

语义推理通常是指根据一组确定的事实或者公理推断出逻辑结论的过程。这里的语义推理特指语义网上的推理（Inference on the Semantic Web or Semantic Web Inference）。推理是语义网的一个关键特性。在语义网上，数据被建模为一组资源之间的（被命名的）关系，推理就是基于这些数据和一套推理规则自动发现或生成新关系的过程[17]。推理规则可通过词表或者规则集来定义，这两种方式都需要用到知识表示技术。通常来说，本体偏重于分类方法，重点定义类和子类以及实例与类之间的关系；而规则偏重于定义在已有关系基础上发现和生成新关系的通用机制[17]。在语义网规范中，RDFS和OWL本体语言用于定义本体，而W3C新推出的规则交换格式RIF则被用于在已有的规则语言间进行规则交换[18]。目前已经出现了多种用于语义网推理的推理机，如RacePro、Pellet、FaCT++等，基本上都采用一阶谓词逻辑进行推理。语义网推理能够通过数据中已知的关系发现未知的新关系，是实现Web智能的关键要素之一。通过推理，还能够发现集成后数据中的不一致或者可能的不一致，是提高Web上数据集成质量的重要手段之一。

（5）语义搜索（Semantic Search）

[②] The Semantic Web Challenge是在“国际语义网会议（ISWC）”中举办的一个竞赛，参赛者提交采用语义网技术开发的面向终端用户的在线应用进行评比，是反映最新语义网研究状况的一个窗口。

语义搜索是指通过理解用户的检索意图和检索词在检索空间的语境意义（即上下文意义）来提高检索精度[19]。语义搜索对于传统的搜索引擎来说是一个巨大的挑战。Google、Bing等搜索引擎巨头纷纷采用各种手段在部分程度上实现对传统Web的语义搜索，主要包括：考虑词的语义变体、同义词检索、概念匹配检索、泛化的和细化的查询以及自然语言问答式检索等[19]。在实施上述方法时，虽然本体常被用来支持对用户查询的分析，帮助理解查询词的含义及相互间的关系，但从本质上来说，上述语义搜索只是对传统关键词检索进行了增强，并非利用语义网进行纯粹的语义搜索。 语义网的出现是缘于人们对传统搜索引擎的不满和对语义搜索的追求。基于语义网机器可读可理解的语义数据模型，能够在更高层次上实现语义检索，目前已经有以下几种实现方式：

- 丰富网页摘要：通过采用RDFa、微格式、微数据标记HTML网页，使搜索引擎能够理解网页上的内容，提高搜索的准确性，并在每条搜索结果下方显示几行文字（结构化信息），帮助用户了解搜索结果是否与搜索内容相关。目前Google、Bing、Yandex、Baidu等著名搜索引擎都在一定程度上支持丰富网页摘要[20]。
- 面向关联数据的自动问答式检索：针对LOD云中的RDF数据集，允许以自然语言提问的形式检索信息。由于底层数据是具有语义的结构化数据，能够进行推理，实现真正的语义检索，但其难点在于：如何将一个自然语言提问准确地转换为一个结构化的SPARQL查询。目前有大量关于此方面的学术研究出现，是一个研究热点[21]。
- 基于知识库的检索：当用户对一些著名人物（如爱因斯坦）进行查询时，Google搜索引擎除了常规地返回一组包含查询词的相关网页，还能够从一个或多个知识库中抽取出关于该人物的结构化信息，经集成后形成人物简介显示在结果页面的左侧。目前已有的知识库包括Freebase、DBPedia、美国中央情报局出版的The World Factbook等，这些知识库其实是一种语义数据库，将各种对象通过语义链接相互关联起来，形成一个语义网络。目前，Google正在打造知识图谱（Knowledge Graph），一个综合已有知识库的更大知识库，为其语义检索提供支持[22]。

（6）语义仓储（Semantic Repository）

语义仓储，也被称为语义网仓储（Semantic Web Repository），是类似于数据库管理系统（DBMS）的一个引擎，允许存储、查询和管理大量的RDF数据，支持SPARQL查询，并对RDFS和OWL表示的Schema和本体进行解释[23]。因此，这类引擎也自然承担了语义网Web服务器的角色。语义仓储具有DBMS的某些功能和特性，但两者的主要区别在于：(1)语义仓储采用本体作为语义模式，能够对数据进行自动的语义推理；(2) 语义仓储采用灵活和通用的物理数据模型（如RDF图），因此很容易理解并实时采用新本体或新元数据方案，也即接受数据结构的改变[23]。换句话说，语义仓储可以看作是推理引擎和列存储两者的相加。语义仓储是语义网应用的基础，其重要性就如同HTTP服务器之于传统的Web应用。具有代表性的语义仓储引擎包括Sesame、Jena SDB、OWLIM[24]、Virtuoso、AllegroGraph等。

