Адаптация технологии фабрик данных к разработке систем визуальной аналитики в области цифровой медицины

С.И. Чуприна [1]

[1] *ФГАОУ ВО «Пермский государственный национальный исследовательский университет», ул. Букирева, 15, г. Пермь, 614068, Россия*

Аннотация

В работе предлагается и обосновывается инновационный подход к расширению сферы применения технологии фабрик данных, которые в настоящее время используются практически исключительно для создания корпоративных хранилищ данных. Подход предусматривает доступ к цифровым данным конкретной персоны не на принципах их консолидации из разнородных источников, а на принципах виртуальной интеграции. Сервисы такой фабрики данных вместо сведений о подразделениях, производимой продукции, оказываемых услугах и пр. обеспечивают доступ по требованию к данным, поступающим с носимых персональных устройств, извлекаемым из бланков лабораторных и клинических исследований, опросников, электронных мед. карт в привязке к соответствующим системам органов человека. Взаимосвязи отражают родовидовые, причинно-следственные и временные зависимости между данными, а также линии «коммуникации», соответствующие основным физиологическим системам человеческого организма. Вводится понятие «персональной фабрики данных» и описывается концепция семантической интеграции, агрегации и визуальной аналитики данных в области цифровой 4П-медицины.

Ключевые слова

Фабрика данных, федерализация данных, онтологический инжиниринг, семантический веб, семантическая интеграция, граф знаний, 4П-медицина, визуальная аналитика.

To Adapt Data Fabric Technology to Visual Analytics Systems Development in the Field of Digital Medicine

S.I. Chuprina [1]

[1] *Perm State University, Perm, Bukireva Str. 15, 614990, Russia*

Abstract

The paper proposes an innovative approach to expanding the application scope of Data Fabric technology, which is currently used almost exclusively for the creation of corporate data warehouses. The approach provides access to digital data of a particular person not on the principles of their consolidation from heterogeneous sources, but on the principles of virtual integration. Services of such Data Fabric make available on-demand access to data coming from wearable personal devices, and data extracting from laboratory clinical research documents, questionnaires, electronic health records instead of information about corporation units, products, services provided, etc. These personal data are rendered related to the certain systems of human organism, and the relationships reflect both paradigmatic, causal and temporal dependencies between the data, and lines of "communication" corresponding to the basic physiological systems of the human body. The concept of a "personal data fabric" is introduced and the application of this concept to data semantic integration, aggregation and visual analytics in the field of digital 4P-medicine is described.

ГрафиКон 2023: 33-я Международная конференция по компьютерной графике и машинному зрению, 19-21 сентября 2023 г., Институт проблем управления им. В.А. Трапезникова Российской академии наук, г. Москва, Россия
EMAIL: chuprinas@inbox.ru (С.И. Чуприна)
ORCID: 0000-0002-2103-3771 (С.И. Чуприна)

Keywords
Data Fabric, data federation, ontology engineering, Semantic Web, semantic integration, knowledge graph, 4P-medicine, visual analytics.

1. Введение

В последние годы активно развиваются подходы к созданию нового поколения хранилищ данных (ХД) с использованием технологий построения современных интеллектуальных фабрик данных (Data Fabric, DF) [1], в которых интеграция данных из разнородных, в т.ч. неструктурированных, источников, включая как ресурсы интернет, так и корпоративные ХД, а также локальные ресурсы предприятий и их подразделений, выполняется не на принципах консолидации, а на принципах виртуальной интеграции (федерализации) без их физического перемещения в единое хранилище.

Фабрика данных может быть определена как интегрированное виртуальное хранилище, архитектура и сервисы которого обеспечивают смысловую взаимосвязь между данными, хранящимися в разнородных источниках вне зависимости от формата их хранения и технологии создания систем – источников данных. При этом данные остаются на исходном месте, а DF способны в онлайн режиме по запросу выдавать нужные данные в нужное время и доставлять их в нужное место, выполняя необходимую предобработку, агрегацию и аналитику по требованию.

Для решения задач интеграции разнородных данных и их визуальной аналитики интеллектуальные фабрики данных проектируются и реализуются преимущественно на принципах управляемой моделями архитектуры (MDA – Model Driven Architecture). Помимо моделей на базе UML и предметно-ориентированных языков, в последнее время для этих целей все активнее используются модели онтологий. В качестве инструментального окружения для создания фабрик данных можно использовать уже имеющиеся отечественные решения, в частности, российскую платформу DataFabric KGL (http://datafabric.cc/), созданную резидентом Сколково ООО "ДАТАФАБРИК". Как отмечается в обзоре аналитического агентства TAdviser (https://www.tadviser.ru/index.php/Статья:ИИ:_от_данных_–_к_знаниям), компания DataFabric создала универсальную платформу корпоративных онтологий DataFabric KGL, которая позволяет унифицировать доступ ко всем данным на предприятии с помощью платформы виртуализации данных с открытым исходным кодом. DataFabric KGL (или Logical DWH) предприятия реализуется на основе графовой (семантической) модели данных в терминах соответствующей предметной области. Она позволяет федеративно обращаться к гетерогенным данным из разных источников без необходимости их предварительного сбора, агрегации и хранения: обращение к данным идет в терминах предметной области через слой абстракции в виде бизнес-глоссария и не требует выполнения каких-либо операций на физическом уровне хранения данных.

Как показано далее, применение технологии фабрик данных для создания современных систем визуальной аналитики в области цифровой медицины и персональных, а не только корпоративных, фабрик данных вполне обосновано. Настройка на специфику конкретных предметных областей и источников данных осуществляется благодаря онтологическому слою знаний, описывающих смысловое содержание данных и семантические связи между ними в терминах устоявшихся понятий соответствующей предметной области. Такой подход позволяет выполнять смысловую интеграцию данных из разнородных источников, унифицировать средства их обработки и анализа, значительно упростить доступ к данным конечных пользователей, предоставляя им возможность единообразного обращения к данным, хранящимся в разных местах и в разных форматах, – в виде привычных запросов на естественном языке (ЕЯ) по аналогии с запросами в поисковых сервисах интернет. Это в значительной степени способствует успешной реализации на практике всех основных составляющих концепции 4П-медицины [2], таких как:

1. Персонализация (индивидуальный подход к каждому пациенту).
2. Предикция (создание вероятностного прогноза здоровья).
3. Превентивность (предотвращение развития заболеваний).
4. Партисипативность (мотивированное участие пациента).

