Вісник Харківського національного університету
Серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»
№ 733, 2006, с.47-57

УДК 004.822

Определение ограничений к системе интеграции информации с использованием онтологий и пример реального приложения

В.Н. Владимиров
Запорожский национальный университет, Украина

In this article we consider existing ontology-based approaches to the integration of autonomous, distributed and heterogeneous information sources. We identify similar pitfalls for this domain and consider existing solutions for them. Basing on these solutions we propose possible system implementation restrictions to make possible development of framework of intelligent ontology-driven information retrieval from distributed, heterogeneous, legally and physically autonomous IR in the frame of the organizational network of the National Higher Education System.

1. Введение в проблематику и постановка задачи

На данный момент в различных сферах промышленности и хозяйства широко используются компьютерные информационных систем. Методы сбора, обработки информации в них являются хорошо разработанными. Современные задачи требуют полного доступа к информации хранящейся в этих информационных системах, являющихся распределенными и несогласованными между собой. Поэтому разработка способов получения информации от распределенных, гетерогенных, физически автономных информационных ресурсов (ИР) является важным направлением научных исследований.

Эти исследования относятся к сфере распределенного интеллектуального извлечения информации, или в более широком смысле – интеллектуальной интеграции информации. В этой сфере последние 10 лет ведутся интенсивные исследования. Например, в Information Society Technologies Key Action Line of the EU FP6 и подобных ей национальных и международных сетях. Другими примерами исследовательских проектов разрабатывающих формальные, алгоритмические, архитектурные инфраструктуры, прототипы программного обеспечения для распределенного интеллектуального извлечения информации из распределенных, разнородных ИР и интеллектуальной интеграции информации являются: BUSTER [1], DOME ([2], [3]), InfoSleuth [4], KRAFT [5], MOMIS [6], OBSERVER [7], Ontobroker [8], PICSEL [30], SIMS [9], TSIMMIS [10], и другие.

Непосредственно в данной работе нас интересуют вопросы, возникающие на этапах анализа требований и проектирования интегрированных гетерогенных, физически автономных систем обмена информацией.

В связи с этим отметим высокую актуальность реальной задачи создания современной системы обмена информацией UnIT-NET IEDI[†].

[†] UnIt-NET: IT in University Management Network. http://www.unit-net.org.ua/ . Описание в [18]

Поэтому целью нашей работы является разработка концепции, позволяющей систематизировать и применять в конкретных приложениях теорию и опыт использования онтологий и проверить эту концепцию в задаче анализа требований для UnIT-NET IEDI.

Нашу задачу поставим так: на базе сложившейся к настоящему времени модели объединения распределенных первоначально несогласованных компьютерных систем [14-16] разработать концептуальную точку зрения, позволяющую согласовать до степени прикладной применимости результаты текущих исследований [11] и проверить эффективность такого подхода путем его приложения к выработке общих требований для проекта UnIT-NET IEDI.

2. Проблема взаимодействия

Проблема объединения несогласованных и распределенных компьютерных систем (КС) известна также как проблема взаимодействия.

[14] предлагает следующую классификацию проблем взаимодействия:

- *системные*: Они возникают в случае интеграции КС с различным аппаратным и программным обеспечением (например, операционными системами);
- *синтаксические*: возникают в случае, когда в интегрируемых КС используются различные языки и представления данных;
- *структурные*: связаны с использованием различных моделей данных;
- *семантические*: связаны со значениями терминов используемых при обмене информацией между интегрируемыми КС.

Группа проблем семантического взаимодействия также известна как проблемы семантической несогласованности [12].

Для достижения семантической согласованности в несогласованной информационной системе, должно быть достигнуто общее для всей системы понимание о смысле и значении передаваемой информации. Семантические конфликты возникают всякий раз, когда два контекста используют различные интерпретации информации.

Выделяют три основных случая семантической неоднородности [15]:

- *конфликты смешения смысла* – возникают в случаях, когда единицы информации имеют одинаковое название в разных случаях, а их значения - различаются. Например, значение выражения «последняя цена на торгах» зависит от времени торгов, указываемом в контексте;

- *конфликты систем измерения* – возникают, когда для измерения значений используются различные справочные системы измерения. Пример – различные валюты;

- *конфликты наименования* – возникают в случаях, когда используются различные схемы информации. Одним из признаков такого конфликта является наличие синонимов и омонимов.

