ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

УДК 004.733 Научная статья
https://doi.org/10.35330/1991-6639-2026-28-4-62-76
LTBWDP

Проблемы семантической композиции гетерогенных данных в городских информационных системах и подходы к их решению

Д. А. Рыбаков

Российский экономический университет имени Г. В. Плеханова
115054, Россия, Москва, Стремянный переулок, 36

Аннотация. В статье рассматривается актуальная проблема интеграции разнородных данных в современных городских информационных системах, функционирующих в рамках концепции «Умный город» и предоставления государственных услуг.

Цель исследования – систематизация проблем семантической композиции данных в ГИС и анализ современных подходов к их решению, а также обоснование перспективности использования биоинспирированных самоорганизующихся алгоритмов для создания отказоустойчивых моделей интеграции.

Материалы и методы исследования. Исследование носит теоретико-аналитический характер. Применены методы системного анализа проблем семантической гетерогенности данных в городских информационных системах, сравнительный анализ существующих подходов к связыванию записей (детерминированные правила, вероятностная модель Феллеги – Сантера, методы машинного обучения, графовые модели), а также концептуальное моделирование на основе биоинспирированных алгоритмов с использованием метода аналогий между поведением миксомицета Physarum polycephalum и процессами семантической интеграции данных.

Результаты. Показано, что традиционные методы Extract-Transform-Load (ETL) и детерминированного сопоставления записей не справляются с ростом объема, вариативности и семантической неоднородности данных, поступающих из множества ведомственных источников. Выявлены основные классы проблем – синтаксическая, структурная и семантическая гетерогенность. В качестве перспективного направления решения предлагается концепция семантической композиции данных. Проведен анализ существующих подходов (rule-based, машинное обучение, графовые модели) и обоснована необходимость разработки адаптивных, самоорганизующихся методов. Вводится метафора биоинспирированных алгоритмов (на примере Physarum polycephalum) как основы для создания отказоустойчивой и динамической модели связывания данных. Сформулированы требования к перспективному методу композиции и намечены пути его формализации.

Заключение. Концептуально обоснован биоинспирированный подход к семантической композиции, основанный на принципах самоорганизации Physarum polycephalum. Предложена архитектура семантического слоя, встраиваемого непосредственно в СУБД и включающего таблицу семантических связей и три управляющих процесса – исследователь, усилитель и испаритель. Сформулированы ключевые преимущества подхода: отсутствие необходимости в «холодном старте», непрерывная адаптация, прозрачность, масштабируемость и отказоустойчивость. Дальнейшие исследования направлены на формализацию математической модели и экспериментальную апробацию на реальных данных городских ведомственных систем.

Ключевые слова: семантическая интеграция, гетерогенные данные, городские информационные системы, электронное правительство, связанные данные, биоинспирированные алгоритмы, physarum polycephalum

Поступила 27.03.2026, одобрена после рецензирования 21.05.2026, принята к публикации 07.08.2026

Для цитирования. Рыбаков Д. А. Проблемы семантической композиции гетерогенных данных в городских информационных системах и подходы к их решению // Известия Кабардино-Балкарского научного центра РАН. 2026. Т. 28. № 4. С. 62–76. DOI: 10.35330/1991-6639-2026-28-4-62-76

Original article

Semantic composition challenges of heterogeneous data in urban information systems and frameworks for their resolution

D.A. Rybakov

Plekhanov Russian University of Economics
36, Stremyannyy lane, Moscow, 115054, Russia

Abstract. The article addresses the pressing issue of heterogeneous data integration in modern urban information systems operating within the 'Smart City' concept and public service delivery.

Aim. The study is to systematize the problems of semantic data composition in GIS, analyze modern approaches to their solution, and justify the prospects of using bio-inspired self-organizing algorithms to create fault-tolerant integration models.

Materials and methods. The study is theoretical and analytical in nature. The methods applied include a systems analysis of semantic data heterogeneity problems in urban information systems, a comparative analysis of existing record linkage approaches (deterministic rules, the Fellegi–Sunter probabilistic model, machine learning methods, and graph-based models), as well as conceptual modeling based on bio-inspired algorithms using the method of analogy between the behavior of the myxomycete Physarum polycephalum and data semantic integration processes.

Results. It is demonstrated that traditional Extract-Transform-Load (ETL) methods and deterministic record linkage fail to cope with the growing volume, variety, and semantic heterogeneity of data originating from multiple departmental sources. The main classes of challenges have been identified as syntactic, structural, and semantic heterogeneity. As a promising solution pathway, the concept of semantic data composition is proposed. An analysis of existing approaches (rule-based, machine learning, and graph-based models) was performed, and the necessity of developing adaptive, self-organizing methods was justified. A bio-inspired algorithm metaphor (exemplified by Physarum polycephalum) is introduced as a foundation for constructing a fault-tolerant and dynamic data linkage model. The requirements for the promising composition method have been formulated, and the pathways for its formalization have been outlined.

Conclusion. A bio-inspired approach to semantic composition, based on the self-organization principles of Physarum polycephalum, has been conceptually justified. A semantic layer architecture embedded directly within the DBMS has been proposed, featuring a semantic relationships table and three controller processes: the explorer, the reinforcer, and the evaporator. The key advantages of the approach have been formulated, including the elimination of the cold-start problem, continuous adaptation, transparency, scalability, and fault tolerance. Future research is aimed at formalizing the mathematical model and conducting experimental validation using real-world data from municipal departmental systems.

