

УДК 004.82

<https://doi.org/10.21869/2223-1560-2024-28-2-114-133>

Разработка онтологической модели предметной области для обработки данных корпоративных хранилищ

Ю.А. Халин ¹, А.А. Афанасьев ² ✉, В.А. Кудинов ²

¹ Юго-Западный государственный университет
ул. 50 лет Октября, д. 94, г. Курск 305040, Российская Федерация

² Курский государственный университет
ул. Радищева, д. 33, г. Курск 305000, Российская Федерация

✉ e-mail: afanasyevkursk@gmail.com

Резюме

Цель исследования. Формализовать онтологическую модель предметной области исследования и разработать на ее основе модели данных для информационной системы извлечения и обработки ожиданий к качеству данных корпоративных хранилищ.

Методы. В данном исследовании были использованы онтологический подход и методы моделирования данных, а также формализована задача реализации модели данных с помощью понятийного аппарата и формального языка реляционной модели данных и теории множеств. В качестве подхода к моделированию использована модель «Сущность-связь» (англ. Entity-Relationship model).

Результаты. В результате исследования были разработаны концептуальная, логическая и физическая модели для реляционной базы данных информационной системы извлечения и обработки ожиданий к качеству данных корпоративных хранилищ на основе онтологии исследуемой предметной области. Это позволяет расширить возможности использования разрабатываемой в дальнейших исследованиях информационной системы в промышленной эксплуатации организациями, осуществляющими работы по созданию корпоративных хранилищ данных.

Заключение. В итоге, результаты исследования показали, что переход к использованию реляционной модели данных с сохранением онтологической семантики возможен и изучался рядом исследователей: наборы классов, отношений между ними и их атрибутов определены в онтологической модели, что делает возможным провести разработку моделей данных для реляционной базы данных без предположения необходимости изменений онтологии, а реляционная база данных, в свою очередь, может хранить экземпляры классов онтологии в отдельных таблицах для каждого класса, полями которых будут являться атрибуты конкретного класса.

Ключевые слова: качество данных; хранилище данных; онтология предприятия; онтологическое моделирование; реляционная база данных; концептуальная модель данных; логическая модель данных; физическая модель данных.

Конфликт интересов: Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Для цитирования: Халин Ю.А., Афанасьев А.А., Кудинов В.А. Разработка онтологической модели предметной области для обработки данных корпоративных хранилищ // Известия Юго-Западного государственного университета. 2024. Т. 28, №2. С. 114-133. <https://doi.org/10.21869/2223-1560-2024-28-2-114-133>.

Поступила в редакцию 28.02.2024

Подписана в печать 17.04.2024

Опубликована 25.06.2024

Development of an ontological model of the subject area for corporate storage data processing

Yuri A. Khalin ¹, Alexander A. Afanasyev ² ✉, Vitaly A. Kudinov ²

¹ Southwest State University

50 Let Oktyabrya str. 94, Kursk 305040, Russian Federation

² Kursk State University

33, Radishcheva str., Kursk 305000, Russian Federation

✉ e-mail: afanasyevkursk@gmail.com

Abstract

Purpose of research. Formalize an ontological model of the research subject area and develop, based on it, data models for an information system for extracting and processing expectations for the quality of corporate data storage.

Methods. In this study, an ontological approach and data modeling methods were used, and the task of implementing a data model was formalized using the conceptual apparatus and formal language of the relational data model and set theory. The Entity-Relationship model was used as a modeling approach.

Results. As a result of the research, conceptual, logical and physical models were developed for a relational database information system for retrieving and processing expectations for the quality of corporate data storage based on the ontology of the subject area under study. This allows us to expand the possibilities of using the information system developed in further research in industrial operation by organizations carrying out work on creating corporate data warehouses.

Conclusion. As a result, the results of the study showed that the transition to using a relational data model while preserving ontological semantics is possible and has been studied by a number of researchers: sets of classes, relationships between them and their attributes are defined in the ontological model, which makes it possible to develop data models for a relational database without providing for the need for ontology changes, and the relational database, in turn, can store instances of ontology classes in separate tables for each class, the fields of which will be the attributes of a particular class.

Keywords: data quality; data warehouse; enterprise ontology; ontological modeling; relational database; conceptual data model; logical data model; physical data model.

Conflict of interest. The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

For citation: Khalin Y. A., Afanasyev A. A., Kudinov V. A. Development of an ontological model of the subject area for corporate storage data processing. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta = Proceedings of the Southwest State University*. 2024; 28(2): 114-133 (In Russ.). <https://doi.org/10.21869/2223-1560-2024-28-2-114-133>.

Received 28.02.2024

Accepted 17.04.2024

Published 25.06.2024

Введение

Одним из перспективных направлений, призванных удовлетворить современным требованиям к аналитиче-

ской работе в организациях, а также курсу на цифровизацию отечественной экономики (в целом) и деятельности предприятий (в частности), является со-

здание корпоративных хранилищ данных [1, 2, 3].

При этом важная проблема при создании корпоративных хранилищ данных заключается в необходимости определения и повышения уровня качества данных в них [4]. К тому же, исследователи выделяют работу над качеством данных как тренд цифровизации экономики¹ [5].

Однако процессы определения и повышения уровня качества данных на текущий момент слабо формализованы, а в наиболее распространенных методологиях под требованиями к качеству данных подразумеваются неструктурированные ожидания конкретных потребителей, что затрудняет процесс расчета и получения метрик качества данных.

Разработанные к настоящему времени подходы к извлечению и обработке ожиданий к качеству данных относятся к методам извлечения знаний. Однако указанные методы, имеющие место в инженерии знаний, еще недостаточно изучены в исследуемой предметной области – качество данных корпоративных хранилищ.

Исследуемая предметная область в отечественной практике построения корпоративных хранилищ является инновационной.

Следовательно, весьма актуальным для совершенствования подходов к из-

влечению и обработке ожиданий к качеству данных является проведение комплексного исследования возможностей автоматизации активных индивидуальных и групповых методов извлечения знаний, на основе чего представлялось бы возможным учитывать как теоретические положения самих методов, так и возможности современных MDM-, DG-систем, с помощью которых при построении корпоративных хранилищ в компаниях осуществляют организацию управления данными [6].

Таким образом, решение проблемы извлечения и обработки ожиданий к качеству данных корпоративных хранилищ требует решения таких задач, как создание алгоритма извлечения и обработки ожиданий к качеству данных корпоративных хранилищ, разработка метода детализации ожиданий к качеству данных до конкретных правил, разработка метода классификации правил к качеству данных.

Для разработки подобных методов и алгоритмов актуальна реализация модели данных на основе онтологии исследуемой предметной области, что требует постановки соответствующей задачи и ее формализации.

Материалы и методы

Для выбора метода описания предметной области исследования целесообразно провести сравнительную экспертную оценку различных подходов и выбрать наиболее оптимальный из них. Оценим по пятибалльной шкале такие методы, как ER-модель, UML-диаграммы

¹ Стратегия цифровой трансформации отрасли науки и высшего образования. 2021 / Министерство науки и высшего образования РФ. URL: https://www.minobrnauki.gov.ru/documents/?ELEMENT_ID=36749 (дата обращения: 05.12.2023).

