Технология интеграции и представления музейных данных в среде Web с помощью онтологий♣

© В.В. Иванов

Казанский (Приволжский) федеральный университет
nomemm@mail.ru

Аннотация

Описаны результаты работ по созданию методов и моделей интеграции музейных электронных коллекций. Предлагаемые методы находятся в русле известных работ по интеграции разнородных данных, но при этом опираются на интенсивное использование онтологических ресурсов двух типов – формальных онтологий и тезаурусов, а также на учет специфики исходных музейных описаний. Освещаются вопросы представления связей между тезаурусом, содержащим терминологию предметной области, и онтологией верхнего уровня. В заключении описан опыт применения указанных методов для создания хранилища данных.

1 Введение

Проблемам, связанным с интеграцией разнородных источников информации, посвящено множество работ как в области технологий баз данных [1], так и в области искусственного интеллекта [2]. Главная цель при решении задачи интеграции данных состоит в обеспечении доступа к множеству разнородных источников на основе общего для всех источников интерфейса запросов. Использование онтологий (в качестве концептуальных моделей предметной области) для решения задач интеграции информации представляется перспективным направлением [3 – 5]. С одной стороны, онтологии предназначены для явного описания понятий и связей между понятиями предметной области, а с другой стороны, они являются разделяемыми ресурсами и наилучшим образом подходят на роль общего интерфейса к разнородным источникам данных. Как правило, интеграция данных производится в рамках некоторой фиксированной предметной области. В данной работе была выбрана предметная область культурного наследия, в частности, музейное дело. Эта предметная область представляет особый интерес, поскольку в современных (отечественных) музейных автоматизированных информационных системах переход от представления данных в виде неструктурированных текстовых документов к хорошо структурированным форматам и схемам еще не завершился. Как правило, имеет место представление данных в виде слабоструктурированных (или полуструктурированных) документов. Описания музейных предметов представляются в виде таблиц, содержащих в ячейках преимущественно текстовые значения. Такая форма представления данных (далее – структурированные текстовые описания) является доминирующей. Однако в последнее время наблюдаются тенденция к формальному описанию схем данных, введению стандартов метаданных, массовое внедрение в музеях и библиотеках информационных систем фактографического типа и переход от традиционных электронных библиотек (с текстами) к мультимедийным, содержащим фото-, аудио- и видеоматериалы (см., например, проект Europeana [6]). При доступе к структурированным источникам данных, созданным и поддерживаемым независимо в разных музейных системах, возникают проблемы, связанные с неоднородностью следующих видов (в зависимости от уровня, на котором производится объединение источников). *Физическая неоднородность* возникает из-за использования разных форматов хранения и обмена данными на физическом уровне. *Структурная неоднородность* порождается наличием большого числа различных схем баз данных. *Семантическая неоднородность* является следствием различий в множествах понятий и отношений предметной области, а также способов их интерпретации, которые применяются в различных компьютерных системах и/или организациях.

Наиболее актуальным направлением исследований является преодоление семантической неоднородности. Здесь возникают проблемы, связанные, в первую очередь, с отсутствием общего взгляда на структуру понятий предметной области (онтологии верхнего уровня), а также с отсутствием единой терминологии (набора понятий, или общего подъязыка предметной области). Многие современные подходы ориентированы либо на обработку структуры источника данных (метаданных, концептуальной схемы), либо на текстовое содержимое. Предлагаемый подход вместе с разнообразием структур данных учитывает и различия в терминологии. В рамках подхода была создана онтология по культурному наследию, формализующая основные понятия и отношения области музейной документации и содержащая более 20 тыс. понятий. Построенная онтология использовалась для автоматической об-

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

работки различных электронных музейных коллекций с целью их интеграции.

В разделе 2 упомянуты общие подходы к интеграции данных, особенности, возникающие при интеграции данных на основе онтологий, а также ресурсы онтологического типа, используемые далее. В разделе 3 описаны особенности подхода к построению онтологии по культурному наследию, изложены основы модели интеграции музейных описаний, а также метод автоматизированного установления семантических соответствий между структурными элементами разных схем данных. В разделе 4 описан опыт применения предложенных методов для создания RDF-хранилища, представляющего данные электронного музейного каталога в среде Linked Data. Связанные результаты опубликованы в других работах: в [7] предложен подход к разрешению лексической многозначности, возникающей при обработке текстовых данных, хранящихся внутри слабоструктурированных документов; в [8] приводятся результаты экспериментов с предложенными методами.

