基于OWL本体论映射的数据库网格语义模式集成研究*)

裘 君 吴朝晖 徐 昭
(浙江大学计算机学院网格计算实验室 杭州 310027)

摘 要 本文提出了一种在数据库网格中OWL本体论映射机制如何用于基于语义的数据库模式集成。方法是首先把关系模式转化为RDF/OWL语义描述以完成局部映射,再通过把局部数据语义与全局共享本体建立联系来完成全局映射。本质是把异构数据库模式的语义通过本体显性地表达出来,并在语义Web层完成模式的集成。特点是实现了在统一的语义层次上进行共享与查询,同时采用了局部映射与全局映射松耦合的构架,其特有的分层结构使得在跨库/单库环境中进行语义查询变得更加灵活。

关键词 数据库网格,DartGrid,本体论映射,网格服务

A Framework of Semantic Schema Integration Based on OWL Ontology Mapping in Dart Database Grid

QIU Jun WU Zhao-Hui XU Zhao
(Grid Computing Lab, College of Computer Science, Zhejiang University, Hangzhou 310027)

Abstract Brings forward a framework on how to integrate database schemas using OWL ontology mapping which is based on semantic in Database Grid. The framework adopts a two-stages mapping method as following, the first step is to implement local mapping, or converting relational schema to RDF/OWL semantic schema; the second step is to implement global mapping, or bridging relationship between local data semantic and global sharable ontology. Essentially, heterogeneous database schema is represented clearly through ontology and integrated on the semantic Web layer. At the same time, it implements sharing and query based on semantic layer and keep both local and global ontology mapping loose coupling which makes semantic query more flexible between across-database and single database.

Keywords Database grid, DartGrid, Ontology mapping, Grid service

1 背景

网格计算代表了新一代面向Internet的分布式计算技术趋势,已经越来越受到人们的关注。数据管理是网格技术平台的一个重要组成部分,而面向数据库资源的网格数据管理问题也已成为其中的一个焦点。数据库网格是一个面向网格的数据库资源管理平台,旨在为实现现有大量位于Internet后台的数据库资源的共享提供一个可接入的环境,为网格应用提供基础结构级的数据库资源访问、发现、整合等一系列问题的通用解决方案。DartGrid[1]是浙江大学网格计算实验室开发的一个基于语义的虚拟组织模型,提供了一整套在数据库网格环境下构建支持大规模数据库资源的共享和管理的解决方案,支持数据库资源的动态化的语义注册、分布式的语义查询与知识级的语义浏览。

DartGrid的最终目标是实现异构数据库在统一语义层次上的知识共享与查询,因此其关键是实现各个数据库模式最终在语义表示层上的统一,本文所要提出的基于OWL本体论映射就能很好地解决这样的问题。方法是将关系模式转化为RDF/OWL语义模式以完成局部本体论映射,并在此基础上通过将各个局部数据语义模式映射至上层全局本体论层,同时构建各个本体论间的关联,从而实现共享本体论以完成本体论映射。本质是将异构数据库模式的语义通过本体论显性地表达出来,并在语义Web层完成模式的集成。其特点是实现了在统一的语义层次上进行共享与查询,同时采用了局部映射与全局映射松耦合的构架,特有的分层结构使得在跨库/单库环境中进行语义查询变得更加灵活。

DartGrid数据库网格本体论映射,简称DGOM,是一种具有良好弹性、延伸性并具有独立性的本体论映射模块,它能满足中医药科技信息数据库群数据资源智能共享和协同共建,适应以中医药TCM[2]信息网格应用为背景的DartGrid飞梭信息网格知识领域中的所有描述,其应用范围十分广泛。

2 体系结构

2.1 服务栈

DGOM是由一系列独立且相互协同工作的服务接口组件构成,其中核心服务包括:

- 本体论服务(Ontology Service):在虚拟组织VO中提供元信息服务;
- 数据库服务(Grid Database Service):提供基于关系型的数据服务,将数据库中的源关系模式自动转成以RDF/OWL为形式的语义模式;
- 语义注册服务(Semantic Registration Service):提供本

*)本文受如下项目资助:中国863研究项目(编号:2001AA113142);中国科技部基础技术和研究项目—基于网格的中医药动态信息资源管理和知识服务子项目。裘 君 硕士生,研究方向:网格计算;吴朝晖 博士生导师,研究方向:分布式人工智能、网格计算、生物认证与嵌入式系统;徐 昭 硕士生,研究方向:网格计算。

