语义门户的框架研究

田晓迪，宋 文

（中国科学院 文献情报中心，北京 100080）

摘 要：本文概要介绍了语义门户的信息集成、框架结构、功能模块以及相应的站点管理等方面的内容，描述了语义门户的概貌。

关键词：语义门户；信息集成；框架模块；站点管理

中图分类号：TP311　　**文献标识码**：A　　**文章编号**：1007－7634(2007)01－0103－05

Research on the Architecture of Semantic Portal

TIAN Xiao－di, *SONG Wen*

(*Library of Chinese Academy of Sciences*, *BeiJing* 100080, *China*)

Abstract: This paper introduces the architecture, information integration and the core modules of Semantic Portal, as well as the website mannagemant.

Key words: semantic portal; information integration; frame architecture; website management

一般网站都有两个主要的属性：它们要集成不同的信息源以及需要一个适合的网站管理系统。SEAL（Semantic portAL）是一个基于本体的概念模型，它可以同时满足上述两个要求，而本体可以较好的支持复杂的网络信息集成以及网站的管理。那么什么是SEAL，简单来说，SEAL是一个基于本体的，用来管理机构网站和网络门户的框架结构，语义门户的核心概念就是用本体的概念关系组织门户中的资源，为机构用户的信息交流提供管理机制和信息的集成，所以语义门户具有了语义浏览和语义查询能力。为了减少设计和维护的难度，语义门户通过一个本体对现有的数据源进行语义集成以及管理和展示站点。语义门户通过采用本体向用户和/或代理商提供了信息获取机制、构建机制和信息共享机制，所以，语义门户结合了两者的优点，为一个门户提供了信息的语义检索以及对建设和维护这个门户。

1 语义门户的特征

语义门户通过使用语义技术来提供语义检索、浏览和内容集成，语义门户是资源的集合，这些资源通过使用一个内容丰富的领域本体（与之相反的是一个关键词列表平台）来进行标引。门户通过挖掘该领域本体的结构来提供资源的搜索与导航，这样在门户所提供的导航与领域语义之间就可能建立了间接的映射，当领域标引稳定和可重用时，门户可能被改组以适应不同的用户需要。目前，这种间接映射已被开发，如：在线课程项目中，一个包含2000个教育概念的词汇本体被应用到教育资源的注释中，获取门户根据当前UK国家课程需要对这些注释资源进行导航。从用户检索或导航词汇到领

收稿日期：2006－03－15

作者简介：田晓迪（1983－），女，山东人，硕士研究生，从事信息资源的组织与建设研究；宋 文（1962－），女，浙江人，研究馆员，从事信息资源的组织与建设.

域本体之间的映射本身就是一个推理过程——类似于在 TAP 语义检索指示器中，免费的文本检索词与领域本体中的属性和类标签相匹配来支持传统关键词检索的语义扩增。

语义门户具有以下特征：

(1) 多维检索和浏览。使用一个明确的、共享的领域本体能够实现多维的分类与浏览体系。使用标准格式来对本体进行编码便于本体的重用，许多项目已经从本体驱动的门户设计中获益。

(2) 信息结构的演化与扩展。随着时间流逝，需求的变化导致信息模型的扩展。语义网主要从两种方面来促进演化。第一，用户界面与提交工具能从公布的本体中产生。第二，RDF 的半结构化数据表示形式允许数据加到一个新的格式中，这样原始格式和扩展格式可以交替使用。

(3) 结构与意见的区域扩展。许多门户支持受限的注释，如评论和等级级别，语义网允许更加广泛的定制。如，在野生动植物多媒体门户中，很明显许多用户团体喜欢专门的数据导航（基于所描述的物种或行为），这点是集中化的门户难以实现的，而通过使用分散的方法、开发特定的导航结构，以此作为门户数据的一套外部 RDF 注释，这就变成了可能的。

(4) 通过集成分散的资源进行门户的维护，传统信息门户的一个问题是它们经常依赖维护人员的工作，因此如果人员的工作行为消失，那么数据也会随之消失。在语义网中，供应群自己存储数据，门户成为一种汇集服务。仍然需要中心组织（如：提供最初的推动力并确保采用恰当的本体和控制词表）。然而，一旦系统达到临界值会比较容易进行自我维护。