2 Состояние дел в области интеграции музейных данных

2.1 Общие подходы к интеграции данных

Выделяют общие направления на основе федеративных БД, медиаторов и хранилищ данных [9]. В [10] отмечается, что важным аспектом при интеграции данных является наличие глобальной концептуальной схемы. Задачи интеграции данных обычно ставятся как задачи сравнения схем баз данных [11, 12], реже – как задачи сравнения содержимого разнородных БД [13]. Для анализа структуры и содержимого используются подходы на основе нейронных сетей [14], машинного обучения [15] и информационного поиска [16]. Между элементами схем устанавливаются соответствия, на основе которых схемы связываются специальным набором отношений. Совокупность соответствий и отношений называется *отображением*. Для решения задачи сравнения схем данных разработано множество подходов, как специфичных для предметной области [17], так и направленных на использование конкретных языков представления схем [18]. Тезаурусы и лексические базы данных (WordNet) используются в качестве наборов синонимов при сопоставлении лексических меток элементов схем [20]. С точки зрения теории разработки систем интеграции с использованием онтологий выделяются три основные направления, описанные в [10, 21 – 23].

2.2 Особенности музейных коллекций

Попытки создать единую схему для описания метаданных музейных предметов предпринимались несколькими исследовательскими группами. Наиболее плодотворным отечественным проектом, направленным на интеграцию в области стандартизации музейной документации, является работа по созданию стандарта краткого описания (т. н. «этикетки музейного предмета») [24]. Цель проекта состояла в том, чтобы выработать рекомендации по общему для всех музеев списку полей описания музейного предмета. В результате был получен следующий список полей описания музейного предмета: «Организация (место хранения)»; «Идентификационный номер предмета»; «Типология»; «Автор»; «Место создания/производства»; «Название предмета»; «Датировка»; «Культурный период»; «Материалы и техника»; «Размеры, вес»; «Ключевые слова»; «Краткое описание предмета»; «Комментарий».

Большинство предложенных полей является атрибутами различных сущностей и только косвенно связаны друг с другом через (неявно подразумеваемые) свойства этих сущностей. Поля «Авторы», «Место создания/производства», «Датировка», «Материалы и техника» характеризуют особенности создания предмета. Поля «Типология», «Ключевые слова», «Культурный период» характеризуют тип предмета, помогают группировать предметы в соответствии с некоторой классификацией периодов, типов и т. п. Поля «Название предмета», «Идентификационный номер предмета» и «Организация» служат для идентификации конкретного объекта из множества всех имеющихся, а также для целей учетно-хранительской деятельности. Предложенный список полей является стандартом «де-факто», и большинство музейных открытых электронных каталогов в России ориентируется на представление своих данных в виде слабоструктурированных документов, построенных на базе этого стандарта. Это с одной стороны приводит к формальному преодолению структурной неоднородности, но на деле усложняет семантическую интеграцию разнородных коллекций, поскольку в каждом поле используется оригинальная для данного музея терминология.

2.3 Онтология CIDOC CRM

В качестве глобальной онтологии выбрана модель CIDOC CRM (Conceptual Reference Model) [25], разработка которой ведется с 2000 года Комитетом по документации (CIDOC) Международного совета музеев (ICOM). В 2007 году модель CIDOC CRM была утверждена в качестве стандарта ISO/CD 21127 и на сегодняшний день является основным международным стандартом для описания информации по культурному наследию. Основным преимуществом онтологии CIDOC CRM является разнообразие свойств, которые, в свою очередь, определяют семантику понятий, входящих в домен или диапазон свойства. Тем не менее, данная онтология является слишком общей: наблюдается существенный разрыв между понятиями CIDOC CRM и понятиями в содержимом реальных музейных описаний.

