

Исследование семантической модели представления объявлений о недвижимости на примере данных площадки «Циан»

И.А. Нестеров, Е.А. Цона, И.В. Исаев, С.В. Клименков

*Университет ИТМО, г. Санкт-Петербург, Россия
электронная почта: nesterov.iv2018@gmail.com*

Аннотация: Интеграция объявлений о недвижимости из разнородных источников порождает два класса трудностей: структурную несогласованность схем атрибутов и семантическое смешение сущностей «объявление» и «объект недвижимости» в одной записи. Существующие доменные онтологии (RealEstateCore, ProDataMarket) не разграничивают эти сущности и не фиксируют происхождение атрибутов; систематическое сравнение моделей хранения для данного класса задач в литературе отсутствует.

Метод. Разработана каноническая семантическая модель с явным разграничением объявления (цена, источник, дата) и объекта недвижимости (тип, площадь, адрес); онтология включает 289 триплетов, 9 классов и механизм фиксации происхождения атрибутов (provenance). Для наполнения модели реальными данными реализован модуль импорта (7 нормализаторов, 25 регулярных выражений для извлечения признаков из текста). На датасете из 11 642 объявлений «Циан» сопоставлены четыре парадигмы хранения — широкая реляционная, нормализованная EAV, JSONB и семантическая — по шести структурно-функциональным критериям. Основные результаты. Полностью подтверждены четыре гипотезы: H1 (корректность) — 100 % структурных критериев, RLA = 78 %; H2 (расширяемость) — 0 миграций у семантической против 20 у широкой при 20 признаках; H3 (разреженность) — 38 714 NULL у широкой против 0; H6 (иерархии) — 0 правок кода против 2–3. Частично подтверждены H4 (фиксация происхождения атрибутов): структурно — да, по нормализованному времени на факт реляционные варианты быстрее в 8 раз; H5 (запросы): семантическая выигрывает на Q2 business, реляционные — на агрегатных запросах.

Обсуждение. Семантическая модель предпочтительна при регулярном расширении схемы, требованиях к трассировке происхождения атрибутов и заполненности признаков ниже 50 %; реляционные модели сохраняют преимущества при фиксированной схеме. Наблюдаемый разрыв в производительности частично обусловлен асимметрией реализаций.

Ключевые слова: интеграция данных, семантическая модель, онтология, граф знаний, происхождение атрибутов, реляционная модель, EAV, JSONB, модуль импорта, недвижимость.

Введение

Агрегация объявлений о недвижимости из множества источников порождает два класса трудностей: структурная неоднородность (несогласованные форматы атрибутов, часть характеристик объекта — в тексте свободного описания) и семантическое смешение сущностей — типичная запись агрегатора объединяет характеристики физического объекта

(площадь, этаж, адрес) и атрибуты публикации (цена, дата, источник). Если одна квартира анонсирована двумя площадками с разными ценами, система без явного разграничения вынуждена либо дублировать данные объекта, либо произвольно смешивать атрибуты разных источников [1, 2].

В области семантической интеграции данных разработаны как общие подходы [3–6], так и доменные онтологии для рынка недвижимости: RealEstateCore [7], модель Shi и Roman [8], ProDataMarket [9] и ряд других [10, 11]. Вместе с тем ни одна из существующих онтологий не предусматривает явного разграничения «объявление / объект недвижимости» и не включает механизм фиксации происхождения атрибутов (provenance), необходимый для разрешения конфликтов при многоисточниковой интеграции. Что касается производительности, Calvanese et al. [5], Gu et al. [6] и Kotiranta et al. [12] показали, что преимущества семантических хранилищ зависят от структуры запроса и схемы хранения, а не от парадигмы в целом.

Настоящее исследование устраняет эти пробелы: предлагается каноническая модель с разграничением «объявление / объект» и встроенным механизмом фиксации происхождения атрибутов, а также проводится первое сравнительное экспериментальное исследование четырёх парадигм хранения для данного класса задач. Рабочая гипотеза: семантическая модель превосходит реляционную по расширяемости схемы, работе с разреженными признаками и иерархическому выводу; реляционные схемы сохраняют преимущества на агрегатных запросах [7, 8].