2 语义门户的框架结构及功能

我们知道，获得相关的语义信息将会给用户间的交流带来很大助益，而语义门户可以将其变为事实。可以看到，这些必须基于一个语义的基础，所以我们需要实际的信息技术将信息整合，而本体就具备这个功能，可以规范定义词条支持智能存取。所以，本体成为了语义门户的基础。语义门户最初起源于 Ontobroker，起初是为了网络信息的语义查询和信息共享而构思的，后来逐渐发展为一个有规模的框架结构，用来支持一个门户网站的信息查询和展示。

2.1 本体

我们知道本体作为语义门户的基础，它的主要功能就是可以支撑语义信息的提供和检索，最大的意义就是促进了交流，提高了信息的利用率，节省了时间和精力，所以我们建设本体的时候必须要考虑到它是用来实现人与人之间，人与软件代理之间交流的，而交流必定就会涉及到符号，概念，逻辑等。一般的交流是由一个三角形来定义的，这三个要素分别是符号或词，概念和世界上的事物。这个三角形揭示了词语并不能完全表达出一个概念或事物的本质，一个词语和事物并不是直接相关的，最合适的过程就是词语调用了一个相应的概念，然后这个概念涉及到某个事物，而这个词语可以代表这个事物。所以我们在定义本体模型的时候就要考虑到这些问题，本体是用来促进人机交流，解决或减少人机交流之中存在的矛盾的，一个易懂的逻辑框架——F - Logic 是非常重要的，它可以保证逻辑严格性，另外这样的本体结构还可以让用户和本体建设者阐明可能的错误理解。

2.2 语义门户的体系结构及其核心模块

2.2.1 体系结构

SEAL 的支柱是知识仓库，比如，数据仓库和 ontobroker 系统。后者的作用相当于一个中间运行系统，主要用于在运行环境变得比较复杂时来协调不同的信息源。首先我们要区分一下三种代理的类型：软件代理，机构内部用户和一般用户，他们都通过网络服务器和系统进行交流，这三种类型的代理分别有三种模式和系统进行交互。

(1) 远程应用（如软件代理）会通过网络处理门户中存储的信息，为了这个目的，RDF 发生器通过 web 服务器生成 RDF 代码，带有 RDF 爬行器的软件代理可以收集这些代码，继而可以对网站存储的语义知识进行直接的存取。

(2) 机构用户和一般用户可以访问网站上的信息，SEAL 支持两种形式的信息存取：通过文档的超连接结构浏览门户和查询信息。门户建设者创建了部分的超连接结构，并通过浏览模块的帮助得以扩展。浏览模块利用推理机的推论功能来构造概念上的超链接结构。查询功能通过查询模块得以实现。除此之外，用户可以通过利用预处理模块“语义个性化”来个性化查询界面，或者通过语义相似性（由后加工模块 semantic ranking 完成）对检索结

果进行排序，查询同样可以有效利用 ontobroker 的推理机制。

(3) 只有机构内部的用户才能提供信息，这些信息主要包括个人信息，研究领域的信息，出版物，活动和其他的一些科研信息，而本体会包含他们所提供的任何一种信息的（至少）一个概念。模板模块可以为输入的信息半自动生成合适的 HTML 格式，机构内部的用户就可以根据模板填写相关信息，并且模板模块会将所填信息存储到知识仓库中[2]。

2.2.2 核心模块

Ontobroker: Ontobroker 系统是一个可扩展的、面向对象的数据库系统，它可在主要存储器和关系数据库（通过 JDBC）上运行，它提供了不同语言的编辑器来描述本体、规则和事实。除此之外，它在这个体系结构中还担任着推理器（服务器）的角色。它基于知识库和本体来读取输入的文档，分析发出的查询，然后结合本体，知识库和查询要求最后得出并向用户返回结果。通过已有的事实和背景知识可获得额外信息的可能性大大提高了知识提供者和查询者的生命力，比如，我们可以阐明一个人如果属于中科院的某一个研究所，那么他也就属于中科院，因此就没有必要再去说明它在中科院和某个研究所的成员资格了。