Использование онтологий для уточнения подразумеваемых значений информации – возможный подход к решению проблемы семантической неоднородности. Достижение взаимодействия является ключевой сферой применения для онтологий [16]. На данный момент разработано множество подходов к интеграции информации на основе онтологий.

Делаем вывод, что для решения задачи интеграции можно использовать техники, подходы и программные парадигмы, которые выделяют похожие сложности в предметной области:

1) Сложности связанные со способом, которым решается проблема семантической несогласованности при интеграции информации с помощью онтологий;

2) Сложностей связанные с вопросами предоставления автономности и динамической сущности элементов открытой системы;

3) Задачи формулирования запросов, эффективной декомпозиции запросов без потери информации, уточнение и объединение результатов запросов.

Применительно к системе UnIT-NET IEDI это позволяет определить вопросы, требующие решения – проблема семантической неоднородности с помощью использования онтологий, обеспечение автономности и динамической сущности элементов системы, формулирование запросов и их эффективная декомпозиция.

3. Походы к решению проблемы семантической несогласованности

Первая группа сложностей связана со способом, которым решается проблема семантической несогласованности при интеграции информации с помощью онтологий. Как выделено в [2], первая группа включает вопросы разработки (подходы снизу-вверх и сверху вниз) и использования онтологий, построения отображений между онтологиями, и установления связей между онтологиями и информационными ресурсами, выступающими в роли поставщиков информации.

3.1 Роль онтологий при интеграции информации

Изначально онтологии представляются как эксплицитная спецификация концептуализации [17]. Поэтому онтологии могут быть использованы при решении задачи интеграции для описания семантики информационных ресурсов (ИР). Онтологии могут быть использованы для идентификации и установления соответствия между семантически сходными компонентами.

Почти во всех подходах к интеграции информации на основе онтологий последние используются для явного описания семантики информационных источников. Большинство проектов применяют один из следующих подходов использования онтологий ([11]): одна онтология (SIMS), несколько онтологий (OBSERVER), гибридный подход (BUSTER, DOME).