Материалы и методы

Цель и задачи исследования

Цель работы — экспериментально исследовать и сравнить четыре парадигмы хранения данных объявлений о недвижимости по структурно-

функциональным критериям, а также разработать каноническую семантическую модель с явным разграничением сущностей как инструмент такого сравнения. Для этого спроектирована онтология с механизмом фиксации происхождения атрибутов, реализован модуль импорта, построен единый стенд и проверены шесть гипотез с выработкой рекомендаций по выбору модели.

Данные и валидация

В качестве источника данных использован публичный датасет московских квартир, опубликованный в рамках Kaggle-соревнования Moscow Housing [13], который содержит 23 285 объявлений с атрибутами площади, этажности, ремонта, балкона, санузла и ряда других. Для имитации многоисточниковой интеграции записи разделены на два подмножества, преобразованных в форматы, характерные для площадок «Циан» и «Авито» соответственно. Для основного эксперимента использована подвыборка в формате «Циан» (11 642 записи), что позволяет сравнивать модели в условиях одного источника, изолируя внутрисхемную неоднородность от межисточниковой. Эталонный набор — 300 записей (80 % типичных, 20 % неполных); 50 записей верифицированы независимо, расхождений не выявлено. Метрика RLA (record-level accuracy) — доля записей, для которых все упомянутые в тексте признаки извлечены верно и без ложноположительных.

Семантическая модель и онтология

Каноническая модель разграничивает объявление (цена, источник, дата, текст) и объект недвижимости (тип, площадь, этаж, адрес); для каждого атрибута фиксируются источник, дата и исходное поле (история происхождения), что позволяет разрешать конфликты без потери исходных значений. Онтология содержит 289 триплетов, 9 классов, иерархию подклассов Feature → Renovation → EuroRenovation, 8 объектных и 12

свойств данных. Модуль импорта выполняет сопоставление полей, нормализацию (7 нормализаторов) и извлечение признаков из текста (25 регулярных выражений). Фрагмент онтологии:

```
realty:Listing a owl:Class . realty:Property a owl:Class .  
realty:Renovation rdfs:subClassOf realty:Feature .  
realty:EuroRenovation rdfs:subClassOf realty:Renovation .  
realty:advertises a owl:ObjectProperty ;  
rdfs:domain realty:Listing ; rdfs:range realty:Property.  
realty:price a owl:DatatypeProperty ; rdfs:domain realty:Listing .  
realty:area a owl:DatatypeProperty ; rdfs:domain realty:Property .
```

Варианты сравнения

Выбор четырёх моделей обусловлен стремлением охватить основные парадигмы хранения атрибутивных данных, актуальные для производственных систем работы с каталогами. Широкая реляционная схема (каждый признак — отдельная колонка [14]) является промышленным стандартом и служит базовым вариантом сравнения: именно она наиболее распространена в действующих агрегаторах недвижимости. Нормализованная EAV (entity-attribute-value) — классическое решение для систем с заранее неизвестной или часто меняющейся схемой, широко применяемое в гибких каталогах товаров и услуг. JSONB-колонка представляет современный гибридный подход, сочетающий реляционное хранение с документной гибкостью в рамках одной СУБД; его популярность резко выросла с появлением нативной поддержки в PostgreSQL. Семантическая модель на основе RDF/OWL [15, 16] — единственная из четырёх, допускающая формальный иерархический вывод через `rdfs:subClassOf` и декларативное описание схемы, что принципиально важно для проверки гипотез Н2, Н4 и Н6. Нативные документные СУБД (MongoDB) и графовые движки на модели «граф свойств» (Neo4j без RDF) в

основной бенчмарк не включены: первые потребовали бы отдельной инфраструктуры, нарушая условие изоляции, вторые реализуют иную графовую парадигму без формального вывода — Neo4j исследован отдельно в качестве дополнительного эксперимента. Три реляционные модели работают на единой СУБД H2 Database, изолируя логические свойства от инфраструктурных различий; семантическая реализация использует итеративный обход in-memory Map, что исключает SQL-разбор и многоверсионное управление параллельным доступом (MVCC) — асимметрия, существенная для интерпретации временных результатов.

