О РАСХОЖДЕНИИ ИНФОРМАЦИОННЫХ ОНТОЛОГИЙ С КОНЦЕПТУАЛИЗАЦИЯМИ ВНЕШНЕГО МИРА*

А.Г. Марчук, В.В. Целищев

Статья посвящена проблемам устранения расхождений между концептуализациями внешнего мира и концептуализациями информационных онтологий. Показано, что при конструировании онтологий для целей извлечения и хранения информации использует неестественные и ad hoc подходы, которые ведут к противоречащим интуиции последствиям. Предлагаются методы устранения указанного противоречия.

Ключевые слова: онтология, информация, интуиция, противоречие.

Построение информационных онтологий явилось шагом на пути к систематизации данных большого объема. Основу для этой работы положили конструирование архивных систем, систематизация данных, организация документов и вообще структуризация больших массивов данных. Подходы к системам записи данных и регистрации фактов были стихийными, в основном без привлечения систем искусственного интеллекта. Однако вскоре стало ясно, что чисто компьютерная обработка массивов данных не дает понимания содержания обрабатываемых документов. Компьютер выступал переносчиком соответствующей информации, но не потребителем. В результате значительное внимание было уделено построению WEB, ставшей архетипом успешного применения вычислительных возможностей в анализе информации. Успех такого подхода был обеспечен несколькими обстоятельствами, среди которых можно назвать наличие всепроникающей инфраструктуры Интернета и стандартных коммуникационных механизмов типа WEB-браузеров. Естественно, что количество усилий в направлении реализации этого проекта обеспечивалось огромной популярностью у потребителей информации, число которых росло практически в экспоненциальном по-

* Исследования, результаты которых отражены в данной статье, поддержаны Междисциплинарным интеграционным проектом СО РАН № 3 (2012–2014 гг.) «Принципы построения онтологии на основе концептуализации средствами логических дескриптивных языков».

рядке. Унифицирующие концепции (линки URL) все более усложнялись, но вскоре стало ясно, что такого рода механизм извлечения информации подходит к своему пределу из-за объема информации или числа данных.

Поиск выходов из положения велся на нескольких направлениях. Одно из них состояло в применении более тонких способов описания информации с использованием дескриптивных метаданных и в приспособлении этих методов к онтологическим структурам. На этом пути возникают проблемы, связанные с естественностью принимаемой онтологии, в том смысле, что она должна быть достаточно реалистичной. В противном случае никто не гарантирован от возникновения противоречий при операциях с онтологиями (погружение одной онтологии в другую, перевод и проч.). Однако более серьезная проблема связана с пониманием контекста, обрабатываемого компьютером.

Сам по себе поиск информации по какой-то конкретной теме является сложным процессом, который начинается с запроса по ключевым словам. Получаемая в результате информация содержит то, что можно назвать непонятными фрагментами, – незнакомые слова и концепции. Повторные обращения по поводу этих непонятных фрагментов расширяют наше видение того, что важно в этой проблематике, и все-таки мы лишены главного, а именно, фонового знания, лежащего в основе проблематики и определяемого интенциональностью сознания. Именно это возражение выдвигал против искусственного интеллекта известный философ науки Х. Дрейфус: «…Человеческий мир заранее структурирован в терминах намерений человека и направленности его интересов, причем таким образом, что именно направленность интересов определяет, что считать объектом или важной особенностью объекта. Здесь машина ничего не может сделать, потому что она в состоянии иметь дело только с уже определенными объектами…» [1].

В своей аргументации Дрейфус в значительной степени полагается на взгляды М. Хайдеггера. Последний считал, что практическая невозможность учета фонового знания является главной причиной невыполнимости «схватывания» ограниченным языком реальности. Действительно, описание того, как направленность человека упорядочивает его жизненный опыт, выделяя в нем соответствующие участки и зоны, мы встречаем в работах Хайдеггера. Дрейфус, следуя Хайдеггеру, утверждает: «У каждого предмета, служащего нам орудием, есть свой участок, свое место, где он лежит “под рукой”… Направленность интересов заключается именно в том, что задается вопрос “где” – где то место, в котором данный предмет окажется “под рукой”» [2].

Формальный язык является в высшей степени тем, что можно назвать ограниченными языковыми средствами. Языку вообще присущ тот процесс «омертвления» реальности, суть которого состоит в концептуализации, в делении процессов в соответствии с категориями. Более тонкое описание включает спецификацию концептуализации, означающей выделение базовых категорий при описании реальности. Именно это было положено в основу предложенного Т. Грубером определения понятия информационной онтологии [3].