Материал сознательно излагается с позиций «человекоцентричности», причем акцент делается не на персоне врача, а на персоне пациента («пациент-центрированность», «пациент-центричность» [2]) и доступности по его требованию (запросу) всех ассоциированных с ним персональных данных

из разных источников в нужное время и в нужном месте («персональная фабрика данных пациента»). С технологической точки зрения описываемый подход вполне соответствует также и концепции «персональной фабрики данных врача», в основе которой для полноценной аналитики и принятия врачебных решений лежит семантическая интеграция разнородных данных о пациентах конкретного врача.

Отметим, что серьезное рассмотрение вопросов информационной безопасности в соответствии с концепцией предлагаемого подхода является предметом отдельного обсуждения и выходит за рамки данной статьи.

2. Обоснование применения DF-технологий для реализации принципов 4П-медицины

Не секрет, что человеческий организм в процессе своей жизнедеятельности интенсивно генерирует большие объемы данных, которые обладают одной из важнейших характеристик Больших Данных (Big Data) – «Variety» (разнообразие, разнородность). Другое дело, что не все эти данные регулярно оцифровываются и сохраняются, а у самой персоны есть проблемы с доступом к своим собственным персональным цифровым данным, хранящимся в разных средах. Например, даже доступные пациенту данные в его личном кабинете некоторой МИС (Медицинской Информационной Системы), во-первых, доступны не в полном объеме, что мешает реализации принципов предикции и превентивности, если потребуется оперативный анализ данных вне рамок этой МИС, во-вторых, их переиспользование в рамках сторонних МИС других государственных или частных медучреждений, а также доступных аналитических платформ как для отдельных исследователей, так и для крупных медучреждений, например, санаторно-курортного лечения, зачастую невозможно или сильно затруднено.

Сложности объясняются тем, что требуется не просто конвертация данных из-за разности в форматах представления, а необходима автоматизированная предобработка и смысловая нормализация, агрегация и трансформация разнородных данных. Эти проблемы решаются в среде традиционных ХД, построенных на принципах консолидации, однако при этом требуется организация постоянного сопровождения единого ХД с целью синхронизации обновлений между источниками и ХД, иначе возникает проблема «свежести» данных. Такое постоянное сопровождение затратно не только для отдельных персон, но и для целых организаций, хотя в последнее время есть надежда, что расширение отечественных сервисов облачных хранилищ приведет к сокращению стоимости их услуг. Тем не менее, важно отметить, что если речь идет о персональном, а не о корпоративном, ХД, то на первое место выступает как раз потребность в предоставлении для аналитики данных именно «по требованию» (а не постоянно). Наравне с ретроспективными данными и результатами прошлых исследований должна быть доступна для мониторинга и анализа и наиболее свежая информация, которая может хранится как в базах данных сторонних МИС, так и на локальном персональном компьютере клиента (например, результаты лабораторных исследований в формате pdf, которые сделаны по инициативе самой персоны без обращения к врачу за направлением на анализы).

Мы считаем, что любой персоне по праву д.б. доступно большинство ее личных цифровых данных, включая цифровые данных из хранилищ сторонних систем с возможностью их виртуальной интеграции в процессе анализа с данными, хранящимися в разных форматах на ее персональном компьютере. Это необходимо для мониторинга состояния здоровья, оперативного анализа и принятия решений, например, чтобы срочно обратиться к врачу или обоснованно оперативно подкорректировать свой рацион питания. Представляется, что ни о какой реальной партисипативности не может быть и речи без того, чтобы персона была полноправным владельцем своих собственных оцифрованных персональных данных с правом их оперативного предоставления третьим лицам, например, сотрудникам санаторно-курортных учреждений или врачам скорой помощи.

С т.зр. объема к категории Больших Данных можно отнести большинство потоков данных свыше 100 Гб в день, что на современном этапе развития ИТ-технологий далеко не всегда характерно для большинства персональных медицинских данных. Однако со временем по мере развития интернета вещей и повсеместных вычислений (Ubiquitous Computing), совершенствования, роста востребованности и доступности носимых персональных устройств, цифровой след все большего числа персон будет постепенно приближаться по размеру к объёмам Больших Данных. Кроме того,

для целей аналитики, например, для учета факторов образа жизни человека, в состав персональных данных пациента, помимо чисто медицинских данных, должны входить также данные из т.н. «дневников здоровья» (ведет сам пациент; обычно хранятся локально на компьютере персоны), сведения из различных опросников и анкет с данными об экологических, психолого-социальных условиях проживания, образе жизни и жизнедеятельности человека, включая профессиональные факторы.

Вторая основная характеристика Больших Данных, Velocity, подразумевает не только скорость генерации/прироста/изменения данных, что также пока не характерно для персональных медицинских данных, но и необходимость их высокоскоростной обработки, что существенно для реализации всех основных принципов 4П-медицины.

Что касается третьей основной характеристики, Variety, где под разнообразием понимается разнородность данных и необходимость одновременной и унифицированной обработки различных типов как структурированных, полу-структурированных, так и неструктурированных данных, то она в полной мере присуща специфике персональных данных человека и соответствует целям, задачам и ключевым особенностям построения фабрик данных. Важно отметить, что эта третья характеристики по всеобщему признанию в настоящее время является основным драйвером развития технологий Больших Данных.

В контексте обоснования применимости DF-технологий к обработке персональных данных для решения задач 4П-медицины подчеркнем также сходство целей и задач, ведь именно автоматическое извлечение из данных новой полезной информации и знаний, в т.ч. неизвестных ранее закономерностей с тем, чтобы автоматически решать аналитические задачи и строить прогнозы, полностью согласуется с приведенными выше принципами предикции и превентивности 4П-медицины. Для мониторинга, интеграции свежих персональных данных пациента с ранее полученными прогнозами с целью оценки рисков здоровью и предотвращения развития заболеваний как нельзя лучше подходят современные DF-технологии. Используемые в них методы и средства, в частности, методы и средства семантического веба (Semantic Web) [1] не требуют построения и сопровождения нового хранилища данных на принципах консолидации.