Экспериментальные сценарии и стенд

Шесть критериев сравнения охватывают три аспекта, актуальных для задач интеграции каталогов: структурные свойства хранилища (H1 — корректность, H2 — расширяемость, H3 — разреженность), функциональные требования предметной области (H4 — происхождение атрибутов, H6 — иерархии) и операционные характеристики (H5 — производительность типовых запросов). Выбор критериев опирается на обзоры Kotiranta et al. [12] и Hoseini et al. [3], в которых аналогичные оси сравнения используются для оценки гетерогенных хранилищ.

Проверяются шесть гипотез (H1–H6) с использованием следующих запросов: Q1 — агрегатное распределение по типам объектов; Q2 business — поиск 20 похожих объектов по типу сделки, ценовому диапазону $\pm 30\%$ и пяти признакам; Q3 — происхождение атрибутов; Q4 — иерархический запрос «все объекты с ремонтом»; Q5 — обоснование значения признака. Стенд: Apple M4 Pro (12-ядерный), 24 ГБ ОЗУ, JDK 24, -Xmx2g; H2 Database (Spring Boot 3.2.5, in-memory, MODE=PostgreSQL); Neo4j 5.18 (Docker, heap/page cache 512 МБ), Bolt; 50 итераций, время в мкс.

Результаты и обсуждение

Корректность модуля интеграции (H1)

На рабочей выборке зафиксировано 11 480 успешных записей, 162 частичных (отсутствие обязательного поля в источнике), ни одной отклонённой и ни одной ошибки нормализации. Полная структурная корректность подтверждена на эталонном наборе из 300 записей. Значение RLA составило 78 % (234/300): все 66 несовпадений обусловлены ложноотрицательными результатами по признаку «ремонт» — текстовые формулировки источника выходят за пределы текущего набора из 8 регулярных выражений.

Таблица № 1

Признак	TP	FP	FN	Точность, %	Полнота, %	F1, %
Ремонт	69	0	66	100	51,1	67,6
Балкон	175	0	0	100	100	100

Точность, полнота и F-мера извлечения признаков из текста (300 записей)

Ложноположительные результаты не зафиксированы; ложноотрицательные по «ремонту» (66 случаев) объясняются выходом формулировок источника за пределы текущих 8 регулярных выражений.

Вывод по H1: согласуется.

Расширяемость схемы (H2)

В широкой схеме новый признак требует миграции БД и правок кода; в EAV, JSONB и семантической — только изменения сопоставления или онтологии.

Таблица № 2

Стоимость расширения схемы

Модель	Признаков	Миграций	Строк кода ¹	Тестов
Широкая	5 / 10 / 20	5 / 10 / 20	20 / 40 / 80	5 / 10 / 20
Нормализ. / JSONB / Семантич. ²	5 / 10 / 20	0	5 / 10 / 20	0

¹ Для семантической модели — декларации в онтологии. ² Императивный код не затрагивается.

Вывод по Н2: согласуется.

Разреженные и частично заполненные признаки (Н3)

Пять исследуемых признаков имеют среднюю заполненность от 8 % (санузел) до 55 % (балкон), что типично для реальных каталогов недвижимости. В широкой реляционной схеме каждая из 11 642 записей содержит ячейки для всех признаков вне зависимости от их наличия, порождая 38 714 NULL-значений по пяти признакам. Нормализованная EAV, JSONB и семантическая модели хранят только фактически присутствующие значения.

