УДК 004.9
DOI: 10.15587/2312-8372.2018.123527

РАЗРАБОТКА МЕТОДОЛОГИИ ПРОЕКТИРОВАНИЯ ХРАНИЛИЩ ДАННЫХ И ЗНАНИЙ КОМПЬЮТЕРНЫХ СИСТЕМ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Сагайда П. И.

1. Введение

В настоящее время необходимы разработка и исследования моделей и методик проектирования эффективных хранилищ данных и знаний (ХДиЗ) современных компьютерных систем для интеллектуальной обработки данных (КС для ИОД) на предприятиях и в организациях [1, 2]. Эти модели и методики должны быть объединены в сквозную методологию на основании математически обоснованного подхода, который обеспечит доказуемость и обоснованность результатов проектирования. При этом нужно учитывать накопленные знания о предметной области (ПрО) у специалистов и инженеров по знаниям, которые не располагают большим опытом проектирования КС, однако могут сформулировать тезаурус или онтологию, описывающие концепты и отношения между ними. Вместе с тем, онтологическая модель, конструируемая в общем виде на основе накопленных сведений о работе предметной области, обладает рядом существенных недостатков. При ее проектировании аналитик не обязан руководствоваться формальными правилами и ограничениями. Несмотря на то, что в настоящий момент разработаны ряд рекомендаций [3, 4], их ценность зависит от эффективности применяемых практик онтологического моделирования и является субъективной. Важнейшей возможностью внести в процесс инженерии знаний математические основы, проверку результатов и доказательную силу топологических шаблонов проектирования обладает теория категорий (ТК). Математический аппарат данной теории позволяет разрабатывать коммутативные диаграммы, в том числе для формализации знаний [5, 6]. Логическое развитие ТК – теория скетчей [7, 8] – дает возможность разрабатывать математические конструкции в составе графов, теорий и категориальных моделей.

Архитектура проектируемых ХДиЗ в составе КС должна обеспечивать требования теории проектирования баз данных (БД) [9], которая подразумевает соответствие отношений реляционной модели данных ограничениям, накладываемым на них нормальными формами. Вместе с тем, навыки и умения членов проектной команды в том числе связаны с графической формализацией результатов системного анализа с помощью соответствующих диаграммных методик. Такая формализация знаний о ПрО дает возможность наглядно представить полученные в ходе анализа сведения, обеспечить эффективное взаимодействие с коллегами и заказчиками, перейти к следующим, более формальным стадиям разработки КС.

Поэтому актуальным является исследование возможностей и недостатков современных подходов к формализации знаний о ПрО, диаграммных методик информационного и даталогического моделирования при разработке ХДиЗ КС для ИОД. Результатом таких исследований является разработка методологии проектирования, обобщающая имеющиеся достижения в этой области на основе единого математического аппарата.

2. Объект исследования и его технологический аудит

Объектом исследования является разработка методологии проектирования эффективных хранилищ данных и знаний современных компьютерных систем для интеллектуальной обработки данных с использованием разработанного автором категориально-онтологического (КО) подхода [10]. Использование КО моделей как мета-метаязыка моделирования предоставляет средство для верификации процесса проектирования ХДиЗ и его результатов. Аналитики ПрО и разработчики ХДиЗ в ходе создания КС для ИОД обычно используют на различных этапах работы разрозненные методики и технологии проектирования и реализации компонентов КС. Эти методики и технологии, основанные на разнородных подходах к формализации и моделированию аспектов ПрО, будут рассмотрены ниже. При этом возникают семантические и лингвистические барьеры при восприятии ПрО, понимании задач и методов их решения, с точки зрения выбора эффективной диаграммной методики при проектировании. Схематически природа и топология возникающих барьеров представлены на рис. 1. Кроме того, проблемой также являются и различия в форматах и технологиях хранения и обмена данными, в моделях построения запросов к ХДиЗ.