Другими словами, поиск информации в рамках простого WEB-подхода с помощью большого числа гиперссылок не учитывает контекста, который определяется не только внешними особенностями, но и внутренними потребностями пользователя. Попытка специфицировать информацию конечным набором подходящих слов не приводит к успеху, поскольку компьютер просто не понимает того, что вы, собственно, от него хотите, и услужливо предлагает все, что у него есть в распоряжении. А при значительной по размерам базе данных в распоряжении есть много чего, что не нужно пользователю. У компьютера имеется доступ к огромному числу фактов, или информации, которая может стать знанием. Знание в этом отношении представляет собой использование концептуальных средств, таких как истина, вера, суждение, методология, технические навыки и пр. Реальный человеческий поиск информации предполагает в том числе использование подобного концептуального аппарата, но только при «включении» человеческой сообразительности по поводу того, какие дополнительные ключевые слова должны быть использованы, какая дополнительная информация должна быть введена для успешного поиска необходимой нам информации.

В рамках простого Web-подхода имеются проблемы и в использовании найденной информации. Информационные страницы ориентированы прежде всего на человеческое восприятие, их содержательная часть перемежается с элементами оформления, страничного группирования нужной и не нужной пользователю информации. При таком подходе затруднено использование программ-ассистентов, поскольку смысл информации выявляется косвенно, через человеческое восприятие, а не через формализованную семантику.

Понимание как когнитивный процесс связано с семантикой, т.е. с тем, как термины языка соотносятся с объектами внешнего мира. Семантическое знание информации и есть, по сути, понимание, и главной задачей в построении успешных систем хранения и извлечения информации является достижение того, чтобы сам компьютер мог совершать

те же модусы вывода, что и человек. Именно такие усилия привели к дополнению WEB семантическим знанием. Родилась концепция Semantic WEB, которая ныне быстро проникает в массу технологических приложений.

Естественно, что сама по себе задача обеспечения «доступа» компьютера к знанию, т.е. доступа к семантике, решается только в том случае, если найден способ представления знания в таком виде, который может быть подвержен компьютерной обработке. Впрочем, этот процесс и называется теперь представлением знания, хотя следовало бы резервировать данный термин для более широких философских контекстов, связанных с общей проблемой ментального представления реальности сознанию [4].

Действительно, поначалу термин «ментальное представление» употреблялся в когнитивных исследованиях как некоторого рода теоретический конструкт в рамках так называемой Вычислительной Теории Ума, согласно которой когнитивные состояния заключаются в преобразовании, хранении в мозгу информационных структур (которые и являются представлениями) той или иной природы. В конечном счете эта концепция восходила к концепции «ментального театра» Декарта, согласно которой уму субъекта представлены мысли и ощущения самого этого субъекта [5]. В этом смысле представление есть объект с семантическими свойствами, которые позволяют соотносить его с предметами, понятием истинности утверждений, содержанием. Тогда нет необходимости соотносить ментальные представления только с компьютерными процессами, и, на самом деле, проблема репрезентации, или ментального представления, стала вполне самостоятельной темой в чистой философии [6]. Тем не менее апелляция к онтологии является частью общей проблемы представления знания.

Если такое представление знания окажется достаточно удачным, тогда компьютер может получать выводное знание, которое будет уже новым. Такой его характер обязан тому, что мы имеем незамкнутую семантическую структуру – в отличие от замкнутого дедуктивного знания. Семантическая структура, о которой идет речь в данном контексте, являет собой семантическую сеть. Даже в относительно простых случаях семантическая сеть – это граф с весьма сложными связями. При анализе осязаемо сложных контекстов взаимосвязь понятий представляет собой нечто совсем невообразимое. Показательным примером может служить небольшой фрагмент из «семантической сети» содержания книги о логике и искусственном интеллекте (рис. 1, [7]).