3. Концепция семантической интеграции на принципах федерализации

Как известно, системы интеграции данных способны обеспечивать интеграцию на различных уровнях: физическом, логическом и семантическом. Интеграция данных на физическом уровне обычно сводится к конвертации данных из различных источников в некоторый единый формат их физического представления. Интеграция данных на логическом уровне обеспечивает возможность доступа к данным, содержащимся в различных источниках, в терминах единой глобальной схемы, которая описывает их совместное представление с учетом структурных, а в случае объектных моделей и поведенческих, свойств без учета семантических свойств данных. Поддержку единого представления данных с учетом их семантических свойств в контексте общего графа знаний о соответствующей предметной области обеспечивает интеграция данных на семантическом уровне. Интерфейс конечного пользователя скрывает все технические аспекты хранения источников данных: местоположение, формат хранения, структуру (если источник данных является структурированным), язык доступа.

На рисунке 1 для сравнения представлена упрощенная схема двух подходов к интеграции данных: традиционного на принципах консолидации (ETL – Extract, Transform, Load) и виртуального на принципах федерализации. В приложениях бизнес-аналитики (BI – Business Intelligence) и OLAP-системах в процессе ETL все необходимые доступные из разных источников данные загружаются в единое хранилище еще до того, как пользователи начнут их запрашивать. Перед загрузкой для приведения разнородных данных к единому представлению ХД используют глобальную эталонную схему данных. В результате выполнения ETL-процессов обеспечивается физическая и логическая интеграция данных. После загрузки данных в единое хранилище доступ к источникам не требуется, однако в корпоративных ХД возникает проблема «свежести» данных.

В отличие от консолидации, федерализация позволяет извлекать данные из различных источников, объединять их и выполнять аналитику в режиме реального времени без необходимости физического перемещения: данные остаются у их владельцев, а доступ к ним осуществляется по требованию (on-demand access), т.е. в ответ на запрос, когда возникает необходимость в процессе

анализа. Таким образом, федерализация поддерживает единое виртуальное пространство для множества разнородных источников данных, обеспечивает их семантическую интеграцию и не требует затрат на создание и поддержку единого физического ХД. Пользователь формулирует запросы на ЕЯ в терминах графа знаний, что позволяет выполнять семантический поиск и смысловую интеграцию данных с учетом описания семантических взаимосвязей между элементами данных, представленных в графе знаний. Все необходимые преобразования данных осуществляются в процессе их извлечения из источников. Важно отметить, что виртуальная интеграция способствует обеспечению правил политики безопасности и лицензионных ограничений, если запрещено прямое копирование данных из исходных систем.

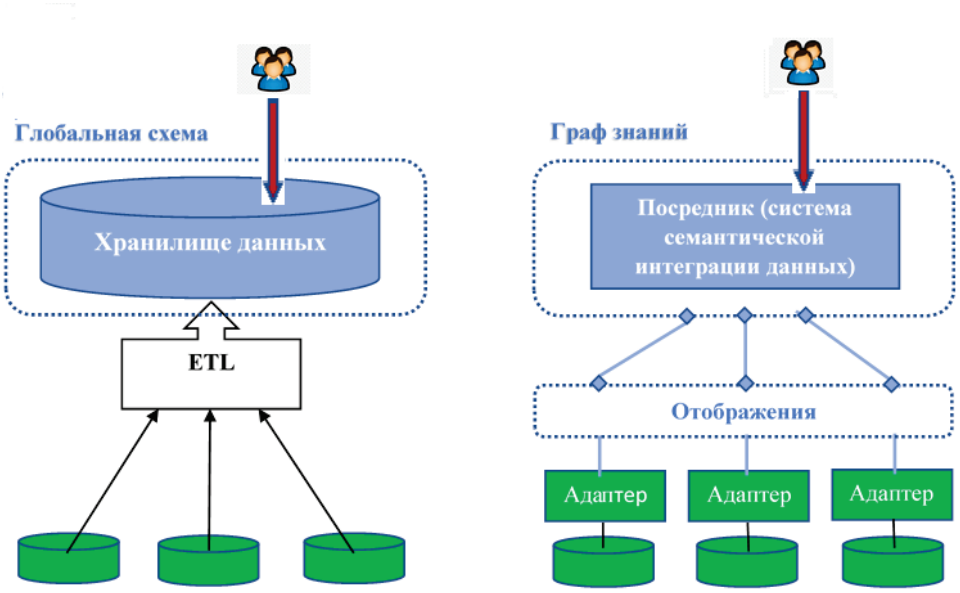

Рисунок 1 – Схемы интеграции на принципах ETL-консолидации (слева) и федерализации (справа)

Логически виртуализация осуществляется за счет дополнительного промежуточного уровня, изолирующего физическое хранение данных от приложений. Центральным компонентом системы интеграции на принципах федерализации является т.н. посредник (mediator), который интегрирует данные, полученные от адаптеров (wrappers). Адаптеры — компоненты, обеспечивающие единообразное взаимодействие посредника с источниками данных (в терминах единой модели). Посредник поддерживает единый пользовательский интерфейс на основе глобального графа знаний, описывающего метаданные и смысловое содержание данных из различных источников, а также осуществляет поддержку отображения между глобальным и локальным представлениями данных. Пользовательский запрос, сформулированный в терминах единого интерфейса, автоматически декомпозируется на множество подзапросов, адресованных к нужным локальным источникам данных. На основе результатов их обработки синтезируется полный ответ на запрос.

Именно благодаря указанным выше преимуществам фабрики данных реализуют семантическую интеграцию на принципах федерализации. Недостатки федерализации могут проявиться в снижении производительности работы аналитических сервисов за счет дополнительных затрат на доступ к многочисленным источникам данных, особенно при условии их большого объема. Однако, как отмечалось выше, специфика персональных данных пациентов такова, что две из трех основных характеристик больших данных (объем и скорость поступления/изменения) не являются их основными характерными чертами, а смысловая связность данных из разных источников гарантирована, как минимум, их принадлежностью к одной и той же персоне, что в значительной степени нивелирует указанную проблему.