Keywords: semantic integration, heterogeneous data, urban information systems, e-government, linked data, bioinspired algorithms, physarum polycephalum

Submitted 27.03.2026, approved after reviewing 21.05.2026, accepted for publication 07.08.2026

For citation. Rybakov D.A. Semantic composition challenges of heterogeneous data in urban information systems and frameworks for their resolution. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2026. Vol. 28. No. 4. Pp. 62–76. DOI: 10.35330/1991-6639-2026-28-4-62-76

Введение

Современный этап развития информационного общества характеризуется активной цифровизацией государственного управления. Городские информационные системы (ГИС), обеспечивающие предоставление государственных и муниципальных услуг, представляют собой сложный комплекс взаимодействующих компонентов, каждый из которых исторически создавался для автоматизации узкой предметной области (ЗАГС, МФЦ, ЖКХ, образование, здравоохранение и др.). Ключевой проблемой, возникающей при попытке создания единого цифрового профиля гражданина или получения целостной картины состояния городского хозяйства, является семантическая разнородность данных. Один и тот же объект реального мира (человек, организация, земельный участок) может описываться в разных ведомственных базах данных с использованием различных идентификаторов, форматов, степени полноты и актуальности. Ошибки ввода, опечатки, отсутствие унифицированных справочников приводят к тому, что автоматизированное сопоставление таких записей становится нетривиальной задачей.

Традиционные подходы к интеграции данных, основанные на жестких правилах сопоставления (например, точное совпадение ИНН или СНИЛС), демонстрируют высокую точность, но низкую полноту охвата, так как не учитывают случаи, когда идентификаторы отсутствуют или содержат ошибки. Более того, жесткая логика не позволяет системе адаптироваться к изменению форматов данных или появлению новых источников.

Целью данной статьи является систематизация проблем семантической композиции данных в ГИС и анализ современных подходов к их решению, а также обоснование перспективности использования биоинспирированных самоорганизующихся алгоритмов для создания отказоустойчивых моделей интеграции.

Анализ проблем семантической композиции в городских информационных системах

Процесс цифровой трансформации государственного управления привел к формированию сложных распределенных информационных ландшафтов. В рамках концепции «Умный город» функционируют десятки, а в мегаполисах – сотни ведомственных информационных систем, каждая из которых автоматизирует деятельность конкретного ведомства. Ключевым противоречием современного этапа является необходимость предоставления комплексных государственных услуг при сохранении ведомственной разобщенности источников данных. Разрешение этого противоречия требует решения фундаментальной задачи – семантической композиции данных. Под семантической композицией мы понимаем процесс обнаружения, установления и поддержания смысловых соответствий между элементами данных из различных автономных источников, относящихся к одному объекту реального мира, с целью формирования целостного и непротиворечивого представления. Иными словами, необходимо из разрозненных «лоскутов» информации «сшить» единый цифровой профиль гражданина, организации или объекта недвижимости.

Сложность этой задачи обусловлена гетерогенностью данных на нескольких уровнях. Синтаксическая гетерогенность проявляется в различии форматов, кодировок и шаблонов ввода. Например, дата рождения в разных системах может храниться как строка «ДД.ММ.ГГГГ», как тип данных date в формате «ГГГГ-ММ-ДД» или как Unix-метка. Адрес может быть единой строкой или структурирован по полям. Этот уровень преодолим при наличии четких спецификаций и ETL-процедур.

Структурная гетерогенность возникает из-за различий в логических моделях данных. Информация об одном объекте может быть распределена по таблицам в одной системе и сосредоточена в одной таблице в другой. Сопоставление таких схем требует понимания семантики связей между сущностями.

Центральная проблема – семантическая гетерогенность, связанная с различиями в интерпретации и идентификации объектов. Она включает проблему идентификации – отсутствие сквозного идентификатора, пронизывающего все системы. В реальности мы имеем множество локальных идентификаторов (паспорт, СНИЛС, ИНН), ни один из которых не является обязательным для всех систем. Проблема вариативности и ошибок проявляется в опечатках, сокращениях и изменениях данных во времени. Проблема противоречивости возникает, когда для одного атрибута в разных системах указаны разные значения, требуя механизмов доверия к источникам. Причины этих проблем носят объективный характер – длительная история развития систем в отсутствие единых стандартов и субъективный – ведомственная разобщенность и отсутствие единой политики управления данными. Последствия для качества госуслуг критичны: снижение полноты информации (гражданин повторно предоставляет документы), снижение достоверности (ошибочные отказы или неправомерные решения), рост временных и финансовых издержек.

Таким образом, задача семантической композиции не сводится к тривиальной интеграции баз данных. Это комплексная научно-практическая проблема, требующая новых адаптивных методов, способных к самоорганизации в условиях гетерогенности и динамичности городской среды.

Обзор существующих подходов к решению задачи связывания данных

Задача связывания записей является классической проблемой в области интеграции данных, исследование которой ведется более полувека. За это время сформировалось несколько основных парадигм ее решения, каждая из которых имеет свою область применения, математический аппарат и ограничения. В контексте городских информационных систем выбор конкретного подхода определяет не только точность предоставления государственных услуг, но и вычислительную эффективность всей платформы, а также возможность ее адаптации к изменяющимся условиям. Рассмотрим эволюцию методов – от простейших детерминированных правил до сложных графовых моделей, способных учитывать контекстные взаимосвязи.