классов, DFD, IDEF, ORM и онтологии, по следующим критериям: простота использования, эффективность представления связей между сущностями, уровень абстракции и формализации, возможности интеграции и семантической связности, гибкость и возможности модификации.

Рассмотрим значения оценок:

1 – эта оценка означает низкую эффективность метода в рассматриваемом критерии. Например, если метод оценен в 1 из 5 по критерию «Уровень абстракции и формализации», это означает, что метод предлагает низкий уровень абстракции и не обладает формализацией в достаточной степени;

2 – отражает ниже среднего уровень по рассматриваемому критерию. Метод может быть оценен так, если он демонстрирует некоторую эффективность, но все еще имеет значительные недостатки;

3 – оценка указывает на средний уровень по рассматриваемому критерию. Это означает, что метод достаточно хорошо справляется с этим критерием, но есть возможности для улучшения;

4 – уровень выше среднего, что говорит о высокой эффективности метода в рассматриваемом критерии, хотя все еще есть возможности для роста или улучшений;

5 – оценка, указывающая на высший уровень эффективности по рассматриваемому критерию. Такая оценка говорит о том, что метод отлично справляется с этим аспектом.

Результат оценки методов представлен в табл. 1.

Исходя из результатов оценки, для описания исследуемой предметной области и разработки базы данных наиболее подходящим методом является использование онтологий. Онтологии обладают следующими преимуществами в контексте разработки баз данных и описания предметных областей [7, 8, 9, 10].

1. Формализация знаний: Онтологии позволяют формализовать знания о предметной области с использованием явных связей, понятий и ограничений, что облегчает анализ и интеграцию данных.

2. Семантическая связность и стандартизация: Онтологии обеспечивают семантическую связность в базе данных, что снижает вероятность дублирования данных и позволяет однозначно интерпретировать значения.

3. Интеграция данных: Онтологии можно использовать для интеграции данных из различных источников, обеспечивая их согласованность и единое представление предметной области.

4. Гибкость и возможности модификации: Онтологии могут быть гибко модифицированы, что позволяет легко адаптировать базу данных к изменяющимся потребностям и требованиям предметной области.

5. Стандартизация и расширяемость: Онтологии предлагают стандартизированный подход к представлению знаний, что облегчает их расширение и поддержку в долгосрочной перспективе.

Таблица 1. Сравнительная экспертная оценка методов описания предметных областей**Table 1.** Comparative expert assessment of methods for describing subject areas

Критерий / Criteria Метод / The method	Простота использования / Ease of use	Эффективность представления связей между сущностями / Efficiency of representation of relationships between entities	Уровень абстракции и формализации / The level of abstraction and formalization	Возможности интеграции и семантической связности / Possibilities of integration and semantic connectivity	Гибкость и возможности модификации / Flexibility and possibilities of modification
ER-модель	4	5	3	3	3
UML-диаграммы классов	4	5	4	4	4
DFD	3	2	2	2	3
IDEF	3	4	4	3	4
ORM	4	5	4	4	4
Онтологии	3	5	5	5	4
Комментарий	ER-модель, UML-диаграммы классов и ORM предлагают относительно простые в использовании способы описания предметных областей.	ER-модель, UML-диаграммы классов, ORM и онтологии эффективно представляют связи между объектами в предметной области	Онтологии и ORM обладают высоким уровнем абстракции и формализации, что делает их более мощными инструментами для описания предметных областей	Онтологии являются лучшим методом для создания семантической связности и интеграции данных	UML-диаграммы классов, IDEF, ORM и онтологии обладают высокой гибкостью и возможностями модификации

Также применение онтологии на раннем этапе исследования было обосновано новизной исследуемой области обеспечения качества данных в корпоративных хранилищах и необходимостью определения исследуемого периметра предметной области для дальнейшей работы. Практическая значимость онтологической модели заключается в выявлении концептуальных классов этой предметной области, исследования в которой сейчас ведутся [11].

В процессе дальнейшего исследования предполагается подробно рассмотреть выделенные на рис. 1 классы и отношения для формализации задачи исследования [12].

Разработанная онтология предметной области исследования [13] отразила

множества классов и отношений. При этом, данная онтология позволяет выявить несколько особенностей, которые затрудняют работу в области оценки качества данных корпоративных хранилищ:

1) необходимость обмена данными между различными информационными системами в области управления данными (База данных, Каталог метаданных, Бизнес-гlossарий);

2) необходимость использования активных индивидуальных и групповых методов извлечения знаний;

3) последующая аналитическая обработка извлеченных знаний (классификация по измерениям качества данных, детализация ожиданий в правила для возможности последующей разработки проверок).

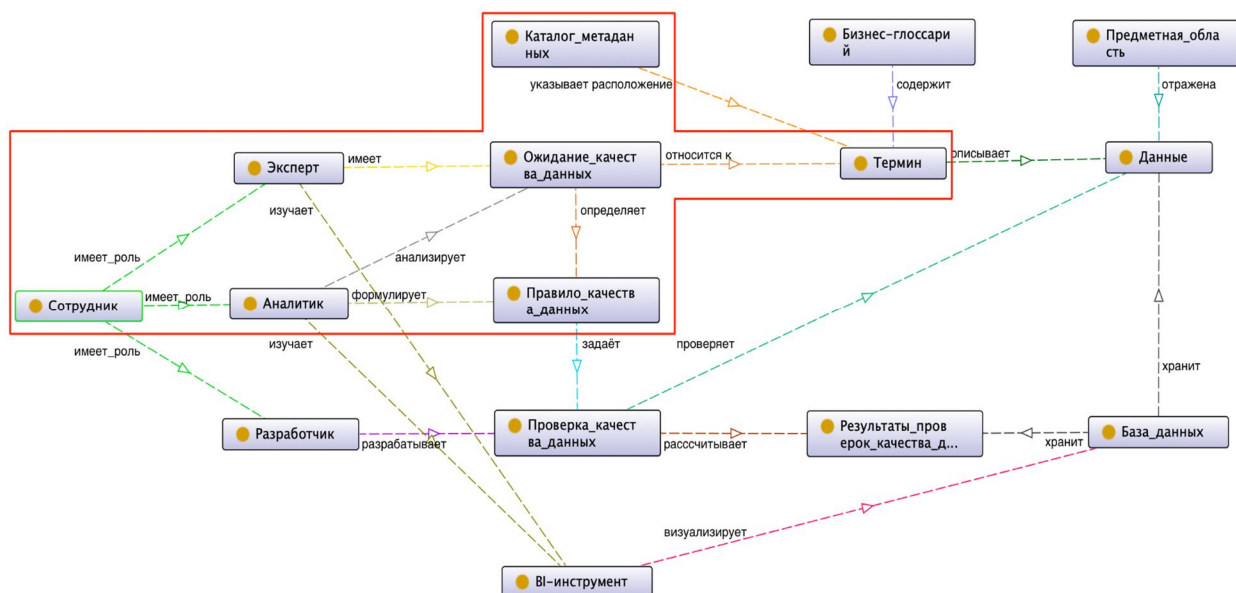


Рис. 1. Выделенная для дальнейшего исследования область онтологии извлечения ожиданий к качеству данных корпоративных хранилищ

Fig.1. The area of the ontology for extracting expectations for the quality of data in corporate warehouses has been identified for further research.