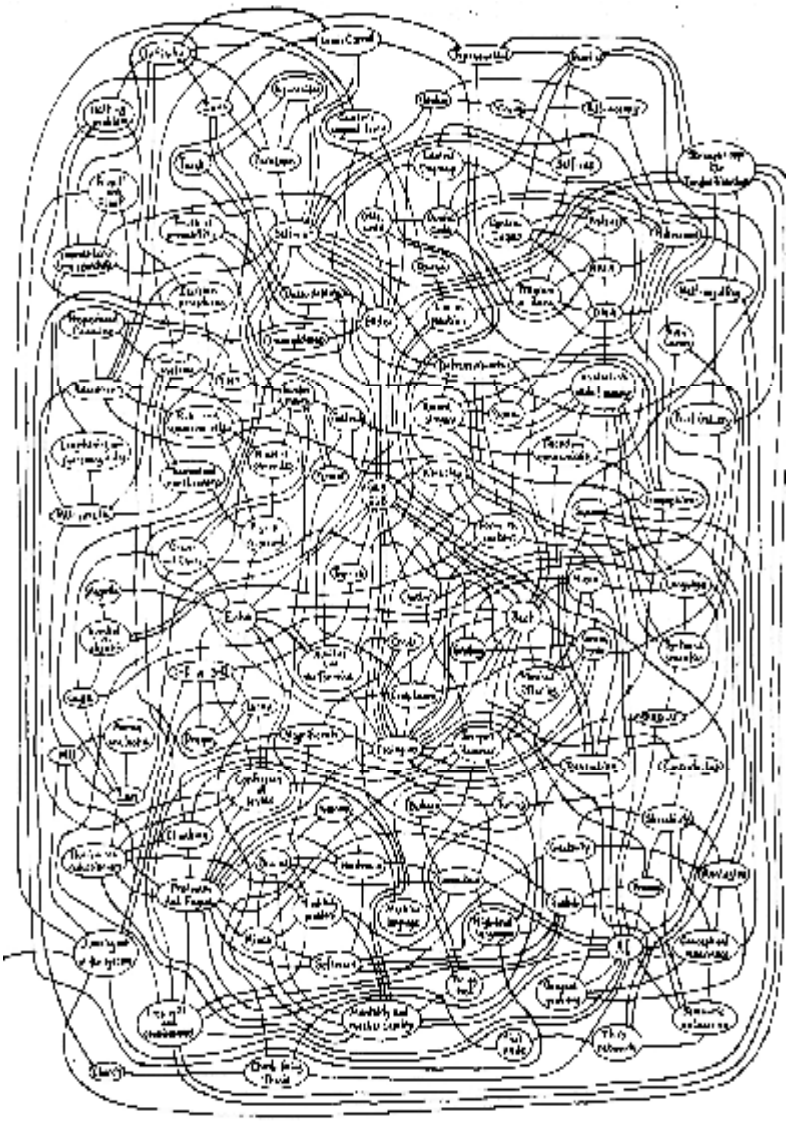

Рис. 1

Представление знаний для технологических целей может иметь менее причудливый, чем в приведенном примере, характер, но при этом должна быть более отчетливой онтология, т.е. система понятий с отношениями над ними. И здесь будет возможен автоматический вывод с получением на компьютере новых знаний. Соответствующая онтология должна быть понятной как производителям, так и пользователям знаний, что влечет за собой необходимость приемлемой для обеих сторон концептуализации. Другими словами, должно быть согласие относительно определенных иерархий вещей, составляющих онтологию. Так, в работе К. Уолтона приведена онтология для фотокамер (рис. 2, [8]).