Понятия верхнего уровня приблизительно соответствуют названиям таблиц и столбцов в музейных БД (см. «этикетку музейного предмета»), а не терминам, описывающим значения в ячейках этих таблиц. С одной стороны, такая ситуация (будем называть ее «терминологическим пробелом») существенно ухудшает выразительность онтологии и, оче-

видно, ведет к потере точности при интеграции информации, поиске и т. п. С другой стороны, наличие терминологического пробела вполне ожидаемо: онтология верхнего уровня не должна описывать все возможные понятия, которыми оперируют пользователи при составлении описаний и формулировке запросов.

2.4 Тезаурус AAT

Подъязык предметной области состоит из связанных друг с другом терминов и может моделироваться с помощью тезаурусов. Использование тезаурусов общей тематики (например, WordNet [27]) не целесообразно – в области культурного наследия имеет смысл ориентироваться на специализированные информационно-поисковые тезаурусы. Одним из таких ресурсов является тезаурус по искусству и архитектуре AAT (Art & Architecture Thesaurus) [28], поддерживаемый обществом П. Гетти (P. Getty).

Помимо того, что тезаурус является англоязычным, существуют ограничения, затрудняющие его прямое использование. Область охвата AAT соответствует терминологии, принятой в Западной Европе. Другой недостаток тезауруса AAT состоит в том, что он разрабатывался в расчете на индексирование вручную. Несмотря на это, тезаурус AAT подходит для автоматической обработки текстов. Положительный момент состоит в том, что AAT является наиболее полным среди тезаурусов в данной предметной области (общее число понятий – около 33 тыс., число терминов (т. е. лексических единиц) – более 130 тыс.). Перевод существенного фрагмента тезауруса на русский язык осуществлен в НИВЦ МГУ [29]. В процессе перевода тезаурус был адаптирован к русскоязычной лексике и существенно расширен синонимичными текстовыми входами, которые извлекались из крупного Тезауруса по общественно-политической тематике.

Для использования и внедрения онтологии CIDOC CRM в российских музеях имена классов и свойств, а также текстовые комментарии к ним были переведены на русский язык. Важно отметить необходимость этого перевода, поскольку семантика классов или свойств описывается именно текстовым комментарием, а не названием.

3 Создание и применение онтологии по культурному наследию для интеграции описаний музейных предметов

3.1 Объединение онтологии CIDOC CRM и тезауруса AAT

Для преодоления терминологического пробела онтологию CIDOC CRM необходимо расширять, подключая к ней специализированные словари по культуре, списки географических названий, имен деятелей культуры и т. п. Подобные источники значительно более детально представляют значения понятий нижнего уровня. Главная трудность здесь состоит в том, чтобы построить прикладную онтологию, адекватно структурирующую понятия предметной области и тесно связанную с подъязыком экспертов предметной области.

Связывание двух ресурсов онтологической природы: онтологии верхнего уровня и лексической онтологии, представляет нетривиальную задачу [30, 31]. Подход к решению задачи связывания онтологии и тезауруса существенно зависит от дальнейшего применения расширенной онтологии. Для подключения тезауруса AAT к онтологии CIDOC CRM был выбран следующий подход.

Связывание осуществляется с помощью определения набора логических ограничений, накладываемых на множества допустимых значений формальных свойств, заданных в онтологии верхнего уровня. В качестве множества допустимых значений некоторого свойства P выступают группы близких понятий тезауруса, которые обычно представляются как фасеты или дескрипторные блоки. Логические ограничения имеют следующий вид:

$$C(y) \equiv \forall x.\ P(y,x) \rightarrow DB(x)$$ (строгая форма)

либо

$$C(y) \equiv \exists x.\ P(y,x) \wedge DB(x)$$ (ослабленная форма),

где C – унарный предикат (класс онтологии CIDOC CRM), P – бинарный предикат (свойство онтологии CIDOC CRM), а DB – унарный предикат (множество экземпляров, объединенных некоторым фасетом тезауруса AAT). В общем случае вместо DB может использоваться предикат, истинный на произвольном подмножестве понятий тезауруса. Предложенный подход позволяет явно уточнять значение класса онтологии CIDOC CRM через подмножество понятий тезауруса AAT. Например, накладывание «ослабленного» ограничения на значение свойства *P45_состоит_из (входит_в_состав)* класса *E22_Рукотворный_Объект* интерпретируется как доопределение семантики класса *E22_Рукотворный_Объект* (т. е. любой объект, созданный человеком, имеет в составе некоторый материал, заданный фиксированным фасетом из AAT).