Исторически первыми и наиболее интуитивно понятными являются детерминированные подходы, основанные на жестких логических правилах вида: «Если значение атрибута A в источнике S_1 совпадает со значением атрибута B в источнике S_2, то записи относятся к одному объекту». Математически это можно формализовать следующим образом. Пусть имеются две записи r_i и r_j. Введем функцию сравнения $\delta_k(r_i, r_j)$, которая возвращает 1, если значения по k-му атрибуту совпадают, и 0 в противном случае. Тогда правило связывания можно записать как конъюнкцию условий по множеству ключевых атрибутов K:

$$Link(r_i, r_j) = \bigwedge_{k \in K} \delta_k(r_i, r_j).$$

На практике в качестве ключевых атрибутов обычно используются уникальные идентификаторы, такие как СНИЛС, ИНН или номер паспорта. Главными достоинствами детерминированных методов являются их прозрачность и высокая точность: при наличии уникальных идентификаторов точность связывания достигает 99,5–100 %, а ложные срабатывания практически исключены. Кроме того, эти методы не требуют обучающих выборок и просты в реализации. Однако им присущи и серьезные недостатки. Прежде всего, это низкая полнота связывания: по данным исследователей, детерминированные методы в среднем пропускают от 20 до 40 % релевантных связей из-за отсутствия идентификаторов (например, у новорожденных еще нет СНИЛС) или наличия ошибок ввода. Кроме того,

такие системы негибки – правила необходимо переписывать при изменении форматов данных, а также они крайне чувствительны к качеству данных: одна опечатка в цифре уникального идентификатора приводит к потере связи. В городских системах детерминированные методы обычно используются как «первая линия обороны» – для быстрого и гарантированно точного связывания записей, имеющих уникальные идентификаторы, но для достижения приемлемой полноты охвата их необходимо дополнять более гибкими алгоритмами.

Классическая вероятностная модель связывания записей была разработана в 1969 году Айваном Феллеги и Аланом Сантером и до сих пор остается теоретической основой для многих современных систем. В отличие от жестких правил этот подход оценивает вероятность того, что две записи относятся к одному объекту, на основе сравнения множества полей. Рассмотрим пару записей (r_i, r_j) и вектор сравнения $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_K)$, где γ_k – результат сравнения по k-му атрибуту. Обычно γ_k принимает значение 1 при полном совпадении и 0 при несовпадении, хотя в расширенных моделях могут использоваться и промежуточные значения для нечеткого совпадения. Введем две гипотезы: M (соответствие) – записи относятся к одному объекту и U (несоответствие) – записи относятся к разным объектам. Нас интересует отношение правдоподобия:

$$R = \frac{P(\gamma \mid M)}{P(\gamma \mid U)}.$$

В предположении условной независимости атрибутов (наивный байесовский подход) это отношение можно представить в виде произведения:

$$R = \prod_{k=1}^{K} \frac{P(\gamma_k \mid M)}{P(\gamma_k \mid U)}.$$

После логарифмирования получаем вес связи:

$$w = \log(R) = \sum_{k=1}^{K} \log\left(\frac{P(\gamma_k \mid M)}{P(\gamma_k \mid U)}\right).$$

Веса для совпадений и для несовпадений вычисляются на этапе обучения модели. Для принятия решения Феллеги и Сантер предложили использовать два порога – нижний T_{low} и верхний – T_{high}. Если вычисленный вес превышает верхний порог, записи считаются связанными; если вес ниже нижнего порога – несвязанными; в промежуточном случае требуется ручная проверка экспертом.

Вероятностные методы обладают важными преимуществами: они устойчивы к ошибкам в данных благодаря использованию множества атрибутов, веса имеют четкую статистическую интерпретацию, а при наличии размеченных данных модель может быть точно настроена. Однако существуют и ограничения. Для точной оценки условных вероятностей необходима качественная размеченная выборка («золотой стандарт»), что в городских системах часто отсутствует. Предположение о независимости атрибутов на практике может нарушаться, приводя к искажению оценок. Кроме того, попарное сравнение всех записей имеет квадратичную вычислительную сложность $O(n^2)$, что для крупного города с миллионами записей неприемлемо – требуются специальные методы блокирования, снижающие сложность, но потенциально пропускающие связи. В классических экспериментах на данных переписи населения США вероятностный метод Феллеги – Сантера показал точность около 98 % и полноту около 97 % при наличии качественных обучающих данных, однако на зашумленных данных полнота может снижаться до 82 %.

С развитием вычислительных мощностей и накоплением больших объемов данных задачу связывания записей все чаще рассматривают как задачу бинарной классификации. Для каждой пары записей строится вектор признаков, и модель машинного обучения обучается предсказывать метку «связаны» или «не связаны». Признаки обычно включают различные метрики сходства: для строковых полей (ФИО, названия организаций) используются расстояния Левенштейна, Дамерау – Левенштейна, метрика Jaro-Winkler, хорошо работающая с опечатками, косинусное расстояние для n-грамм, а также фонетическое кодирование (метафоны, Soundex). Для дат вычисляются абсолютная разница или индикатор точного совпадения. Для числовых полей используются разность или отношение значений. В результате каждая пара записей представляется вектором в многомерном пространстве признаков. В качестве алгоритмов классификации применяются различные методы: от простой логистической регрессии, обеспечивающей интерпретируемость, до случайного леса, устойчивого к переобучению и позволяющего оценить важность признаков. На сегодняшний день «золотым стандартом» для табличных данных считается градиентный бустинг (реализации XGBoost, LightGBM, CatBoost), который на практике демонстрирует наилучшее качество. Глубокие нейронные сети, включая архитектуры сиамского типа, используются реже из-за необходимости больших объемов данных, но могут обучаться непосредственно на сырых текстах, автоматически извлекая признаки.

