МЕТОДЫ АНАЛИЗА И ОЦЕНКИ ПОКАЗАТЕЛЕЙ КАЧЕСТВА ПАТЕНТНЫХ ДАННЫХ, ИСПОЛЬЗУЕМЫХ ПРИ ФОРМИРОВАНИИ И РАЗВИТИИ РАСПРЕДЕЛЕННЫХ ПАТЕНТНЫХ ИНФОРМАЦИОННЫХ ФОНДОВ

Сиротюк В.О.

Институт проблем управления им. В.А. Трапезникова РАН,
Россия, г. Москва, ул. Профсоюзная, д.65

vsirotyuk@ipu.ru

Аннотация. Сформулированы требования к качеству патентных данных распределенного патентного информационного фонда (РПИФ). Рассмотрены структура РПИФ, особенности формирования баз данных (БД) патентной документации РПИФ, доступа к данным и их использования, факторы, влияющие на качество проведения патентных поисков международного типа и экспертизы заявок на изобретение. Описаны показатели качества патентных данных БД РПИФ. Предложены основные критерии качества патентной информации. Для согласования критериев оценки качества БД РПИФ, предъявляемых их разработчиками и пользователями, разработана структура базы метаданных РПИФ.

Ключевые слова: патентная информация, патентная база данных, распределенный патентный информационный фонд, патентный поиск, показатели качества патентных данных, полнота патентных баз данных, достоверность патентной информации, база метаданных.

Введение

Патентные информационные фонды (ПИФ) формируются патентными организациями для проведения экспертизы заявок на изобретение, а также выполнения хозяйствующими субъектами научных и патентных исследований с целью создания конкурентоспособных товаров и услуг, технологии и техники, обоснования принимаемых решений.

Эффективность и качество проведения исследований технического уровня и тенденций развития объектов хозяйственной деятельности, их патентоспособности и конкурентоспособности определяются, в первую очередь, качеством данных патентных баз данных (ПБД) ПИФ, наличием развитых информационных технологий и средств доступа пользователей к ПБД.

Для повышения полноты и качества патентного поиска, обеспечения соответствия его требованиям патентного поиска международного типа, необходимо предоставление доступа к данным не только локальных ПБД ПИФ, создаваемых национальными или региональными патентными ведомствами, но и к данным внешних удаленных ПБД, формируемых другими патентными ведомствами и организациями (службами) – провайдерами патентной информации.

Эти факторы обусловливают необходимость создания распределенного ПИФ (РПИФ), инфраструктура которого должна обеспечивать доступ к локальным и внешним ПБД через единый пользовательский интерфейс, поиск и обработку данных, логическую интеграцию информации и ее использование.

Высокое качество патентных данных локальных ПБД РПИФ достигается разработкой моделей и методов анализа и синтеза оптимальных структур данных по различным эксплуатационным критериям эффективности функционирования ПБД, удовлетворяющих требованиям обеспечения неизменности характеристик данных, сохранения семантических свойств данных, информационных и функциональных связей между ними, исключения дублируемости и несогласованности данных, а также разработкой формализованных методов оценки показателей качества ПБД, выработкой на основе анализа результатов оценки соответствующих мероприятий по повышению качества информации на всех этапах сбора, хранения и обработки данных.

В работе рассмотрены особенности формирования ПБД РПИФ, факторы и причины, снижающие их качество. Предложены формализованные методы оценки качества патентных данных ПБД. Сформирована структура базы метаданных ПБД и разработаны методы ее построения, используемая при согласовании критериев качества данных разных категорий пользователей РПИФ и оценке показателей качества ПБД.

Особенности формирования РПИФ и характеристики ПБД. Причины снижения качества ПБД

Современный РПИФ представляется в виде виртуального распределенного патентно-информационного хранилища, предоставляющего доступ к локальным (внутренним) ПБД и к внешним патентным информационным ресурсам, доступным по каналам связи [1,2].

РПИФ должен содержать (предоставлять доступ) [1,3]:

- фонд патентной документации стран, входящих в минимум документации PCT,
- фонд региональной и национальной патентной документации,
- фонд непатентной литературы по списку, рекомендованному Международным бюро ВОИС,
- фонд патентно-ассоциированной документации.

Центральное место в РПИФ занимает патентная документация. Патентные документы содержат уникальную информацию по различным аспектам научно-технических, экономических, социальных, культурных и других видов знаний. Патентная информация характеризуется уникальностью, полнотой, новизной, оперативностью, достоверностью, унифицированностью, персонифицируемостью, доступностью [1].