Подходы к реализации систем семантической интеграции разнородных данных на основе посредников в начале 90-х годов базировались на использовании реляционных моделей данных: основная метамодель – ODMG-93; затем с конца и до 2004-2005 гг. для этих целей в основном применяли онтологии: основная метамодель – языки описания онтологий; начиная с 2004 года акцент сместился на применение технологий семантического веба: основная метамодель – RDF. Семантический веб поднял роль онтологий на новый, можно сказать, «всемирный» уровень, т.к. именно они играют ключевую роль в автоматической смысловой разметке и семантической

интеграции данных из разных ресурсов интернет вне зависимости от того, на каком естественном языке представлен тот или иной веб-ресурс.

Развитие технологий Semantic Web дало толчок, в первую очередь, виртуальной интеграции, что сделало этот подход более перспективным. Хотя нельзя не отметить, что в последнее время имеет место сближение разных подходов к интеграции данных [3]. При реализации современных корпоративных фабрик данных используют интеграцию на принципах федерализации, где в качестве графа знаний выступают онтологии. Именно этот подход мы и предлагаем использовать, расширив его применение для поддержки принципов 4П-медицины и создания персональных фабрик данных пациентов с аналитическими сервисами на базе семантической интеграции разнородных персональных данных.

4. Технологические аспекты построения персональных фабрик данных для систем визуальной аналитики

Как отмечалось выше, фабрики данных строятся на принципах семантической виртуальной интеграции разнородных источников данных на основе их согласованной модели, представленной в виде графа знаний. Сервисы DF позволяют конечным пользователям задавать ЕЯ-запросы в терминах этого графа знаний ко всему виртуальному пространству данных, не заботясь о месте их реального хранения. А т.к. эти данные могут хранится и в текстовом виде, и в форматах баз данных (БД), и в виде веб-ресурсов, то для их унифицированной обработки методами NLP (Natural Language Processing) граф знаний содержит метаданные о всех интегрируемых источниках данных, включая как сведения о формате хранения и инструментальном окружении, так и спецификацию их смыслового и прагматического содержания, а также сведения о синонимах (с учетом контекста), аббревиатурах и сокращениях.

Если говорить не о корпоративных, а о персональных фабриках данных, с акцентом на область медицины, то построение графа знаний здесь осложняется тем, что в данной области много не общепринятых, но устоявшихся и постоянно циркулирующих в том или ином сообществе профессионалов терминов, аббревиатур и сокращений. Причем одинаковым обозначениям могут соответствовать разные понятия. Например, «СР» может означать «боль в груди» при введении данных кардиологом или врачом первичной помощи, но может означать и «церебральный паралич» при введении данных невропатологом или педиатром.

Наш подход предполагает, что помимо результатов анализа данных, хранящихся в среде разнообразных ИС, например, государственных или частных МИС, доступен также анализ на основе интеграции с данными из интернет, из частного или публичного облака и из личных файлов персоны, хранящихся локально (на компьютере пользователя, на flash-носителе и т.п.) зачастую в виде «плоских» текстовых файлов или электронных таблиц с персональными данными мониторинга давления, дневного рациона питания, физической нагрузки и т.п. Акцент на персоноцентричность с обязательностью предполагает в случае необходимости автоматической обработки личных файлов персоны ориентацию и на бытовую лексику пользователя, не являющегося профессионалом в области медицины: при описании жалоб на плохое самочувствие персона может использовать общеупотребительную, а не профессиональную лексику. Поэтому проблеме автоматического установления смыслового соответствия и снятия многозначности в автоматической интерпретации ЕЯ-текстов уделяется особое внимание. Помимо традиционных NLP-сервисов для этих целей, в частности, для учета синонимов в контексте, мы используем методы и средства онтологического инжиниринга в купе с доступными медицинскими цифровыми словарями и справочниками. На рисунке 2 схематично представлены разнородные источники, в т.ч. носимые устройства, генерирующие персональные данные, и некоторые инструменты персональной фабрики данных.

Основываясь на нашем опыте прошлых разработок [4] и совместных проектов с ООО БИОГЕНОМ [5–7], мы предлагаем в качестве графа знаний использовать ни одну единую большую онтологию, а согласованную совокупность (семейство) онтологий, имеющих общую модель. Онтологии хранятся в т.н. смарт-репозитории, работа которого управляется метаонтологией, содержащей метаданные о самих онтологиях (дату, авторство, местоположение, привязку к категории описываемых данных типа «Группа заболеваний», «Медкарта», «Опросник», «Анкета», «Справочник», «Дневник здоровья» и др.), используемые для семантической индексации при поиске и слиянии онтологий. Различные методы и средства объединения онтологий хорошо изучены и активно используются на практике [8].

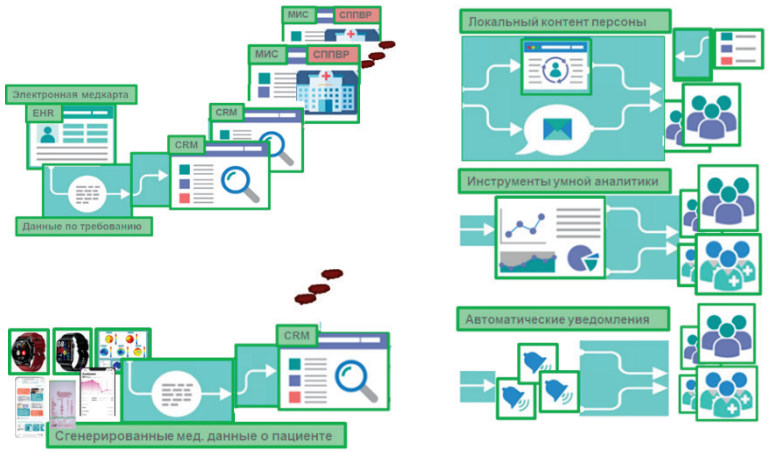

Рисунок 2 – Разнородные источники персональной фабрики данных пациента

В процессе парсинга ЕЯ-запросов посредник сначала находит в репозитории нужные онтологии и при необходимости устанавливает между ними соответствие, которое сохраняется в виде спецификации отображений на принципах OBDA-мэппинга [9]. Такое соответствие в случае неизменности онтологий устанавливается 1 раз и сохраняется в репозитории как метаданные для дальнейшего использования. В случае внесения изменений в соответствующие онтологии, монитор обновления онтологий отслеживает этот факт и автоматически запускает повторный мэппинг. Далее посредник при помощи адаптеров организует доступ к необходимым данным и их агрегацию в ответ на запрос. На рисунке 3 представлен фрагмент онтологии анкетных данных персоны, часть которых хранится в формате БД, а часть – в текстовом виде и в формате xml, т.е. в разнородных источниках.