Перечисленные особенности делают процесс оценки уровня качества данных трудоемким, что увеличивает время от момента извлечения знаний до получения результатов проверок качества данных – агрегированной статистики с количеством проверенных объектов и ошибок, а также наборов данных с детальными результатами (идентификаторами проверенных объектов и описаний найденных ошибок в данных). Это снижает степень актуальности разработанных проверок или вовсе затрудняет их разработку из-за изменяющихся ожиданий.

Эти факторы позволяют сделать заключение об обоснованности поиска перспективных путей автоматизации извлечения и обработки ожиданий к качеству данных корпоративных хранилищ, которые позволят максимально сократить количество времени от извлечения ожиданий к качеству данных до разработки проверки, которая сможет отразить уровень качества данных в виде количественных характеристик и перечня ошибок, требующих исправления со стороны пользователей систем, поставляющих данные в корпоративное хранилище.

Задачу реализации модели данных можно формализовать с помощью реляционной модели данных и теории множеств. Перечисленные далее (табл. 2) атрибуты классов онтологической модели необходимы для представления классов и их экземпляров в виде отношений R с

точки зрения реляционной модели данных. Поскольку классы «Эксперт» и «Аналитик» являются подклассами «Сотрудник», отразим атрибут «Роль» этого класса для указания ролевой принадлежности сотрудника.

Так, класс «Термин» и его экземпляры представим в виде отношения R_T (рис. 2) как подмножества декартова произведения $T_1 \times T_2 \times T_3$ с заголовком H_T : {(Идентификатор, T_1), (Название, T_2), (Класс, T_3)} и телом B_T , определяемом как множество кортежей K_t , в котором каждый кортеж K_t , соответствующий заголовку H_T , является множеством упорядоченных пар $\langle T_j, t_i \rangle$, где T – название атрибута; j – индекс атрибута в заголовке; t – допустимое значения из соответствующего атрибуту множества T_j ; i – количество экземпляров класса.

Аналогично представим другие классы онтологической модели и их экземпляры в виде отношений R_M , R_E , R_P , R_W (рис. 3).

Так как классы онтологической модели, на основе которых определены перечисленные отношения, концептуально связаны, формализуем виды связей между ними (табл. 3).

Тело отношения R_T можно представить как множество экземпляров класса «Термин», определяемое как:

$$T = \{K_t \mid \forall K_t \in R_T\} :\Leftrightarrow \{\langle t_{11}, t_{12}, t_{13} \rangle, \langle t_{21}, t_{22}, t_{23} \rangle, t_{31}, t_{32}, t_{33} \rangle, \dots, \langle t_{i1}, t_{i2}, t_{i3} \rangle\}, \quad (1)$$

где i – мощность отношения R_T .

Таблица 2. Атрибуты классов онтологической модели**Table 2.** Attributes of ontological model classes

Класс / Class	Атрибут / Attribute	Множество значений / Set of values
Термин	Идентификатор	T_1
	Название	T_2
	Класс	T_3 (при этом $T_3 = \{\text{Термин}\}$)
Каталог метаданных	Идентификатор	M_1
	Метаданные	M_2
	Идентификатор термина	M_3
	Класс	M_4 (при этом $M_4 = \{\text{Каталог метаданных}\}$)
Ожидание качества данных	Идентификатор	E_1
	Название	E_2
	Описание	E_3
	Идентификатор термина	E_4
	Идентификатор эксперта	E_5
	Класс	E_6 (при этом $E_6 = \{\text{Ожидание качества данных}\}$)
Правило качества данных	Идентификатор	P_1
	Название	P_2
	Описание	P_3
	Идентификатор ожидания	P_4
	Идентификатор аналитика	P_5
	Измерение качества данных	P_6
	Класс	P_7 (при этом $P_7 = \{\text{Правило качества данных}\}$)
Сотрудник	Идентификатор	W_1
	Роль	W_2 (при этом $W_2 = \{\text{Эксперт, Аналитик}\}$)
	Класс	W_3 (при этом $W_3 = \{\text{Сотрудник}\}$)

Аналогично, тела отношений R_M , R_E , R_P , R_W представим как множества M , E , P , W соответственно:

$$M = \{K_m \mid \forall K_m \in R_M\} : \Leftrightarrow \{ \langle m_{11}, m_{12}, m_{13}, m_{14} \rangle, \langle m_{21}, m_{22}, m_{23}, m_{24} \rangle, \langle m_{31}, m_{32}, m_{33}, m_{34} \rangle, \dots, \langle m_{k1}, m_{k2}, m_{k3}, m_{k4} \rangle \}, \quad (2)$$

где k – мощность отношения R_M ;

$$E = \{K_e \mid \forall K_e \in R_E\} : \Leftrightarrow \{ \langle e_{11}, e_{12}, e_{13}, e_{14}, e_{15}, e_{16} \rangle, \langle e_{21}, e_{22}, e_{23}, e_{24}, e_{25}, e_{26} \rangle, \langle e_{31}, e_{32}, e_{33}, e_{34}, e_{35}, e_{36} \rangle, \dots, \langle e_{x1}, e_{x2}, e_{x3}, e_{x4}, e_{x5}, e_{x6} \rangle \}, \quad (3)$$

где x – мощность отношения R_E ;

$$P = \{K_p \mid \forall K_p \in R_P\} : \Leftrightarrow \{ \langle p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}, p_{17} \rangle, \langle p_{21}, p_{22}, p_{23}, p_{24}, p_{25}, p_{26}, p_{27} \rangle, \langle p_{31}, p_{32}, p_{33}, p_{34}, p_{35}, p_{36}, p_{37} \rangle, \dots, \langle p_{y1}, p_{y2}, p_{y3}, p_{y4}, p_{y5}, p_{y6}, p_{y7} \rangle \}, \quad (4)$$

где y – мощность отношения R_P ;

$$W = \{K_w \mid \forall K_w \in R_W\} : \Leftrightarrow \{ \langle w_{11}, w_{12}, w_{13} \rangle, \langle w_{21}, w_{22}, w_{23} \rangle, \langle w_{31}, w_{32}, w_{33} \rangle, \dots, \langle w_{z1}, w_{z2}, w_{z3} \rangle \}, \quad (5)$$

где z – мощность отношения R_W .

Таким образом, каждый выделенный в онтологической модели класс будет рассматриваться в качестве множества кортежей, где каждый кортеж – экземпляр конкретного класса онтологической модели. Тогда выделенный на рис. 1 участок онтологической модели как формальная онтология будет иметь вид:

$$O = \langle T, M, E, P, W \rangle, \quad (6)$$

где T – конечное счетное множество кортежей (терминов Бизнес-гlossария);

M – конечное счетное множество кортежей (метаданных БД);

E – конечное счетное множество кортежей (ожиданий к качеству данных);

P – конечное счетное множество кортежей (правил качества данных);

W – конечное счетное множество кортежей (сотрудников).