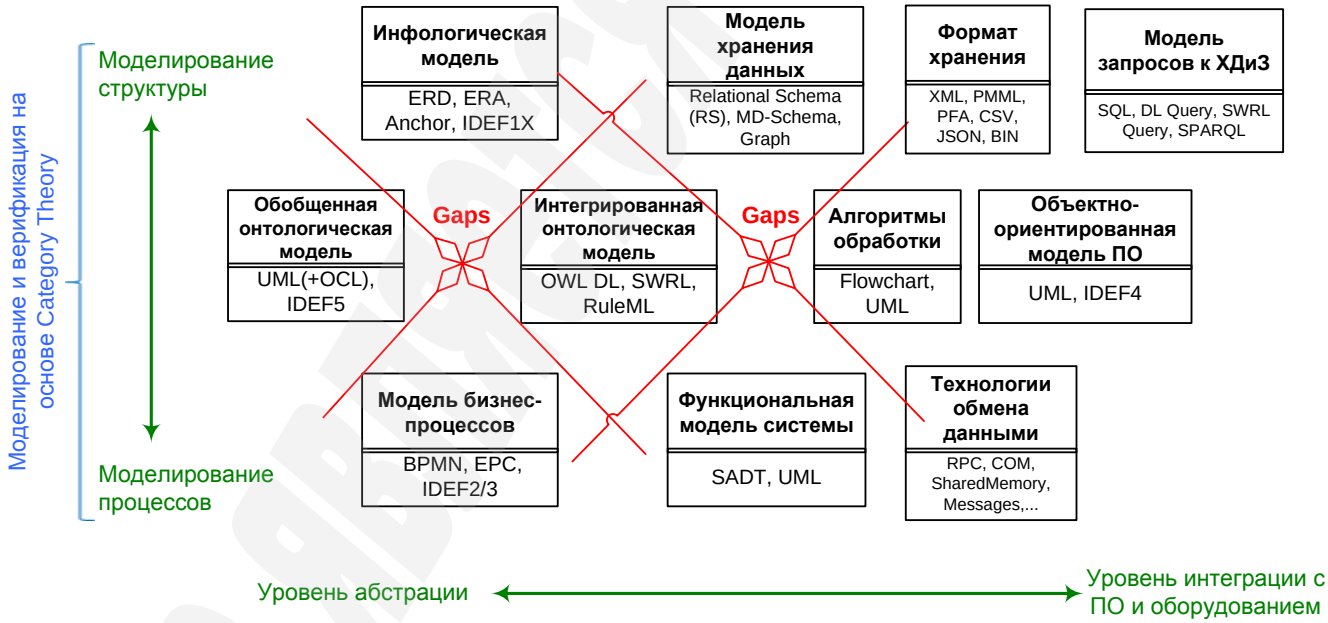

Рис. 1. Решение проблемы семантических и лингвистических барьеров (gaps) на этапах реализации и в процессе функционирования компьютерных систем для интеллектуальной обработки данных на предприятиях и в организациях

Главным недостатком нынешней теории и практики проектирования ХДиЗ является то, что у заказчиков, аналитиков и разработчиков нет возможности использовать одну сквозную методологию, пригодную для коллективной работы и обмена знаниями. Такая методология должна быть понятной и объективной с точки зрения доказуемости и обоснованности моделей, быть пригодной для полного комплекса исследовательских и проектных работ. Модели, создаваемые на различных этапах проектирования и с помощью разных методик, должны, в рамках такой методологии, отображаться друг в друга без потерь. Этим требованиям в настоящее время не удовлетворяет ни одна применяемая методика сама по себе. Необходима их интеграция на основе математической теории и с использованием методов инженерии знаний, в частности, онтологического подхода.

3. Цель и задачи исследования

Целью данной работы является совершенствование процесса проектирования ХДиЗ путем разработки методологии их проектирования на основе онтологического подхода к моделированию предметной области автоматизации обработки и анализа данных.

Для достижения поставленной цели необходимо:

1. Проанализировать существующие модели и методики формализации знаний о ПрО работы КС для ИОД и проектирования ХДиЗ, выполнить их классификацию, определить особенности и возможности интеграции.
2. Разработать методологию проектирования ХДиЗ и ее формальную категориально-онтологическую (КО) модель, на основе которой обосновать требуемый порядок применения методик, вошедших в единую методологию.

4. Исследование существующих решений проблемы

Традиционные подходы к концептуальному моделированию и проектированию ХДиЗ могут быть классифицированы как такие, что формализуют, в основном, отношения в ПрО (Relation-oriented Approach – RoA), атрибуты ПрО (Attribute-oriented Approach – AoA) и объекты (Object-oriented Approach – OoA). Причем большинство технологий проектирования имеют смешанную природу. Так, технология с использованием Entity-Relationship-Attribute (ERA) Diagrams [11] и технология Anchor [12] представляют комбинированный (RoA-AoA) подход, и только проектирование на основе концепции функциональных зависимостей (Functional Dependencies Diagrams) является ориентированным только на атрибуты (AoA). Онтологическая модель этой группы технологий, более ориентированных на представление атрибутов в моделях, представлена на рис. 2.

Более высокоуровневыми технологиями моделирования ПрО, используемыми также и при проектировании хранилища знаний (ХЗ), однако использующими комбинированный (RoA-AoA) подход, являются модель «сущность-связь», Entity-Relationship Diagrams (ERD) [4, 9], и основа Semantic Web, фреймворк Resource Description Framework (RDF) [13]. Онтология этих технологий моделирования ПрО при проектировании ХДиЗ, более

ориентированных на представление отношений в моделях, представлена на рис. 3.