体论向底层数据的映射功能和基于本体论的资源分类功能；

·语义查询服务(Semantic Query Service)：接受语义查询请求，并进行分布式查询规划和查询分解与分配；

另外还包括其他一些辅助服务，即监控服务(Monitor Service)、代理服务(Broker Service)、本体论验证服务(Ontology Verify Service)、本体论融合服务(Ontology Fusion Service)等。

这些服务构成了特有的DGOM服务栈，如图1所示。

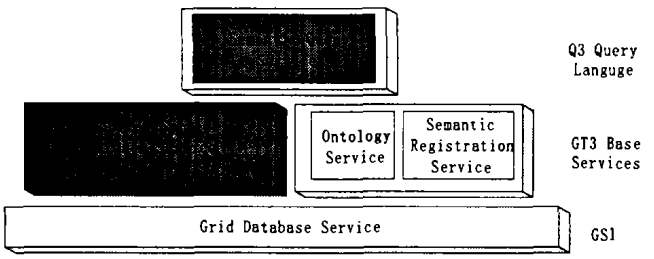

图1 DGOM服务栈

在这个堆栈中的最底层是数据库服务层，是数据库网格服务中的最基础服务，它封装了数据库元数据到以RDF/OWL形式的语义映射。在中间层服务是本体论服务、语义注册服务以及一些辅助服务的融合，定义并注册了为最上层所服务的本体论，包含了从公共路径或者语义规则来获得验证过程的信息。当一系列底层服务处理完成以后，这个栈就进入了最上层的服务处理，即为语义查询服务，它通过对分布式查询的规划与查询分解，并将各个分解后的查询请求通过进行调用中间层的本体论服务得到各个查询的服务句柄，最后提交给数据库服务，对相应的数据库进行查询，因此语义查询服务将各个服务贯穿起来，形成了特有的本体论映射服务流。

2.2 映射层次总体构架

在DGOM服务栈的实现过程中构建了本体论映射的总体构架，如图2所示。这一构架的实现采用了Globus toolkit3.0为网格开发平台，局部、全局本体论以RDF/OWL本体论语言表示，以JDBC和Jena2.0为映射开发工具，并使用了特有的Q3[1]语义查询语言。