3.2 Интеграция музейных описаний на основе онтологии

После создания онтологии по культурному наследию становится возможным описывать с ее помощью факты, извлекаемые из различных источников данных, формировать хранилище данных, схемой которого является онтология CIDOC CRM. Идея данного подхода к интеграции данных с помощью CIDOC CRM имеет много общего с идеей генерации централизованных хранилищ данных, описанной, например, в [19]. Для построения хранилища данных требуется определить отображение между исходной схемой S (например, заданной «этикеткой музейного предмета») и результирующей схемой – онтологией CIDOC CRM. Особенностью нашего подхода является метод полуавтоматической генерации отображения, основанный на сопоставлении текстовых значений и понятий тезау-

руса. Общая модель процесса интеграции описана в [8]. В [26] указывается на возможность декомпозиции задачи преодоления неоднородности на две подзадачи: задачу поиска соответствий между элементами схем и задачу определения сложного отображения, использующего найденные соответствия. Остановимся подробно на задаче поиска соответствий между элементами схем. Введем понятие элементарного соответствия и поставим задачу поиска элементарных соответствий [32]. Предлагаемые для ее решения методы позволяют автоматически находить семантические соответствия между элементами схемы источника (*S*) и результирующей схемой *T*.

Элементарным соответствием между классами из схем *S* и *T* назовем семерку

$$\langle C_S, Subject_T, Property_T, Object_T, \delta, type, w \rangle,$$

где C_S – класс из схемы *S*, $Subject_T$, $Object_T$ – классы из схемы *T*, связанные свойством $Property_T$ из схемы *T*, δ – основа для построения данного соответствия, *type* – тип связи между классами и w – вес данного элементарного соответствия.

Каждое элементарное соответствие задает связь между классами C_S и $Object_T$. Параметр *type* – отношение между C_S и $Object_T$ на домене интерпретации (например, отношение включения или эквивалентности). Параметры $Subject_T$ и $Property_T$ определяют контекст в схеме T, в котором множества экземпляров C_S и $Object_T$ могут быть связаны отношением *type*. Параметр δ указывает, на основе каких компонентов значения построено данное элементарное соответствие между классами C_S и $Object_T$, например, δ может представлять регулярное выражение или набор ключевых слов, содержащихся в текстовых представлениях экземпляров класса C_S и класса $Object_T$. В случае семантической интеграции δ представляет собой список понятий тезауруса, которые описывают экземпляры класса C_S в исходной схеме S и допустимые значения свойства $Property_T$ класса $Subject_T$ в результирующей схеме *T*. Множество элементарных соответствий определяет отображение между схемами *S* и *T*, которое далее называется *частичным отображением*.

Задача построения частичного отображения. Пусть даны исходная схема *S* и результирующая (глобальная) схема *T*. Для заданного числового порога $0 \le \theta \le 1$ необходимо построить частичное отображение ϕ, содержащее элементарные соответствия, для каждого из которых выполняются условия:

1) C_S^I и $Object_T^I$ связаны отношением *type* при $I = (\delta, (\square)^I)$;

2) $\delta = C_S^I \cap Object_T^I \neq \varnothing$; 3) $w \ge 0$.

Аналогичным образом определяются элементарные соответствия между бинарными предикатами (свойствами) исходной и результирующей схем и ставится *задача построения частичного отображения бинарных предикатов* из исходной схемы на бинарные предикаты из результирующей.