Патентная информация становится доступной общественности после ее официальной публикации и загрузки в соответствующие ПБД. Патентные базы данных содержат информацию о заявках на изобретение, патентах, сведения о правовом статусе патентов и др. информацию. По содержанию ПБД являются документальными политематическими БД, т.к. в них хранится патентная информация о различных областях знаний (тематиках). Как правило, ПБД создаются на основе патентной документации одной страны подачи заявок и выдачи патентов (например, ПБД РФ, ПБД США, ПБД ЕПВ, ПБД Японии и т.д.). Требования, предъявляемые к их структуре более высокие, чем к традиционным библиографическим или фактографическим БД.

Состав, структура и функциональные сервисы ПБД должны проектироваться с учетом требований и рекомендаций ВОИС и цифровых библиотек интеллектуальной собственности (ЦБИС, IPDL) [3,4,5]. Каждая ПБД формируется путём обработки данных определённого источника патентной информации. Учитывая, что источники данных являются разнородными и владельцами (собственниками) ПБД могут использоваться различные доступные им методы и средства обработки, представления и хранения патентной информации, патентные БД РПИФ носят, несмотря на их единое назначение, гетерогенный характер. Они различаются друг от друга не только составом образующих их записей, но и структурой данных, формой представления, форматами данных, протоколами и интерфейсами доступа, сервисными средствами.

Качественное проектирование и разработка ПБД требует от исследователей и разработчиков не только профессиональной подготовки в области теории и практики создания БД, но и знаний в области патентной информации (особенностей ее подготовки, хранения, обработки и представления), источников ее получения и возможностей их многоаспектного использования при проведении тематических патентных поисков и патентных исследований. Для проектирования ПБД могут использоваться как структурные, так и объектно-ориентированные методы. Следует отметить, что объектно-ориентированные методы наиболее полно отражают технологию формирования и хранения ПБД в РПИФ и позволяют учитывать особенности и характеристики источников патентной информации, а формируемая с их использованием объектно-ориентированная ПБД может использоваться также при оценке полноты ПБД [6,7].

Основными процедурами обработки ПБД являются операции сбора патентных данных, загрузки информации в ПБД, хранения данных в системе хранения данных (СХД), поиска и выдачи данных. Операции корректировки данных могут выполняться для информационных элементов, касающихся правового статуса патента и поддержания его в силе, сведений о заявителях, авторах, патентообладателях.

Повышение качества патентных данных ПБД является ключевым аспектом принятия обоснованных решений и эффективного функционирования и развития патентных организаций и хозяйствующих субъектов.

Основными факторами и причинами, влияющими на качество ПБД РПИФ при их использовании, являются:

- наличие большого количества пользователей РПИФ и решаемых ими задач с различными требованиями к качеству данных (полноте, достоверности, своевременности, доступности данных);
- хранение в ПБД большого объема разнородной информации, требующей эффективных мер по обеспечению непротиворечивости, доступности, достоверности сохранности и безопасности хранимых данных. Источниками возникновения ошибок в хранимой в ПБД информации могут являться массивы данных, используемые при пополнении, обновлении и корректировке ПБД (наличие в них пропусков, пробелов, отсутствие отдельных элементов и структур данных), а также сотрудники службы администратора ПБД (низкая квалификация, небрежность,

неправильная интерпретация данных и т.д.), ограниченная надежность технических средств подготовки данных и носителей информации;

- усложнение структур хранения ПБД и средств доступа к данным в распределенной информационно-управляющей структуре РПИФ;
- хранение в ПБД как собственно данных (объектов и элементов предметной области) и связей (отношений) между ними, так и метаданных РПИФ, что обусловливает необходимость комплексного рассмотрения проблемы повышения достоверности информации в ПБД;
- обеспечение возможности параллельного доступа различных категорий пользователей к данным. Ошибки в спецификациях запросов пользователей могут быть связаны с неправильным применением языков описания данных и манипулирования данными, языков запросов, ошибками в алгоритмах поиска и выборки данных, отсутствием средств контроля и защиты данных.

На качество патентной информации РПИФ также влияют ошибки, возникающие на этапах проектирования ПБД. Рассмотрим основные причины этих ошибок.

На этапе анализа информационных требований пользователей и структуризации предметной области ПБД РПИФ ошибки появляются при наличии не выявленных синонимов и омонимов информационных элементов, отдельных неучтенных элементов и связей между ними, при неверной трактовке семантики взаимосвязей между информационными элементами; неучтенных ограничений предметной области, а также ограничений неизменности и корректности данных и связей между ними и др. причинах. Наличие ошибок при анализе информационных требований пользователей приводит к ошибкам в проектируемой канонической структуре ПБД.

При реализации последовательных этапов проектирования канонической, логической и физической структур ПБД, рассматриваемых как процессы отображения одной структуры в другую, основными ошибками являются неадекватное и неполное отображение или потеря информационных элементов и связей, недопустимое объединение элементов в записи или файлы, в структурах которых становится невозможной реализация отдельных путей доступа. Поэтому одним из основных требований, предъявляемых к структурам данных при их отображении, является требование обеспечения максимальной полноты сохранения выделенных в предметной области типов информационных элементов, объектов данных и связей между ними [7].