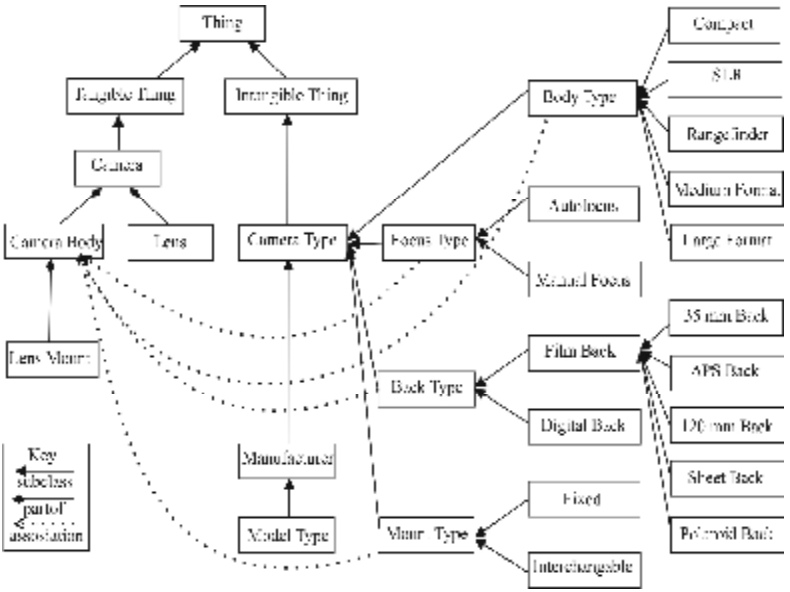

Рис. 2

Во главе иерархии стоит понятие вещи, а второй уровень онтологии включает в себя ощутимые (tangible) и неощутимые (intangible) вещи. К первым относятся материальные предметы, а ко вторым – скорее то, что можно назвать пропозициональными установками. Более конкретно, к третьему уровню иерархии, к ветви с ощутимыми вещами относится материальный предмет, а именно, камера, которая состоит из деталей,

составляющих остальные уровни онтологической иерархии. Что касается вещей неощутимых, то к ним относятся тип камеры, производитель, форматы съемки, тип фокусирования и проч.

На этом примере хорошо видно различие между «реальной» онтологией и онтологией «суррогатной». С точки зрения реальной онтологии, например, тип фокусирования определенно включает в себя «осязаемые» вещи, но в данной системе они относятся к вещам «неосязаемым». Дело в том, что здесь принята такая концептуализация, которая удовлетворяет определенным целям. Вполне возможно, что подобная концептуализация решила частную проблему, но «онтология» осталась, и этот стихийный процесс создания онтологий привел к беспорядку, препятствующему созданию подлинно фундаментальной онтологии для больших баз данных. Наличие большого числа «суррогатов» приводит ко многим несогласованиям и, возможно, противоречиям при объединении онтологий. Кроме того, использование отношений «подкласс», «часть», «ассоциирование» зачастую при переходе к другим онтологиям приводит к неестественным результатам, о которых будет сказано ниже.

Понятие онтологии, заимствованное из философии, относится к характеристике того, что существует реально. Вопрос о существовании с точки зрения логики является сложным [9], но для наших целей можно считать, что существует все, что может быть представлено. Формальная онтология определяет множество объектов и отношения между ними, и в этом смысле она есть просто утверждение логической теории. Специфичность ранее принятых онтологий могла быть преодолена в попытках интеграции только с появлением основополагающей статьи об онтологии. Интересно отметить, что сама статья, разъясняющая концепцию Semantic WEB, появилась не в специализированном издании, а в научно-популярном «Scientific American».

Определение онтологии осуществляется в различных языках, среди которых особенно распространенными являются RDF, RDFS и OWL, использующие несколько синтаксисов, в том числе XML, с разной выразительной силой. В настоящее время эти стандарты представления данных являются доминирующими, и в этом смысле разговоры о едином информационном пространстве стали реалистичными. В частности, RDF используется как формат для представления структурирования данных, а OWL – как формат для представления собственно онтологий. В 2004 г. на эти форматы были приняты стандарты, а в 2008 г. форматом SPARQL был стандартизирован язык запросов к RDF.

Как уже отмечалось, с некоторых пор создание формализованных спецификаций (онтологий) носит стихийный характер. Очень многие онтологии содержат в себе определения и предполагают использование сущностей традиционной природы: персон, организаций, событий, географических расположений и т.д. При этом либо соответствующие спецификации таких сущностей вставляются в онтологические построения, либо используются уже существующие. Эта ситуация с неупорядоченным «онто-творчеством», контрастирует с процессами стандартизации, происходившими в 90-х годах, когда международные группы тщательно выверяли решения, ориентированные на общее, межсистемное использование. Например, одно из наиболее известных формализованных определений набора полей описания метаданных Dublin Core, расширяющее классический библиографический подход, формировалось с 1995 г., и его формирование до сих пор не закончено [10].

При этом за первые 10 лет удалось стандартизировать лишь 15 элементов набора. В такой замедленности процесса создания системы базовых понятий и определений, видимо, кроется принципиальный момент – отсутствие единого и эффективного подхода к формированию онтологических построений даже для «обыденной» картины мира.