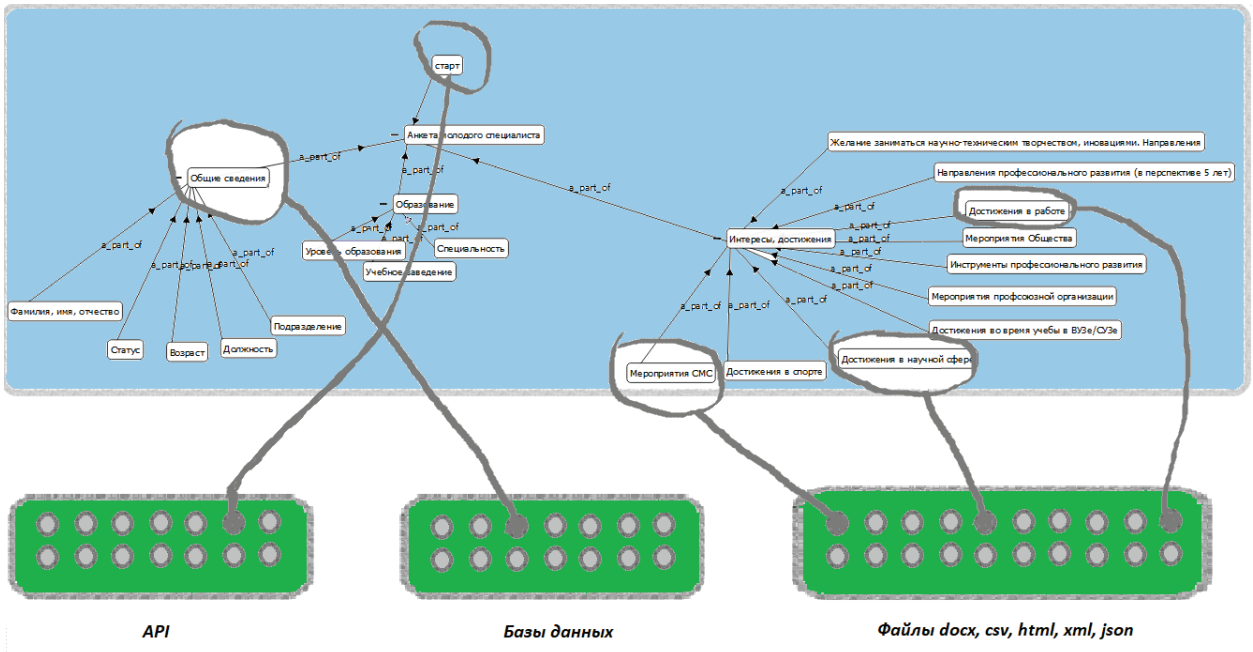

Рисунок 3 – Фрагмент онтологии анкетных данных персоны для семантической интеграции разнородных данных

Одной из особенностей нашего подхода является то, что мы разрабатываем т.н. легковесные (lightweight) онтологии [10], модель которых включает в себя тезаурус понятий предметной области и базовый набор поддерживаемых типов связей, но без аксиоматики. Т.е. аксиоматика не специфицируется в явном виде как отдельный компонент модели, а интерпретируется программными механизмами системы. Однако в нашем подходе эти механизмы интерпретации поддерживаемых типов связей и ограничений являются онтологически управляемыми, т.к. специфицируются в атрибутах вершин и дуг онтологии и автоматически учитываются ризонером (англ., reasoner) в ходе логического вывода. Это позволяет при необходимости расширять поддерживаемый набор базовых типов связей. Причем такие спецификации имеют json-подобный формат, что позволяет унифицированным образом описывать как путь к библиотечной функции,

интерпретирующей семантику некоторого типа связи (в спецификации указан путь для вызова функции после ключевого слова «path»), так и другие категории метаданных (спецификация каждой категории начинается с указания ее уникального идентификатора). Ризонер в ходе логического вывода обходит граф онтологии и, обнаружив непустой атрибут текущего узла/дуги, интерпретирует спецификацию метаданных в соответствии с их категорией. Например, если в описании атрибутов присутствует спецификация «path», то автоматически запускается на исполнение соответствующая функция из репозитория системы, аргументы которой также известны из спецификации и извлекаются из онтологии.

Обращение к другим сервисам системы выполняется аналогично. Например, если при решении какой-то аналитической задачи ризонер находит в атрибутах вершин/дуг текущей прикладной онтологии спецификации для запуска методов машинного обучения, то происходит вызов соответствующих библиотечных функций, хранящихся в репозитории системы. В примере на рисунке 3 в атрибутах вершины «старт» содержится обращение к API для запуска ризонера. Несмотря на то, что наш подход предполагает при необходимости использование вместо одной онтологии целого семейства взаимосвязанных онтологий, все они имеют общую модель, что позволяет применять один интерпретатор для разных онтологий. В частности, для интерпретации взаимосвязей типа «a_part-of» («часть-целое») используется одна и та же функция для разных онтологий (см. рисунок 3).

Ниже описана модель онтологии, используемая нами для реализации онтологически управляемых средств визуального анализа, легко адаптируемых к персональным предпочтениям: источнику данных и его местоположению; типу графика, отображающему результат анализа данных; структуре данных (описанию соответствия полей структуры данных элементам графика с указанием, какие измерения данных откладываются по осям абсцисс, ординат и аппликат, если график трёхмерный, каков отображаемый интервал и т. п.); цветовой схеме, используемой при визуализации.

Предметом описания онтологии, фрагмент которой приведен на рисунке 4, являются программные сущности, используемые для визуализации, их свойства и взаимодействие, а также источники данных (на рисунке 4 в качестве источников данных указаны файлы формата csv со статистическими данными о росте заболеваний в Китае и США). Использование для реализации визуального анализа данных инструментальных средств, описанных в этой онтологии, хорошо зарекомендовало себя на практике при реализации адаптируемых средств визуализации с использованием сервисов платформы SciVi [4]. Множество концептов предметной области в этой онтологии включает в себя такие понятия, как «Data Source» (источник данных), «Data Field» (поле данных), «Chart» (диаграмма), «Chart View» (область построения диаграммы), «Color» (цвет) и т. п., часть таксономии видов диаграмм в NChart3D, конкретные экземпляры источников данных и диаграмм.