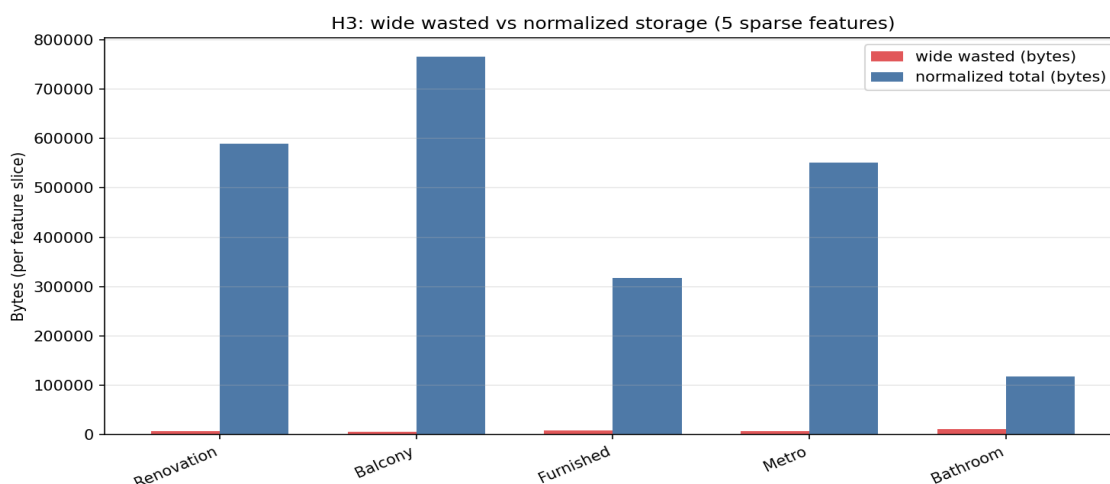


Рис. 1. – Потери широкой схемы (красный, байт) и объём нормализованного хранения (синий, байт) по пяти признакам (n = 11 642)

Вывод по H3: согласуется.

Происхождение атрибутов (H4)

Семантическая модель хранит историю происхождения атрибута непосредственно как аннотации ребра графа ($\approx 1,5$ МБ на датасет); реляционные варианты требуют отдельной таблицы `AttributeOrigin` — 100 994 строки, $\approx 64,7$ МБ. При полном запросе Q3 семантическая возвращает 39 012 фактов за 1 384 мкс; нормализованная — столько же фактов за 169 мкс. Нормализованное время составляет 0,035 против 0,004 мкс/факт соответственно.

Таблица № 3

Запросы о происхождении (медиана / p95, мкс; 50 итераций)

Запрос	Широкая	Нормализ.	JSONB	Семантич.
Q3 полный (медиана/p95,	162 / 181	169 / 289	33 / 76	1 384 / 1 671

мкс)				
мкс/факт	0,007	0,004	0,003	0,035

Вывод по Н4: частично согласуется: структурное преимущество подтверждено, производительность на единицу факта — нет.

Типовые предметные запросы (Н5)

Наиболее показательным является запрос Q2 business: семантическая модель выполняет его за 473 мкс против 3 029 мкс у широкой реляционной. Однако эту разницу следует интерпретировать с учётом архитектурной асимметрии: реализация в оперативной памяти лишена SQL-разбора, планировщика и MVCC-оверхеда H2 Database. В диагностическом сценарии worst case (без предфильтрации) и на агрегатном запросе Q1 реляционные варианты быстрее.

Таблица № 4

Запрос	Широкая	Нормализ.	JSONB	Семантич.
Q1: распределение	3 001 / 3 240	3 / 4	3 491 / 3 807	811 / 1 668
Q2: business	3 029 / 4 242	76 149 / 78 457	2 397 / 2 591	473 / 1 269
Q4: иерархический	32 / 33	32 / 32	30 / 39	788 / 2 355

Q5: обоснование ¹	<1 / 1	<1 / 1	<1 / <1	159 / 184
---------------------------------	--------	--------	---------	-----------

Время выполнения ключевых запросов (медиана / p95, мкс; 50 итераций)