Уровень безопасности данных также важен с точки зрения повышения качества патентной информации, поскольку определенная часть ее носит закрытый характер и задачей владельца ПИФ является обеспечение ее конфиденциальности, имманентности (неизменности) и доступности. При наличии потенциальных угроз безопасности с учетом роста количества пользователей и предоставляемых им услуг несоблюдение требований информационной безопасности информационных ресурсов РПИФ может нарушить нормальный режим его функционирования. А к открытой патентной информации, относящейся к категории Big Data, требуются специальные меры по обеспечению ее сохранности, достоверности и доступности.

Критерии и показатели качества патентной информации ПБД РПИФ

Качество патентных данных является обобщенным понятием, отражающим степень их пригодности к решению определенной задачи (экспертизы заявок, проведения патентного поиска, публикации патентов, формирования аналитических отчетов, принятия решений и т.д.).

Критерии эффективности и качества РПИФ характеризуют соответствие состава, содержания, структуры, эксплуатационных и сервисных характеристик ПБД РПИФ спецификациям требований, предъявляемых стандартами ВОИС и ЦБИС (IPDL) к фондам патентной и непатентной документации, а также рекомендациям и требованиям экспертов, проводящих патентные поиски международного типа, к эксплуатационным и сервисным характеристикам их использования.

Оценка качества патентных данных и действия по его повышению являются необходимым этапом жизненного цикла управления данными и функционирования патентных информационных систем. Некачественные данные ПБД могут привести либо к неработоспособности систем, либо к некорректным результатам при их обработке, снижающим эффективность и качество принимаемых решений. Приведение исходных («сырых») данных в соответствие с требуемыми критериями качества, определяемыми спецификой решаемой задачи, осуществляется на этапе предобработки данных. Предобработка данных предполагает выполнение двух процедур - очистку данных и оптимизацию структур данных. Очистка производится с целью исключения ряда негативных факторов, снижающих качество данных: устранение дублируемых элементов и противоречий, восстановление и заполнение пропусков, исправление аномальных значений и др. В процессе

очистки восстанавливаются также нарушения структуры, полноты и целостности данных, преобразуются некорректные форматы данных. Оптимизация структур данных в отличие от очистки обеспечивает повышение эффективности данных при решении конкретных прикладных задач (поиска, экспертизы, анализа и т.п.).

В соответствии со стандартами серии ISO 8000 и ISO 9000:2015 основными критериями (показателями) качества данных являются полнота, достоверность, точность, согласованность, доступность и своевременность [8,9].

Для патентной информации с учетом особенностей ее формирования и использования наиболее важными являются показатели полноты, достоверности, защищенности и доступности данных, т.к. они напрямую влияют на качество выполняемых пользователями РПИФ патентных и научных исследований [1,4].

Рассмотрим подробнее данные показатели.

Показатель полноты РПИФ характеризует соответствие состава, структуры, содержания, эксплуатационных и сервисных характеристик ПБД спецификациям требований к составу и структурам данных, контенту, поисковым, сервисным и функциональным требованиям эталонной ПБД. Эталонная ПБД (ЭПБД) формируется на основе общих требований и рекомендаций стандартов ВОИС и ЦБИС (IPDL). Она является общей (типовой), и ее описание используется для оценки полноты локальной ПБД ПИФ путем сравнения. В упрощенном варианте в качестве эталонных ПБД могут использоваться описания ПБД источника патентной информации, например, ПБД РФ, ПБД США и др., а также ПБД всемирных патентных информационных систем типа Patentscope или Espacenet, содержащих коллекцию политематических ПБД [3-5]. Полноту РПИФ следует рассматривать комплексно и различать *показатели структурной, функциональной, структурно-функциональной и информационной полноты* [10].

Показатель достоверности характеризует степень соответствия данных об объектах, зафиксированных в ПБД ПИФ, реальным объектам в данный момент времени. Изменение показателя достоверности может быть связано с ошибками при вводе данных по заявкам на изобретения, некорректными записями о правовом статусе патентов, не обновлёнными данными о поддержании патентов в силе, о заявителях, авторах, патентообладателях. Показатель достоверности рассчитывается как отношение количества правильных записей, документов и структурных частей патентного документа (реферат, формула и др.) к их общему числу в ПБД.

Показатель уровня защищенности данных ПИФ характеризует уровень обеспечения конфиденциальности, имманентности (неизменности) и доступности информационных материалов по заявкам на изобретения и патентов, официальных изданий, а также других информационных ресурсов ПБД.