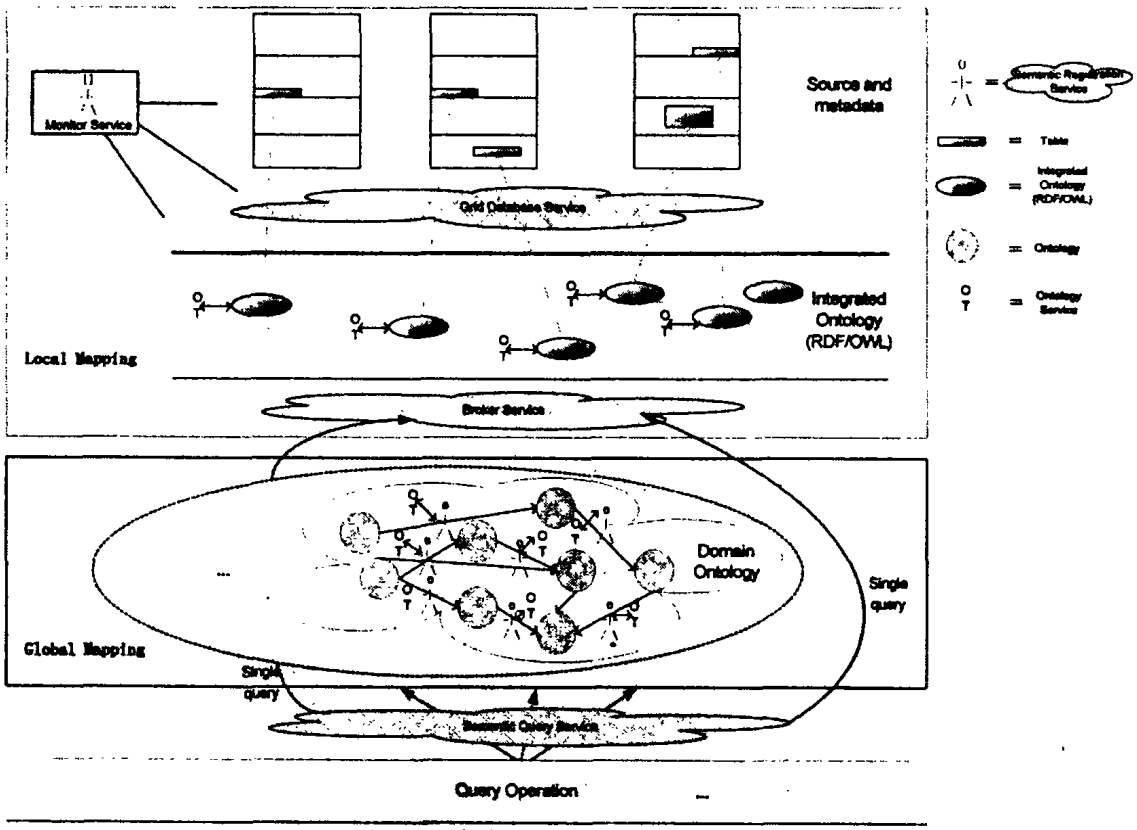

图2 本体论映射

这种层次架构一方面在跨异构数据库进行语义查询时，以一种统一的语义视图来访问；另一方面，由于DGOM具有松耦合的特性，在对单数据库进行语义查询时即向局部映射生成的局部本体论提交Q3语义查询。

DartGrid中的本体论映射可分为以下若干阶段。

2.2.1 *数据采集* 在进行本体论映射之前，根据提出的一系列解决数据库模式映射转换至语义模式的转换规则原型[3]，首先提取各个数据库的元数据，包括数据模式，即获取数据库、表、字段及表之间的关系(表间映射关系如1∶1，1∶n，n∶1，库、表以及字段间的继承关系)和存在的概念模型，将所有的源数据资源进行分类，获得数据库模式的元数据信息，协助本体论的构建。

2.2.2 *局部本体论映射* 在元数据采集之后，通过HP Jena API开发包在各个分布式数据库与语义模式之间自动创建语义映射，将各个数据库中采集后的元数据解析成以RDF/OWL为形式的语义模式，最终使得源数据库模式以语义的形式表现出来，映射对应关系如图3所示，其中{ns}为域命名空间，假设{ns}=http://zju.edu.cn，{dn}为数据库名，{tn}为表名，{fn}为表中字段名，{PKn}为主键名和{FKn}为外键名。其映射算法参见3.2节映射算法中的定义1所示。

局部映射是基于多服务接口的，其中首先数据库服务(Grid Database Service)负责将数据库中的源关系模式自动转成以RDF/OWL为形式的语义模式；同时监控服务(Monitor Service)，用来监控数据库关系模式中是否有更新的数据库结构，若有，则通知数据库服务也相应地进行更新；随后通过调用语义注册服务(Semantic Registration Service)负责提

供本体论向底层数据的映射功能和基于本体论的资源分类及其注册由数据库服务转化成的语义模式功能；本体论服务(Ontology Service)指的是通过调用语义注册后在 VO 中为上层的共享与查询服务提供元数据访问接口；代理服务(Broker Service)，向上层负责收集本地的语义模式，作为统一至全局本体论的中间代理，同时向下层路由至各个以 RDF/OWL 表示的语义模式。

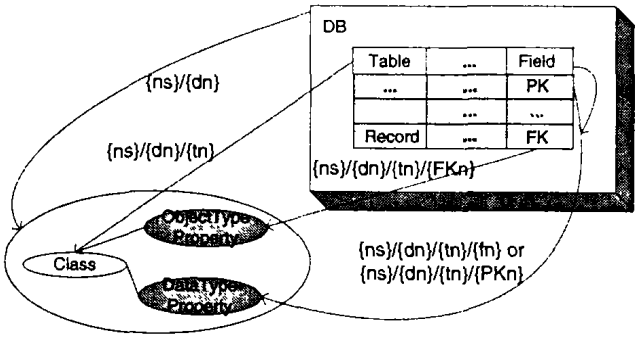

图 3　局部映射对应关系图

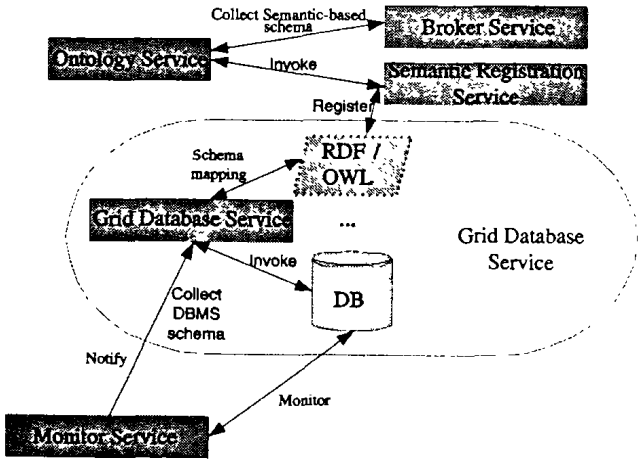

图 4　局部本体论映射

另一方面，局部映射模块提供单独为在单个元数据上进行操作的访问接口，而屏蔽中间层的全局本体论层，即将生成的语义模式通过调用语义注册服务注册至本地的语义库中，并调用本体论服务，从而使得在 VO 中为直接共享与查询单个元数据提供访问接口。示意图见图 4 所示。