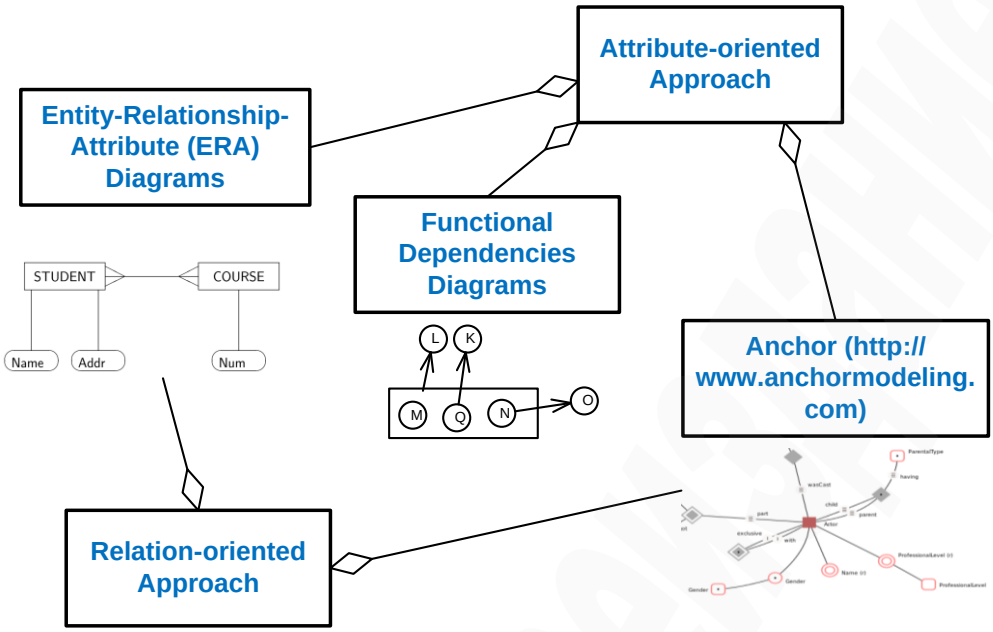

Рис. 2. Онтология технологий моделирования предметной области при проектировании хранилищ данных и знаний, более ориентированных на представление атрибутов

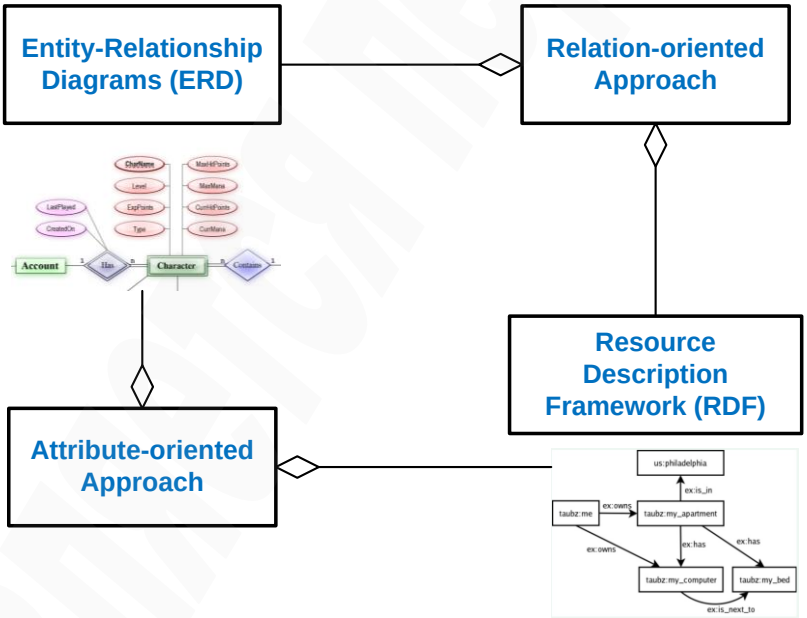

Рис. 3. Онтология технологий моделирования предметной области при проектировании хранилищ данных и знаний, более ориентированных на представление отношений

Исторически правильно одной из первых технологий моделирования ПрО при проектировании ХД следует считать Object Role Modeling (ORM). На основе ORM затем была развита (в сторону сокращения представляемых в моделях знаний) технология Integration DEFinition for information modeling

(IDEF1X), являющаяся составной частью семейства диаграммных методик IDEF для формализации различных аспектов ПрО. Результатом слияния большого количества однотипных подходов к моделированию и формализации, а также к проектированию аппаратно-программных комплексов явилась технология моделирования Unified Modeling Language (UML) [14], предоставляющая широкие возможности по моделированию классов, их экземпляров и отношений между ними. Однако для представления бизнес-правил и алгоритмов взаимодействия объектов в ПрО понадобилась разработка специализированного языка на основе UML-диаграмм – Object Constraint Language (OCL) [14]. Использование OCL позволило использовать для данной цели выражения, близкие по выразительности логике первого порядка, однако специфичные и сложные с точки зрения специалистов в ПрО. Онтология рассмотренных технологий моделирования ПрО при проектировании хранилища данных, ориентированная на представление объектов и отношений, приведена на рис. 4.