(1) 知识仓库。知识仓库是一个以 F－Logic 格式存储数据的知识库。从推论的角度看它和 ontobroker 的区别可以忽略，但是从系统的维护角度看，它和本体定义之间的区别以及它的示例就非常有用了。知识库是围绕一个关系数据库来组织的，这个数据库的事实和概念以具体化的格式存储，它以一阶目标来说明关系和概念，因此在面对本体的某些变化时非常灵活。

(2) 浏览模块。除了针对领域知识的树型等级超链接结构外，基于本体的概念关系，浏览模块还可以支持复杂的基于图形的语义链接。这种超级链接都是根据语义相关的概念关系从一个页面链接到另一个页面，如 member of 或者 has part，也或者是根据相关的属性关系，如 has Name。因此，知识库中的实例可以自动的和其他相关的实例链接。比如，在个人信息页面上会有相应的超链接连接到其所在的科研组织，科研领域和工程的页面。

(3) 查询模块。查询模块提供了一个易用的查询界面，它的查询能力基于 ontobroker 的 F－Logic 查询界面。门户建设者为特殊的查询需要提供了页面模型，如查询工程或人。为了这个目的，选择列表会限制提供给用户的查询权限，选择列表是由本体中的知识和知识库编辑而成的。比如说，查询界面允许用户依照科研组织来查询科研人员，科研组织的列表由 F－Logic 查询动态组成然后以一个下拉列表的形式展示给用户。

(4) 模板模块。为了便于机构用户提供信息，模板模块为每个用户可能用到的例示概念提供了 HTML 格式。比如说，由人的概念定义而生成的一个输入模板，这个模板中的数据接着就会被浏览模块使用来生成相应的个人信息页面。

为了减少需要输入的信息量，门户建设者特别指定了哪些属性和关系可以从其他模板中获得，比如，门户建设者会特别说明工程的成员资格是在工程模板中定义的。如果一个工程的合作者输入了某个人是工程的参与者的信息，那么在利用相应反转关联的 F－Logic 规则生成个人网页时就会使用到这个信息。所以，在个人模板中就没有必要输入这个信息了[1]。

2.3 语义门户的技术框架

SEAL 的技术构建来源于 KAON（Karlsruhe Semantic Web and Ontology Infrastructure）的体系结构，这个框架的组成部分实现了上述部分中各项功能，KAON 的框架结构的组成部分可以分为三个层次。

(1) 数据和远程服务层。代表了可选择的外部服务，可以用于上部层，如，针对推论和查询的推理服务，或对 Edutella Peer－To－Peer 网的连接器，和可以选择的数据仓库中的数据存储机制。

(2) 中间层。提供了一个高水平的 API 来利用本体和关联的数据，并且对所有的客户屏蔽了实际的存储方式和交流方式。因此，客户不能分辨出他们是在本地文件系统上（由 RDF API 提供）工作还是在一个多用户服务器上（将数据存储在一个关系数据库中）工作。中间层同样为 QEL 提供了界面，在 Edutella 网络内使用的查询语言不仅只在 Peer－To－Peer 网内用来交流疑难，同样也可以查询数据仓库。

(3) 应用和服务层。将使用下部层服务的应用聚合在了一起。现在这只是一个方面，单机的应用建设通过使用 Ont－O－Mat 应用框架或门户的建设通过使用 KAON 的门户软件，为第三部分中所讨论的功能提供帮助。Ont－O－Mat 应用是作为插件程序而建设的，它是 Ont－O－Mat 应用框架的一部

分[7]。

最后，KAON 的核心就是领域本体本身，它是由 RDF Schema（在语义网中用来表示本体的数据模型）形式来展示的，它提供了基本的类和属性层级以及类和事物之间的关系。从历史观点来说，SEAL 平衡了 RDF Schema 模型和 F－Logic 之间的映射，提供了一个逻辑公理形式的观点和查询机制，这使得我们可以依赖于 OntoBroker 的推理服务。