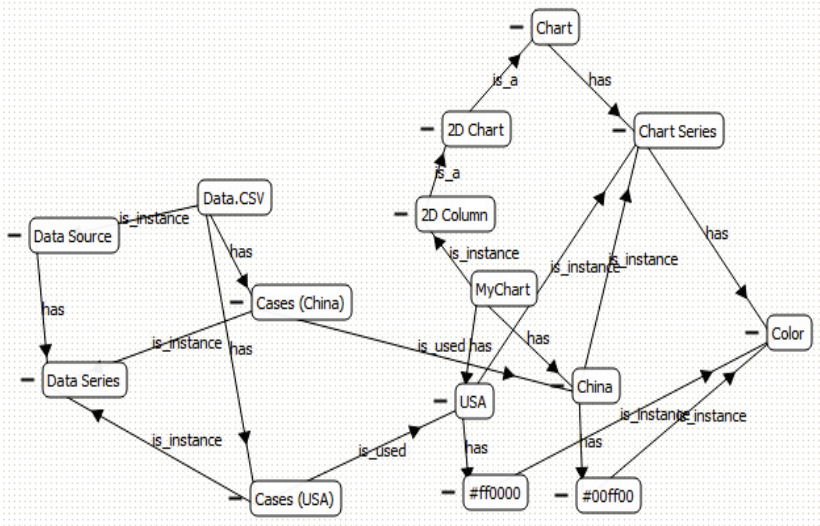

Рисунок 4 – Фрагмент прикладной онтологии для реализации онтологически управляемого визуального анализа данных

Модель включает следующее множество поддерживаемых типов связей между концептами:

1. is_a – связь типа «подкласс-класс», описывающая отношение между родительскими и дочерними понятиями с поддержкой наследования свойств родительского класса.
2. has – связь для описания одного понятия как свойства или атрибута другого.
3. is_instance – связь, описывающая отношение принадлежности некоторого экземпляра (индивида) некоторому классу (отношение типа «экземпляр-класс»).
4. is_used – связь, описывающая использование одним экземпляром другого посредством вызова в процессе работы программы.

Тип связи «is_used» очень важен при разработке онтологически-управляемых программных решений в задачах визуальной аналитики, так как именно он определяет фактические взаимодействия между экземплярами программных сущностей, описанных онтологией, и влияет на ход управления работой системы. Как отмечалось выше, множество описываемых аксиом в используемой нами модели онтологий является пустым, поскольку, как показал наш опыт использования онтологически управляемых сервисов платформы визуальной аналитики SciVi, для целей адаптации средств визуализации к специфике анализируемых данных и персональным предпочтениям пользователей достаточно прикладных легковесных онтологий.

Известно, что средства интерактивной и когнитивной компьютерной графики способствуют большей вовлеченности человека в процесс анализа, чем простое созерцание графического изображения. В контексте обсуждаемой тематики такие средства в большей степени работают на реализацию принципа партисипативности 4П-медицины. Для настройки средств визуализации на персональные предпочтения как конечных пользователей, так и разработчиков, мы используем методы и средства платформы визуальной аналитики SciVi, расширяя ее репозиторий моделями визуализации медицинской направленности. Какой способ визуализации для какого рода данных и знаний предпочитает та или иная персона, сохраняется в метаданных в привязке к ее онтологическому профилю и онтологиям соответствующих источников данных.

В качестве примера модели визуализации, которая удобна, по нашему мнению, для реализации в среде фабрик данных интерактивных средств когнитивной графики в составе сервисов визуального анализа и интеграции разнородных медицинских данных, может быть использована модель человеческого тела в виде схемы метро (Underskin Body Map), созданная Сэм Ломен (Sam Loman). На этой схеме (см. рисунок 5) вместо привычных линий метро отображены основные физиологические линии человеческого тела, маркированные различными цветами, например, красным цветом обозначены человеческие артерии, синим – венозная система. Пояснения к этим и другим цветам для маркировки мускульной системы человека, центральной нервной системы, дыхательной системы, системы лимфообращения, выделительной и пищеварительной систем содержатся в легенде для рисунка 5.

Основные органы перечисленных физиологических систем человеческого организма выделены отдельно в нечто подобное станциям на схеме линий метрополитена. Индустрия цифрового здравоохранения могла бы использовать методы визуализации, такие как «анатомические карты метро», чтобы лучше донести медицинские концепции до пациентов, пояснить проблемы с их текущем состоянием здоровья, а врачам – оказать помощь в анализе разнородных данных о пациенте. Такой способ визуализации физиологических систем человеческого организма хорошо масштабируется и подходит для рендеринга на экраны мобильных устройств. Реализация управляемых онтологиями интерактивных средств визуальной аналитики с использованием механизма семантических фильтров SciVi и описанных выше средств семантической интеграции позволит выводить доступную из виртуального пространства персональную, а также агрегированную информацию в привязке к соответствующим органам и физиологическим системам человеческого организма.

Размер пиктограмм может соответствовать объему доступных данных о соответствующем органе, интенсивность цвета линий «метро» – о наличии результатов анализа влияния тех или иных медицинских данных, генетических и экологических факторов, данных о питании и образе жизни человека на определенные системы органов человеческого организма. Соответствующий анализ выполняется либо методами машинного обучения, либо методами инженерии знаний на основе базы знаний системы, учитывающей родовидовые, причинно-следственные, временные и другие зависимости между данными для оценки рисков развития заболеваний и здоровому долголетию. Такого рода визуальная интеграция на одном экране большого объема разнородных данных на принципах когнитивного сжатия информации вполне естественна, наглядна и понятна не только профессионалам. Средства интерактивного взаимодействия (клик на пиктограмму

соответствующего органа или «линию метро») позволяют, помимо исходной и агрегированной информации, увидеть объяснение сделанных выводов о рисках здоровью (реализация выполняется по типу компоненты объяснения в экспертных системах).