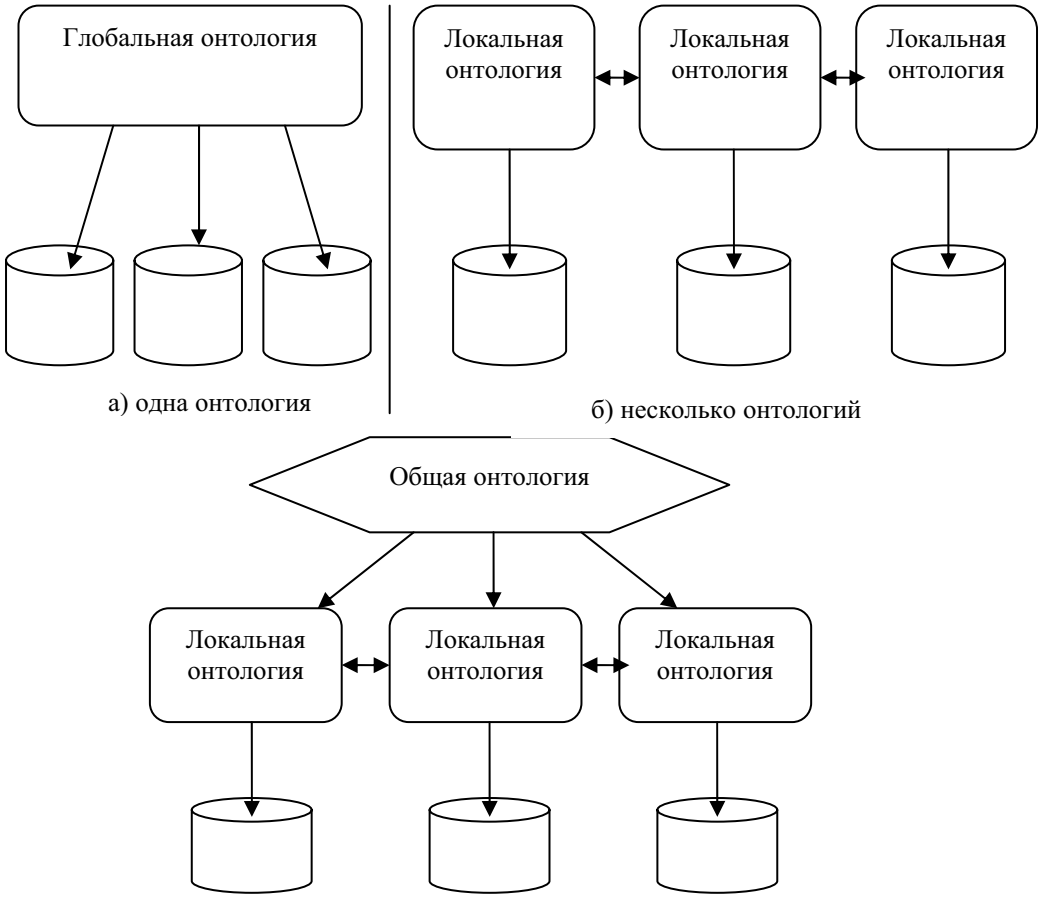

Рисунок 1. Подходы к использованию онтологий

Таблица 1. Сравнение различных подходов к интеграции ИР с помощью онтологий

	Одна онтология	Множество онтологий	Гибридный
Усилия на внедрения и разработку	Небольшие	Большие	разумные
Степень поддерживаемой семантической неоднородности между информационными ресурсами	Низкая (похожие представления о предметной области)	Поддерживает неоднородные представления	Поддерживает неоднородные представления
Действия необходимые для добавления/удаления ИР	Требует изменений в глобальной онтологии	Требует 1) построение и предоставление новой онтологии ИР;	построение и предоставление новой онтологии ИР

		2) установление связей между новой онтологии ИР и уже существующих онтологий ИР	
Сравнение нескольких онтологий	Не требуется	Является сложной задачей из-за отсутствия общего словаря	Является простой задачей, так как онтологии ИР используют общий словарь
Примеры реализации подходов	SIMS, ONTOLINGUA	OBSERVER	COIN, MECOTA, BUSTER

Возвращаясь к системе UnIT-NET IEDI можно сказать, что для решения задачи интеграции информации на основе онтологий в данном случае наиболее подходит гибридный подход к использованию онтологий.

3.2 Отображения между онтологиями и ИР

Связи между онтологиями и схемами информационных ресурсов устанавливаются с помощью отображений между элементами онтологий и элементами схем данных. Как указано в [2], причинами для построения этих отображений являются:

- Определения схем данных не всегда являются подходящим источником знаний о предметной области для людей запрашивающих информацию в системе, обычно эти схемы выполняют только техническую роль;
- Запросы, сформулированные к системе, выражаются на онтолого-ориентированном языке запросов без использования схемы данных. Поэтому построение и использование отображений между элементами онтологий и элементами схем данных позволяет автоматически выполнять пользовательские запросы в системе.

Как правило, информация в ИР храниться в базе данных. Онтология ИР может быть связана как со схемой данных, так и с отдельными терминами из базы данных ИР.

Выделяют [11] следующие подходы к построению отображений между онтологиями и ИР:

- *Использование структурного сходства*. Онтология строится путем точного копирования структуры базы данных и представления ее с помощью языка поддерживающего автоматический вывод. В этом случае выполняется интеграция с копией модели данных ИР. Этот подход реализован в системе TSIMMIS [20]и медиаторе системы SIMS [19];
- *Определение терминов.* Онтология используется для описания семантики терминов из базы данных, определяя соответствующие им

концепты. В этом случае онтология ИР уже не зависит от структуры базы данных ИР. Это подход реализован в системе BUSTER [21];

- *Расширение структуры.* Этот подход является наиболее распространенным. В нем комбинируются два предыдущих подхода. Логическая модель строится подобной структуре информационного ресурса, а также содержит дополнительная определения концептов. Этот подход использован для интеграции информации в системах OBSERVER [22], KRAFT [23], PICSEL [24] и DWQ[25];
- *Мета-аннотация.* Поход заключается в использовании мета-аннотаций, которые добавляют информацию о семантике в информационный ресурс. Этот подход является типичным для интеграции информации представленной в World Wide Web, где аннотирование является исходным и общепринятым способом для добавления семантики. Этот подход реализован в системах Ontobroker [26] и SHOE [27].

В конкретном случае, учитывая разнородную структуры баз данных ИР входящих в систему UnIT-NET IEDI, использование расширения структуры в качестве способа построения отображений является наиболее подходящим.