（7）社交语义网（Social Semantic Web）

Web 2.0应用在过去取得了巨大的成功，它的一个重要特征是促进用户间的合作与共享。社交网（Social Web）是一个用于描述具有高社交性、会话性和参与性的一类Web交互的术语[25]，所代表的就是Web 2.0的这一特性。在当前社交网中存在着一个很大局限，即社交站点之间是相互隔离的，犹如海洋中的一个个孤岛，在各自封闭的世界和独立的数据仓中运行[12]。导致这一现象出现的主要原因是：目前大多数社交网应用或社区没有共同的知识和信息交换标准，不支持互操作，因此把用户限制在某一站点，使其无法在不丢失信息、联系和历史的情况下迁移到另一站点[25]。语义网的出现为上述难题的解决带来了契机，为定义灵活、可扩展的信息交换和互操作标准提供了必要工具。

语义社交网是语义网和社交网两者相结合的产物，集成了语义网、社交软件和Web 2.0的技术、测量与方法[26]。语义网和社交网的结合在于两方面：一方面，基于社交本体（如FOAF和SIOC），对社交数据采用统一的数据模型表示，使得在应用之间进行互操作和迁移变得更加容易；另一方面，利用Web 2.0 中的群体智慧可以创建大量的语义网数据，譬如，社交站点用户通过大众分类（folksonomies）已经并正在创建大量词表和语义丰富的标注[27]。因此，在语义社交网中，社交网和语义网不仅能够互补，而且能产生超过两者之和的更大优势。孤立的社交网能够通过语义技术进行互联，而在用户生成内容中蕴含的大量知识又起到了对语义网进行增强的作用。总的来说，社会语义网为自动化程度的提高和信息传播的增强提供了很多可能，诸如，从相

关的社交空间获得相关信息，允许用户跨站点收集其贡献和个人信息，避免用户在多个社交空间重复多次表达同样的信息，将Web作为一个剪贴板在各种合作应用中进行信息交换，提供对内容进行个性化和创建智能用户界面的新方法，利用语义从内容和嵌入的元数据中抽取更多的信息等等，而这些功能在当前的社交软件中是很难实现的[25]。目前，语义社交网已经有了语义维基、语义博客、语义微博、语义社会化书签、语义社交网络等多种应用。

（8）语义Web服务（Semantic Web Services or Semantic Services）

Web服务是一个设计用于支持网络上计算机之间进行交互的软件系统[28]。因为其具有松散耦合、即插即用等优点，便于异构系统间的互连、共享和组合，得到了广泛的关注与使用。但是现有的基于XML的Web服务规范没有提供足够的手段来描述Web服务，要想将Web上各种类型的Web服务加以组合和利用还需要大量手工操作，这极大地限制了Web服务的使用[29]。语义网技术的出现为解决Web服务的这些缺陷提供了方案。语义Web服务（简称语义服务）是传统Web服务和语义网技术相结合的产物，它的出现使服务描述可以带有语义信息，通过一种统一的、计算机可读可理解的方式来和其他语义Web服务进行交互[29]。实现语义Web服务自动匹配的关键步骤是对Web服务进行语义描述，目前主要有3种方法：OWL-S（Semantic Markup for Web Services）[30]，WSMO（Web Service Modeling Ontology）[31]和SAWSDL（Semantic Annotations for WSDL and XML Schema）[32]，这些方法都是利用本体来描述Web服务，然后通过这些带有语义信息的描述来实现服务的自动发现、调用和组合，但是它们各自所用到的本体不仅仅在语义上有所区别，而且在表达能力上也各不相同。语义Web服务仍然是一个不断发展的领域，虽然其在互操作和自动化方面具有很大的优势，但是它目前的能力还很有限，譬如无法提供推理能力来帮助用户决定想要哪个服务，而且支持语义Web服务的工具也不多，因此对语义Web服务的研究还将是一项长期的任务。