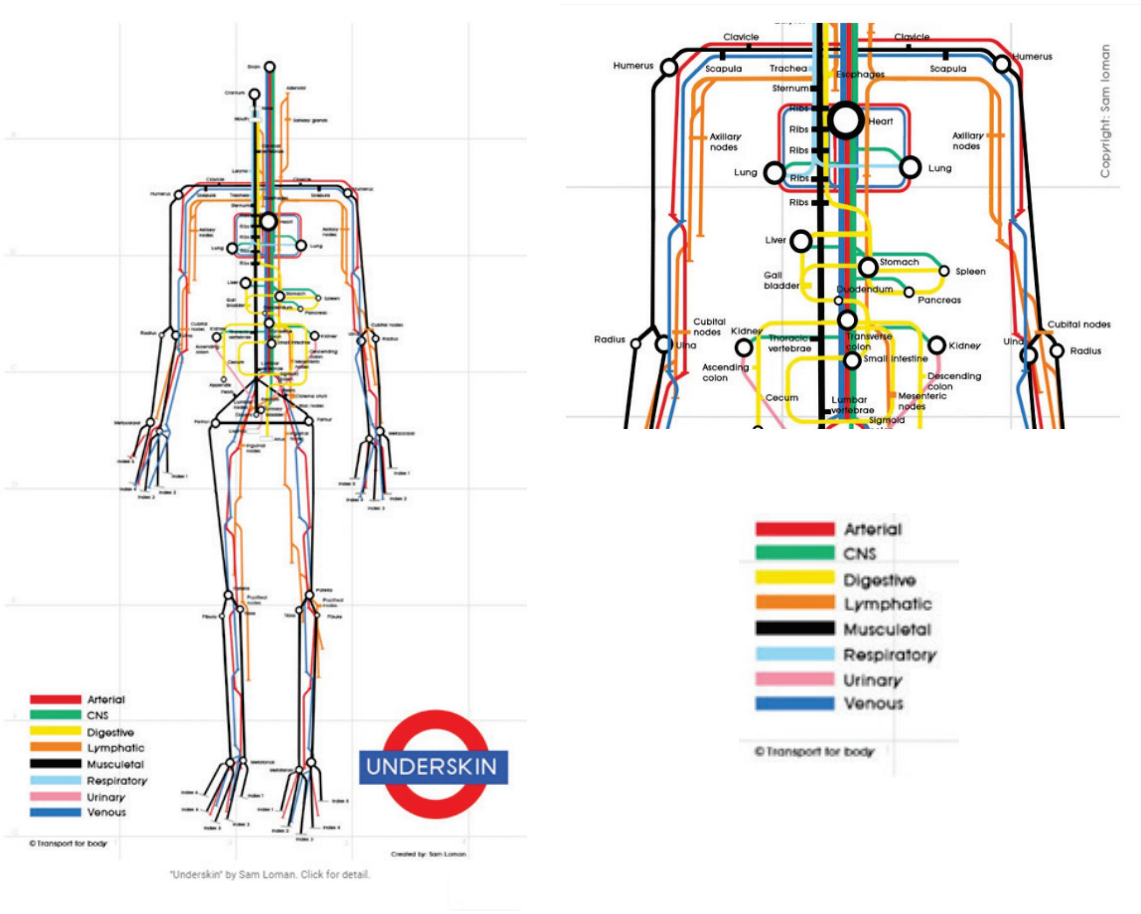

Рисунок 5 – Визуализации физиологических систем человеческого организма в виде схемы метро (https://vizworld.com/2010/03/sam-lomans-underskin-visualization/)

Использование подобных моделей визуализации в значительной мере способствует реализации таких принципов 4П-медицины, как предикция и превентивность, т.к. интерактивное взаимодействие с такого рода графикой при условии полноты и непротиворечивости анализируемых данных способно помочь заблаговременно выявить у персоны совокупность тех ключевых факторов, которые определяют развитие того или иного заболевания, и сформулировать рекомендации по коррекции образа жизни и/или необходимости обращения к специалистам для решения задач, которые требуют участия врача. Рекомендации по коррекции образа жизни и сохранение текущих и агрегированных результатов в персональной фабрике данных способны помочь в сопровождении пациента на пути к изменению ежедневного поведения и формированию новых привычек в части сна, питания, движения, управления стрессом и др.

Управляющие семантической интеграцией, анализом и визуализацией данных онтологии разрабатываются нами в среде визуального редактора ОНТОЛИС (ONTOLIS), который предназначен для создания легковесных онтологий с автоматической интерпретацией в ходе логического вывода содержимого атрибутов вершин и дуг графа онтологии. В отличие от большинства других редакторов онтологий, он ориентирован в том числе и на неподготовленного пользователя, а не только на инженеров по знаниям [11]. Сам визуальный редактор также реализован как онтологически управляемое решение и позволяет легко адаптировать предоставляемые им средства визуализации под индивидуальные предпочтения конечных пользователей.

Как показывает практика, применение lightweight онтологий позволяет снизить порог вхождения новичков в область онтологического инжиниринга, уменьшить трудоемкость и сократить время на разработку онтологий [10]. Важно подчеркнуть, что одно из главных преимуществ разработки онтологически управляемых решений заключается в том, что изменение поведения программной

системы и ее адаптация к изменениям в предметной области решаемых задач достигается не за счет внесения изменений в исходный код программной платформы, а за счет изменения онтологий. Это принципиально важно при создании фабрик данных как для адаптации к новым источникам данных и их семантической интеграции на принципах федерализации, так и для автоматизации разработки DF и ее сервисов. Кроме того, за счет использования методов онтологического инжиниринга упрощается настройка функционирования аналитических сервисов фабрик данных на персональные предпочтения пользователей, что также направлено на реализацию одного из важнейших принципов 4П-медицины – персонализацию. Особенно эффективно указанные принципы реализуются, когда фабрики данных имеют микросервисную архитектуру.