3.3 Отображения между онтологиями

В случаях, когда архитектура онтологий системы включает в себя несколько онтологий, объединенных “горизонтально” (как в подходе, использующем несколько онтологий) или “вертикально” (в гибридном подходе), необходимо построить отображения между онтологиями.

Отображения онтологий в системе предоставляют связи между элементами эквивалентными или родственными элементами онтологий, таким образом, обеспечивается повторное использование онтологий.

Выделяют [11] следующие подходы к построению отображений между онтологиями используемы в системах для интеграции информации:

- *Предопределенные отображения.* Этот подход реализован в системе KRAFT [28], где трансляция между онтологиями выполняется специальными программными агентами-медиаторами, использующими предварительно заданные правила трансформации. При реализации этого подхода используются различные виды отображений – от простых отображений один-к-одному между классами и значениями в онтологиях до отображений между областями онтологий, заданных составными выражениями. Этот подход является гибким, но не позволяет проверять сохранность семантической информации – пользователь имеет возможность задать такие правила отображения, которые не будут иметь смысла либо порождать конфликты;
- *Лексические связи.* Этот подход заключается в расширении общей модели описательной логики добавлением связями между онтологиями заимствованными из лингвистики. В системе

OBSERVER [22], в которой реализован этот подход, используются связи типа синоним, омоним, пересечение, включение и отсечение;

- *Связь с онтологией верхнего уровня.* Реализуется путем наследования концептов из общей онтологии верхнего уровня локальными онтологиями. Таким образом, устанавливаются связи между локальными онтологиями и онтологией верхнего уровня, что позволяет разрешать конфликты и неоднозначности, т.е. сравнивать онтологии между собой. Но не позволяет устанавливать прямые связи между концептами локальных онтологий. Пример реализации – система DWQ [25]
- *Семантические соответствия.* Заключается в определении семантических соответствий заданных с определенной системой. Например, используя [29] общий словарь для определения концептов различных онтологий. Позволяет избежать построения произвольных отображений между концептами и избежать проблем связанных с неоднозначностью, возникающей при построении непрямых отображений с помощью онтологии верхнего уровня, выполняемых при использовании предыдущего подхода.

В применении к системе UnIT-NET IEDI, в рамках которого предполагается использовать гибридный подход к использованию онтологий, мы можем сделать вывод, что для построения отображении между онтологиями ИР следует использовать подход, заключающийся в установлении связей с онтологией верхнего уровня.

4. Требования автономности информационных ресурсов и открытости системы

Требования автономности информационных ресурсов и открытости системы в целом являются другими причинами для использования отображений между элементами онтологий и схемами данных информационных ресурсов.

Вторая группа сложностей рассматривает вопросы предоставления автономности и динамической сущности элементов открытой системы. В качестве решений, в этой области предлагается использовать одну из медиаторных архитектур: централизованную или децентрализованную. Централизованная медиаторная архитектура подразумевает единственный центр, который содержит всю информацию об онтологиях, информационных ресурсах, их отображения, а также контролирует формулирование и выполнение запросов. Такой подход реализован в системе TSIMMIS.

Децентрализованная медиаторная архитектура подразумевает наличие отдельного агента/враппера для каждого информационного ресурса, который хранит отображения глобальной/общей онтологии (-иях) и обслуживает информационные ресурсы (InfoSleuth, SIMS, KRAFT).

Следовательно, в отношении системы UnIT-NET IEDI использование централизованной медиаторной архитектуры с центральным медиатором позволяет обеспечить целостность системы с одновременным предоставлением автономности входящим в систему ИР.

5. Задачи формулирования запросов, эффективной декомпозиции запросов

Третьей группой подзадач подходов к интеграции информации вызывающих возможные сложности являются задачи формулирования запросов, эффективной декомпозиции запросов без потери информации, уточнение и объединение результатов запросов.

Анализируя [19, 22, 24], можно выделить сложившиеся к настоящему времени подходы к решению этих задач:

- использование знаний содержащихся в онтологиях (отношения hypernym/hyponym) для переформулирования запросов, содержащих термины которые не определены в онтологии (-иях), для конструирования запросов без потери информации (OBSERVER);
- использование некоторых техник переписывания и отображения для получения запросов к информационным ресурсам, которые почти полностью соответствуют исходному запросу (PICSEL);
- использование глобальной онтологии при формулирования запросов в качестве глобальной модели для построения запросов. В этом случае пользователь формулирует запрос в терминах глобальной онтологии, а затем запрос трансформируется в подзапросы к ИР в терминах онтологий ИР (SIMS).