Существующее множество онтологий, используемое для спецификации накапливаемых формализованных данных, уводит от решения одной из главных задач, провозглашенных в концепции Semantic WEB, – задачи интеграции данных. Дело в том, что конкретные онтологии дают описание предметной области каждая в своей системе понятий, а сводимость одной системы к другой отсутствует. И формальная система установления таких соответствий слишком скудна, для того чтобы можно было бы определять такое соответствие для типовых содержательных случаев. Предложенный стандартом OWL набор средств для установления отношений между онтологиями состоит из утверждений типа «эквивалентный класс», «эквивалентное свойство», «подкласс», «подсвойство». Впрочем, этот раздел рекомендаций консорциума WWW быстро развивается.

Понятно, что устанавливать отношения между онтологиями «каждый с каждым» неэффективно, а авторитетного общего «эсперанто», т.е. базовой онтологии, не существует. Пока авторитетность онтологии формируется по массовости использования данных, структурированных с ее помощью, а не по качеству онтологического построения. В качестве примера такой популярной онтологии, рассмотрим некоторые моменты, связанные с онтологией DBpedia Ontology [11].

Информационный ресурс (база данных) dbpedia является одним из самых больших в мире описателей данных (энциклопедических) общего назначения. На текущий момент база данных содержит более 4 млн сущностей разных классов, включая 832 тыс. персон, 639 тыс. географических объектов, 372 тыс. творческих произведений (аудио, видео, игры), 209 тыс. организаций. Эта база данных является переработкой Википедии (http://www.wikipedia.org/), но переработкой не ручной, а автоматической. В силу этого обстоятельства в базе данных зафиксирована не вся полезная информация из текстов энциклопедических статей, а только ее небольшая часть, содержащаяся в формализованных «абстрактах» к статьям. Всего в базе данных в учетом языковых вариантов почти 2,5 млрд RDF-триплетов. DBpedia используется большим количеством баз данных для «привязки» «своих» данных к общему информационному полю. Онтология DBpedia_3.9 – последний на настоящий момент времени вариант онтологии, используемой в базе данных и во множестве других информационных ресурсов. Онтология выражена средствами OWL, содержит описание 529 классов, 2333 видов свойств.

В данной онтологии есть немалое количество понятий – классов и свойств, не очевидных по семантике и возможностям относительно применения. Рассмотрим некоторые из них.

В большом количестве случаев дерево классов сущностей используется для группирования по какому-то вторичному признаку. Например, от класса «персона» наследуются конкретные профессии или социальный статус, такие как «военный», «священник» и др. Принципиальное обилие подобных градаций, не позволяет онтологически определить все существующие варианты, поэтому в онтологии присутствуют лишь популярные группы, такие как «политик», «футболист», «живописец», но нет большинства «малозначимых» профессий или видов деятельности. В ряде случаев в онтологии вводятся понятия уровня базы знаний, такие как класс «имя», класс «год» и др. Такое введение является фрагментарным, и непонятно, как использовать подобные определения. В некоторых, часто важных случаях отношение вместо прямой ссылки определяется как текстовое свойство (DatatypeProperty), как в случае определения участника для события. Совсем неуместными выглядят определения широко известных понятий, но делаются они для очень частных случаев. Например, свойство «эпоха» определено для планет, а «конечный пункт» (EndPoint) – только для водных каналов. Имеется очень много определений, специфич-

ных для конкретной культуры (США) понятий, которые неактуальны или имеют в других культурах иной смысл (usSales, football).

Еще один вариант группирования, используемый в онтологии DBpedia, также вызывает неприятие. Для класса «событие» определены два отношения – previousEvent и followingEvent, отсылающие от события к предыдущему и последующему событиям соответственно. Семантическая проблема, связанная с таким построением, заключается в том, что событие может принадлежать лишь к одной цепочке следования, иначе возникает путаница. При использовании данной схемы структуризации делается попытка решить задачу группирования совместно с задачей упорядоченности элементов группы. Подобные построения встречаются и в других онтологиях. Например, могут быть введены отношение «следующий чемпион мира по шахматам» и ему обратное.

Типовая проблема, касающаяся почти всех используемых онтологий, – отсутствие атрибутов у отношений. Например, у отношения «работник» (А работает в организации Б) отсутствуют данные о том, с какого и по какое время человек работал, в какой должности. Получается, что человек всегда работал и до сих пор работает там. Если указывается, что персона где-то проживает, то аналогично невозможно проследить временную упорядоченность мест проживания. Проблема упирается в то, что используемые в RDF/OWL отношения ObjectProperty не содержат в себе атрибутов, а намеченный в RDF механизм reification логически не очевиден и редко используется в структурных построениях.