Для решения поставленной задачи необходимо сравнить интерпретации элементов из схем *S* и *T*, т. е определить $I = (\delta, (\square)^I)$ для каждого возможного соответствия. Сравнение может выполняться экспертом, понимающим значение, стоящее за символами классов и свойств в схемах, но для автоматизации этого процесса необходимо моделировать интерпретацию *I*. Допущение, лежащее в основе данного подхода к моделированию интерпретации, состоит в том, что совокупность текстовых выражений элементов экстенсионала определяет значение (интенсионал) этого класса. Это значение используется для поиска семантически близких классов в результирующей схеме *T*. Для реализации подхода достаточно сделать следующее: для каждого класса C_S из исходной схемы *S* построить список, содержащий те понятия тезауруса, которые встретились в лексическом выражении экстенсионала класса C_S. Таким образом, интерпретация определяется операционально – через процедуру индексирования текстовых значений с помощью понятий тезауруса. Список понятий определяет интерпретацию класса C_S в терминах информационно-поискового языка тезауруса. Связи между классами и понятиями тезауруса, заданные при создании онтологии, используются для автоматического выделения в схеме *T* классов $Object_T$, семантически близких классу C_S. Значения параметров $Subject_T$ и $Property_T$ берутся из соответствующего логического ограничения. Подобная процедура интерпретации имеет очевидные недостатки, поскольку опирается на текстовое содержимое источника данных, а оно может быть неточным, неполным или, наоборот, избыточным (в т. ч. многозначным). Но, с другой стороны, процедура интерпретации позволяет исключить соответствия, не имеющие смысла для содержимого (экстенсионала) данного источника. Для каждого структурного элемента в исходной схеме возможно построение нескольких элементарных соответствий, при этом каждая из альтернатив подтверждается конкретным примерами вхождений понятий тезауруса в текстовое выражение элемента схемы *S*.

4 Представление данных электронного каталога музея в среде Linked Data (проект Open Kunstkamera Data)

В 2010 году Лабораторией математической и компьютерной лингвистики НИИММ Казанского университета совместно с компанией КАМИС и Музеем антропологии и этнографии РАН имени

Петра Великого (Кунсткамера) – далее МАЭ РАН – выполнен проект Open Kunstkamera Data (OKD). Цель проекта OKD состояла в том, чтобы представить данные каталога МАЭ РАН в открытом и стандартизированном виде в среде Web.

Задача заключалась в том, чтобы построить RDF-хранилище, которое соответствует рекомендациям Международного Совета Музеев (ICOM) и стандартам среды Web. В хранилище представлены взаимосвязанные данные о предметах, персоналиях, событиях создания и сбора этнографических материалов, датах и т. п. Список полей описания фотографии оказался очень близок к «этикетке музейного предмета» (рис. 1).

Имя поля	**Значение**
Учетный номер	№ 3196-79
Название	Голубиная башня в окрестностях Герата
Этническая принадлежность	афганцы
Время съемки	1924
Место съемки	Афганистан, провинция Герат, г. Герат
Автор	Букинич Дмитрий Демьянович
Собиратель	Букинич Дмитрий Демьянович, Почвовед, инженер-ирригатор, археолог
Экспедиция	Экспедиция Н.И. Вавилова в Афганистан (1924)
Тематическая принадлежность	Занятия, животноводство, птицеводство
Жанр	Место, в отношении построек, культовых сооружений (объектов)

Рис. 1. Пример описания музейного фотографии из фонда МАЭ РАН

При выполнении проекта OKD использовались элементы описанного выше подхода. В качестве схемы хранилища выступала онтология CIDOC CRM, представленная на языке OWL DL. В качестве подключаемой терминологии вместо тезауруса AAT были использованы иерархические справочники внутренней БД Комплексной автоматизированной музейной информационной системы (КАМИС).

Справочники были преобразованы в формат SKOS, рекомендованный W3C для представления ресурсов подобного рода. Далее было построено и выполнено отображение схемы описания музейного предмета на онтологию CIDOC CRM. Фрагмент построенного отображения приведен ниже (рис. 2).

Исходная база данных фотоиллюстративного фонда МАЭ РАН содержит более 40 тыс. единиц описания. Полученное хранилище объемом более 5 млн. RDF-триплетов было загружено в специализированную СУБД OpenLink Virtuoso.