Показатель доступности данных характеризует возможность получения и обработки данных из внешних удаленных ПБД РПИФ. Показатель доступности патентных данных определяется, в первую очередь, *степенью представления входящих в ПБД РПИФ документов в электронной форме,* что делает возможным организацию удаленного доступа к открытой информации ПБД. Этот показатель рассчитывается как отношение количества документов ПИФ, представленных в электронном виде, к общему количеству документов, хранящихся в ПИФ.

Модели и методы анализа и оценки полноты и достоверности патентных данных

Как отмечалось ранее, для оценки показателей полноты данных ПБД РПИФ используется понятие эталонной патентной БД (ЭПБД) (в английском варианте – Master Data Base (MDB)).

ЭПБД (MDB) служат для определения степени общности структур ПБД и ЭПБД и расчета численных значений показателей полноты данных. Оценка структурной и функциональной, а также структурно-функциональной полноты ПБД осуществляется с помощью методов анализа общности (сходства) канонических структур ПБД и ЭПБД путем вычисления функций подобия ПБД ПИФ и ЭПБД по хранящимся в них объектам данных, информационным элементам, связям и процедурам поиска и обработки данных. Информационная полнота ПБД рассчитывается как отношение количественных характеристик объектов данных и информационных элементов ПБД РПИФ к количественным характеристикам ЭПБД.

Исходными данными для расчета показателей полноты ПИФ являются формализованные описания и характеристики канонических структур ПБД и ЭПБД, представляемые в виде графов $G_v(D_v, R_v)$ и $G_{кс}^{об}(O,\Delta)$ соответственно [1,11], где $D_v = \{d_\varepsilon \ / \ \varepsilon \in L_v^{ob}, L_{ob}^v \subseteq L_v\}$ – множество классов (объектов) данных v-*й* ПБД, R_v – множество взаимосвязей (отношений) между элементами; $O=$

$\{O_\varepsilon / \varepsilon = \overline{1, \varepsilon_0}\}$ - множество объектов предметной области, а $\Delta = \{\delta_{\varepsilon\varepsilon'} / \varepsilon, \varepsilon' = \overline{1, \varepsilon_0}\}$ - множество связей (отношений) между объектами. Формализовано граф $G_v(D_v, R_v)$ описывается матрицей смежности $W_v = \|w^v_{\varepsilon\varepsilon'}\|$, элементы которой $w^v_{\varepsilon\varepsilon'} = 1$, если между объектами d_ε и $d_{\varepsilon'}$ имеется информационная или функциональная взаимосвязь и $w^v_{\varepsilon\varepsilon'} = 0$, в противном случае. Граф $G^{об}_{кс}(O, \Delta)$ описывается булевой матрицей смежности $B^{об}_{кс} = \|b_{\varepsilon\varepsilon'}\|$ между объектами, составами объектов $H(O_\varepsilon) = \{d_l, \ldots, d_L, (d_l, d_j), \ldots, (d_k, d_l), \{f^\varepsilon_r\}\}$.

Для оценки элементной полноты v-*й* ПБД РПИФ используется нормированный показатель подобия, вычисляемый по формуле:

$$\varepsilon_{эл} = \frac{1}{2}\left(\frac{p_{11}}{p_{11} + p_{10}} + \frac{p_{11}}{p_{11} + p_{01}}\right) = \frac{1}{2}\left(\frac{p_{11}}{|O|} + \frac{p_{11}}{|D_v|}\right) = \frac{\alpha + \beta}{2},$$

где $p_{11} = |O \cap D_v|$, $p_{10} = |O| - p_{11}$, $p_{01} = |D_v| - p_{11}$.

Для оценки полноты ПБД РПИФ по связям (отношениям) между структурными элементами используется мера подобия, вычисляемая по формуле:

$$\varepsilon_{св} = \frac{p_{11}}{p_{11} + p_{10} + p_{01}},$$

Здесь $p_{11} = \sum_i |F(d_i) \cap F(d_{i'})|$, $p_{10} = \sum_i |F(d_i)| - p_{11}$, $p_{01} = \sum_{i'} |F(d_{i'})| - p_{11}$, $\forall d_i, d_{i'} \in D_{ko}$, где $F(d_i) = \{d_j\}, F(d_{i'}) = \{d_{j'}\}$- множества достижимости, соответственно, для элементов $d_i \in O$ и $d_{i'} \in D_v$ $(d_i, d_{i'} \in D_{ko})$, $D_{ko} = O \bigcap D_v$.

Структурная полнота ПБД РПИФ $\varepsilon_{стр}$ определяется как $\varepsilon_{стр} = \varepsilon_{эл} + \varepsilon_{св}$.