Множества E и P могут быть пустыми и в рамках исследования предполагается, что:

$$E = \emptyset \wedge P = \emptyset. \quad (7)$$

	Идентификатор, T_1 / Id, T_1	Название, T_2 / Name, T_2	Класс, T_3 / Class, T_3
K_{t1}	t_{11}	t_{12}	t_{13}
K_{t2}	t_{21}	t_{22}	t_{23}
...	...	...	...
K_{ti}	t_{i1}	t_{i2}	t_{i3}

Рис. 2. Отношение R_T (термины бизнес-гlossария)

Fig. 2. R_T Relationship (Business Glossary Terms)

Отношение R_M				
	Идентификатор, M_1	Метаданные, M_2	Идентификатор термина, M_3	Класс, M_4
K_{m1}	m_{11}	m_{12}	m_{13}	m_{14}
K_{m2}	m_{21}	m_{22}	m_{23}	m_{24}
...	...	...	...	...
K_{mk}	m_{k1}	m_{k2}	m_{k3}	m_{k4}

Отношение R_E						
	Идентификатор, E_1	Название, E_2	Описание, E_3	Идентификатор термина, E_4	Идентификатор эксперта, E_5	Класс, E_6
K_{e1}	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}	e_{16}
K_{e2}	e_{21}	e_{22}	e_{23}	e_{24}	e_{25}	e_{26}
...	...	...	...	...	...	...
K_{ek}	e_{k1}	e_{k2}	e_{k3}	e_{k4}	e_{k5}	e_{k6}

Отношение R_P							
	Идентификатор, P_1	Название, P_2	Описание, P_3	Идентификатор ожидания, P_4	Идентификатор аналитика, P_5	Измерение качества данных, P_6	Класс, P_7
K_{p1}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}
K_{p2}	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}
...	...	...	...	...	...	...	...
K_{p7}	p_{71}	p_{72}	p_{73}	p_{74}	p_{75}	p_{76}	p_{77}

Отношение R_W			
	Идентификатор, W_1	Роль, W_2	Класс, W_3
K_{w1}	w_{11}	w_{12}	w_{13}
K_{w2}	w_{21}	w_{22}	w_{23}
...	...	...	...
K_{w7}	w_{71}	w_{72}	w_{73}

Рис. 3. Отношения R_M (метаданные), R_E (ожидания к качеству данных), R_P (правила качества данных), R_W (сотрудники)

Fig. 3. Relationships R_M (metadata), R_E (data quality expectations), R_P (data quality rules), R_W (employees)

Таблица 3. Связи между отношениями R_T , R_M , R_E , R_P , R_W

Table 3. Connections between the relations R_T , R_M , R_E , R_P , R_W

СВЯЗЬ ОТНОШЕНИЙ / Relationship relationship	Характер связи / The nature of the connection	Условия связи (атрибуты как первичные и внешние ключи) / Communication conditions (attributes as primary and foreign keys)
$R_T \xrightarrow{N} R_M$	Один ко многим	$T_1 = M_3$
$R_T \xrightarrow{U} R_E$	Один к одному	$T_1 = E_4$
$R_E \xrightarrow{N} R_Q$	Один ко многим	$E_1 = P_4$
$R_W \xrightarrow{U} R_E$	Один к одному	$W_1 = E_5 \wedge W_2 = \text{Эксперт}$
$R_W \xrightarrow{U} R_P$	Один к одному	$W_1 = P_5 \wedge W_2 = \text{Аналитик}$

Целью дальнейших исследований станет наполнение этих множеств элементами, соответствующими элементам множеств T , M , W , с помощью разрабатываемых моделей и алгоритмов.

Рассмотрим следующие ситуации:

1. При $T = \emptyset$, $M = \emptyset$, $W = \emptyset$ – в этом случае дальнейшее исследование не имеет смысла, так как отсутствуют термины глоссария, метаданные, сотрудники.

2. При $T \neq \emptyset$, $M = \emptyset$, $W = \emptyset$ или $T \neq \emptyset$, $M \neq \emptyset$, $W = \emptyset$ или $T = \emptyset$, $M \neq \emptyset$, $W = \emptyset$ дальнейшее исследование также не имеет смысла, поскольку множество сотрудников пусто, что делает невозможным наполнение множества E элементами.

3. При $T \neq \emptyset$, $M = \emptyset$, $W \neq \emptyset$ – дальнейшее исследование возможно, но отсутствие метаданных затруднит детализацию ожиданий к качеству данных.

Таким образом, в рамках исследования считается, что множества T , W не должны быть пустыми:

$$T \neq \emptyset \wedge W \neq \emptyset. \quad (8)$$

Так как если они пусты, то наполнение множества E невозможно из-за отсутствия терминов и сотрудников, которые могут иметь ожидания к качеству данных в соответствии с онтологической моделью.

Результаты и их обсуждение

Проведенное исследование подходов к моделированию данных на основе онтологии предметной области качества данных корпоративных хранилищ показывает, что использование реляционных баз данных целесообразно для хранения

онтологий [11, 14, 15, 16, 17, 18, 19, 8, 20, 12, 21, 22].

Поскольку в процессе формализации онтологической модели использовались понятия реляционной модели данных для формального отражения отношений между классами, множества элементов каждого выделенного класса рассматриваются в качестве множеств кортежей, где каждый кортеж – экземпляр конкретного класса онтологической модели [17, 18]. В связи с этим, в ходе дальнейшего исследования актуально разработать модели данных для служебной реляционной базы данных информационной системы извлечения и обработки ожиданий к качеству данных корпоративных хранилищ с целью сохранения онтологических связей между выделенными классами (см. рис. 1).

Переход к реляционной модели данных в дальнейшей работе обусловлен повышением степени практической значимости исследования в целом, так как специалистами по управлению данными более изучены реляционные модели данных, а не онтологические. Так как наборы классов, отношений между ними и их атрибутов определены в онтологической модели, это делает возможным провести разработку моделей данных для реляционной базы данных без предположения необходимости изменений онтологии. Так, реляционная база данных может хранить экземпляры классов онтологии в отдельных таблицах для каждого класса, полями кото-

рых будут являться атрибуты конкретного класса. Такая реализация позволит сделать взаимодействие с системой более доступным как для пользователей, так и для интеграции с необходимыми системами для получения данных об экземплярах, например, Бизнес-гlossариями и Каталогами метаданных. Таким образом, данный переход повышает возможности использования результатов исследования в промышленной эксплуатации организациями, осуществляющими работы по созданию корпоративных хранилищ данных [11].