通过上述的局部映射我们将不同的 TCM 概念分类成八类父类，每一父类包含若干子类，这种继承关系就构成了一棵类层次树。

2.2.3　全局本体论映射　将局部映射过程中已经生成的语义模式映射至上一层即统一的领域全局本体论层，作为中心本体论，不但作为本地语义模式的补充，而且作为信息的集中式视图，通过定义的映射构成了信息的共享语义，并在语义注册服务的基础上最终使得在语义层上将原本独立的数据源得到了统一与共享。具体映射算法参见 3.2 节映射算法中的定义 2-4 所示。

同样全局本体论映射也是基于多服务接口的，其中语义注册服务(Semantic Registration Service) 负责提供本体论向底层数据的映射功能和基于本体论的资源分类功能，并同时负责在发现异构本体论之间相关联的映射关系，对本体论间的关系进行一定的关联或调整，包括本体论验证、本体论融合、本体论对齐等，对关联后的本体论进行语义注册，从而为上层语义查询提供了统一访问的接口，这里的本体论验证服务(Ontology Verify Service)指的是验证本体论是否符合信息网格特定领域本体论的定义规范；本体论融合服务(Ontology Fusion Service)指的是允许将两个相关本体进行语义层次上的合并，同时也可以视作本体论标准可扩充的一个标志。最后通过调用本体论服务(Ontology Service)将重新构造的本体论层次结构发布。因此在跨多库环境中调用语义查询服务(Semantic Query Service)将优化分解后的查询分发至全局本体论层即可实现对异构数据库的语义查询操作。示意图见图 5 所示。

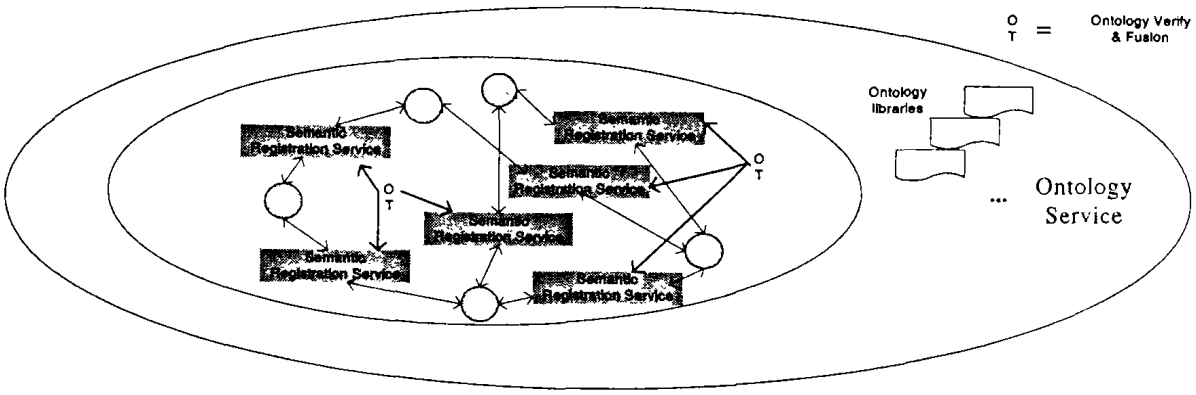

图 5　全局本体论映射

3　映射实现

3.1　映射维度

在 DGOM 本体论映射过程中定义了一种适用于本体论映射的三维原语 D-R-E，具体来说：

· D 即 Discover，发现维，表示手工、自动或者半自动发现已定义的本体论间的关系。

I．设定任务：发现本体论之间相关联的概念或属性，以及它们之间的关系；

I．局部自动映射：由于手工进行模式匹配是一件耗时、易错且费力的过程，加上网络中的数据资源的快速增长和电子商务集成度的提高，导致了本体论映射数目增加，因此有必要对部分局部的本体论定义语义规则，从而在一定程度上促使自动映射的实现。