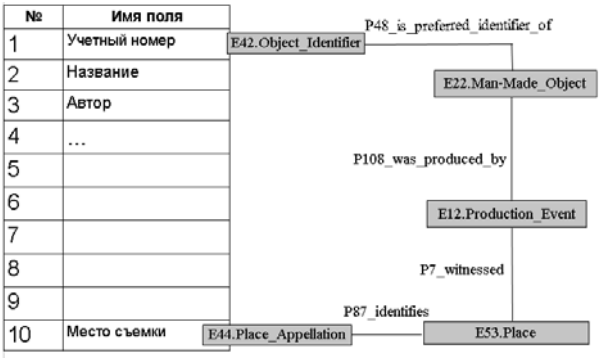

Рис. 2. Фрагмент отображения, построенного на двух элементарных соответствиях: [«Учетный номер» → E42.Object_Identifier (Идентификатор Объекта)] и [«Место съемки» → E44.Place_Appellation (Описание Местоположения)]

Для поддержания актуальности хранилища данных разработан и реализован механизм регулярных обновлений. Построенное хранилище позволяет реализовывать новые сервисы и решать новые задачи. Для реализации прототипа геоинформационного сервиса потребовалось выполнить связывание терминов из справочников с внешним набором данных: geonames.org, что позволило отображать на глобальной карте как совокупности сущностей из музейного каталога, например, места сбора материалов, так и отдельные предметы. Для поддержки выполнения сложных запросов к хранилищу открыта точка доступа по протоколу SPARQL. Следствием чего стало включение каталога МАЭ РАН в среду Linked Data – общемировое распределенное хранилище взаимосвязанных знаний. В ближайшее время запланировано применение предлагаемой технологии к данным электронного каталога Этнографического музея Казанского университета, что позволит говорить об интеграции музейных баз данных в масштабах среды Web.

Литература

[1] Lenzerini M. Data integration: A theoretical perspective // ACM PODS Conference. – 2002. – P. 233-246.

[2] Calvanese D., De Giacomo G. Data integration: a logic-based perspective // AI Magazine. – 2005. – V. 26, No 1. – P. 59-70.

[3] Baader F., McGuinness D., Nardi D., Patel-Schneider P. The description logic handbook: theory, implementation and applications. – Cambridge: Cambridge University Press, 2003.

[4] Doan A., Madhavan J., Domingos P., Halevy A. Ontology matching: a machine learning approach // Handbook on Ontologies in Information Systems / Ed. by S. Staab and R. Studer. – Springer-Verlag, 2004. – P. 397-416.

[5] Doerr M., Hunter J., Lagoze C. Towards a core ontology for information integration // J. of Digital Information. – 2003. – V. 4. – Issue 1.

[6] Europeana portal, 2010. – http://www.europeana.eu/portal/.