По отношению к системе UnIT-NET IEDI использование глобальной онтологии при формулировании запросов и использование техники переписывания запросов с использованием отображений для формулирования, обработки и выполнения запросов, является наиболее эффективной.

6. Выводы

В данной работе представлена авторская концепция системного подхода к использованию онтологий для решения задачи интеграции распределенных, гетерогенных, физически автономных информационных ресурсов, позволившая выяснить фактическую применимость этого подхода при разработке требований и проектных решений в рамках создания современной системы обмена информацией UnIT-NET IEDI.

Это позволило обнаружить, что некоторые из проанализированных проблем имеют только частичные решения, например, проблема семантического взаимодействия, как правило, частично решается путем поручения некоторого рода соглашений узлам-участникам, тем самым, предоставляя инфраструктуру для семантических представлений. Эти частичные решения, очевидно, ограничивают сферу приложения и функциональность реализуемых программных прототипов выполняющих задачу распределенного интеллектуального извлечения информации.

Прикладное значение настоящей работы демонстрируется тем, что по результатам исследования были определены ограничения, накладываемые в случае реализации на систему UnIT-NET IEDI - программную инфраструктуру, позволяющую обмениваться электронными данными между университетами и государственными учреждениями Украины. Сложность задачи связана с тем, что субъекты обмена физически удаленны между собой, принадлежат

различным юридическим владельцам, а также их информационные ресурсы являются семантически разнородными [18]. Найденные ограничения можно свести к следующим требованиям

- IEDI строится на принципах медиаторной архитектуры с центральным медиатором
- IEDI использует гибридный подход [11] для представления знаний
- IEDI использует регистрацию информационных ресурсов перед тем как ресурс станет доступным для запросов
- IEDI не предоставляет полную автоматизацию для процессов построения отображений и выравнивания онтологий
- Компоненты IEDI используют технику переписывания с использованием отображений для формулирования, обработки и выполнения запросов

В рамках этих ограничений разработка формальных методов для поддержки получения информации из гетерогенных, распределенных, автономных источников, а именно механизма декомпозиции и выполнения запросов к распределенным автономным семантически гетерогенным информационным источникам является новой задачей, не имеющей готовых решений в данном контексте. Разработка адекватных решений может служить направлением дальнейших исследований.

ЛИТЕРАТУРА

1. Stuckenschmidt H., Wache H., Voegele T., Visser U.: Enabling technologies for interoperability. In: (Visser, U., Pundt H. Eds.) Workshop on the 14th International Symposium of Computer Science for Environmental Protection, Bonn, Germany, 2000, pp. 35-46.
2. Cui, Z., Jones, D., O'Brien, P.: Issues in Ontology-based Information Integration. In: (A. Gomez-Perez, M. Gruninger,H. Stuckenschmidt,M. Uschold) Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle, USA, August 4-5, 2001, pp.141-146.
3. Cui, Z., Jones, D., O'Brien, P.: Semantic B2B Integration: Issues in Ontology-based Applications. SIGMOD Record, Vol.31, No.1, March 2002. Pp.43-48
4. Bayardo et al.: InfoSleuth: Semantic Integration of Information in Open and Dynamic Environment. In Proceedings of the 1997 ACM International Conference on the Management of Data (SIGMOD), Tucson, Arisona, May 1997.
5. Gray. P., Preece A., Fiddian N., Gray W., Bench-Capon T., Shave M., Azarmi N., Wiegand M.: KRAFT: Knowledge Fusion From Distributed Databases and Knowledge Bases. In: Proc. 8th Intl. Worckshop on Database and Expert System Applications (DEXA-97), IEEE Press, pp. 682-691.
6. Bergamaschi, S., Castano, S., De Capitani di Vimercati, S., Montanari, S. Vincini, M.: An Intelligent Approach to Information Integration. In: Proc. Of Formal Ontology in Information Systems (FOIS-98), June, 1998.
7. Kashyap V., Sheth A.: Information Brokering across Heterogeneous Digital Data: A Metadata-based Approach. Kluwer Academic Publishers, 2000