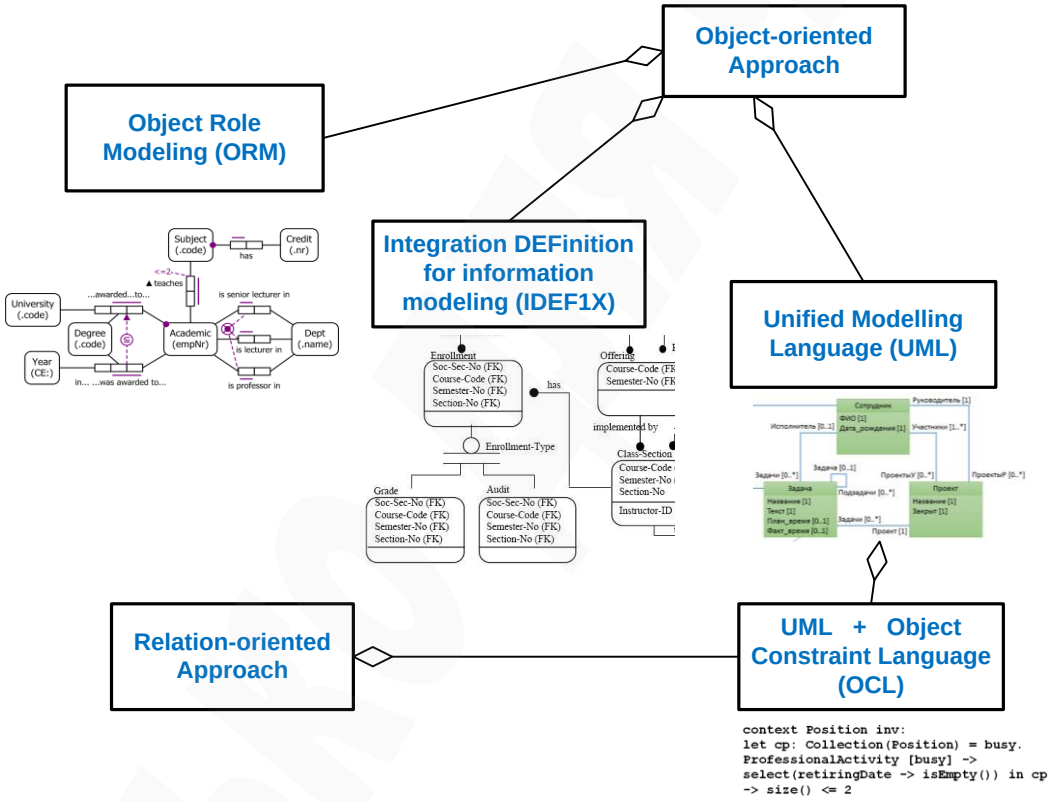

Рис. 4. Онтология технологий моделирования предметной области при проектировании хранилищ данных, ориентированных на представление объектов и отношений

Для повышения уровня абстракции при моделировании и автоматизации разработки КС, в последние годы разрабатывается и внедряется инициатива мета-объектных средств (Meta-Object Facility – MOF) [15]. MOF предполагает использование стандарта для разработки, управляемой моделями, подготовленного группой Object Management Group (OMG) на основе UML.

Анализ показывает, что данный подход является мета-моделью над UML и других полезных свойств для решаемых в данной работе задач не имеет.

Предпочтительными в качестве основы для разрабатываемой методологии являются технологии, основанные на категориальном подходе к моделированию ПрО, так как такой подход позволяет выполнить верификацию информационных моделей, обычно выражающих субъективную точку зрения их авторов, на основе формального математического аппарата. К данным технологиям, базирующимся как на представлении атрибутов, так и на подход теории категорий, относятся Olog Diagrams [6] и Sketches, являющиеся развитием технологии ERA [11]. Авторами данных технологий разработаны языки и средства автоматизации получения схем баз данных из диаграмм, соответственно Algebraic Query Language (AQL) и EASIC (graphical modeling of EA sketches and views). Онтология технологий моделирования ПрО при проектировании ХДиЗ, ориентированных на представление атрибутов и подход теории категорий, представлена на рис. 5.

Недостатком этих технологий является отсутствие в их составе средств представления правил для сложных ПрО работы предприятий и организаций.

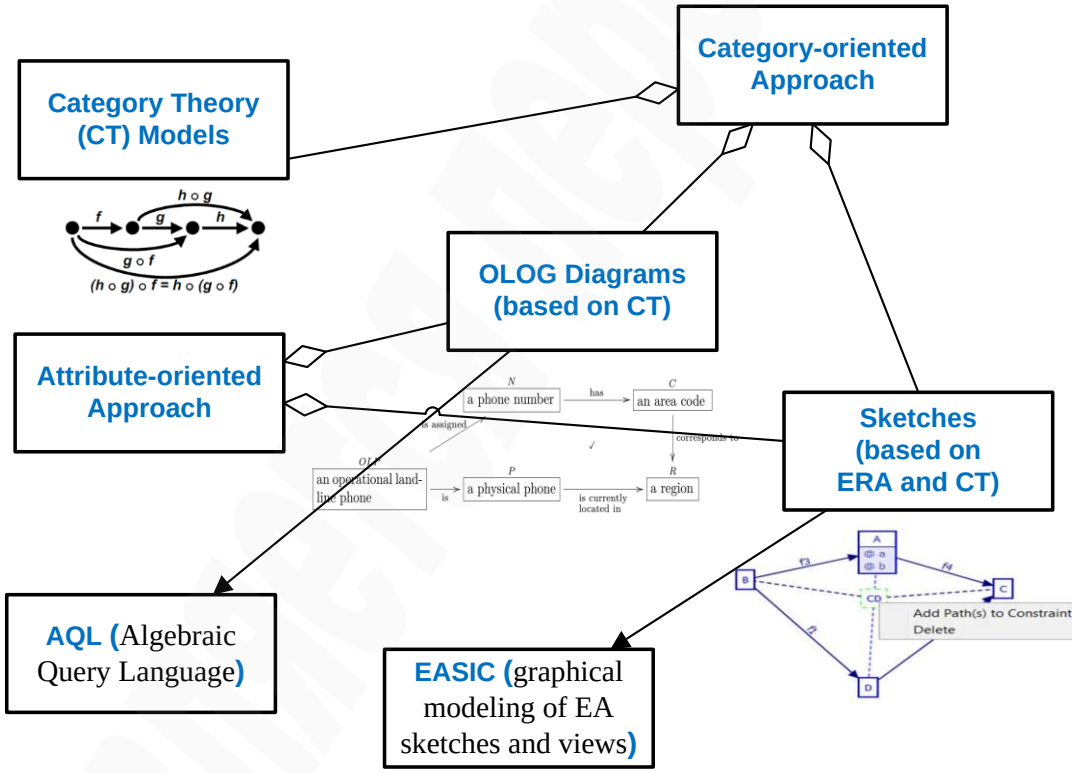

Рис. 5. Онтология технологий моделирования предметной области при проектировании хранилищ данных и знаний, ориентированных на представление атрибутов и подход теории категорий