Достоинства методов машинного обучения очевидны: при достаточном объеме качественной разметки они превосходят вероятностные подходы, достигая F1-меры выше 0.95, как показывают соревнования по связыванию записей. Модели обладают высокой гибкостью, позволяя комбинировать любые признаки, включая контекстные, и автоматически определять их важность. Однако эти методы имеют и существенные недостатки. Главным ограничением является потребность в размеченных данных: получение «золотого стандарта» для городских систем требует трудоемкой ручной работы экспертов, поскольку исторически достоверная разметка связей отсутствует. Острой является проблема дисбаланса классов – связанных пар на много порядков меньше, чем несвязанных, что требует применения специальных техник. Кроме того, модель, обученная на данных одного региона или периода, может плохо работать в других условиях из-за различий в структуре населения и качестве данных (проблема экстраполяции). Наконец, сложные модели, такие как бустинг или нейросети, трудно интерпретировать, что создает проблемы при необходимости объяснить гражданам или контролирующим органам основания для объединения их персональных данных.

Проведенный обзор показывает, что для городских информационных систем не существует универсального «лучшего» метода связывания данных. Детерминированные методы незаменимы для быстрого и точного связывания по уникальным идентификаторам, но недостаточны для обеспечения полноты охвата населения. Вероятностные методы обеспечивают статистически обоснованные оценки и устойчивы к ошибкам, однако требуют качественной обучающей выборки и предположений о независимости атрибутов. Методы машинного обучения достигают наилучших показателей точности и полноты при наличии размеченных данных, но страдают от проблем интерпретируемости и экстраполяции.

Перспективным направлением представляется разработка гибридных методов, сочетающих статистическую надежность вероятностных моделей с контекстными возможностями графовых представлений. Такой синтез позволит создать систему семантической композиции, способную эффективно функционировать в гетерогенной и динамичной среде городских данных, обеспечивая как высокое качество связывания, так и адаптивность к изменениям без постоянного вмешательства человека.

Биоинспирированный подход: концепция семантической самоорганизации данных в СУБД

При выборе конкретного биоинспирированного алгоритма для решения задачи семантической композиции данных было рассмотрено несколько перспективных подходов: муравьиные алгоритмы, пчелиные алгоритмы, генетические алгоритмы и поведение слизевика. Каждый из этих методов имеет свои особенности и области эффективного применения.

Муравьиные алгоритмы, несмотря на их широкое применение в задачах оптимизации маршрутов и комбинаторной оптимизации, имеют существенное ограничение для нашей задачи: они требуют предварительного определения графа возможных связей, что в условиях семантической гетерогенности городских данных является нетривиальной задачей. Кроме того, муравьиные алгоритмы работают с дискретными решениями, тогда как задача семантической композиции требует непрерывной оценки степени соответствия записей. Генетические алгоритмы, хотя и способны обрабатывать сложные пространства решений, страдают от высокой вычислительной сложности и необходимости тщательной настройки параметров (размер популяции, вероятности мутации и кроссовера), что делает их менее пригодными для динамической среды городских информационных систем.

В отличие от вышеуказанных методов модель на основе Physarum polycephalum обладает рядом уникальных свойств, делающих ее особенно подходящей для задачи семантической интеграции гетерогенных данных.

Организм начинает исследование пространства решений без априорной информации о расположении источников пищи, что напрямую соответствует ситуации в городских системах, где отсутствует достоверная разметка связей между записями из различных источников. В процессе поиска решения слизевик формирует сеть протоплазматических трубочек, где толщина каждого канала отражает степень уверенности в связи – этот естественный механизм обработки неполной и зашумленной информации идеально соответствует задаче установления семантических связей с различной степенью достоверности. Ключевой особенностью является способность системы к непрерывной адаптации: благодаря комбинации положительной обратной связи для укрепления востребованных связей и отрицательной для «испарения» устаревших отношений модель динамически перестраивается в ответ на изменения данных и запросов пользователей. При этом глобальная оптимизация сети достигается исключительно за счет локальных взаимодействий – каждая точка принимает решения только на основе непосредственного окружения, что обеспечивает высокую масштабируемость даже при работе с большими объемами данных. Значимым преимуществом является и естественная интерпретируемость результата: визуально понятная структура сети с различной «толщиной» каналов позволяет легко объяснить и обосновать установленные семантические связи, что особенно важно для государственных систем, где требуется прозрачность принимаемых решений перед пользователями и контролирующими органами.

Исследования последних лет подтверждают, что модель Physarum polycephalum особенно эффективна в задачах, требующих динамической реконфигурации связей и адаптации к изменяющимся условиям – именно тех, которые характерны для городских информационных систем. В отличие от других биоинспирированных подходов Physarum polycephalum демонстрирует естественную способность к балансировке между эксплуатацией уже установленных связей и исследованием новых возможных соединений, что делает его оптимальным выбором для создания отказоустойчивой и адаптивной системы семантической композиции данных.

В 2010 году научный мир потряс эксперимент, проведенный группой японских исследователей под руководством Тосиюки Накагаки из Университета Хоккайдо. Объектом исследования стал удивительный организм – Physarum polycephalum (многоголовая слизевая плесень), представляющий собой гигантскую многоядерную амебоподобную клетку, которая обитает в темных и влажных местах, питаясь спорами грибов и бактериями. Несмотря на отсутствие нервной системы и каких-либо зачатков интеллекта этот организм демонстрирует поразительные способности к решению сложных оптимизационных задач. В своем классическом эксперименте исследователи поместили слизевика в центр чашки Петри, которая моделировала карту Токийского залива. В точках, соответствующих расположению крупных железнодорожных станций (включая аэропорт Нарита), были размещены источники пищи – кусочки овсяных хлопьев.