Функциональную полноту ПБД ПИФ целесообразно оценивать с помощью показателя меры подобия, учитывающей число схожих элементов (процедур поиска, доступа и обработки данных) в канонических структурах ПБД РПИФ и ЭПБД:

$$\varepsilon_{пр} = \frac{2P'_{11}}{2P'_{11} + P'_{10} + P'_{01}},$$

где P'_{11} – число общих процедур (методов) в требованиях пользователей ПБД РПИФ и ЭПБД, P'_{10} – число процедур, присутствующих в требованиях пользователей ЭПБД, но отсутствующие в требованиях пользователей ПБД РПИФ; P'_{10}– число процедур, присутствующих в требованиях пользователей ПБД РПИФ, но отсутствующие в требованиях пользователей ЭПБД.

Интегрированный показатель структурно-функциональной полноты ПБД ПИФ ε вычисляется по формуле: $\varepsilon = \varepsilon_{стр} + \varepsilon_{пр} = \varepsilon_{эл} + \varepsilon_{св} + \varepsilon_{пр}$.

Информационная полнота v-*й* ПБД определяется как отношение количества экземпляров записей данной ПБД к количеству экземпляров записей аналогичной ЭПБД. Поскольку, как отмечалось ранее, ПБД являются политематическими БД, содержащими информацию из разных областей знаний, то создание (или нахождение первоисточника) аналогичной по составу ЭПБД проблематично. Поэтому информационную полноту РПИФ целесообразно оценивать по полноте тематической патентной БД (ТПБД), формируемой на основе проведения тематических поисков в ПБД РПИФ.

Для определения информационной полноты ТПБД, характеризуемой количеством отобранных из ПБД РПИФ в ТПБД документов заданной тематики, введем следующие параметры и характеристики тематических поисковых запросов.

Число документов, находимых в результате реализации множества тематических запросов $Q = \{q_k\}$, $k = \overline{1, K_0}$ на графе объектной канонической структуры ЭПБД $G^{об}_{к.с}(O, \Delta)$, формализовано

представим в виде множества $W_k = \{w_k\},\ k = \overline{1, K_0}$, где w_k – число документов, находимых в ПБД стран минимума РСТ и непатентной литературы по списку ВОИС при реализации *k*-го тематического запроса.

Число документов, находимых в результате реализации множества тематических запросов на графах канонических структур ПБД РПИФ $G_v(D_v, R_v)$, $v = \overline{1, V_0}$ формализовано представим в виде множества $M_v = \{m_v^k\},\ v = \overline{1, V_0},\ k = \overline{1, K_0}$, где m_v^k – число документов, находимых в *v*-й базе данных РПИФ при реализации *k*-го тематического запроса.

Тогда информационная полнота тематической базы данных РПИФ при реализации *k*-го тематического запроса определится из выражения:

$$\varepsilon_{\text{инф}} = \frac{1}{w_k} \sum_{v=1}^{V_0} m_v^k .$$

Для оценки достоверности структур данных при отображении предметной области пользователей РПИФ в каноническую структуру ПБД используется отношение числа типов и экземпляров данных и типов связей и экземпляров связей между ними, зафиксированных в канонической структуре ПБД, к их общему числу в множестве информационных требований пользователей. Методы расчета данного показателя достоверности данных приведены в [11].

Определим показатель достоверности данных при отображении канонической структуры ПБД в логическую структуру ПБД. Логическая структура ПБД представляется графом G(N,L), где $N=\{n_j/j=\overline{1,J}\}$ - множество логических записей, $L=\{(n_j,n_{j'})/j,j'=\overline{1,J}\}$ - множество внешних ключей, отражающих взаимосвязи между записями [11]. Каждая логическая запись $n_j \in N$ состоит из подмножества включенных в нее объектов данных и информационных элементов, т.е. информационный состав *j-й* записи $\mathrm{E}(\mathrm{n_j})$ определяется из выражения $\mathrm{E}(\mathrm{n_j}) = \left\{\mathrm{d_1}\ /\ \mathrm{d_1} \in \mathrm{n_j}\right\}$. Обозначим полное множество путей доступа на логической структуре через $\mathrm{Y} = \left\{\mathrm{y_\theta} \mid \theta = \overline{1, \theta_0}\right\}$, где $\mathrm{y_\theta}$ - θ–й путь доступа.

Логическая структура ПБД характеризуется следующими параметрами:

- вектором количества экземпляров логических записей $\mathrm{R}_л = \left\{\mathrm{r}_1^л, \mathrm{r}_2^л, \ldots, \mathrm{r}_j^л, \ldots, \mathrm{r}_J^л\right\}$, где $\mathrm{r}_j^л$ -количество экземпляров $\mathrm{n_j}$ –й логической записи;

вектором количества экземпляров связей между записями, вошедшими в пути доступа $\mathrm{S}_л = \left\{\mathrm{s}_1^л, \mathrm{s}_2^л, \ldots, \mathrm{s}_\theta^л, \ldots, \mathrm{s}_{\theta_0}^л\right\}$, где $s_\theta^л$ - количество экземпляров связей между записями.