3 语义网技术在数字图书馆中的应用

在语义网发展的早期，图书馆领域就对语义网技术给予了关注。虽然本体语言很早就被用来对元数据方案进行规范化描述，产生了如BIBO等元数据本体[33]，但是一直没有大规模的语义网实践出现。随着SKOS语言的产生和关联数据的兴起，许多图书馆和相关机构渐渐意识到了语义网技术在解决数字图书馆的语义互操作、信息集成、智能检索等方面的巨大潜力，大力推广语义网技术在数字图书馆中的应用。

2010年5月28日，W3C成立了图书馆关联数据孵化小组（W3C Library Linked Data Incubator Group），专门探讨如何利用现有的图书馆基石（如元数据模型、元数据方案，以及各种标准和协议）推动图书馆数据在互联网上的关联与全球互操作，并为其他领域所用。2010年8月，IFLA成立了“语义网兴趣小组（Semantic Web Special Interest Group）”，其目标是详细制定语义网相关的标准和准则，增强和传递语义网方面的图书馆专门知识，提高图书馆界对语义网技术与图书馆的相关性和应用潜力的认识。2011年6月，在美国旧金山举办了国际图书馆、档案馆和博物馆关联开放数据峰会（The International Linked Open Data in Libraries, Archives, and Museums Summit，简称LOD-LAM），超过85个团体参加了本次峰会，该峰会的宗旨是“促进关联开放数据公布途径的实用性和可行性”。

（1）语义元数据与语义检索

从某种意义上来说，图书馆是关于元数据的科学，旨在采用书目元数据对文献资源进行描述、组织和检索。同时，语义网的基石RDF数据模型本质上是一种元数据语言，因此将其应用于图书馆书目元数据具有天然的契合性。采用OWL本体语言形式化地描述元数据方案，能够更加精确地定义元数据元素的语义和相互关系。基于元数据本体，能够将不同类型、不同格式的元数据转换为统一的以RDF格式表示的语义元数据。相比普通元数据，语义元数据具有以下优点：（1）为书目元数据提供了一种统一的语义表达形式，能够在原本基于不同元数据标准的元数据间实现语义互操作；（2）可进行语义检索，不仅能够在检索中实现概念匹配，还能够基于元数据本体进行一定程度的语义推理；（3）是实现图书馆关联数据的基础，使同一数字图书馆系统中的不同数据集合，或者不同数字图书馆系统中的数据集合，实现集成与关联。

（2）关联数据与数据集成

图书馆拥有并一直在不断生成大量高质量的结构化数据，譬如书目数据、知识组织数据等，这些数据的

发布、集成、发现是图书馆的核心工作之一，因此图书馆具有成为关联数据实践者和提供者的天然特性，可以利用关联数据发布资源，扩展资源发现服务，进行数据融合，促进异构关联数据的开放与复用，实现数字图书馆系统之间以及与其他信息系统之间的集成等。

图书馆采用关联数据发布最多的是知识组织资源。在LOD云中，具有代表性的词表数据有美国国会图书馆发布的美国国会图书馆标题表LCSH[34]，联合国粮农组织发布的多语言农业词表AGROVOC[35]，OCLC发布的部分杜威十进制分类法DDC[36]，欧盟研究项目TELplus发布的法国国家图书馆主题词表RAMEAU[37]，德国国家经济图书馆发布的经济学词表STW[38]等。这些关联数据化的词表通常采用标准SKOS语言和（或）SKOS标签扩展（SKOS-XL）语言进行表示，采用RDF存储器进行存储，支持基于HTML和RDF浏览器的浏览和通过SPARQL终端进行查询。

图书馆发布的第二大类关联数据是书目数据，代表性项目是瑞典国家图书馆将瑞典联合书目LIBRIS发布为关联数据[39]，是首个实现图书馆书目数据关联数据化的实例。2012年6月，OCLC将WorldCat.org①中的书目元数据发布为关联数据，是目前Web上最大的关联书目数据[40]。此外，RDF Book Mashup提供了一个虚拟的书目数据关联数据化的发布和访问模式[41]。该项目是将来自多个不同Web APIs的书目信息集成到一个语义网界面中，其实质是通过构建一个包装器使得需要用户通过各个不同Web APIs访问的书目信息能够统一地以关联数据的虚拟形式进行访问。