Таким образом, для преодоления рассматриваемых семантических и лингвистических барьеров, интеграции разнородных методик проектирования на основе онтологических моделей, верифицированных на основе теории категорий, для построения эффективной, теоретически обоснованной методологии, на основе:

– возможностей теории категорий и скетчей [5–8];

– подхода к категориально-онтологическому моделированию ПрО и процессов, которые в них протекают [10];

– методики взаимного отображения диаграмм ER и FDs [4] для совместного использования информационных и даталогических моделей при проектировании БД;

– методики представления запросов к базам данных и знаний, обычно формализуемых на разнородных языках запросов (SQL, DL Query, SPARQL, SQueryWRL) [13], с помощью предметно-ориентированного языка на базе категориально-онтологических моделей.

5. Методы исследований

В данной работе для моделирования ПрО использован категориально-онтологический подход, при котором частная онтологическая модель представляется математически строго, путем наложения на представляемые концепты и связи ограничений объектов и морфизмов теории категорий. Представленная на рис. 6 диаграмма является коммутативной диаграммой теории категорий, построенной в соответствии с графом в теории скетчей [16]. Так как стрелки в теории категорий являются функциями-преобразованиями (отображениями) одних объектов в другие, то коммутативность означает равенство путей (композиций морфизмов), с помощью которых достигаются результаты преобразований.

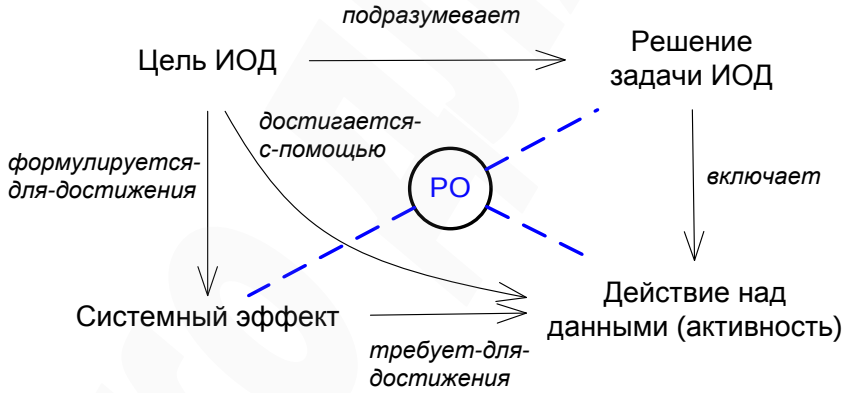

Рис. 6. Коммутативная диаграмма теории категорий для фрагмента онтологической модели решения задачи интеллектуальной обработки данных

Наложение требований коммутативности и использование понятий теории категорий позволило выявить для данной диаграммы топологический шаблон, математически описанный и обоснованный как необходимо присущий для данной семантики предметной области, а именно – pushout, обозначенный на диаграмме как «РО». Формулировка таких ограничений соответствует в теории скетчей разработке конусов и коконусов для диаграммы и их вводу в состав скетча [17].

Обобщенная схема разработанной автором методологии проектирования ХДиЗ на основе категориально-онтологических моделей приведена на рис. 7.

На рис. 7 представлены последовательно этапы преобразования частной онтологической модели в модель категориально-онтологическую (КО), т. е. модель ПрО, верифицированную на основе теории категорий. На основе такой КО модели выполняются дальнейшие проектные процедуры: разрабатываются совместные информационная (в виде диаграммы Entity-Relationship – ER) и даталогическая (в виде диаграммы функциональных зависимостей Functional Dependencies – FDs) модели. Такой совместный подход позволяет дополнить и обосновать обе диаграммы с использованием положений теории проектирования реляционных БД. А также построить на их основе концептуальную модель (реляционную схему) БД, близкую к оптимальной с точки зрения отсутствия проблем, ведущих к нарушению целостности данных при эксплуатации.