Реляционная модель данных – это наиболее распространенный подход к моделированию баз данных, который основывается на использовании таблиц для хранения данных. Данный подход был предложен Эдгаром Коддом в 1970 году [21]. Реляционная модель данных имеет простую структуру, обеспечивает эффективный поиск и изменение данных, а также обеспечивает целостность данных. В качестве подхода к моделированию целесообразно использовать модель «Сущность-связь» (англ. Entity-Relationship model) – это подход к моделированию реляционных баз данных, который основывается на использовании сущностей и их связей для описания данных. Данный подход был предложен Питером Ченом в 1976 году [22]. Преимущество этого подхода в рамках данного исследования заключается в том, что он имеет много общего с онто-

логическим подходом и позволяет связать понятия онтологии с понятиями реляционной модели данных.

Первым этапом разработки моделей данных для реляционной базы данных является разработка концептуальной модели данных. Питер Чен дал определение концептуальной модели данных: «Концептуальная модель данных – это модель, которая описывает сущности (англ. entities) в предметной области, связи между ними, атрибуты этих сущностей и ограничения на данные» [22]. ER-модель позволяет описывать отношения между сущностями, такие как «один-к-одному», «один-ко-многим» и «многие-ко-многим», а также определять ограничения на данные, такие как «уникальность» и «необходимость».

На рис. 4 продемонстрирована концептуальная модель данных об извлечении ожиданий к качеству данных корпоративных хранилищ, разработанная на основе онтологической модели [19, 9]. Сущности концептуальной модели соответствуют классам онтологии, также на модели отражены соответствующие онтологии связи между сущностями. В соответствии с нотацией Р. Баркера, связи между сущностями описываются в «обе» стороны. Так, связь между сущностями «Ожидание качества данных - Относится - Термин» дополняется обратным описанием «Термин - Должен соответствовать - Ожидание качества данных».

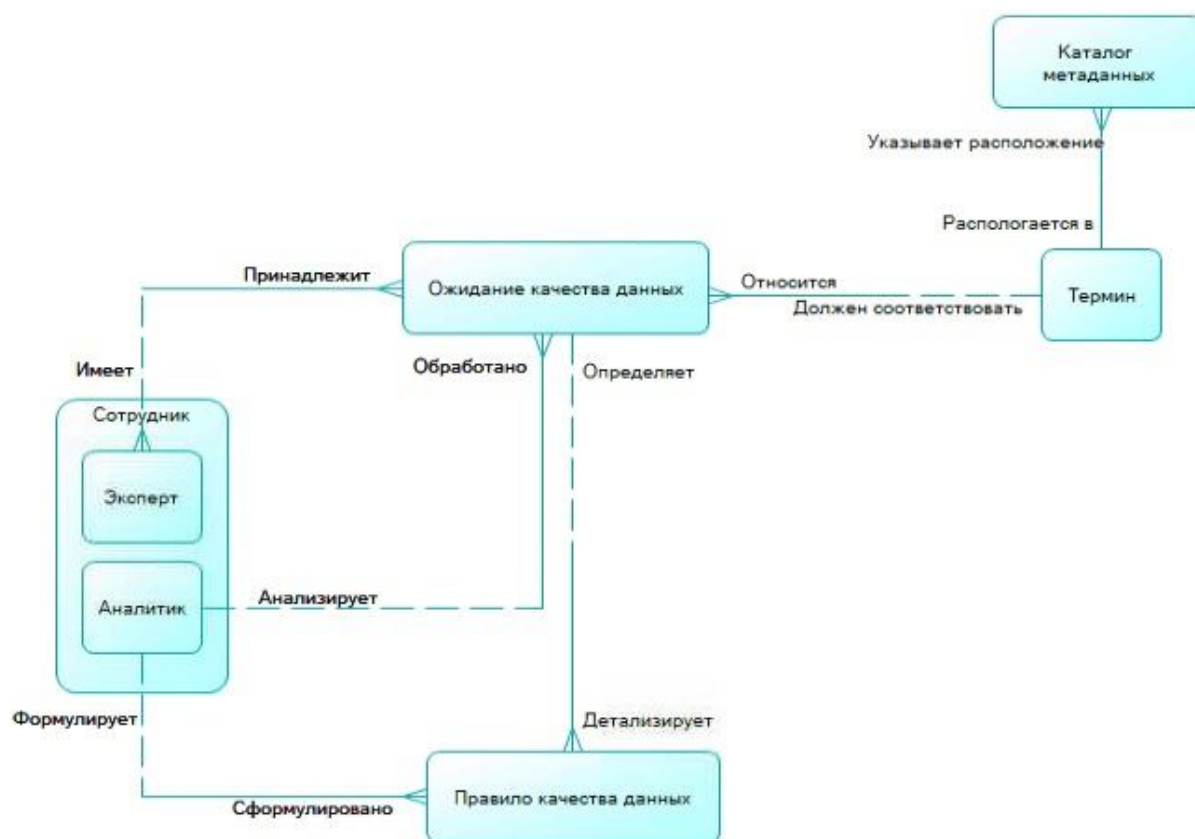


Рис. 4. Концептуальная модель данных об извлечении ожиданий к качеству данных корпоративных хранилищ

Fig. 4. Conceptual Data Model for Extracting Expectations for Enterprise Data Quality Data

Также на концептуальной модели данных графически отражены типы отношений между сущностями, такие как «один-к-одному» (англ. one-to-one), «один-ко-многим» (англ. one-to-many) и «многие-ко-многим» (англ. many-to-many) и обязательность связи между ними – непрерывная линия до середины связи при обязательной, пунктирная линия при необязательной.

На следующем этапе проводится разработка логической модели данных. Логическая модель данных дополняет концептуальную, переводя ее описание в форму, которая может быть использо-

вана для создания физической базы данных. Логическая модель данных описывает, как данные будут храниться и организовываться в базе данных, включая описание таблиц, полей и связей между таблицами по конкретным полям, которые будут использоваться в качестве первичных и внешних ключей. В целом, логическая модель данных является более детальной и конкретной, чем концептуальная модель данных, и представляет собой более точное описание того, как данные будут храниться и использоваться в базе данных.

На рис. 5 продемонстрирована логическая модель данных об извлечении ожиданий к качеству данных корпоративных хранилищ, разработанная на основе концептуальной модели. Модель дополнена атрибутами сущностей, которые будут описывать экземпляры кон-

кретной сущности. Также детализирована связь между сущностями путем выделения первичных и внешних ключей и устранена связь «многие-ко-многим» путем выделения сущности «Связь эксперта с ожиданием качества данных».