8. Decker, S., Erdmann, M., Fensel, D., and Studer, R.: Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In R. Meersman et al. (eds.): Semantic Issues in Multimedia Systems. Proceedings of DS-8. Kluwer Academic Publisher, Boston, 1999, 351-369.
9. Arens, Y., Knoblock, C.A., Shen, W.: Query Reformulation for Dynamic Information Integration. Journal of Intelligent Information Systems. 1996.
10. Garcia-Molino, H. et. al.: The TSIMMIS Approach to Mediation: Data Models and Languages. In Proceedings of the NGITS (Next Generation Information Technologies and Systems), June 1995.
11. Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hubner, S.: Ontology-Based Integration of Information - A Survey of Existing Approaches. In: (A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, M. Uschold) Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing,
12. Won Kim and Jungyun Seo. Classifying schematic and data heterogeinity in multidatabase systems. IEEE Computer, 24(12):12–18, 1991. problem classification of semantic heterogeneity.
13. V. Kashyap and A. Sheth. Schematic and semantic semilarities between database objects: A context-based approach. The International Journal on Very Large Data Bases, 5(4):276–304, 1996.
14. A.P. Sheth. Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kottman (eds.) Interoperating Geographic Information Systems, Kluwer.
15. Cheng Hian Goh. Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources. Phd, MIT, 1997.
16. Mike Uschold and Michael Gr.uninger. Ontologies: Principles, methods and applications. Knowledge Engineering Review, 11(2):93–155, 1996.
17. Tom Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220, 1993.
18. Bulat, A., Ermolayev, V., Gray, E., Keberle, N., Plaksin, S., Shapar, V., Vladimirov, V., Zholtkevich, G. (2004) The Infrastructure for Electronic Data Interchange. Reference Architecture Specification. Version 1.0. UNIT-NET Deliverable No D2.2.D.1, Feb. 2004.
19. Yigal Arens, Chun-Nan Hsu, and Craig A. Knoblock. Query processing in the sims information mediator. In Advanced Planning Technology. AAAI Press, California, USA, 1996.
20. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The tsimmis project: Integration of heterogeneous information sources. In Conference of the Information Processing Society Japan, pages 7–18, 1994.
21. Heiner Stuckenschmidt and Holger Wache. Context modelling and transformation for semantic interoperability. In Knowledge Representation Meets Databases (KRDB 2000). 2000.
22. E. Mena, V. Kashyap, A. Sheth, and A. Illarramendi. Observer: An approach for query processing in global information systems based on interoperability between

pre-existing ontologies. In Proceedings 1st IFCIS International Conference on Cooperative Information Systems (CoopIS ’96). Brussels, 1996.
23. A.D. Preece, K.-J. Hui, W.A. Gray, P. Marti, T.J.M. Bench-Capon, D.M. Jones, and Z. Cui. The kraft architecture for knowledge fusion and transformation. In Proceedings of the 19th SGES International Conference on Knowledge-Based Systems and Applied Artificial Intelligence (ES’99). Springer, 1999.
24. Francois Goasdoue, V.eronique Lattes, and Marie-Christine Rousset. The use of carin language and algorithms for information integration: The picsel project,. International Journal of Cooperative Information Systems (IJCIS), 9(4):383 – 401, 1999.
25. Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. Description logics for information integration. In Computational Logic: From Logic Programming into the Future (In honour of Bob Kowalski), Lecture Notes in Computer Science. Springer- Verlag, 2001
26. D. Fensel, I. Horrocks, F. Van Harmelen, S. Decker, M. Erdmann, and M. Klein. Oil in a nutshell. In 12th International Conference on Knowledge Engineering and Knowledge Management EKAW2000, Juanles- Pins, France, 2000.
27. Jeff Heflin and James Hendler. Semantic interoperability on the web. In Extreme Markup Languages 2000, 2000.
28. A.D. Preece, K.-J. Hui, W.A. Gray, . Marti, T.J.M. Bench-Capon, D.M. Jones, and Z. Cui. The kraft architecture for knowledge fusion and transformation. In Proceedings of the 19th SGES International Conference on Knowledge-Based Systems and Applied Artificial Intelligence (ES’99). Springer, 1999.
29. Holger Wache. Towards rule-based context transformation in mediators. In S. Conrad, W. Hasselbring, and G. Saake, editors, International Workshop on Engineering Federated Information Systems (EFIS 99), Kuhlungsborn, Germany, 1999. Infix-Verlag.
30. Lattes V., Rousset M.-C.: The Use of CARIN Language and Algorithms for Information Integration: The PICSEL System. International Journal of Cooperative Information Systems, Vol.9, No.4, 2000, pp.383-401.