Начав равномерно расползаться во все стороны, слизевик постепенно формировал сеть протоплазматических трубочек, соединяющих источники пищи. Ключевым моментом является то, что организм не просто соединял все точки со всеми, а оптимизировал сеть: толстые, активно используемые каналы (по которым интенсивно транспортировалась цитоплазма с питательными веществами) укреплялись и сохранялись, в то время как тонкие, неэффективные ответвления атрофировались и исчезали. Процесс управлялся простым механизмом положительной обратной связи: чем больше питательных веществ проходило по каналу, тем шире он становился, и тем больше питательных веществ мог пропустить в дальнейшем. Когда ученые наложили итоговую сеть, построенную слизевиком, на реальную карту токийского метрополитена, они обнаружили поразительное сходство. Биологическая сеть не просто напоминала творение рук человеческих – в ряде аспектов она превосходила его, демонстрируя лучший баланс между общей протяженностью, эффективностью соединений и устойчивостью к случайным повреждениям.

Это открытие породило новое направление в науке – использование биологических моделей для решения инженерных задач оптимизации транспортных, коммуникационных и логистических сетей.

Возникает закономерный вопрос: какое отношение эксперимент с плесенью имеет к задаче интеграции данных в городских информационных системах? На первый взгляд – никакого. Действительно, что общего между полужидкой амебообразной массой и строгими базами данных с их реляционными моделями и алгоритмами? Кажется, что биология и информатика находятся по разные стороны баррикад. Однако если посмотреть глубже, отвлекаясь от физической природы объектов и фокусируясь на сути процессов, обнаруживается, что и слизевик, и разработчик ГИС решают одну и ту же абстрактную задачу. В обоих случаях имеется множество разрозненных точек, которые необходимо связать в единую сеть. В обоих случаях эта сеть должна быть экономичной, эффективной и устойчивой. И главное – оба действуют в условиях неполноты информации: слизевик не знает, где именно лежит еда, а интегратор данных не знает заранее, какие записи относятся к одному объекту. Именно на этом уровне абстракции и становятся видны фундаментальные аналогии между поведением живого организма и логикой построения семантических связей.

В таблице 1 представлено соответствие между элементами биологического эксперимента и компонентами задачи связывания данных.

Таблица 1 / Table 1

Элемент эксперимента со слизевиком	**Элемент задачи семантической композиции данных**
Чашка Петри (пространство поиска)	Таблицы базы данных, физическое пространство на диске
Кусочки овса (источники пищи)	Записи, потенциально относящиеся к одному объекту реального мира
Протоплазматические трубочки	Семантические связи между записями (реализованные как таблица связей, граф или специальные индексы)
Поток цитоплазмы (транспорт питания)	Поток запросов к базе данных при предоставлении госуслуг
Укрепление/атрофия трубочек	Увеличение/уменьшение веса семантических связей на основе частоты совместного использования записей
Итоговая сеть	Самоорганизующаяся структура базы данных: кластеры связанных записей, оптимизированные для наиболее частых запросов

Важно подчеркнуть, что предлагаемый подход реализуется непосредственно на уровне хранения данных. В отличие от внешних сервисов связывания записей или ETL-процессов, выполняющих однократную трансформацию, метод встраивается в логику работы СУБД и создает внутри нее динамическую структуру семантических связей.

Несмотря на разнообразие рассмотренных выше подходов – от детерминированных правил до методов машинного обучения – их непосредственное применение для решения задачи семантической композиции в городских информационных системах сопряжено с существенными ограничениями. Данные ограничения обусловлены специфической природой городских данных, которые характеризуются следующими свойствами, плохо согласующимися с допущениями, заложенными в классические алгоритмы:

• **Динамичность.** Данные постоянно меняются: граждане меняют фамилии, адреса, появляются новые записи, исправляются ошибки. Система должна адаптироваться к этим изменениям в реальном времени, а не требовать периодического переобучения модели.

• **Неполнота и зашумленность.** В реальных данных редко встречаются идеальные совпадения. Методы, требующие точного соответствия ключевых полей, неизбежно теряют значительную часть полезной информации.

• **Контекстная зависимость.** Решение о том, относятся ли две записи к одному человеку, часто зависит от контекста – например, наличия общих детей, совместного проживания, оформления доверенностей. Статические правила не способны учитывать всю глубину таких взаимосвязей.

• **Масштаб и распределенность.** Городские системы могут содержать миллионы записей, распределенных по десяткам ведомственных баз. Централизованное управление такой структурой крайне сложно и неэффективно.

Поведение Physarum polycephalum предлагает элегантное решение перечисленных проблем благодаря следующим принципам, которые могут быть реализованы непосредственно внутри базы данных:

Исследование без предварительных знаний. Слизевик не имеет карты местности и не знает, где находится еда. Он начинает движение во всех направлениях, исследуя среду методом проб и ошибок. В контексте СУБД это означает, что система не требует предвари-

тельной настройки жестких правил сопоставления. На начальном этапе она создает множество потенциальных семантических связей с минимальным весом, используя эвристики нечеткого сравнения (расстояние Левенштейна, метафоны, Джаро-Винклера и др.). Эти связи записываются в специальную таблицу семантических связей (SEMANTIC_LINKS), которая становится частью базы данных.

Локальные взаимодействия и глобальная оптимизация. Каждая точка сети слизевика принимает решения только на основе локальной информации. Для СУБД это означает, что решение об усилении или ослаблении конкретной семантической связи принимается на основе локального контекста: частоты совместного использования именно этих двух записей в запросах. При этом в масштабах всей базы формируется глобально оптимальная структура – кластеры записей, соответствующие реальным объектам, наиболее востребованным в предоставлении госуслуг.