· R 即 Representation，表示维，由 3.2 节映射算法中提出的本体论映射算法语言来表示本体论之间的关系。这里涉

及到两方面技术，其一为本体论验证，它检查已有的本体论是否符合信息网格特定领域本体论的定义规范；其二即为本体论融合，若两个原本单独的本体论进行语义匹配算法后若在语义层次上表示相似，则将它们合并。

·E 即 Execution，执行维，表示将源本体论的实例转换成目标本体论的实例，通过元数据采集、局部映射和全局映射三个阶段来实现本体论映射。

3.2 映射算法

由于本体论映射过程中对于类、类属性之间的各种映射关系，因此我们有必要提出如下若干基本的映射算法。

定义 1 通用语义映射对应关系通常分为两步，其一为设定对应实体的全局唯一标识 URI，其二通过将关系模式映射至语义模式的类与属性关系，具体来说包括，

设 $URI_{ni}=(http://zju.edu.cn/)$，则

1. $M_{urn\_d_i}\{DB_1,DB_2,\cdots,DB_n\}=URI_{d_i}$，则 $URI_{d_n}=URI_{ni}+URI_{d_i}$；

2. $M_{urn\_n}\{Table_1,Table_2,\cdots,Table_n\}=URI_{tn}$，则 $URI_{tn}=URI_{d_n}+URI_{tn}$；

3. $M_{urn\_f_i}\{Field_1,Field_2,\cdots,Field_n\}=URI_{f_i}$，则 $URI_{f_n}=URI_{tn}+URI_{f_i}$；

4. $M_{urn\_PK_i}\{PK_1,PK_2,\cdots,PK_n\}=URI_{PK_i}$，则 $URI_{PK_n}=URI_{tn}+URI_{PK_i}$；

5. $M_{urn\_FK_i}\{FK_1,FK_2,\cdots,FK_n\}=URI_{FK_i}$，则 $URI_{FK_n}=URI_{tn}+URI_{FK_i}$；

6. $M_{ci}(Talbe_1,Table_2,\cdots,Table_n)=Class_i$；

7. $M_{pij}(Field_{i1},Field_{i2},\cdots,Field_{in})=Property_j$；

8. $M_i=\langle M_{ci},M_{pi1},M_{pi2},\cdots,M_{pin}\rangle$，$M_i$ 即为语义映射；

9. 若 M_i 为语义映射，则在 $Table_i$ 中的任一条记录 $Record_i$ 被直接映射至属于 $Class_i$ 的一个实例 $Instance_i$。

定义 1 建立了源数据库与语义模式之间自动映射的对应关系，并通过语义注册服务将源数据以本体论的形式加入 VO 中，使在语义视图中进行语义查询。

定义 2 假设五维组 $L=(I,T,I_r,P_r,R_r)$，其中

I 是实例集合；

T 是类型集合；

I_r 是实例间的分类关系，即为集合 I 和 T 之间的二维关系；

P_r 是属性的分类关系，在一定程度上反映了层次间的继承关系；

R_r 是结果关系，即类型集合中的二维关系。

其中有两类局部逻辑尤为重要，其一是三维组 (I,T,I_r)，即为实例局部逻辑，二维关系 I_r 决定了实例集 I 对应类型集 T 的分类。例如，xI_ra 表示 $x\in I$ 是由类型 $a\in T$ 来分类的。其二是二维组 (T,R_r)，即类型局部逻辑，它由一组序列 $\{\Gamma,\Delta\}$ 组成，假设 $\Gamma,\Delta\subseteq T$。Γ 是关联集，Δ 是非关联集，因此假如 x 是 Γ 中的任一类型，对于实例 $x\in I$ 满足序列 $\{\Gamma,\Delta\}$，那么 x 就是 Δ 中的某一类型，同时 $\Gamma R_r\Delta$，序列 $\{\Gamma,\Delta\}$ 即为局部逻辑的约束。

P_r 即为 OWL 中的属性类型关系，其中包括传递属性关系 (transitive property)，即属性标识 P，$\exists\ P(x,y)$ 且 $P(y,z)\Rightarrow P(x,z)$；对称属性 (symmetrical property)，$\exists\ P(x,y)\Rightarrow P(y,x)$；等价属性 (equivalent property)，$P(x)=P(y)$。