除了词表数据和书目数据，一些科技论文数据也被语义网实践者们以关联数据的形式发布为数据之网的一部分。德国柏林自由大学和汉诺威大学的研究者们采用D2R服务器将著名的计算机科技文献书目数据库DBLP发布为关联数据[42,43]。英国南安普顿大学的研究者们采用RKB Explorer将DBLP发布为关联数据[44]。RKB Explorer是欧盟ReSIST项目开发的一个能够将来自多种异质数据源的数据进行集成并在语义网上统一发布的工具。除了DBLP，PKB Explorer还能够发布来自Citeseer、ACM、NSF和部分IEEE会议的学术资源。由爱尔兰和英国的研究者们共同开发的Semantic Web Dog Food是一个以关联数据形式发布的语义网学术会议资料库[45]。在该项目中，开发者采用OWL语言构建会议本体，并依据会议本体将近200个语义网会议和专题讨论会的元数据采用RDF格式进行表示，最后采用Jena的RDF存储器和Joseki SPARQL服务器存储并发布RDF/XML格式的会议元数据[45]。

关联数据除了是一种结构化数据的网络发布方式，还是一种有效的数据集成手段。通过在不同数据集中建立RDF语义链接，能够实现文献资源与知识组织资源等相关资源的集成，使图书馆内部的各种资源构成一个有机联系的统一整体。此外，还能够实现图书馆内部资源与外部资源（如DBPedia）的无缝连接，使图书馆数据成为整个LOD云的一部分，从而促进图书馆资源的发现与利用[16]。

(3) 社会语义网与社会语义数字图书馆

语义数字图书馆是采用了语义网规范和技术的数字图书馆。相对于普通数字图书馆，语义数字图书馆有两个主要优点：(1) 提供了对信息空间新的搜索范式，如基于本体的搜索/分面搜索；(2) 提供了数据层面的互操作，如集成各种不同来源的元数据，在不同的数字图书馆系统之间建立连接[46]。目前具有代表性的语义数字图书馆项目有JeromeDL、SIMILE和Bricks。

SIMILE是麻省理工学院、万维网联盟（W3C）和HP实验室联合研制的一个数字图书馆项目，其目的是支持和扩展DSpace数字资源管理系统，提高它对分布存储在不同地点和环境中的各类数字资产、概念体系（包括词表和本体等）、元数据之间语义互操作的支持[47]。通过对RDF和语义网技术的应用，SIMILE提供了一系列用于转换、浏览、检索和映射异质元数据的工具，首先针对不同类型的元数据构建元数据本体并在它们之间建立映射关系，然后依据各个本体对相应的元数据类型进行语义化转换，最后通过元数据本体间的映射关系实现不同元数据间的互操作[47]。此外，SMILE还将不同类型的数据（包括数字资产的元数据、OCLC人名规范文档、维基百科中的人物生平信息）进行了关联，可以看作是关联数据的雏形，但是因为没有采用可参引的HTTP URI地址将关联的数据在Web上发布，还不能看作是真正的关联数据[47]。

BRICKS是一个欧盟研究项目，目的是建立分布式文化遗产数字图书馆网络基础结构并实现互操作[48]。Bricks与SMILE实现元数据语义互操作的方法大致相同，都是采用元数据本体间相互映射的方法，但是

① WorldCat.org是OCLC的全球图书馆和其他资料的在线编目联合目录，是世界最大的联机书目数据库。

Bricks是采用OAI-PMH协议[①]在不同数字图书馆系统之间实现互操作，而SIMILE则是在同一数字图书馆系统内部实现不同元数据间的互操作。

Bricks和SIMILE都还仅是语义数字图书馆，侧重于有意义信息的检索，而非给用户提供知识共享的机会，而JeromeDL则是一个社会语义数字图书馆。社会语义数字图书馆是由爱尔兰DERI研究所（Digital Enterprise Research Institute）的Kruk等人首先提出的一个概念，是建立在传统数字图书馆、语义网、社会网络和人机交互研究之上的一个新事物[46]。社会语义数字图书馆系统将传统图书馆中的知识组织系统与语义网和社会网络技术相结合，支持对信息的语义标注和与其他信息系统间的语义互操作，并允许用户参与到信息标注和知识共享中来，使信息发现变得更加容易。JeromeD是波兰Gdansk理工大学图书馆与爱尔兰DERI研究所合作进行的一个社会语义数字图书馆项目，它采用一个共享的书目本体MarcOnt作为中介实现不同类型元数据（即Dublin Core、BibTeX和MARC21）的语义化转换以及它们之间的互操作，从而在同一个数字图书馆内部实现对各种资源的语义搜索和浏览[49]。此外，JeromeDL还采用社会化书签（social bookmarking）技术实现对信息资源的社会化语义标注以及对标注的共享，并在此基础上通过社会化语义协同过滤（social semantic collaborative filtering）技术实现信息推荐[49]。