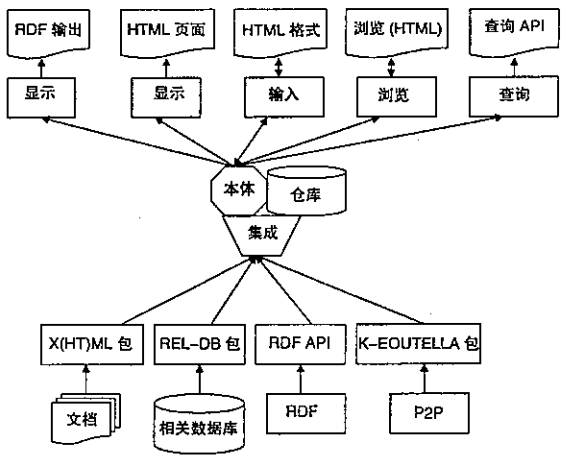

图 1 语义门户的概念框架

3 语义门户的信息集成

SEAL 概念上的体系结构在这个方面给出了一个方案，在图 1 中我们可以看到这个体系结构，图中上半部分的结构主要侧重于站点的管理，而下半部分则主要侧重于信息的集成，而 SEAL 利用中间部分的本体将两者结合在了一起。SEAL 有两大作用，一是信息集成，为语义网提供信息，另一个是站点的管理。

要建设一个具备精确数据的站点，其中一个最大的挑战就是多种类的网络信息集成，近几年来我们在管理多来源语义的信息上有了很大的进步。针对不同来源信息之间的语义融合出现了一个较完善并得到一致公认的概念模型－本体，它主要有三个层次结构。

（1）不同种类的数据源（如：数据库，xml，还有在 HTML 中发现的数据）

（2）将这些不同来源的信息集合的一个数据模型（如：RDF）

（3）集成模型（动态情况下的数据协调者），可将不同语义的数据源融合集成[4]。

因此，集成/协调任务的复杂性就大大减少了。

对于集成信息，一般的目标就是将所有不同的信息源集合在一个共同的数据模型中，比如 RDF。另外，本体也是不同种类输入信息的一个语义模型。就像在图 1 中所示的概念体系，不同种类的数据源以这样的方式输入：首先，网络在很大程度上是由静态的半结构化的 HTML 网页构成的，包括表格，列表等。基于本体的 HTML wrapper 通过半自动注释建立。因此，基于一系列预先确定的手工注解的 HTML 页面，新的 HTML 页面的结构就会被分析，和已注释的 HTML 页面及从这些页面中提取的相关信息进行比较。HTML wrapper 现在已经可以处理 XML 类型的文档。第二，XML wrapper 的概念就是这些 xml 文档都涉及到一个从本体中生成的 DTD，因此就可以得到从 xml 到数据模型自动映射。第三，精确数据的利用尤其依赖关系数据库。一个关系数据库的 wrapper 是指关系数据库中的 shema 和本体映射，这样就形成了关系数据库自动生成的 RDF 语句的语义基础。第四，在一个理想的条件下，内容提供者注册并同意依据共享本体利用基于 RDF 的元数据来描述和丰富他们的内容。在这种情况下，通过自动执行信息集成的过程就可以完成集成信息的任务。如果内容提供者还没有注册，但是在他们的网页上提供了基于 RDF 的元数据，可以使用着眼于本体的元数据发现和爬行技术来查探相关的 RDF 语句[3]。

4 语义门户的站点管理

语义门户的一个难点是集成不同种类的数据源，这些资源可能属于不同的机构成员或外部团体并满足着不同的需求，因此这些资源在结构和设计上都各不相同。一般的机构站点都需要对不同水平的信息进行整合，当信息的整合能够满足一个本体所提供的整体结构要求时，对站点的建设和维护就非常重要了[5]，下面将从几个角度来介绍相关的工具和思想。

（1）从表达的角度：首先，语义门户不仅满足了软件代理的需要，并且产生了机器可以理解的 RDF，这对于语义网来说是一个贡献。其次，语义门户为代理商提供了 HTML 页面，尤其是对数据仓库的内容查询通过对内容的选择规定了表达的角度，同样也会用到对图解的查询，如对表格的标题的标注。

（2）从输入的角度：维护一个门户并保持它的

内容始终是活跃的，那么必须对门户进行经常性的更新，不仅仅要对各种来源的信息进行集成，同样也要有专家的参与。输入的观点是由对图解的查询来规定的，也就是说，对本体自身的查询，这和在本体外产生表格来支持知识获取任务是相似的。表格依照本体不断的搜集数据，最终会被存储在数据仓库中。