Определим через $D_{кл} = \mathrm{D}_{кс} \cap \mathrm{E}$ подмножество общих информационных элементов в канонической и логической структуре ПБД, а через $\mathrm{M}_{кл} = \mathrm{M}_{кс} \cap \mathrm{Y}$ - общие пути доступа, реализованные в канонической и логической структурах БД.

Тогда показатель достоверности информации при отображении канонической структуры ПБД в логическую структуру ПБД определяется из выражения:

$$\mathrm{P}_{кл} = \frac{\sum\limits_{\mathrm{n_j} \in \mathrm{D}_{кл}} \mathrm{r}_j^л + \sum\limits_{s_\theta^л \in \mathrm{M}_{кл}} s_\theta^л}{\sum\limits_{d_\varepsilon \in \mathrm{D}_{кс}} r_\varepsilon + \sum\limits_{s_\mu \in \mathrm{M}} s_\mu} .$$

Показатель достоверности хранимой в ПБД информации определяется из выражения:

$$P_{ПБД} = \frac{\sum_{\delta=1}^{\delta_0} p_\delta \cdot r_\delta^з}{\sum_{\delta=1}^{\delta_0} r_\delta^з},$$

где p_δ - достоверность δ–й физической записи $(0 \le p_\delta \le 1)$, $r_\delta^з$ – количество экземпляров δ–й записи. При этом достоверность записи p_δ можно определить как произведение достоверности информационных элементов ($p_\delta^{эл}$), входящих в состав записи, и достоверности выбора пути доступа ($p_\delta^{д}$) к ней, т.е. $p_\delta = p_\delta^{эл} \cdot p_\delta^{д}$.

Методы повышение конфиденциальности и доступности патентных данных

Главной целью защиты патентной информации является обеспечение конфиденциальности, неизменности и доступности информационных материалов по заявкам на изобретения и патентов, публикаций и официальных изданий, других информационных ресурсов РПИФ [12].

Средства защиты ПБД РПИФ включают в себя организационные, процедурные, структурные, аппаратные и программные методы.

Рассмотрим структурные методы защиты, повышающие качество патентных данных.

Механизм защиты $M(G_v)$ канонической структуры *v*-й ПБД есть отображение $\{(u_k, \pi_k, a_j, d_\varepsilon, \phi_i)\} \to \{0,1\}$, где $u_k \in U$, $\pi_k \in \Pi$, $a_j \in A$, $d_\varepsilon \in D_v$, $\varphi_i \in \Phi$. Случай «1» соответствует правомочности доступа типа a_j k-го пользователя, имеющего уровень полномочий π_k, к объекту данных d_ε, который имеет степень секретности φ_i, а случай «0» соответствует запрету такого доступа. Механизм защиты $M(G_v)$ представляет собой средство установления правомочности действий пользователей по отношению к типам объектов данных канонической структуры ПБД.

Эффективность $M(G_v)$ определяется наличием на сформированной канонической структуре ПБД разрешенных путей доступа ко всем данным, требуемым для удовлетворения множества санкционированных запросов пользователей.

Механизм защиты $M(G_л)$ логической структуры ПБД G(N,L) есть отображение $\{(u_k, \pi_k, a_j, (n_j, n_{j'}), \hat{\varphi}_{jj'}, n_j, \hat{\varphi}_j)\} \to \{0,1\}$. Значение «1» означает, что пользователь $u_k \in U$ с уровнем полномочий $\pi_k \in \Pi$ обладает правом доступа типа $a_j \in A$ в отношении элементов логической структуры ПБД (связи и логической записи) $(n_j, n_{j'})$ и n_j, которые имеют степени секретности $\hat{\varphi}_{jj'} \in \tilde{\Phi}$ и $\hat{\varphi}_j \in \hat{\Phi}$ соответственно. Значение «0» соответствует неправомочности такого доступа.

Механизм защиты логической структуры ПБД обеспечивает возможность идентификации правомочности доступа к защищенным типам логических записей и взаимосвязям между ними со стороны всех пользователей. Критериями эффективности при решении задачи синтеза оптимального механизма защиты логической структуры ПБД являются минимум суммарного числа подсхем, используемых пользователями, минимум суммарной длины путей доступа к данным, минимум суммарного интерфейса между ПБД [11,12].

Механизм защиты $M(G_ф)$ физической структуры ПБД разрабатывается на этапе формирования структуры хранения данных ПБД, формально представляемой графом $G_ф$ ($D^Ф$, $W^Ф$), где $D^Ф$-множество физических записей (блоков), $W^Ф$-множество связей (отношений) между записями.