[7] Иванов В.В., Соловьев В.Д. Применение онтологий для разрешения лексической многозначности в структурированных источниках данных // Третья межд. конф. по когнитивной науке. – М.: Художественно-издательский центр, 2008. – Т. 2. – С. 577-580.
[8] Иванов В.В., Иванов В.А. Модели и методы интеграции структурированных текстовых описаний на основе онтологий // Труды казанской школы по когнитивной и компьютерной лингвистике (под ред. О.А. Невзоровой, В.Д. Соловьева, Д.Ш. Сулейманова). – Казань: Изд-во Казан. ун-та, 2009.
[9] Fundamentals of data warehousing / Ed. by M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis. – Springer-Verlag, 1999.
[10] Wache H., Vogele T., Visser U., Stuckenschmidt H. et al. Ontology-based integration of information – a survey of existing approaches // Proc. of the IJCAI-2001 Workshop: Ontologies and Information Sharing. – Seattle, WA, 2001.
[11] Madhavan J., Bernstein P.A., Doan A.H., Halevy A. Corpus-based schema matching // Proc. of Int. Conf. on Data Engineering (ICDE). – 2005.
[12] Do H.H., Rahm E. COMA – a system for flexible combination of schema matching approach // Proc. of Int. Conf. on Very Large Databases (VLDB). – 2002.
[13] Doan A.H., Madhavan J., Domingos P., Halevy A. Learning to map between ontologies on the Semantic Web // Proc. of Int. Conf. World Wide Web (WWW). – 2002.
[14] Li W.S., Clifton C., Liu S.Y. Database integration using neural networks: implementation and experiences // Knowledge and Information Systems. – 2000. – V. 2. – No 1.
[15] Berlin J., Motro A. Database schema matching using machine learning with feature selection // Proc. of Int. Conf. Advanced Information Systems Engineering (CaiSE). – 2002.
[16] Cohen W. Integration of heterogeneous databases without common domains using queries based on textual similarity // Proc. of ACM SIGMOD Int. Conf. Management of Data. – 1998. – P. 201–212.
[17] Bergamaschi S., Castano S., Vincini M., Beneventano D. Semantic integration of heterogeneous information sources // Data and Knowledge Engineering. – 2001. – V. 36, No 3. – P. 215-249.
[18] Miller R.J. et al. The CLIO project – managing heterogeneity // ACM SIGMOD Record. – 2001. – V. 30, No 1. – P. 78-83.
[19] Doerr M., Iorizzo D. The dream of a global knowledge network – a new approach // ACM J. on Computers and Cultural Heritage. – 2008.
[20] Embley D.W., Jackmann D., Xu L. Multifaceted exploitation of metadata for attribute match discovery in information integration // Proc. of Int. Workshop on Information Integration on the Web (WIIW). – 2001.
[21] Levy A.Y., Rajaraman A., Ordille J.J. Querying heterogeneous information sources using source descriptions // Proc. of Int. Conf. on Very Large Databases (VLDB). – Bombay, 1996.
[22] Calvanese D., De Giacomo G., Lenzerini M. Ontology of integration and integration of ontologies // Description Logics. – 2001.
[23] Levy A.Y., Mendelzon A.O., Sagiv Y., Srivatstava D. et al. Answering queries using views // Proc. of PODS. – San Jose, CA, 1995.
[24] Кузьмина Е.С., Ноль Л.Я., Черненко В.В., Кощеева Е.Л. и др. Краткое описание музейного предмета: информационно-лингвистическое обеспечение. – Псков; М., 2001.
[25] Crofts N., Doerr M., Gill T., Stead S. Definition of the CIDOC conceptual reference model. – http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.0.pdf.
[26] Euzenat J., Shvaiko P. Ontology matching. – Heidelberg: Springer, 2007. – 340 p.
[27] Miller G. Nouns in WordNet // WordNet – an electronic lexical database / Ed. by C. Fellbaum. – Cambridge: The MIT Press,1998.
[28] Art and architecture thesaurus (Research at the Getty). – http://www.getty.edu/research/conducting_research/vocabularies/aat/.
[29] Добров Б.В., Лукашевич Н.В., Соловьев В.Д. Тезаурус по архитектуре и искусству как средство формализации описаний музейных предметов // Электронный журнал FCCL. – 2006. – http://fccl.ksu.ru/issue_spec/docs/aat_index.doc.
[30] Нариньяни А.С. Кентавр по имени ТЕОН: тезаурус+онтология // Межд. семинар по компьютерной лингвистике и ее приложениям «Диалог'2001». – 2001. – Т. 1. – С. 184-188.
[31] Нариньяни А.С. ТЕОН-2: от Тезауруса к Онтологии и обратно // Межд. семинар по компьютерной лингвистике и ее приложениям «Диалог'2002». – Протвино, 2002. – Т. 1. – С. 307-313.
[32] Иванов В.В. Онтологический подход к созданию информационной системы по культурному наследию // Учёные записки Казанского государственного университета. Серия физико-математические науки. – 2007. – Т. 149, Кн. 2. – С. 73-92.

Ontology-based techniques for integration and representation of museum collections on the Web

Vladimir Ivanov

The paper describes a model and ontology-based methods for integrating of heterogeneous museum descriptions. A method for linking an upper level ontology and thesauri is proposed. In conclusion an experience of creating large RDF-store in order to represent museum collections as Linked Data is discussed.

♣ Работа выполнена при финансовой поддержке РФФИ (проекты 09-07-97007-р_поволжье, 10-07-00445 и 09-07-12059-офи_м)