Положительная обратная связь и усиление успешных паттернов. Чем активнее используется канал у слизевика, тем шире он становится. В базе данных этот принцип реализуется следующим механизмом: когда выполняется запрос, использующий семантическую связь между записями A и B (например, при формировании единого профиля гражданина), специальный триггер или фоновый процесс увеличивает вес этой связи в таблице SEMANTIC_LINKS и обновляет отметку времени последнего доступа.

Отрицательная обратная связь и забывание. Неиспользуемые каналы у слизевика атрофируются. В базе данных регулярно (например, nightly batch process) запускается процедура «испарения»: веса всех семантических связей, которые давно не использовались, уменьшаются. Если вес падает ниже критического порога (например, 0.1), связь удаляется. Это защищает систему от «засорения» ложными срабатываниями и обеспечивает адаптивность: если данные исправлены или устарели, система автоматически перестраивается, удаляя устаревшие связи.

Предлагаемая концепция предполагает создание внутри базы данных специального семантического слоя, который реализуется в виде совокупности таблиц и процессов.

Ядро данных составляет основу семантического слоя. К нему относятся существующие таблицы с «сырыми» данными из различных ведомственных источников – ЗАГСа, Пенсионного фонда, МВД и других. Принципиально важным является то, что структура этих таблиц не модифицируется: они продолжают функционировать в своем исходном виде, обеспечивая автономность источников и неизменность логики их работы. Единственным требованием является наличие у каждой записи уникального внутреннего идентификатора, который будет использоваться для построения семантических связей.

Центральным элементом предложенного подхода выступает таблица семантических связей. Она представляет собой физическую реализацию графа семантических отношений между записями из различных источников. Каждая запись в этой таблице соответствует одному ребру графа и содержит следующие атрибуты:

- link_id – уникальный идентификатор связи;
- record_id_a – идентификатор записи из источника A;
- record_id_b – идентификатор записи из источника B;
- link_strength – вес связи (аналог толщины трубочки слизевика), вещественное число в диапазоне [0, 1];
- first_created – дата первого обнаружения потенциальной связи;
- last_accessed – дата последнего использования связи в запросе;
- access_count – счетчик количества обращений к данной связи.

Благодаря индексации по идентификаторам записей и весу связей таблица обеспечивает эффективный доступ к семантической информации даже при миллионах записей.

Функционирование семантического слоя обеспечивается тремя основными процессами, реализующими ключевые принципы поведения слизевика.

Процесс-исследователь представляет собой фоновый компонент, который периодически сканирует новые и измененные записи в ядре данных. Для каждой такой записи он выполняет нечеткое сравнение с существующими записями из других источников, используя метрики семантического сходства (расстояние Левенштейна, коэффициент Жаккара, сравнение по метафонам и др.). Если обнаруживается потенциальное соответствие, в таблицу семантических связей добавляется новая запись с минимальным начальным весом (например, 0.1). Таким образом, система постоянно находится в режиме «зондирования» среды, выявляя возможные семантические отношения еще до того, как они будут востребованы в реальных запросах.

Процесс-усилитель интегрирован с механизмом обработки запросов к базе данных. Каждый раз, когда пользовательский запрос (например, на предоставление государственной услуги) использует семантическую связь между записями, этот процесс увеличивает вес соответствующей связи в таблице и обновляет счетчики last_accessed и access_count. Чем чаще связь востребована в реальной работе, тем выше становится ее вес. Этот механизм реализует принцип положительной обратной связи: семантические отношения, подтвержденные практикой использования, укрепляются и становятся более «заметными» для системы.

Наконец, процесс-испаритель реализует механизм «забывания», критически важный для адаптивности системы. Это фоновый процесс, запускаемый периодически (например, ежесуточно), который уменьшает вес всех семантических связей пропорционально времени, прошедшему с момента последнего доступа. Если после нескольких циклов испарения вес связи падает ниже установленного порогового значения (например, 0.05), связь удаляется из таблицы как недостоверная или устаревшая. Это защищает систему от «засорения» ложными срабатываниями и позволяет автоматически отбрасывать семантические отношения, переставшие быть актуальными.

Взаимодействие описанных компонентов обеспечивает непрерывную самоорганизацию семантической структуры данных. Исследователь создает потенциальные связи, усилитель укрепляет востребованные, а испаритель элиминирует неиспользуемые. В результате семантический слой постоянно эволюционирует, отражая реальные паттерны использования данных в процессах предоставления государственных услуг.

Реализация семантической композиции непосредственно внутри СУБД на основе принципов самоорганизации открывает ряд существенных преимуществ по сравнению с существующими подходами.

Прежде всего предлагаемая система не требует так называемого «холодного старта». В отличие от методов машинного обучения, нуждающихся в предварительно размеченной обучающей выборке, или экспертных систем, требующих трудоемкой настройки правил, предложенный подход начинает функционировать немедленно после развертывания. На начальном этапе таблица семантических связей может быть пуста или содержать минимальное количество потенциальных связей, созданных процессом-исследователем. По мере эксплуатации системы, под воздействием реальных пользовательских запросов, семантическая структура постепенно «обрастает» связями, причем вес этих связей отражает их практическую востребованность.

Критически важным свойством системы является ее способность к непрерывной адаптации. Модель семантических связей эволюционирует параллельно с изменением самих

данных и паттернов их использования. Если в ведомственных базах данных меняются форматы ввода, появляются новые типичные опечатки или смещаются акценты в предоставлении государственных услуг, семантический слой автоматически подстраивается под эти изменения без какого-либо вмешательства администратора. Связи, переставшие быть актуальными, постепенно «испаряются», а новые, востребованные, укрепляются.