За редкими исключениями, подобные сомнительные свойства онтологических построений не носят технологического характера. Это и понятно: система структуризации – бинарные предикаты первого порядка, представляющие собой универсальный формализм. Например, проблему атрибутирования свойств можно решить, вводя вместо большинства прямых отношений составные отношения или псевдосущности, структурно являющиеся сущностями и отсылающие прямыми ссылками к соответствующим объектам. Это было сделано в онтологии BONE (Basic Ontology on Nonspecific Entities) [12]. В ней, например, имеется составное отношение participation, отсылающее ссылкой participant к субъекту отношения и ссылкой in-org к организационной системе, в которой участвует данный персонаж. А сама псевдосущность participation атрибутируется и временными точками, и видом участия (первое лицо, организатор, участник, гость и др.), и ролью.

Для случаев, когда требуется группирование или упорядоченное группирование, это несложно сделать либо с помощью множества исхо-

дящих ссылок (отношений), либо с помощью множества входящих ссылок (отношений), либо посредством использования RDF-контейнеров rdf:Bag и rdf:Seq, либо вводя классы коллекций и отношения «член коллекции», как это сделано в BONE. Создается впечатление о некомпетентности создателей некоторых онтологий. Возможно, такие онтологии сначала создавались в узкоутилитарных целях (например, описать формализованную часть Википедии), а общезначимость ресурса придает этим построениям более общий смысл. Фактически, чисто инженерные разработки становятся по чистой случайности стандартом в конструировании онтологий. Естественно, что при этом теряется соответствие между «естественными» концептуализациями внешнего мира, и «искусственными» концептуализациями мира артефактов.

Вопрос о соотношении двух типов концептуализации является довольно сложным, поскольку вторжение в деятельность человека так называемых симулякров – «копий», не имеющих оригиналов в реальности, ведет к представлению о гиперреальности, или же виртуальной реальности [13]. В какой степени гиперреальность будет существенным фактором в технологическом развитии, предсказать трудно. Однако существенное расхождение между концептуализациями, которые в той или иной степени соотносятся с внешней средой, и концептуализациями, которые опираются на все большее введение в оборот искусственных онтологий, вызывает озабоченность, поскольку перенаселенность артефактов или симулякров гиперреальности может оказаться фактором, отнюдь не способствующим прогрессу ни в технологической, ни в социальной сфере.

Примечания

1. *Дрейфус Х.* Чего не могут вычислительные машины: Критика искусственного интеллекта. – М., 1978. – С. 228.

2. Там же. – С. 244.

3. *Gruber T.* Ontology // Encyclopedia of Database Systems / Eds. Ling Liu, etc. – Springer Verlag, 2008.

4. См., например: *Field H.* Mental representations // Erkennntnis. – 1978. – V. 13. – P. 9–61.

5. См.: *Рорти Р.* Философия и зеркало природы. – Новосибирск, 1997.

6. См.: *Mental* representation. – URL: www.plato.stanford.edu.

7. См.: *Hofstadter D.* Godel, Escher, Bach. – N.Y., 1979. – P. 370.

8. См.: *Walton C.* Agency and the Semantic WEB. – Oxford University Press, 2007. – P. 4.

9. См., например: *Целищев В.В.* Логика существования. – М., 2010.

10. *Dublin* Core Metadata Initiative. – URL: http://dublincore.org.

11, *The DBpedia* Ontology (3.9). – URL: http://wiki.dbpedia.org/Ontology39?v=g9b

12. См.: *Марчук А.Г., Марчук П.А*. Архивная фактографическая система // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XI Всерос. науч. конф. (RCDL-2009). – Петрозаводск, 2009.

13. Термин введен Ж. Бодрийяром (см.: *Бодрийяр Ж*. Симулякры и симуляция. – Тула, 2013).

Дата поступления 11.11.2013 г.

Институт систем информатики
СО РАН, г. Новосибирск

Институт философии и права
СО РАН, г. Новосибирск

mag@iis.nsk.su
director@philosophy.nsc.ru

Marchuk, A.G. and V.V. Tselishchev. **On divergence of information ontologies and conceptualizations of the external word**

The paper deals with the problems of elimination of divergences between conceptualizations of the external world and those of information technologies. It shows that when constructing ontologies in the purpose of data mining and storage unnatural and ad hoc approaches are used which result in consequences contradicting intuition. The authors suggest methods how to eliminate this contradiction.

Keywords: ontology; information; intuition; contradiction