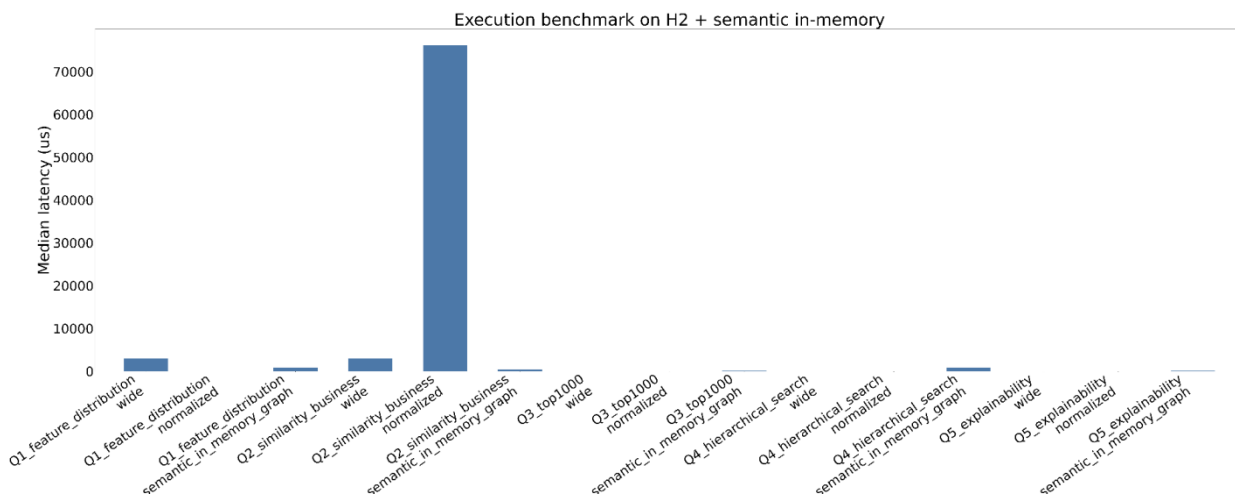


Рис. 2. – Время выполнения ключевых запросов (медиана, мкс, логарифмическая шкала; n = 11 642; База данных H2 + семантическая модель в оперативной памяти)

Вывод по Н5: частично согласуется; предпочтительность определяется классом запроса и архитектурой движка.

Иерархии и выводимые факты (Н6)

При добавлении нового подкласса (например, EuroRenovation как разновидности Renovation) реляционные модели требуют правок SQL-запросов и императивной логики. Семантическая модель не требует изменений кода запроса — достаточно добавить три декларативных правила `rdfs:subClassOf`; иерархия материализуется при загрузке данных.

При добавлении подкласса EuroRenovation широкая схема требует 3 правки кода (SQL + логика), нормализованная и JSONB — 2; семантическая модель не требует правок вовсе — достаточно 3 деклараций `rdfs:subClassOf` в онтологии. Результат запроса идентичен у всех моделей (2 650 объектов).

Вывод по Н6: согласуется.

Внешний графовый движок

Дополнительно проведён замер на Neo4j 5.18 (Docker). Разрыв относительно in-методу реализации составил $\times 10$ на Q1 и $\times 41$ на Q4, что объясняется Bolt-оверхедом сетевого протокола, транзакционными границами и отсутствием прогрева кэша при неоптимизированной конфигурации. Результат подтверждает необходимость единой инфраструктуры для корректного сравнения парадигм.

Сводка проверки гипотез

Результаты проверки всех шести гипотез обобщены в таблице 5.

Таблица № 5

Гипотеза	Статус	Ключевой результат
Н1 — корректность	Согласуется	100 % структурных; RLA 78 % (ремонт 51 %, балкон 100 %)
Н2 — расширяемость	Согласуется	0 миграций против 20 (20 признаков)
Н3 — разреженность	Согласуется	38 714 NULL у широкой против 0
Н4 — фиксация происхождения атрибутов	Частично	Структурно: 1,5 МБ против 64,7 МБ; время/факт: 0,035 против 0,004 мкс
Н5 — запросы	Частично	Семантич. лучше на Q2 business; реляционные — на Q1, Q4
Н6 — иерархии	Согласуется	0 правок кода против 2– 3

Сводка проверки гипотез

Семантическая модель демонстрирует устойчивые структурные преимущества по четырём из шести критериев; частичное подтверждение Н4 и Н5 объясняется архитектурной асимметрией стенда, а не принципиальными ограничениями парадигмы.

Ограничения

Обобщаемость результатов ограничена одной площадкой («Циан»); PostgreSQL и документные СУБД (MongoDB) не тестировались. Асимметрия архитектур (in-memory vs. H2 с MVCC) частично завышает преимущества семантической модели по времени. Полнота текстового извлечения оценена лишь по двум признакам; разметка выполнена одним аннотатором без вычисления к Коэна. Поведение моделей при 100К–1М записях не исследовано эмпирически.