4 结语

近5年来，语义网在企业界的应用呈井喷式增长，各种面向终端用户的应用系统，如语义门户、语义知识管理系统、语义推荐系统、语义搜索引擎等，如雨后春笋般出现，语义数字图书馆是其中非常重要的一员。随着知识经济的兴起，数字图书馆不仅要作为信息库而存在，更重要的是要成为人类知识的巨大宝库和人类信息交互与共享的平台，能够为用户提供决策支持、专家咨询、智能信息检索、知识管理、信息推荐等多种功能，语义网技术以及其他新信息技术的出现和发展使建立更加智能的数字图书馆系统成为可能，它们在数字图书馆领域的应用具有非常广阔的空间。

随着大数据时代的到来，将语义网技术和大数据相结合是一种新的发展趋势。 一方面，大数据的崛起有利于语义网技术的采用。采用RDF数据模型描述大数据，能够使数据具有机器可读可理解的形式化语义，不仅丰富了大数据的语义，而且使大数据具有更好的互操作性[50]。采用关联数据连接来源多样的大数据，能够基于旧数据产生新数据、发现新知识，从而支持更多的服务，甚至新的商业模式[51]。另一方面，大数据技术也为Web上关联数据的急剧增长保驾护航，提供发布工具、RDF仓储解决方案、并行查询和挖掘的实现手段以及各类管理工具等[51]。可以预见，大数据将为语义网的发展和应用提供更广阔的空间，当关联数据的数据量增长到一定程度，大规模的语义数据在未来一定会引起Web的质变，从而迎来真正的智能Web时代。

参考文献

[1] ANTONIOU G, HARMELEN F. A Semantic Web Primer [M]. 2nd ed. The MIT Press, 2008.

[2] BERNERS-LEE T. Design Issues: What the Semantic Web can represent [EB/OL]. [2014-03-03]. http://www.w3.org/DesignIssues/RDFnot.html.1998.

[3] BERNERS-LEE T, HENDLER J, LASSILA O. The Semantic Web [J]. Scientific American, 2001 [2014-03-03]. http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html.

[4] BERNERS-LEE T. Design Issues: Linked data [EB/OL]. [2014-03-03]. http://www.w3.org/DesignIssues/LinkedData.html.2006.

[5] BIZER C, JENTZSCH A, CYGANIAK R. State of the LOD Cloud (version 0.3) [EB/OL]. [2014-03-03]. http://www4.wiwiss.fu-berlin.de/lodcloud/state/.

[6] HERMAN I, ADIDA B, SPORNY M, et al. RDFa 1.1 Primer [EB/OL]. W3C Working Group Note 22 August 2013. [2014-03-03]. http://www.w3.org/TR/xhtml-rdfa-primer/.

[7] About Microformats [EB/OL]. [2014-03-03].http://microformats.org/about.

[8] HTML Microdata [EB/OL]. [2014-03-03]. http://www.w3.org/TR/microdata/.

[9] What is Schema.org? [EB/OL]. [2014-03-03]. https://schema.org/.

[10] BONTCHEVA K, CUNNINGHAM H. Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation [G]//

[①] OAI-PMH：全称Open Archives Initiative Protocol for Metadata Harvesting，用于收割基于XML的描述性元数据记录，实现不同信息系统间互操作的协议标准。

DOMINGUE J, FENSEL D, HENDLER J. A Handbook of Semantic Web Technologies. Heidelberg, Berlin: Springer, 2011: 77-116.

[11] HEFLIN J, HENDLER J. Dynamic Ontologies on the Web [C]// Proceedings of the 7th National Conference for Artificial Intelligence. Menlo Park, CA: AAAI/MIT Press, 2000: 443-449.

[12] KOGUT P, HOLMES W. AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web [C/OL]// Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the 1st International Conference on Knowledge Capture, 2001 [2014-03-03]. http://km.aifb.kit.edu/ws/semannot2001/positionpapers/AeroDAML3.pdf.

[13] RDFizers [EB/OL]. [2014-03-03]. http://simile.mit.edu/wiki/RDFizers.

[14] D2R Server: Accessing database with SPARQL and as Linked data [EB/OL]. [2014-03-03]. http://d2rq.org/d2r-server.