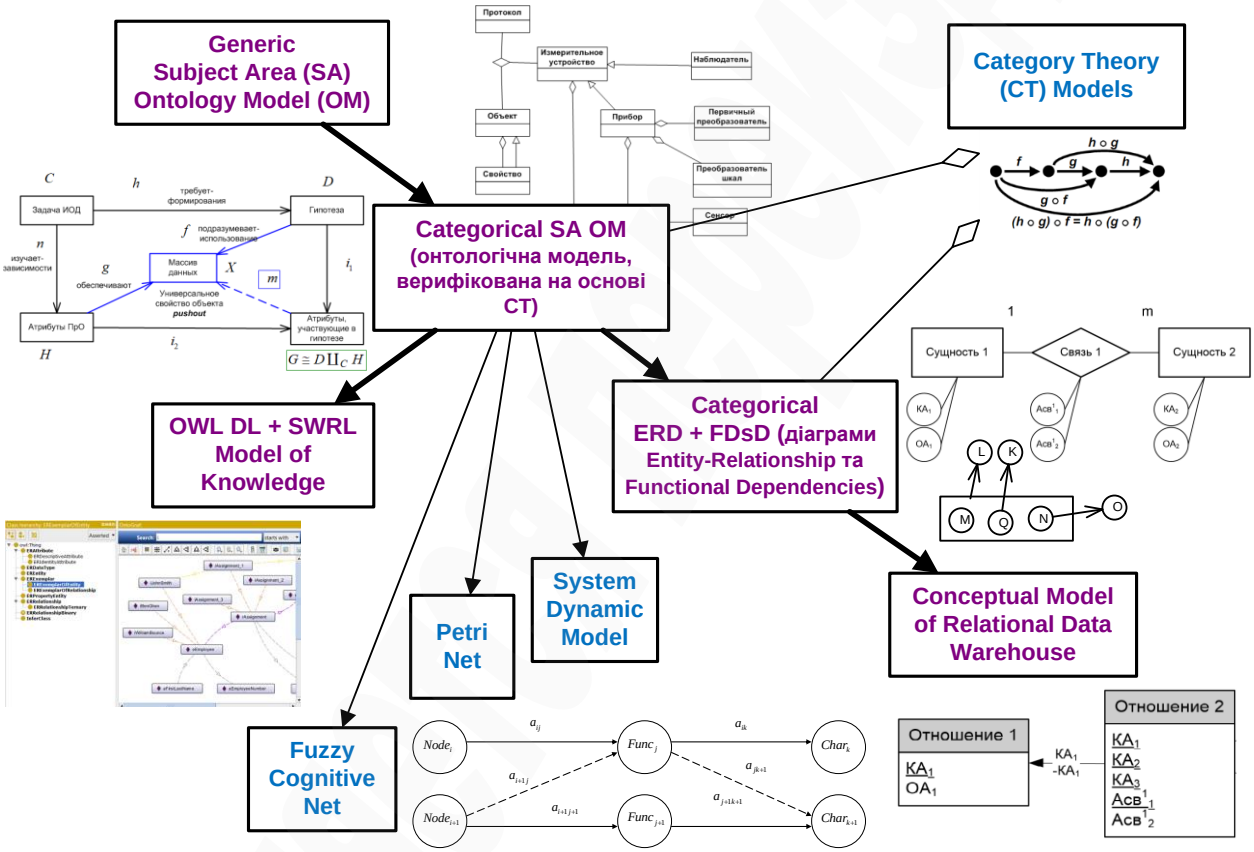

Рис. 7. Обобщенная схема методологии проектирования хранилищ данных и знаний для компьютерных систем а основе категориально-онтологических моделей

Кроме того КО модель ПрО позволяет непосредственно получить модель базы знаний ПрО для проектируемой КС, в виде аксиом OWL DL и правил на языке SWRL [13]. Получение концептуальных моделей ХДиЗ на основе КО модели ПрО гарантирует полноту и корректность представляемых сущностей (классов) и связей (отношений) ПрО. Общая схема методологии включает в себя также дополнительную возможность получения на основе КО модели редуцированных моделей для имитационного моделирования работы ПрО

(системно-динамических моделей, сетей Петри, нечетких когнитивных карт). Технология имитационного моделирования с помощью нечетких когнитивных карт разработана в [18].

6. Результаты исследований

Для категориально-онтологического представления разработанной методологии и последовательности преобразований, выполняемых над данными, разработана соответствующая модель, которая представлена на рис. 8.