Заключение

В работе предложена каноническая семантическая модель представления объявлений о недвижимости с явным разграничением сущностей «объявление» и «объект недвижимости» и механизмом фиксации происхождения атрибутов, а также проведено сравнительное экспериментальное исследование четырёх парадигм хранения на реальном датасете из 11 642 объявлений площадки «Циан». Четыре гипотезы из шести подтверждены полностью — семантическая модель превосходит реляционные аналоги по расширяемости схемы, работе с разреженными признаками, компактному хранению истории происхождения атрибутов и поддержке иерархического вывода без правок кода. Преимущество по скорости сценарно-зависимо и частично обусловлено архитектурной асимметрией стенда.

Семантическую модель целесообразно применять при регулярном расширении схемы, требованиях к трассировке истории происхождения атрибутов, а также при заполненности признаков ниже 50 %; при фиксированной схеме реляционные модели остаются предпочтительными. Направления развития: расширение словаря для «ремонта», верификация на PostgreSQL, масштабирование на 100К–1М записей.

Литература (References)

1. Putrama I.M., Martinek P. Heterogeneous data integration: Challenges and opportunities. *Data in Brief*. 2024. V. 56. Art. 110853. DOI: 10.1016/j.dib.2024.110853.
2. Halevy A.Y., Rajaraman A., Ordille J.J. Data Integration: The Teenage Years. *Proceedings of the 32nd International Conference on Very Large Data Bases*. Seoul: VLDB Endowment. 2006. pp. 2–12.
3. Hoseini S., Theissen-Lipp J., Quix C. A survey on semantic data management as intersection of ontology-based data access, semantic modeling and data lakes. *Web Semantics*. 2024. V. 81. Art. 100819.
4. Van Assche D., Delva T., Heyvaert P. et al. Declarative RDF graph generation from heterogeneous (semi-)structured data: A systematic literature review. *Journal of Web Semantics*. 2023. V. 75. Art. 100753. DOI: 10.1016/j.websem.2022.100753.
5. Calvanese D., Lanti D., Ozaki A. et al. Accessing scientific data through knowledge graphs with Ontop. *Patterns*. 2021. V. 2. №10. Art. 100346.
6. Gu Z., Liu Y., Wang J. et al. Ontology-based data federation and query optimization. *Knowledge-Based Systems*. 2025. V. 329. Art. 114216.
7. Hammar K., Wallin E.O., Karlberg P., Hälleberg D. The RealEstateCore Ontology. *The Semantic Web — ISWC 2019. Lecture Notes in Computer Science*. Cham: Springer. 2019. V. 11779. pp. 130–145.
8. Shi L., Roman D. Ontologies for the Real Property Domain. *Proc. of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018)*. Novi Sad: ACM. 2018. Art. 14. DOI: 10.1145/3227609.3227661.
9. Shi L., Sukhobok D., Nikolov N., Roman D. The ProDataMarket Ontology. *Territorio Italia*. 2017. №2. pp. 11–35.
10. Hu Z., Wang Y., Liu Q. et al. Knowledge Graph-Based Housing Market Analysis. *CEUR Workshop Proceedings*. 2021. V. 2873. pp. 30–35.

11. Silvennoinen H., Kurvinen E., Heinonen J. OntoZoning: An Ontology for Urban Zoning Regulations. *Journal of Urban Management*. 2023. V. 12. №2. pp. 151–167. DOI: 10.1016/j.jum.2023.02.002.
12. Kotiranta P., Junkkari M., Nummenmaa J. Performance of Graph and Relational Databases in Complex Queries. *Applied Sciences*. 2022. V. 12, №13. Art. 6490.
13. Moscow Housing Dataset. Kaggle. URL: kaggle.com/datasets/azharc/moscow-housing.
14. Codd E.F. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*. 1970. V. 13. №6. pp. 377–387.
15. Baader F., Calvanese D., McGuinness D., Nardi D., Patel-Schneider P. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge: Cambridge University Press. 2003. 574 p.
16. Angles R., Gutierrez C. Survey of Graph Database Models. *ACM Computing Surveys*. 2008. V. 40. №1. Art. 1.