[15] 夏翠娟,刘炜,赵亮,等.关联数据发布技术及其实现:以Drupal为例[J].中国图书馆学报,2012,38(1):49-57.

[16] 欧石燕,胡珊,张帅.本体与关联数据驱动的图书馆信息资源语义整合方法及其测评[J].图书情报工作,2014,58(2):5-13.

[17] W3C -Standarads-Semantic Web-Inference [EB/OL]. [2014-03-03]. http://www.w3.org/standards/semanticweb/inference.

[18] BOLEY H, HALLMARK G, KIFER M, et al. RIF Core Dialect (second edition) [EB/OL]. W3C Recommendation 5 February 2013 [2014-03-03]. http://www.w3.org/TR/2013/REC-rif-core-20130205/.

[19] JOHN T. What is Semantic Search and how it works with Google search [EB/OL]. [2014-03-03]. http://www.techulator.com/resources/5933-What-Semantic-Search.aspx.

[20] 关于丰富网页摘要和结构化数据[EB/OL]. [2014-03-03]. https://support.google.com/webmasters/answer/99170?hl=zh-Hans.

[21] LOPEZ V, NIKOLOV A, SABOU M, et al. Scaling up Question-Answering to Linked Data [C]// Proceedings of the 17th international conference on Knowledge engineering and management by the masses. Heiderlberg, Berlin: Springer, 2010: 193-210.

[22] Google Knowledge Graph [EB/OL]. [2014-03-03]. http://www.google.com/insidesearch/features/search/knowledge.html.

[23] KIRYAKOV A, DAMOVA M. Storing the Semantic Web: Repositories [G]// DOMINGUE J, FENSEL D, HENDLER J. A Handbook of Semantic Web Technologies. Heidelberg, Berlin: Springer, 2011: 231-297.

[24] KIRYAKOV A, OGNYANOV D, MANOV D. OWLIM- a pragmatic semantic repository for OWL [C]// Proceedings of the 2005 International conference on Web Information Systems Engineering. Heiderlberg, Berlin: Springer, 2005:182-192.

[25] BRESLIN J G, PASSANT A, VRANDEČIĆ D. Social Semantic Web [G]// DOMINGUE J, FENSEL D, HENDLER J. A Handbook of Semantic Web Technologies. Heidelberg, Berlin: Springer, 2011: 467-506.

[26] WELLER K. Knowledge Representation in the Social Semantic Web [M]. Berlin: Walter de Gruyter, 2010.

[27] BRESLIN J G, PASSANT A, DECKER S. The Social Semantic Web [M]. Heiderlberg, Berlin: Springer, 2010.

[28] HAAS H, BROWN A. Web Services Glossary [EB/OL]. W3C Working Group Note 11 February 2004 [2014-03-03]. http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/#webservice.

[29] HEBELER J, FISHER M, BLACE R, et al. Semantic Web Programming [M]. Indianapolis, In: Wiley Publishing, Inc., 2009.

[30] MARTIN D, BURSTEIN M, HOBBS J, et al. OWL-S: Semantic Markup for Web Services [EB/OL]. W3C Member Submission 22 November 2004 [2014-03-03]. http://www.w3.org/Submission/OWL-S/.

[31] BRUIJN J, BUSSLER C, DOMINGUE J, et al. Web Service Modeling Ontology (WSMO) [EB/OL]. W3C Member Submission 3 June 2005 [2014-03-03]. http://www.w3.org/Submission/WSMO/.[32] FARRELL J, LAUSEN H. Semantic Annotations for WSDL and XML Schema [EB/OL]. W3C Recommendation 28 August 2007 [2013-03-03]. http://www.w3.org/TR/sawsdl/.

[33] D'ARCUS B, GIASSON F. The Bibliographic Ontology [EB/OL]. [2014-03-03]. http://bibliontology.com/.

[34] SUMMERS E, ISAAC A, REDDING C, et al. LCSH, SKOS and Linked Data [C]// Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications. Singapore: Dublin Core Metadata Initiative, 2008: 25-33.

[35] CARACCIOLO C, STELLATO A, SACHIT R, et al. Thesaurus Maintenance, Alignment and Publication as Linked Data: The AGROVOC Use Case [C]// Proceedings of the 5th International Conference on Metadata and Semantics Research. Heidelberg: Springer, 2011:489-499.