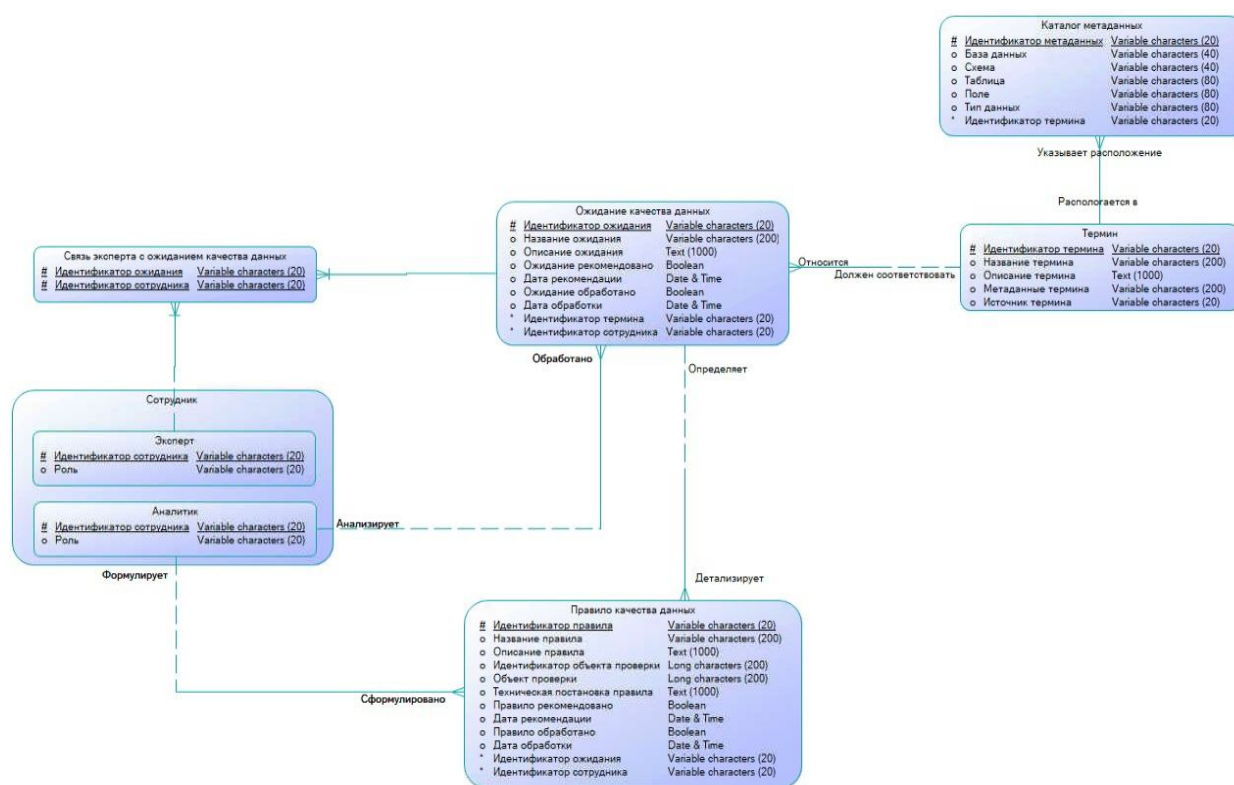


Рис. 5. Логическая модель данных об извлечении ожиданий к качеству данных корпоративных хранилищ

Fig. 5. Logical Data Model for Extracting Expectations for Enterprise Data Quality Data

Для данной модели не требуется дополнительная нормализация отношений, так как ее отношения уже находятся во второй нормальной форме: они находятся в первой нормальной форме и каждый их неключевой атрибут непри- водимо зависим от первичного ключа.

На следующем этапе проводится разработка физической модели данных. Физическая модель данных (рис. 6)

представляет собой перечень таблиц и полей в пригодном для реляционных баз данных виде [14].

Названия таблиц и полей переводятся на латиницу для генерации кода на языке SQL с целью создания таблиц и ограничений в базе данных. Полученная физическая модель позволяет создать служебную реляционную базу данных информационной системы из-

влечения и обработки ожиданий к качеству данных корпоративных хранилищ с сохранением онтологических связей между выделенными классами. Справедливо отметить, что необходимость иметь две таблицы “EXPERT” и “ANALYST” с одинаковым набором атрибутов кажется избыточной, однако это целесообразно из-за устранения связи «многие-ко-

многим» путем выделения сущности “EXPERT_DQ_REQUIREMENT”, так как на практическом опыте у ожидания к качеству данных может быть несколько заинтересованных экспертов, например, при использовании таких групповых методов извлечения знаний, как «круглый стол» или «мозговой штурм» [11].

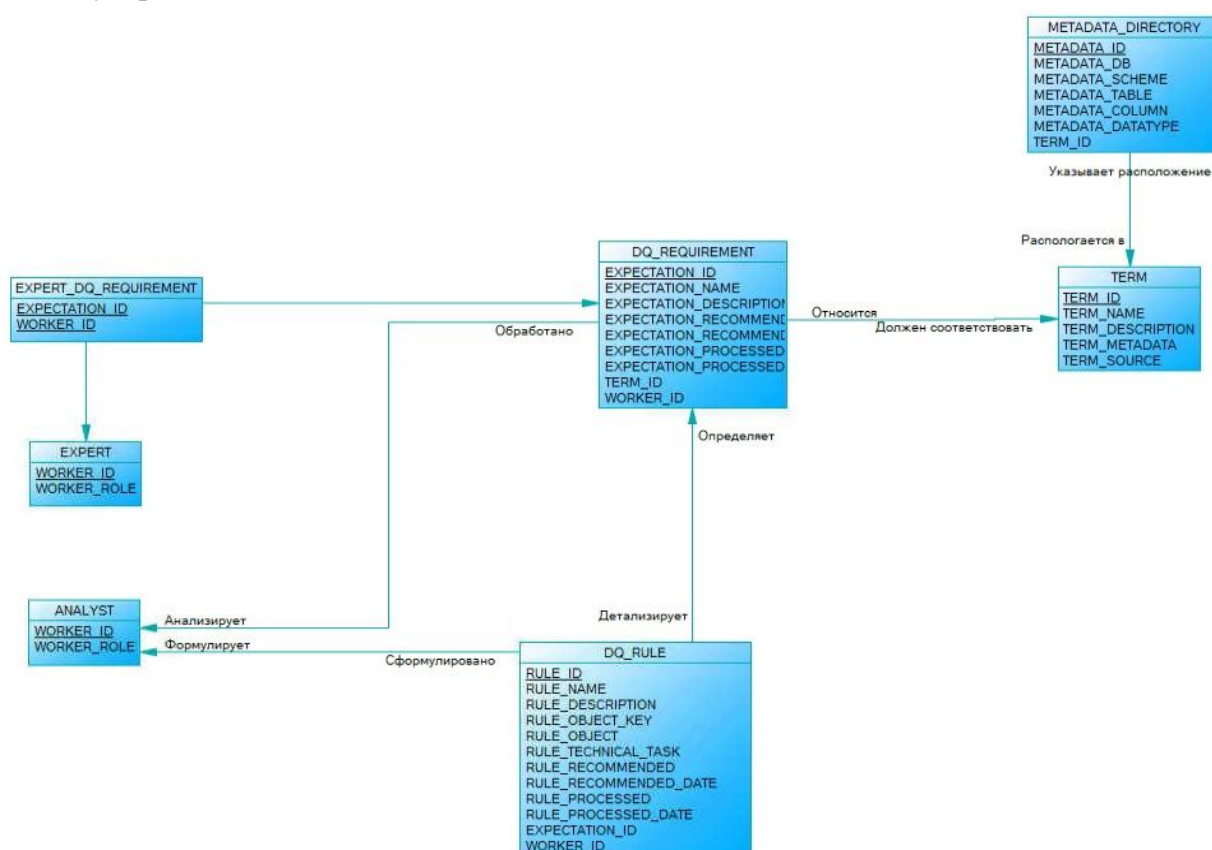


Рис. 6. Физическая модель данных об извлечении ожиданий к качеству данных корпоративных хранилищ

Fig. 6. Physical data model of extracting expectations for data quality of enterprise warehouses

Выводы

В ходе работы проведена формализация разработанной онтологической модели предметной области исследования с помощью реляционной модели данных и теории множеств. В процессе исследова-

ния утверждалось, что актуальным направлением совершенствования подходов к извлечению и обработке ожиданий к качеству данных является автоматизация активных индивидуальных и групповых методов извлечения знаний и использование MDM-, DG-систем. Для

дальнейших исследований в данном направлении актуальна реализация модели данных на основе онтологии исследуемой предметной области, что потребовало постановки соответствующей задачи, ее формализации и выполнения.