Механизм защиты физической структуры ПБД позволяет идентифицировать правомочность доступа пользователей к различным компонентам физической структуры ПБД. Механизм защиты $M(G_ф)$ физической структуры ПБД есть отображение $\{(u_k, \pi_k, a_j, v_p, \varphi_i)\} \to \{0,1\}$, где $v_p \in V$ - множество компонентов физической организации ПБД. При этом «1» означает для пользователя $u_k \in U$ с уровнем полномочий $\pi_k \in \Pi$ возможность доступа типа $a_j \in A$ к элементам $v_p \in V$ физической организации ПБД, которые имеют степени секретности $\varphi_i \in \Phi$, а «0» означает невозможность такого доступа.

Критериями оптимальности при решении задачи синтеза механизма защиты ПБД могут служить максимум информационной независимости пользователей ПБД, минимум затрат на разработку и эксплуатацию системы защиты данных, минимум суммарных потерь от несанкционированного доступа к конфиденциальной информации ПБД.

Рассмотренные модели и механизмы защиты структур данных ПБД обеспечивают повышение качества патентных данных и их доступность, а также эффективность функционирования ПБД при обслуживании множества тематических запросов пользователей.

Методы формирования объектно-ориентированной базы метаданных РПИФ

На основе анализа свойств и характеристик патентных данных (метаданных) формулируются критерии оценки и повышения их качества. Критерии качества данных должны определяться не только разработчиками патентных информационных систем, но и пользователями РПИФ. Это требует согласования, координации и стандартизации их деятельности. Кроме того, изменения в предметной области и в ПБД РПИФ, обусловленные процессами их сопровождения и развития, могут приводить к серьезным проблемам обеспечения надлежащего качества патентных данных.

Для решения этих проблем предлагается использовать единую интегрированную базу метаданных (БмД) репозитория РПИФ.

БмД РПИФ создается с целью:

- информационной поддержки управления процессами проектирования, функционирования и развития ПБД, оптимизации затрат на создание, эксплуатацию и развитие РПИФ;
- оценки показателей качества патентных данных, хранящихся в различных ПБД РПИФ;
- стандартизации описания типов данных и используемой терминологии на основе ведения общесистемных языковых средств (классификаторов, тезаурусов, словарей) для ПБД разных типов;
- конвертирования, трансляции и интеграции различных типов и структур ПБД;
- обеспечения согласованности требований к качеству патентных данных и координации работ по выбору критериев качества данных и планированию качества патентных данных между различными категориями пользователей (разработчиками ПБД РПИФ и внешними пользователями).

БмД репозитория РПИФ создается в виде объектно-ориентированной БД (ООБмД). Она должна удовлетворять следующим основным требованиям:

- хранить метаданные всех стадий жизненного цикла управления данными ПБД РПИФ и предоставлять возможность коллективной работы с ними;
- обеспечить возможность просмотра метаданных ПБД на трех уровнях - концептуальном, логическом и физическом в виде гипертекстовых документов;
- обеспечить администрирование моделей данных, управление версиями и изменениями моделей метаданных на различных уровнях представления ПБД и предоставлять интерфейсы интеграции между ними;
- содержать, помимо классов, описывающих метамодель, ряд системных классов для управления и работы с мета-метаданными;
- поддерживать соглашения по терминологии предметной области РПИФ, а также именованию информационных элементов и объектов данных;
- извлекать и анализировать метаданные из множества патентных источников РПИФ, что гарантирует хранение в БмД актуальной информации об объектах ПБД;
- реализовывать механизм разграничения прав доступа на хранимые данные и метаданные;
- обладать свойством независимости от источников патентной информации.

Рассмотрим структуру и характеристики БмД РПИФ, представляющие интерес с точки зрения обеспечения качества патентных данных и информации.

БмД РПИФ содержат адресно-справочную, содержательную и словарную (лингвистическую) информацию о предметных областях пользователей РПИФ, объектах и элементах данных ПБД, патентных информационных системах и технологиях. В общем виде структура БмД РПИФ с точки зрения обеспечения качества патентных данных должна содержать следующие основные разделы:

- раздел «Описание бизнес-процессов». Включает метаданные для описания условий и последовательностей выполнения бизнес-процессов, связанных с ними информационных потоков и ограничений;

- раздел «Описание понятий». Содержит метаданные для представления и классификации информации, объектов и информационных элементов. Составной частью данного раздела является тезаурус терминов предметной области, стандарты в области патентной информации, классификатор МПК;
- раздел «Общие типы данных». Содержит метаданные, обеспечивающие унификацию типов данных для моделей предметной области и структур ПБД различных уровней представления данных (канонического, логического, физического);
- раздел «Описание объектной модели данных». Содержит метаданные для описания объектных моделей требований пользователей, а также объектной канонической структуры ПБД;
- раздел «Описание схемы базы данных». Включает метаданные для описания данных, поддерживаемых в объектно-ориентированных ПБД.