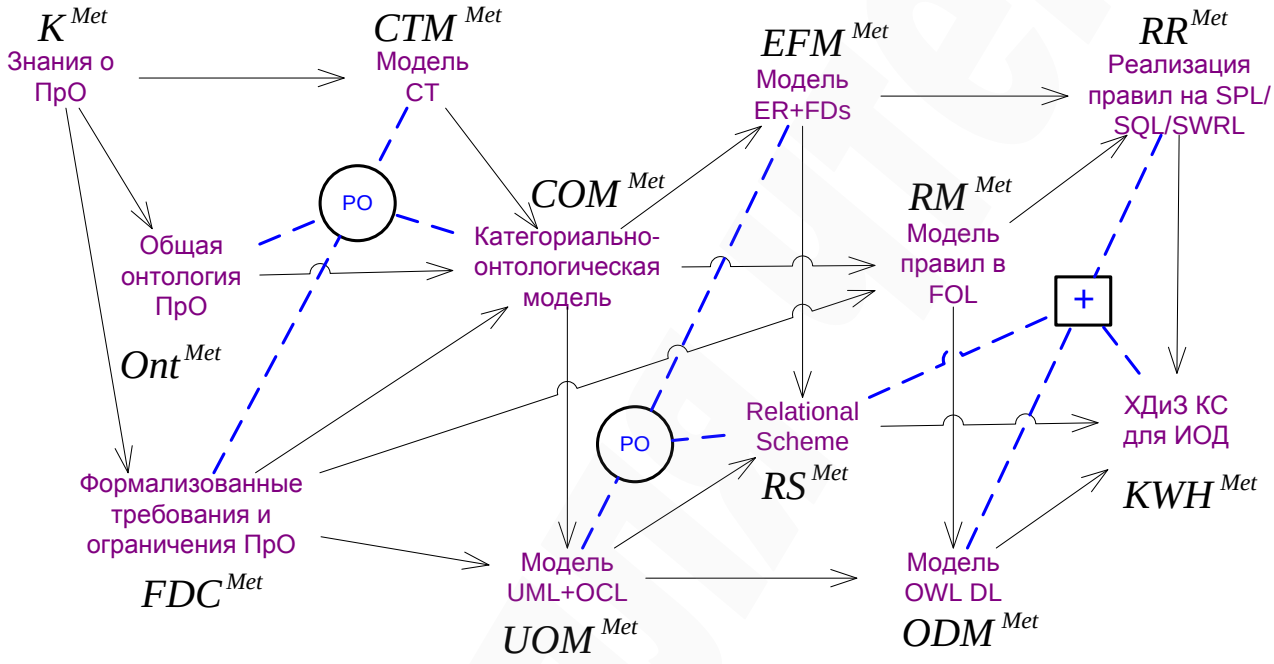

Рис. 8. Категориально-онтологическое представление разработанной методологии и последовательности преобразований, выполняемых над данными

Данная модель также представляет формализацию процедурных моделей ПрО (бизнес-логики) в ходе проектирования КС для ИОД, правил функционирования ПрО, требований и ограничений, накладываемых на содержимое ХДиЗ. Модель отображает следующие особенности реализации разработанной методологии. Категориально-онтологическая модель COM^{Met} является результатом вычисления объекта pushout (PO):

$$COM^{Met} \cong Ont^{Met} \coprod_{K^{Met}} CTM^{Met} \coprod_{K^{Met}} FDC^{Met}, \qquad (1)$$

который моделирует процесс отображения знаний о ПрО K^{Met} в модель COM^{Met} путем построения частной онтологии Ont^{Met}, верифицированной на основе модели теории категорий CTM^{Met} с учетом требований и ограничений (бизнес-логики) ПрО FDC^{Met}. Также модель на рис. 8 представляет процесс получения реляционной схемы хранилищ данных проектируемой КС, близкой к

оптимальной. Этот процесс реализован на основе совместного использования модели «сущность-связь» (ER) и модели на основе концепции функциональных зависимостей (FDs) EFM^{Met} и с использованием модели UML+OCL UOM^{Met}. Реляционную схему хранилищ данных получают из EFM^{Met} и UOM^{Met} на основе COM^{Met}, путем вычисления соответствующего объекта PO:

$$RS^{Met} \cong EFM^{Met} \coprod_{COM^{Met}} UOM^{Met}. \qquad (2)$$

Другие квадраты-грани куба в правой части модели не являются объектами PO, так как модель правил ПрО в виде утверждений логики первого порядка RM^{Met}, конструируется не только на основе модели UOM^{Met}, но и непосредственно на основе FDC^{Met}. Модель RM^{Met} затем используется при построении ХЗ, в частности, с использованием правил на языке SWRL или в виде алгоритмов обработки, реализуемых на Stored Procedure Language. Это ограничение объясняется тем, что не вся бизнес-логика ПрО может быть выражена средствами языка Object Constraint Language в модели UOM^{Met}.

Результатом применения разработанной методологии является построение хранилищ данных и знаний проектируемой КС со структурой, близкой к оптимальной, путем вычисления объекта coproduct:

$$KWH^{Met} \cong ODM^{Met} \coprod RR^{Met} \coprod RS^{Met}. \qquad (3)$$

Результат FDC^{Met} является завершающим этапом построения ХДиЗ для информационного обеспечения функционирования КС для ИОД, структура которых, хранящиеся в них, формализованная онтологическая модель ПрО и бизнес-правила сбора и обработки данных представляют накопленные знания о ПрО. Данный результат получен на основе категориально-онтологической модели COM^{Met} как узлового этапа моделирования и проектирования в рамках разработанной методологии.

7. SWOT-анализ результатов исследований

Strengths. Сильными сторонами исследования является то, что в результате разработки и применения данной методологии преодолены семантические и лингвистические барьеры при проектировании КС. У заказчиков, аналитиков и разработчиков появилась возможность использовать сквозную методологию, пригодную для коллективной работы и обмена знаниями, понятную и объективную с точки зрения доказуемости и обоснованности моделей. Данная методология может быть использована для полного комплекса исследовательских и проектных работ, независимо от форматов представления данных и диаграммных методик формализации знаний о ПрО. Категориально-онтологический подход к моделированию и проектированию дает возможность формально обосновать субъективные результаты инженерии знаний и

использовать объекты теории категорий в виде шаблонов проектирования на высоком уровне абстракции.

Weaknesses. Слабыми сторонами результатов данного исследования является повышение требований к уровню абстрактного мышления аналитиков, занимающихся онтологическим моделированием ПрО. При использовании разработанной методологии они должны овладеть математическими основами теории категорий и скетчей и навыками использования их положений при представлении результатов инженерии знаний. Как и при использовании любой методики формализации знаний, качественные результаты могут быть получены только на основе опыта длительного практического применения разработанной методологии.

Opportunities. Дополнительные возможности, возникающие при достижении цели данного исследования, заключаются в решении задач интеллектуализации компьютерных систем, повышении качества разработки и эффективности использования таких систем на основе методов инженерии знаний. Это позволило разработать производительные КС для ИОД, выполняющие задачи оперативной и релевантной обработки данных, эффективной инженерии знаний и извлечения адекватных моделей из массивов данных. Компоненты таких КС внедрены на ряде промышленных предприятий, в научную деятельность и учебный процесс.