Каждый выделенный в онтологической модели класс рассмотрен в качестве множества кортежей, где каждый кортеж – экземпляр конкретного класса онтологической модели. В формализованном виде целью дальнейших исследований является наполнение множеств E , P элементами, соответствующим элементам множеств T , M , W с помощью разрабатываемых моделей и алгоритмов.

Таким образом, в процессе исследования удалось на основе разработанной ранее онтологической модели разработать концептуальную, логическую и физическую модель для реляционной базы данных. Обзор исследований отечественных и зарубежных ученых под-

твердил возможность подобного подхода. Применение онтологического подхода на раннем этапе исследования позволило определить исследуемый периметр предметной области для дальнейшей работы ввиду ее новизны, а переход к реляционной модели данных позволит повысить степень практической значимости исследования в целом и сделать его результаты более доступными для промышленной эксплуатации организациями, осуществляющими работы по созданию корпоративных хранилищ данных. В частности, достичь этого возможно за счет хранения экземпляров классов онтологии в реляционной базе данных, что позволит сделать взаимодействие с результатами исследования более доступным как для пользователей, так и для интеграции с необходимыми системами для получения данных об экземплярах, например, Бизнес-гlossариями и Каталогами метаданных.

Список литературы

1. Лучкин Н. А. Создание модели хранилища данных корпоративной информационно-аналитической системы предприятия // Объектные системы. 2011. №1 (3). С. 36–39.
2. Основные тренды цифровой трансформации экономики / под ред. д.э.н. Н.Н. Масюк. Владивосток: изд-во ВГУЭС, 2021. 200 с.
3. Салий В.В., Кухаренко Л.В., Ищенко О.В. Цифровая трансформация экономики и внедрение хранилищ данных на основе больших данных в инфраструктуру компании // Вестник Академии знаний. 2021. №3 (44). С. 208–214.
4. Ильин А. А. Некоторые проблемы построения корпоративных Хранилищ данных // Программные продукты и системы. 2005. № 3. С. 29-32.
5. Цифровая трансформация: ожидания и реальность / Г. И. Абдрахманова, С. А. Васильковский, К. О. Вишневский, М. А. Гершман, Л. М. Гохберг; рук. авт. кол. Рудник П. Б. [и др.] // докл. к XXIII Ясинской (Апрельской) междунар. науч. конф. по проблемам развития экономики и общества. М.: Изд. дом Высшей школы экономики, 2022. 221 с.

6. DAMA-DMBOK: Свод знаний по управлению данными. 2-е изд. М.: Олимп–Бизнес, 2020. 828 с.
7. Модели представления и обработки знаний в информационно-аналитических системах / В. П. Добрица, Е. А. Титенко, Ю. А. Халин, А. И. Катыхин. Курск, 2023. 172 с.
8. Загорюлько Ю. А. Современные средства формализации семантики областей знаний на основе онтологий // Информационные и математические технологии в науке и управлении. 2018. №3 (11). С. 27–35.
9. Когнитивное моделирование информационного обеспечения игрового автоматизированного обучения / Халин Ю.А., Катыхин А.И., Зинкин С.А., Шилин А.А. // Известия Юго-Западного государственного университета. 2022. Т. 26, №4. С. 117-131. <https://doi.org/10.21869/2223-1560-2022-26-4-117-131>.
10. Gruber T.R. A translational approach to portable ontology specification // Knowledge Acquisition. 1993. Vol. 5. № 2. Pp. 199-220.
11. Афанасьев А.А., Кудинов В.А. Исследование подходов к моделированию данных на основе онтологии предметной области качества данных корпоративных хранилищ // Интеллектуальные информационные системы: Теория и практика: сб. науч.ст. по материалам IV Всерос. с междунар. участием конф. / отв. ред. А.А. Халин. Курск, 2023. С. 188-200.
12. Khan S. A., Qadir M. A. A Genaric Framework for Evaluation of Ontology to Relational Database Transformation Process // 2010 International Conference on Information and Emerging Technologies (ICIET). Karachi, 2010. pp. 1-5.
13. Афанасьев А.А., Кудинов В.А. Использование онтологического подхода для извлечения ожиданий к качеству данных корпоративных хранилищ // Экономика. Информатика. 2022. № 49(3). С. 566–574. <https://doi.org/10.52575/2687-0932-2022-49-3-566-574>.
14. Olivier Curé, Guillaume Blin. RDF Database Systems. Morgan Kaufmann, 2014. 256 с.
15. Yongqi Huang, Gaoyan Deng. Research on Storage of Geo-ontology in Relational Database // 2nd International Symposium on Information Engineering and Electronic Commerce (IEEC). Ternopil, 2010. Pp. 1-4.
16. Ahmed W., Aslam M. A., Jun Shen, Jianming Yong. A Light Weight Approach for Ontology Generation and Change Synchronization between Ontologies and Source Relational Databases // 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Lausanne, 2011. pp. 208-214.
17. Онтологии и базы данных - взаимная дополнительность при использовании научных данных / А. О. Еркимбаев, В. Ю. Зицерман, Г. А. Кобзев, А. В. Косинов // Мониторинг. наука и технологии. Центр сопряженного мониторинга окружающей среды и природных ресурсов. 2015. №3 (24). С. 41-50.

18. Вехорев М. Н., Пантелеев М. Г. Построение хранилищ онтологических баз знаний // Программные продукты и системы. 2011. №3. С. 1-6.
19. Олейник П. П., Грегер С. Э. Применение и реализация онтологий при разработке приложений баз данных // Прикладная информатика. 2016. №3 (63). С. 76-102.
20. Автоматизация разработки онтологий научных предметных областей на основе паттернов онтологического проектирования / Ю. А. Загорулько, Е. А. Сидорова, Г. Б. Загорулько, И. Р. Ахмадеева, А. С. Серый // Онтология проектирования. 2021. №4 (42). С. 500–520.
21. Codd E. F. A Relational Model of Data for Large Shared Data Banks // Communications of the ACM. 1970. Vol. 13 (6). P. 377–387.
22. Chen P. P.-S. The Entity-Relationship Model: Toward a Unified View of Data // ACM Transactions on Database Systems. 1976. Vol. 1 (1). Pp. 9–36.