(3) 从浏览的角度：要实现对数据仓库进行浏览，语义门户利用对图表和内容的联合查询自动产生浏览的结构。首先，通过使用不同类型的层级(如：is－a，part－of)关系来实现顶层的浏览结构，以此为用户提供对本体不同的浏览角度。第二，本体的每个展示部分都在数据仓库中有相应的内容。因此即使是对门户并不熟悉的用户也可以自如的浏览图表和相应的内容[6]。

5 结 语

当今社会是一个网络信息社会，网络上的各种信息资源纷繁复杂，用户很难寻找到真正需要的系统的信息，资源的利用率不高，在这种情况下，信息资源的整合服务成为一个大趋势，无论是科研教育机构或是企业政府等都在意念和实际行动上表现出这种需要。但是，它的实现务必要有相应的先进理念来支撑，结合实际需要，利用相应技术来实现。WWW 的创始人提姆·柏纳李（Tim Berners－Lee）在 2001 年提出“语义网”，指出语义网是扩展当前的 WWW，使得网络中所有的信息都具有语义，便于人和计算机之间的交互与合作，而语义信息的传递也有利于信息提供者和需求者之间的高质量信息交流。

所以，作为语义网重要的组成部分，语义门户的作用愈显重要，区别于以往对网站资源简单聚集的门户，语义门户是以一个本体为支撑，用户之间、用户与互联网上的其他用户可以通过门户交流，用户可以在门户上发布和下载信息，也可以浏览和查询信息，交流是双向的，可以为领域内的研究机构和研究人员提供交流的平台，提供该研究领域的各种资源和信息，供研究机构和研究人员发表研究成果和相关信息，同时也适应 e－science 科研环境的发展趋势，有利于变革科研人员信息交流方式。

参考文献

1 S.S.;3Alexander Maedche,1;2Steffen Staab,1Nenad Stojanovic,1;2;3Rudi Studer, and 1York Sure. Semantic Portal－The seal approach[EB/OL]. http://scholargoogle.com,2005－09－08.

2 S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, H.－P. Schnurr, R. Studer, and Y. Sure. Semantic community web portals. Proc. of WWW9[J]. Computer Networks,2000,33(1－6):473－491.

3 Alexander Maedche,Steffen Staab,1Rudi Studer,York Sure,1Raphael Volz. SEAL－Tying Up Information Integration and Web Site Management by Ontologies[EB/OL]. http://scholargoogle.com,2005－09－08.

4 Jens Hartmann,York Sure,Raphael Volz,Rudi Studer. Extended OntoWeb. org Portal[EB/OL]. http://ontoweb. aifb. uni－karlsruhe. de/Members/hartmann/OntoWeb _ Del _ 6－4. pdf, 2005－06－20.

5 Anna V. Zhdanova. The People's Portal: Ontology Management on Community Portals[EB/OL]. http://scholargoogle.com,2005－10－08.

6 Steffen Staab,Rudi Studer,York Sure,Raphael Volz. SEAL－a SEmantic portAL with content management functionality[EB/OL]. http://www. aifb. uni－karlsruhe. de/Publikationen,2005－10－08.

7 E. Bozsak,Marc Ehrig,Siegfried Handschuh,Andreas Hotho,Alexander Maedche, Boris Motik, Daniel Oberle, Christoph Schmitz, Steffen Staab, Ljiljana Stojanovic, Nenad Stojanovic, Rudi Studer, Gerd Stumme, York Sure, Julien Tane, Raphael Volz,Valentin Zacharias. KAON－Towards a large scale Semantic Web[EB/OL]. http://www. aifb. uni－karlsruhe. de/Publikationen,2005－10－09.

8 宋 炜,张 铭.语义网简明教程[M].北京:高等教育出版社,2004.40－98.

9 王 洁,刘 南,刘仁义.基于本体的知识门户[J].计算机应用研究,2003,(5):40－42.

10 李 景,钱 平.构建基于 Ontology 的知识门户[J].现代图书情报技术,2004,(2):18－21.

(责任编辑:刘凤勤)