Таким образом, «персональная фабрика данных» содержит только те данные, которые эффективно работают на конкретную персону и ее глобальные сквозные (по всей совокупности источников данных «персональной фабрики») аналитические запросы, что позволяет значительно сократить время на всесторонний анализ информации, получаемой о пациенте из различных сторонних источников за счет современных методов ИИ. От традиционных фабрик данных корпоративного уровня «персональные фабрики» наследуют, в основном, следующее:

1. Источники данных, используя возможности современных графических интерфейсов (графические API), получают сквозную интеграцию.
2. Вместо единого блока программных платформ используются микросервисные архитектуры.
3. Наравне с локальными источниками данных (персональные коллекции документов, а также относящиеся к персоне данные из отдельных МИС как государственных, так и частных клиник) в среде «персональной фабрики данных» используется наибольшее число возможных облачных решений.
4. Информационные потоки оркестрируются (на основе специальных сервисов).
5. Качество информации повышается за счет унификации и виртуализации.
6. Качество анализа, мониторинга, проверки гипотез, оценки рисков и прогнозов повышается за счет использования продвинутых средств научной визуализации и визуального когнитивного анализа данных методами ИИ.
7. Независимо от типа и объема источника данных к нему предоставляется быстрый доступ (к сайтам, реляционным базам данных, хранилищам данных, озерам данных и т. д.).
8. Обеспечение безопасного и разграниченного доступа разным категориям пользователей.

5. Заключение

Предлагаемый в статье подход, названный «персональная фабрика данных», хорошо масштабируется, а использование микросервисной архитектуры позволяет эффективно расширять функциональность системы за счет унифицированного способа подключения к платформе DF новых сервисов и ресурсов, включая файлы с локального компьютера или мобильного устройства персоны. При этом не требуется внесения изменений в уже существующую инфраструктуру источников данных: между сервисами DF и источниками данных добавляется дополнительный семантический слой на базе онтологий, который занимается управлением метаданными и доступом к разнородным данным. Сервисы DF помогают не только оптимизировать хранение, обработку и анализ данных, повысить качество обслуживания информационных ресурсов и аппаратных средств, но и позволяют построить онтологический профиль состояния здоровья владельца «персональной фабрики данных».

Этот профиль автоматически пополняется и обновляется, являясь доступным для дальнейшего семантического анализа, мониторинга здоровья и контекстного поиска, например, с целью получения советов психолога или рекомендаций по питанию и физическим нагрузкам. Представленная в онтологическом профиле агрегированная из различных источников информация в привязке к различным системам органов человека и интерактивные средства визуального анализа позволяют наглядно отследить динамику изменения состояния здоровья и своевременно рекомендовать обратиться к врачу нужной специальности.

Описанный подход к управляемой онтологиями интеграции данных имеет синергетический эффект, т.к. позволяет за счет семантической интеграции расширить возможности поисковых сервисов по сравнению с возможностями современных поисковиков в интернет. В частности, поисковый сервис становится более экспертным за счет способности выдавать ответы типа

«ДА»/«НЕТ», автоматически определяя необходимость именно такого варианта ответа. Кроме того, возрастает пертинентность поисковой выдачи в случае, когда ни один ресурс не содержит полного ответа на запрос и требуется автоматическая смысловая агрегация результатов поиска с разных ресурсов, чтобы вместо выдачи списка ссылок на отдельные ресурсы, содержащие лишь некоторую часть требуемого ответа, сгенерировать на принципах семантической интеграции единый и наиболее полный текст ответа на запрос. Тем самым открываются новые возможности персонифицируемой ИТ-индустрии и не только применительно к области 4П-медицины.

Благодарности

Автор выражает признательность ООО «БИОГЕНОМ» и лично его генеральному директору А.Н. Дубасову за поддержку исследований по адаптации технологии фабрик данных к области 4П-медицины.

Список источников

[1] Patel A., Debnath, N.C., Bhushan, B. (Eds.). Semantic Web Technologies: Research and Applications (1st ed.). CRC Press, 2022. 404 p. DOI:10.1201/9781003309420

[2] Состояние классических средств информатизации здравоохранения и организационная модель медицинской помощи: возможности развития / А. В. Мартюшев-Поклад, Д.С. Янкевич, С.Н. Пантелеев, И.В. Пряников, Я.И. Гулиев // Медицинские информационные системы. 2020. № 5. С. 6-16. DOI:10.37690/1811-0193-2020-5-6-16

[3] Business Intelligence and Analytics: On-demand ETL over Document Stores / M. Souibgui, F. Atigui, S. B. Yahia, S. S.-S. Cherfi // In book: Research Challenges in Information Science. 2020. Vol. 385. P. 556–561. DOI:10.1007/978-3-030-50316-1_38

[4] Ryabinin K., Belousov K., Chuprina S. Novel Circular Graph Capabilities for Comprehensive Visual Analytics of Interconnected Data in Digital Humanities // Scientific Visualization. 2020. Vol. 12(4). P. 56–70. DOI:10.26583/sv.12.4.06

[5] Свидетельство о гос. регистрации прогр. для ЭВМ № 2020615833. BioGenom 2.0: Распознавание бланков лабораторных анализов / А.С. Минин, А.О. Мироненко, С.И. Чуприна, А.Н. Дубасов; правообладатель ООО «БИОГЕНОМ». № 2020615024; заявл. 29.05.2020; зарегистр. 03.06.2020.

[6] Свидетельство о гос. регистрации прогр. для ЭВМ № 2020615911. BioGenom 2.0: Анализ реестров счетов за медицинские услуги / И.С. Постаногов, Т.А. Леонтьева, С.И. Чуприна, А.Н. Дубасов; правообладатель ООО «БИОГЕНОМ». № 2020615008; заявл. 29.05.2020; зарегистр. 04.06.2020.

[7] Чуприна С.И. Технологии фабрик данных как основа интеллектуальных аналитических платформ цифровой среды превентивной и курортной медицины // Вопросы курортологии, физиотерапии и лечебной физической культуры. 2023. Т. 100(3), вып. 2. С. 219.

[8] Noy, N.F. Ontology Mapping // Handbook on Ontologies. Springer, Berlin. 2009. P. 573–590. DOI:10.1007/978-3-540-92673-3_26

[9] Chuprina S., Postanogov I., Nasraoui O. Ontology Based Data Access Methods to Teach Students to Transform Traditional Information Systems and Simplify Decision Making Process // Procedia Computer Science. 2016. Vol. 80. P. 1801–1811. DOI:10.1016/j.procs.2016.05.458

[10] Davies J. Lightweight Ontologies // Theory and Applications of Ontology: Computer Applications. 2010. P. 197–229. DOI:10.1007/978-90-481-8847-5_9

[11] Chuprina S., Nasraoui O. Using Ontology-Based Adaptable Scientific Visualization and Cognitive Graphics Tools to Transform Traditional Information Systems into Intelligent Systems // Scientific visualization. 2016. Vol. 8(1). P. 23-44.