References

1. Luchkin N. A. Creating a data warehouse model for a corporate information and analytical system of an enterprise. *Ob'ektnye sistemy = Object Systems*. 2011; 1 (3): 36-39. (In Russ.)
2. Masyuk N.N. Main trends in the digital transformation of the economy. Vladivostok, izd-vo VGUES; 2021. 200 p. (In Russ.)
3. Saliy V.V., Kukharenko L.V., Ishchenko O.V. Digital transformation of the economy and the introduction of data warehouses based on big data into the company's infrastructure. *Vestnik Akademii znaniy = Bulletin of the Academy of Knowledge*. 2021; (3): 208-214. (In Russ.)
4. Ilyin A. A. Some problems in building corporate data warehouses. *Programmnye produkty i sistemy = Software Products and Systems*. 2005; 3: 29-32. (In Russ.)
5. Abdrakhmanova G. I., Vasilkovsky S. A., Vishnevsky K. O., Gershman M. A., Gokhberg L. M. [et al.], hands auto count Rudnik P. B. Digital transformation: expectations and reality. In: *XXIII Yasinskoy (April'skoy) mezhdunar. nauch. konf. po problemam razvitiya ekonomiki i obshchestva = XXIII Yasinsk (April) international. scientific conf. on problems of economic and social development*. Moscow: Izd. dom Vysshei shkoly ekonomiki; 2022. 221 p. (In Russ.)
6. DAMA-DMBOK: Data Management Body of Knowledge. 2nd ed.]. Moscow: Olimp-Business; 2020. 828 p. (In Russ.)
7. Dobritsa V. P., Titenko E. A., Khalin Y. A., Katykhin A. I. Models of knowledge representation and processing in information-analytical systems. Kursk, 2023. 172 p. (In Russ.)
8. Zagorulko Yu. A. Modern means of formalizing the semantics of knowledge areas based on ontologies. *Informacionnye i matematicheskie tekhnologii v nauke i upravlenii = Information and mathematical technologies in science and management*. 2018; (3): 27-35. (In Russ.)

9. Khalin Y. A., Katykhin A. I., Zinkin S. A., Shilin A. A. Cognitive Modeling of Information Support for Game-Based Automated Learning. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta = Proceedings of the Southwest State University*. 2022; 26(4): 117-131 (In Russ.). <https://doi.org/10.21869/2223-1560-2022-26-4-117-131>.
10. Gruber T.R. A translational approach to portable ontology specification. *Knowledge Acquisition*. 1993; 5 (2): 199-220.
11. Afanasyev A.A., Kudinov V.A. Research of approaches to data modeling based on the ontology of the data quality domain of corporate storages. In: *Intellectual'nye informacionnye sistemy: Teoriya i praktika: sb. nauch.st. po materialam IV Vseros. s mezhdunar. uchastiem konf. = Intelligent information systems: Theory and practice: collection. scientific article based on materials of the IV All-Russian. with international participation of the conf.* Kursk; 2023. P. 188-200. (In Russ.)
12. Khan S. A., Qadir M. A. A Genaric Framework for Evaluation of Ontology to Relational Database Transformation Process. *2010 International Conference on Information and Emerging Technologies (ICIET)*. Karachi, 2010. P. 1-5.
13. Afanasyev A.A., Kudinov V.A. Using an Ontology Approach to Extract Quality Expectations for Enterprise Data Warehouses. *Ekonomika. Informatika = Economy. Computer Science*. 2022; 49(3): 566–574. (In Russ.).
14. Olivier Curé, Guillaume Blin. *RDF Database Systems*. Morgan Kaufmann, 2014, 256 p.
15. Yongqi Huang, Gaoyan Deng. Research on Storage of Geo-ontology in Relational Database. *2nd International Symposium on Information Engineering and Electronic Commerce (IEEC)*. Ternopil, 2010. P. 1-4.
16. Ahmed W., Aslam M. A., Jun Shen, Jianming Yong. A Light Weight Approach for Ontology Generation and Change Synchronization between Ontologies and Source Relational Databases. *15th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. Lausanne, 2011. P. 208-214.
17. Erkimbaev A. O., Zitserman V. Yu., Kobzev G. A., Kosinov A. V. Ontologies and databases - mutual complementarity when using scientific data. *Monitoring. nauka i tekhnologii = Monitoring. Science and technology. Center for Associated Monitoring of the Environment and Natural Resources*, 2015; (3): 41-50. (In Russ.)
18. Vekhorev M. N., Panteleev M. G. Construction of repositories of ontological knowledge bases. *Programmnye produkty i sistemy = Software Products and Systems*. 2011; (3): 1-6. (In Russ.)
19. Oleinik P. P., Greger S. E. Application and implementation of ontologies in the development of database applications. *Prikladnaya informatika = Applied informatics*. 2016; (3): 76-102. (In Russ.)

20. Zagorulko Yu. A., Sidorova E. A., Zagorulko G. B., Akhmadeeva I. R., Sery A. S. Automation of the development of ontologies of scientific subject areas based on ontological design patterns. *Ontologiya proektirovaniya = Ontology of Design*. 2021; (4): 500-520. (In Russ.)

21. Codd E. F. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*. 1970; 13 (6): 377–387.

22. Chen P. P.-S. The Entity-Relationship Model: Toward a Unified View of Data. *ACM Transactions on Database Systems*. 1976; 1 (1): 9–36.

Информация об авторах / Information about the Authors

Халин Юрий Алексеевич, кандидат технических наук, доцент, доцент кафедры программной инженерии, Юго-Западный государственный университет, г. Курск, Российская Федерация, e-mail: yur-khalin@yandex.ru, ORCID <https://orcid.org/0000-0002-7020-8515>

Yuri A. Khalin, Cand. of Sci. (Engineering), Associate Professor, Associate Professor of the Software Engineering Department, Southwest State University, Kursk, Russian Federation, e-mail: yur-khalin@yandex.ru, ORCID <https://orcid.org/0000-0002-7020-8515>

Афанасьев Александр Александрович, аспирант, Курский государственный университет, г. Курск, Российская Федерация, e-mail: afanasyevkursk@gmail.com, ORCID <https://orcid.org/0009-0002-5862-2059>.

Alexander A. Afanasyev, Post-Graduate Student, Kursk State University, Kursk, Russian Federation, e-mail: afanasyevkursk@gmail.com, ORCID <https://orcid.org/0009-0002-5862-2059>.

Кудинов Виталий Алексеевич, доктор педагогических наук, профессор, профессор кафедры программного обеспечения и администрирования информационных систем, Курский государственный университет, г. Курск, Российская Федерация, e-mail: kudinovva@yandex.ru, ORCID <https://orcid.org/0009-0008-7623-5494>.

Vitaly A. Kudinov, Dr. of Sci. (Pedagogy), Professor, Professor of the Software and Information Systems Administration Department, Kursk State University, Kursk, Russian Federation, e-mail: kudinovva@yandex.ru, ORCID <https://orcid.org/0009-0008-7623-5494>.