[36] OCLC. Dewey Summaries as Linked Data [EB/OL]. [2014-03-03]. http://www.oclc.org/dewey/webservices/default.htm.

[37] MEIJ L, ISAAC A, ZINN C. A web-based repository service for vocabularies and alignments in the cultural heritage domain [C]// Proceedings of the 7th European Conference on the Semantic Web: Research and Applications -Volume Part 1. Heidelberg: Springer,

2010: 394-409.

[38] NEUBERT J. Bringing the "thesaurus for economics" on to the web of linked data [C/OL]// Proceedings of the WWW 2009 Workshop on Linked Data on the Web. CEUR-WS.org, 2009 [2014-03-03]. http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf.

[39] MALMSTEN M. Making a Library Catalogue Part of Semantic Web [C]// Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications. Singapore: Dublin Core Metadata Initiative, 2008: 146-152.

[40] WorldCat linked data [EB/OL]. [2014-03-03]. http://www.oclc.org/data.html.

[41] BIZER C, CYGANIAK R, GAUSS T. The RDF book mashup: From web APIs to a web of data [C/OL]// Proceedings of the 3rd Workshop on Scripting for the Semantic Web. CEUR-WS.org, 2007 [2014-03-03]. http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf.

[42] D2R server publishing the DBLP bibliography database [EB/OL]. [2014-03-03]. http://www4.wiwiss.fu-berlin.de/dblp/.

[43] D2R server publishing the DBLP bibliography database, hosted at L3S research center [EB/OL]. [2014-03-03]. http://dblp.l3s.de/d2r/.

[44] GLASER H, MILLARD I, JAFFRI A. RKBExplorer.com: A knowledge driven infrastructure for linked data providers [C]// Proceedings of the 5th European Conference on the Semantic Web: Research and Applications. Heidelberg, Berlin: Springer, 2008: 797-801.

[45] MOLLER K, HEATH T, HANDSCHUH S, et al. Recipes for semantic web dog food - the ESWC and ISWC metadata projects [C]// Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference. Heidelberg, Berlin: Springer, 2007: 802-815.

[46] KRUK S R, MCDANIEL B. Goals of semantic digital libraries [G]// KRUK S R, MCDANIEL B. Semantic Digital Libraries. Heidelberg, Berlin: Springer, 2009: 71-76.

[47] BUTLER M H, GILBERT J, SEABORNE A, et al. Data Conversion, Extraction and Record Linkage Using XML and RDF Tools in Project SIMILE [R]. HP Labs Technical Report HPL-2004-147. Bristol: HP Laboratories, 2004: 2-15.

[48] HASLHOFER B, HECHT R. Metadata Management in a Heterogeneous Digital Library [C]// Proceedings of the eChallenges 2005, Amsterdam: IOS Press, 2005: 1251-1558.

[49] KRUK S R, CYGAN M, CZELLA A, et al. JeromeDL- The social semantic digital library [G]// KRUK S R, MCDANIEL B. Semantic Digital Libraries [M]. Heidelberg: Springer, 2009: 139-150.

[50] 刘炜,夏翠娟,张春景.大数据与关联数据:正在到来的数据技术革命[J].现代图书情报技术,2013(4):2-9.

[51] Fujisu UK. Linked data connecting and exploiting big data [EB/OL]. White Paper: Linked Data. 2012 [2014-03-03]. http://www.fujitsu.com/uk/Images/Linked-data-connecting-andexploiting-big-data-(v1.0).pdf.

作者简介

欧石燕（1971-），女，南京大学信息管理学院教授。E-mail: oushiyan@nju.edu.cn
胡珊（1989-），女，南京大学信息管理学院硕士研究生。

Main Functionalities of Semantic Web and Their Applications in Digital Libraries

Ou Shiyan / School of Information Management, Nanjing University, Nanjing, 210093
Hu Shan / School of Information Management, Nanjing University, Nanjing, 210093

Abstract: Since the birth of the Semantic Web, its developing process is always under continuous changes and adjustments. New Semantic Web standards are frequently recommended, and the functionalities and applications of the Semantic Web are constantly expanded in depth and breadth. This paper first gives a review and analysis of the establishing and developing process of the Semantic Web, then summarizes the main functionalities of the Semantic Web by carrying out a survey on Semantic Web applications, and finally provides an analysis and description on the applications of these functionalities in digital libraries.

Keywords: Semantic web, Linked data, Digital libraries

（收稿日期：2014-03-04）