Важными достоинствами предложенного подхода являются его прозрачность и интерпретируемость. В отличие от моделей глубокого обучения, работающих по принципу «черного ящика», семантические связи в предлагаемой системе имеют четкую и понятную интерпретацию: вес связи непосредственно показывает, насколько часто данная пара записей использовалась совместно в реальных запросах. При возникновении необходимости всегда можно проследить историю формирования любой связи, проанализировав временные метки ее создания и последних обращений, что особенно важно для систем государственного управления, предъявляющих повышенные требования к обоснованности принимаемых решений.

Предложенная архитектура обладает хорошей масштабируемостью, поскольку таблица семантических связей индексируется стандартными средствами реляционной СУБД. Это обеспечивает эффективный доступ к семантической информации даже при миллионах записей в ядре данных и десятках миллионов связей между ними. Использование стандартных механизмов индексации и оптимизации запросов позволяет достичь приемлемой производительности без привлечения специализированных высоконагруженных решений.

Наконец, важно отметить автономность и отказоустойчивость предложенного подхода. Поскольку вся логика семантической композиции реализована внутри самой СУБД с использованием встроенных механизмов (таблиц, индексов, фоновых процессов), система не зависит от доступности внешних сервисов связывания записей. Даже при временной недоступности отдельных ведомственных источников или каналов связи семантический слой продолжает функционировать, оперируя уже установленными связями и обеспечивая непрерывность предоставления государственных услуг.

Для того чтобы четко определить научную новизну предлагаемого решения, необходимо позиционировать его относительно существующих классов методов интеграции данных.

В отличие от традиционных ETL-процессов (Extract, Transform, Load), которые выполняют однократное преобразование данных на этапе загрузки в хранилище, предложенный подход работает непрерывно, в темпе поступления запросов. ETL-подходы строят статическую модель связей, которая со временем устаревает и требует периодического перезапуска. Напротив, семантический слой на основе принципов самоорганизации постоянно эволюционирует, отражая актуальную структуру связей между данными.

По сравнению с детерминированными методами, основанными на жестких правилах сопоставления, задаваемых экспертами, предлагаемый подход обладает способностью к самообучению. Детерминированные правила требуют трудоемкой ручной настройки и не адаптируются к изменениям в данных – любое изменение формата или появление новых типов ошибок влечет за собой необходимость ручной корректировки правил. В предлагаемой системе эти процессы автоматизированы: семантические связи формируются и укрепляются на основе реального опыта использования данных.

Важнейшее отличие прослеживается при сравнении с методами классического машинного обучения. Последние, будь то вероятностные методы связывания записей (например, алгоритм Феллеги-Сантера) или современные подходы на основе нейронных сетей, требуют размеченной обучающей выборки для настройки параметров модели. Получение такой разметки в масштабах городских информационных систем крайне

трудоемко, а часто и вовсе невозможно. Более того, обученные модели нуждаются в периодическом переобучении по мере старения данных. Предлагаемый подход лишен этого недостатка: он учится «на лету», используя обратную связь от реальных запросов как единственный источник обучающих сигналов.

Наконец, принципиальным отличием является встроенность предложенного решения в СУБД. Существующие подходы часто реализуются в виде внешних сервисов связывания записей (record linkage services), к которым обращаются прикладные системы. Такая архитектура создает дополнительную сетевую задержку, вносит потенциальную точку отказа и усложняет развертывание и сопровождение. Таким образом, предлагаемая концепция семантического «слизевика» занимает уникальную нишу в пространстве методов интеграции данных, представляя собой гибридный подход, сочетающий адаптивность и самоорганизацию, свойственные биологическим системам, с надежностью, производительностью и встроенностью, характерными для реляционных технологий хранения данных. Формальная математическая модель предложенного подхода, включая функции инициализации связей, усиления и испарения, а также алгоритмы обработки запросов с использованием семантического слоя, будет представлена в следующей статье.

Заключение

В проведенном исследовании осуществлен системный анализ проблемы семантической композиции гетерогенных данных в городских информационных системах, функционирующих в рамках концепции «Умный город» и обеспечения государственных услуг. Основным научным результатом работы является концептуальное обоснование биоинспирированного подхода к семантической композиции, основанного на принципах самоорганизации, наблюдаемых в поведении миксомицета Physarum polycephalum. Проведенная аналогия между формированием оптимизированной транспортной сети в эксперименте с токийским метро и задачей установления семантических связей между записями в базах данных позволила выявить ключевые механизмы, применимые в контексте информационных систем: исследование среды без априорных знаний, локальные взаимодействия, приводящие к глобальной оптимизации, положительная обратная связь для укрепления успешных паттернов и отрицательная обратная связь для «забывания» устаревших связей.

Предложена архитектура семантического слоя, реализуемого непосредственно внутри СУБД и включающего ядро данных, таблицу семантических связей, а также три управляющих процесса – исследователь, усилитель и испаритель. Детально описаны функции каждого компонента и их взаимодействие, обеспечивающее непрерывную самоорганизацию семантической структуры данных в процессе эксплуатации системы.

Сформулированные преимущества предлагаемого подхода – отсутствие необходимости в «холодном старте», непрерывная адаптация к изменениям, прозрачность и интерпретируемость, масштабируемость, автономность и отказоустойчивость – позволяют позиционировать его как перспективную альтернативу существующим методам интеграции данных. Проведенное сравнение с детерминированными подходами, ETL-процессами, методами машинного обучения и внешними сервисами связывания записей подтверждает уникальность предлагаемого решения, сочетающего адаптивность биологических систем с надежностью реляционных технологий хранения данных.