Threats. Сложности во внедрении полученных результатов в практику проектирования КС связаны со следующим фактором. Специалисты, применяющие какую-либо методику или технологию проектирования длительное время и получающие удовлетворительные результаты, могут иметь соответствующие предубеждения к разработанной методологии и сомневаться в ее практической полезности. Однако растущая сложность КС для ИОД и решаемых с их помощью задач, повышение конкуренции в области информационных технологий, требуют освоения новых подходов к проектированию, основанных на современных достижениях научной мысли.

8. Выводы

1. Выполнен сравнительный анализ и онтологическое моделирование подходов и методик, применяемых в ходе проектирования хранилищ данных и знаний в составе компьютерных систем. Определено, что модели, создаваемые на различных этапах проектирования и с помощью разных методик, не позволяют выполнить их отображение без потерь. Требуется разработка и применение сквозной методологии, пригодной для коллективной работы и обмена знаниями, интегрирующей существующие методики и основанной на едином математическом аппарате.

2. Разработана методология проектирования хранилищ данных и знаний для решения задач обработки и анализа данных на основе категориально-онтологических моделей. Данная методология объединяет, в отличие от существующих методик, проектирование с помощью информационных и даталогических моделей с использованием различных диаграммных методик и языков моделирования. Такой подход позволил устранить недостатки и

дополнить достоинства различных подходов к проектированию и получить структуру хранилищ, близкую к оптимальной.

В результате разработки и применения данной методологии преодолены семантические и лингвистические барьеры, возникающие между членами проектной команды в ходе проектирования КС. Использование категориально-онтологического подхода к моделированию и проектированию дает возможность формально обосновать субъективные результаты инженерии знаний и использовать объекты теории категорий в виде шаблонов проектирования на высоком уровне абстракции.

Литература

1. Data mining: practical machine learning tools and techniques. Ed. 2 / ed. by Witten I. H., Eibe F. Burlington: Morgan Kaufmann Publishers, 2005. 525 p.
2. Sahaida P. I. Modelirovaniye problemnoy oblasti komp'yuterizirovannykh informatsionnykh sistem dlya intellektual'noy obrabotki dannykh s ispol'zovaniyem inzhenerii znaniy // Naukovi pratsi DonNTU. Seriya: Obchislyuval'na tekhnika ta avtomatizatsiya. 2017. Vol. 1 (30). P. 78–87.
3. Palagin A. V., Kryvyy S. L., Petrenko N. G. Ontologicheskiye metody i sredstva obrabotki predmetnykh znaniy: monograph. Lugansk: ENU named after V. Dalya, 2012. 324 p.
4. Sahaida P. I. Ontologicheskiy podkhod k proyektirovaniyu baz dannykh informatsionnykh system: proceedings // Sovremennoye obrazovaniye i integratsionnyye protsessy. Kramatorsk: DGMA, 2012. P. 313–318.
5. Walter R. F. C. Categories and Computer Science. Cambridge: Cambridge Universities Press, 1991. 166 p.
6. Spivak D. I. Category theory for the sciences. MIT Press, 2014. 435 p.
7. Barr M. Models of sketches // Cashiers Topologie Geom. Differentielle. 1986. Vol. 27. P. 93–107.
8. Wells C. A generalization of the concept of sketch // Theoretical Computer Science. 1990. Vol. 70, No. 1. P. 159–178. doi:10.1016/0304-3975(90)90158-e
9. Date C. J. An Introduction to Database Systems. Ed. 8. Pearson, 2003. 1024 p.
10. Sahaida P. I. Kategorial'no-ontologicheskoye modelirovaniye intellektual'noy obrabotki dannykh dlya matematicheskogo obosnovaniya rezul'tatov inzhenerii znaniy // Vimiryuval'na ta obchislyuval'na tekhnika v tekhnologichnikh protsesakh. 2017. Vol. 4. P. 149–158.
11. Johnson M., Rosebrugh R., Wood R. J. Entity-relationship-attribute designs and sketches // Theory and Applications of Categories. 2002. Vol. 10, No. 3. P. 94–112.
12. About Anchor Modeling. URL: http://www.anchormodeling.com (Last accessed: 25.12.2017).
13. Ontology Management: Semantic Web, Semantic Web Services, and Business Applications / ed. by Hepp M. et al. Springer, 2007. 293 p.

14. Larman, C. Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development. Addison Wesley Professional, 2004. 736 p.

15. ISO/IEC 19502:2005 (Information technology – Meta Object Facility (MOF). 2005. URL: http://webstore.iec.ch/preview/info_isoiec19502%7Bed1.0%7Den.pdf (Last accessed: 25.12.2017).

16. Wells C. Sketches: Outline with References. 2009. URL: http://www.cwru.edu/artsci/math/wells/pub/pdf/Sketch.pdf (Last accessed: 25.12.2017).

17. Wojtowicz R. L. A Categorical Approach to Knowledge Management. Computational Category Theory Workshop. National Institute of Standards and Technology. September 29, 2015. URL: http://www.bakermountain.org/talks/nist.pdf (Last accessed: 25.12.2017).

18. Sahaida P. I. Informatsionnaya tekhnologiya i programmno-metodicheskiy kompleks dlya modelirovaniya slozhnykh obiektov proyektirovaniya s ispol'zovaniyem nechetkikh kognitivnykh kart // Visnik Donbas'koi derzhavnoi mashinobudivnoi akademii. 2013. Vol. 2. P. 50–58.