定义 2 为实现构建本体论映射提供了理论依据。

定义 3 本体论六维组 $O=(C,R,\leqslant,\perp,|,\sigma)$，其中

C 是概念有限集合；

R 是关系有限集合；

$\leqslant$ 是 C 上的偏序关系；

$\perp$ 是 C 上的对称关系，是非关联的；

$|$ 是 C 上的对称关系，且是可覆盖的；

σ 为关系标识的函数，即 $\sigma(R)=\langle C_1,C_2,\cdots,C_n\rangle$，$C_i\in C$，$i\in 1,\cdots,n$。

我们得出一个结论，即：$\exists\ A,B\in\tilde{O}$，$\tilde{O}$ 为本体论，$\Rightarrow P(A\cap B)/P(A\cup B)=P(A,B)/\{P(A,B)+P(A,B^*)+P(A^*,B)\}$，对于 $A\subseteq B$，若 B 出现的频率越高且 $P(A|B)$ 的值也越高，从而相似值 $MSP(A,B)$ 就越高，其中 $MSP=$ Most Specific Parent 最相似的父节点，将出现频率高的 B 设为层次的父节点，类似地依次向下扩展。这个结论用于构造本体论映射的层次结构，使得层次具有复用性，上层本体论为下层本体论提供良好的抽象复用接口。

定义 3 提出了构建本体论映射的六元素，是概念集在一系列关系关联的基础上所形成的，为本体论融合提供了推理基础。

定义 4 本体论六维组 $\tilde{O}=(C,R,\leqslant,\perp,|,\sigma)$，$O=(C,R,\leqslant,\perp,|,\sigma)$，其中 $C=(X,C,P_c)$，$R=(X^+,R,P_r)$，如果对于所有的 $x_1,x_2,\cdots,x_n\in x,c,d\in C$，并且 $\sigma(x)=\{c_1,\cdots,c_n\}$，那么

1)若 $x\ P_cC\ c$ 且 $c\leqslant d$，那么 $x\ P_cC\ d$；

2)若 $x\ P_cC\ c$ 且 $c\perp d$，那么 $x\ \tilde{P}_cC\ d$，其中 $\tilde{P}_cC$ 表示 P_cC 符号的逆向；

3)若 $c|d$，那么 $x\ P_cC\ c$ 或者 $x\ P_cC\ d$；

4)若 $\{x_1,x_2,\cdots,x_n\}P_r\ R\ r$ 那么 $x_i\ P_r$ 方 Rc_i，其中 $i\in 1,\cdots n$。

定义 4 为进行本体论验证时提供了一定的验证规则。

总结与展望 本文提出了一种适应于在数据库网格中用于基于语义的数据库模式集成的 OWL 本体论映射工具，在映射过程中采用局部本体论映射与全局本体论映射相结合，同时定义了一组本体论映射的映射维原语与映射算法，使得本体论映射在单库或跨库的平台上都能有很好的扩展，并适应以中医药信息网格应用为背景的 DartGrid 飞梭信息网格知识领域中的所有描述。今后我们将致力于本体论映射推理技术的研究，使规则推理、案例推理等能够在 DartGrid 本体论映射中得到实际的应用；设计和实现 DartGrid 各个节点之间通用的、抽象的本体论映射各服务间的通信协议，也是我们将来的一个工作重点。

参考文献

1 Huang Chang, Wu Zhaohui, Wu Xiaojun, Zheng Guozhou. Dart: A Framework for Database Resource Access and Discovery. In: Proc. Intl. Workshop on Grid and Cooperative Computing (GCC 2003), to be published in Lecture Notes in Computer Science, Springer. Shanghai, China, Dec. 2003

2 Zhou Xuezhong, Wu Zhaohui, Yin Aining, et al. Ontology Development for Unified Traditional Chinese Medical Language System. Special issue "AIM in China" of the International journal of Artificial Intelligence in Medicine, mid-2004 (in press)

3 Chen Huajun, Wu Zhaohui, Zheng Guozhou, Mao Yuxing, Huang Chang, et al. Semantic Query Processing in Dart Database Grid

4 Schorlemmer M. Channel Theory, 2003

5 Bowers S, Delcambre L. Knowledge Transformation for the Semantic Web, chapter On Model Conformance for Flexible Transformation over Data Models, IOS Press, 2003

6 Jakoniene V. Ontology Integration, 2003

7 de Bruijn J. Semantic information integration in the cog project. COG white paper. Available at : http://www.cogproject.org/publications/sii-wp.pdf, 2003

8 de Vergara J E L, Villagra V A, et al. Semantic Management: Application of Ontologies for Integration of Management Information Models. In: Proc. of the Eighth IFIP/IEEE Intl. Symposium on Integrated Network Management (IM'2003), 2003

9 www.w3.org