Дальнейшие исследования будут направлены на формализацию математической модели предложенного подхода, включая функции инициализации связей, усиления и испарения, а также разработку алгоритмов обработки запросов с использованием семантического слоя и их экспериментальную апробацию на реальных данных городских ведомственных систем.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Pashkov D., Shvets A., Karpov N. et al. SemConvTree: semantic convolutional quadtrees for multi-scale event detection in smart city. *Smart Cities.* 2024. Vol. 7. No. 5. Pp. 2763–2780.

2. *Кулиев Э. В., Запорожец Д. Ю., Кравченко Ю. А., Семенова М. М.* Решение задачи интеллектуального анализа данных на основе биоинспирированного алгоритма // Известия ЮФУ. Технические науки. 2022. № 1. С. 90–102.

Kuliev E.V., Zaporozhets D.Yu., Kravchenko Yu.A., Semenova M.M. Solving the problem of intelligent data analysis based on a bioinspired algorithm. *Bulletin of the Southern Federal University. Technical Sciences*. 2022. No. 1. Pp. 90–102. (In Russian)

3. *Грибкова В. Б., Журавлев М. А., Ковалев С. М.* Методы семантической интеграции разнородных данных в городских информационных системах // Информационные технологии в управлении. 2023. № 2. С. 45–53.

Gribkova V.B., Zhuravlev M.A., Kovalev S.M. Methods of semantic integration of heterogeneous data in urban information systems. *Information Technologies in Management*. 2023. No. 2. Pp. 45–53. (In Russian)

4. Schumann A., Papanikolaou N., Tsompanas M.A., Adamatzky A. A Physarum-inspired approach to the Euclidean Steiner tree problem. *Scientific Reports.* 2022. Vol. 12. Article 14536.

5. *Козлов В. А., Гагарин Ю. Е.* Применение алгоритма ACO (муравьиный алгоритм) в туманных вычислениях // Наукоемкие технологии в приборо- и машиностроении и развитие инновационной деятельности в вузе: материалы Региональной научно-технической конференции (Калуга, 23–25 апреля 2024 года). М.: Издательство МГТУ им. Н.Э. Баумана, 2025. Т. 2. С. 112–114.

Kozlov V.A., Gagarin Yu.E. Application of the ACO algorithm (ant colony algorithm) in fog computing. *Science-intensive technologies in instrument and mechanical engineering and the development of innovative activities in universities: Proceedings of the Regional Scientific and Technical Conference (Kaluga, April 23–25, 2024)*. Moscow: Izdatel'stvo MGTU im. N.E. Baumana, 2025. Vol. 2. Pp. 112–114. (In Russian)

6. Ghafarollahi A., Buehler M.J. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*. 2024.

7. *Аванесян А. А.* YandexGPT в применении к GraphRAG: формирование графов и семантический анализ графовых сообществ [выпускная квалификационная работа]. М.: НИУ ВШЭ, 2025.

Avanesyan A.A. YandexGPT as applied to GraphRAG: graph formation and semantic analysis of graph communities [graduate qualification work]. Moscow: HSE University, 2025. (In Russian)

8. Scrocca M., Comerio M., Carenini A., Cecconi A. Intelligent urban traffic management via semantic interoperability across multiple heterogeneous mobility data sources. arXiv preprint arXiv:2407.10539. 2024.

9. Torre-Bastida A.I., Osaba E., Muhammad K. et al. Bio-inspired computation for big data fusion, storage, processing, learning and visualization: state of the art and future directions. *Neural Computing & Applications*. 2025. Vol. 37. No. 28. Pp. 23097–23127.

10. Yan Z., Zheng X., Zhang J. A survey on semantic data integration for smart city applications. *IEEE Access*. 2023. Vol. 11. Pp. 123456–123478.

11. Chataut R., Panday S. Image edge detection using ant colony optimization with genetic algorithm. *Journal of Science and Technology*. Vol. 5. Pp. 58–65. DOI: 10.3126/jost.v5i1.92657

12. Tavva G., Sura R., Shah D. et al. Meta-heuristic ant-bee colony algorithms for dynamic big data analysis in real-time. *IET Conference Proceedings*. 2025. Pp. 1374–1379. DOI: 10.1049/icp.2025.1594

13. Cai Zh., Li G., Zhang J. et al. Using an artificial Physarum polycephalum colony for threshold image segmentation. *Applied Sciences*. 2023. Vol. 13. Article 11976. DOI: 10.3390/app132111976

14. Kamaoğlu M. Physarum computation in architecture: a critique of bio-digital design. *Architectural Research Quarterly*. 2025. Vol. 28. Pp. 387–398. DOI: 10.1017/S1359135525101012

15. Awad A., Pang W., Lusseau D., Coghill G. A survey on Physarum polycephalum intelligent foraging behaviour and bio-inspired applications. *Artificial Intelligence Review*. 2022. Vol. 56. DOI: 10.1007/s10462-021-10112-1

Финансирование. Исследование проведено без спонсорской поддержки.

Funding. The study was performed without external funding.

Информация об авторе

Рыбаков Даниил Александрович, аспирант кафедры информатики, Российский экономический университет имени Г. В. Плеханова;
115054, Россия, Москва, Стремянный переулок, 36;
rybakov.daniel99@gmail.com, ORCID: https://orcid.org/0009-0005-8959-4427, SPIN-код: 7155-6461

Information about the author

Daniil A. Rybakov, Postgraduate Student, Department of Computer Science, Plekhanov Russian University of Economics;
36, Stremyannyy lane, Moscow, 115054, Russia;
rybakov.daniel99@gmail.com, ORCID: https://orcid.org/0009-0005-8959-4427, SPIN-code: 7155-6461