Рассмотрим методы построения объектной модели БмД (ООБмД) репозитория РПИФ. ООБмД репозитария состоит из спецификаций метаданных, экземпляры которых составляют конкретные схемы, хранимые в репозитарии. Метаданные описывают процедуры (методы) обработки, свойства объектов (классы, атрибуты и связи), интерфейсы, в т.ч. связи между объектами, обеспечивающие поддержку ограничений целостности по ссылкам.

Исходными для построения ООБмД являются объектные модели требований пользователей РПИФ, формально представляемые в виде мультиграфов $G_k^{об}(D_k, U_k)$ с одним типом вершин и двумя типами дуг, где $D_k = \{d_l^k \ / \ l = \overline{1, L_k}, L_k \subseteq L\}$ - множество информационных элементов и объектов данных, $U_k = U_k^{эл} \bigcup U_k^{пр}$ - множество дуг, где $U_k^{эл}$ - множество дуг, характеризующих структуру взаимосвязей между данными, а $U_k^{пр}$ - множество дуг, характеризующих технологию обработки данных для *k-го* пользователя в виде реализации совокупности процедур поиска и непосредственной обработки данных. Методы их построения рассмотрены в [11].

На следующем этапе вершины и дуги мультиграфов $G_k^{об}(D_k, U_k)$ описываются соответствующими метаданными с использованием специализированных словарей/справочников и тезаурусов. В результате выполняется отображение объектных моделей требований пользователей на множество метаданных предметной области РПИФ:

$$G_k^{об}(D_k, U_k) \rightarrow F_k^{md}(D_k^{md} \bigcup U_k^{md}),$$

где F_k^{md} - множество метаданных, описывающих требования *k*-го пользователя, D_k^{md} - подмножество метаданных, описывающих информационные элементы и объекты *k*-го пользователя, U_k^{md} - подмножество метаданных, описывающих отношения (связи) между объектами, информационными элементами и процедурами поиска и обработки данных *k*-го пользователя.

Построение обобщенной объектной модели ООБмД осуществляется путем последовательного наложения мультиграфов $G_k^{об}(D_k, U_k)$ друг на друга и объединения множеств F_k^{md}. Разработанная процедура основана на совмещении идентичных информационных элементов независимо от уровня их размещения на графах $G_k^{об}(D_k, U_k)$ и исключении дублированных и избыточных элементов множеств метаданных F_k^{md}.

Результатами выполнения указанных процедур являются интегрированный мультиграф модели классов $G^{md}(D,U)$ с одним типом вершин $D=\{d_l \ / \ l = \overline{1, L}\}$, соответствующих множеству объектов данных и информационных элементов, и двумя типами дуг: $U^{эл}$ - множество информационных взаимосвязей между элементами $d_l \in D$ и $U^{пр}$ - множество технологических взаимосвязей между информационными элементами и объектами данных, $U = U^{эл} \bigcup U^{пр}$, и объединенное множество метаданных $F^{md} = \bigcup_{k=1}^{K} F_k^{md}$.

Формализовано модель ООБмД представляется графом $G^{md}(D,U)$, описывающим структуру ООБмД, и множеством F^{md}, описывающим состав метаданных ООБмД. Основными характеристиками модели ООБмД являются:

- вектор $Z_\nu = \{z_l^\nu\}$ информационных весов вершин графа $G^{md}(D,U)$, где z_l^ν - информационный вес вершины $d_l \in D$, $z_l^\nu \in \{0,1,...,N\}$, который характеризует степень потребности множества пользователей в данном элементе. Чем больше значение z_l^ν, тем более важным и необходимым является элемент d_l для удовлетворения информационных потребностей пользователей, и, следовательно, тем более высокие требования должны предъявляться к качеству данного элемента и описывающим его метаданным;
- вектор $Z_\mu = \{z_l^\mu\}$ технологических весов вершин графа $G^{md}(D,U)$, где z_l^μ - технологический вес d_l-*го* элемента (вершины), $z_l^\mu \in \{0,1,2...,N\}$. Чем больше значение z_l^μ, тем больше данный элемент подвергается локальной обработке при реализации бизнес-процессов, что обусловливает необходимость повышения качества методов, алгоритмов и процедур его обработки;
- вектор $Z_\theta = \{z_{ll'}^\theta\}$ информационных толщин дуг графа $G^{md}(D,U)$, где $z_{ll'}^\theta$ - информационная толщина дуги $(d_l, d_{l'})$, $z_{ll'}^\theta \in (0,1,2...N)$, которая характеризует степень принадлежности d_l и $d_{l'}$ множеству элементов, описывающему состояние класса объектов. Чем больше значение $z_{ll'}^\theta$, тем более семантически связаны (ассоциированы) элементы d_l и $d_{l'}$, что подтверждено подмножеством пользователей в их информационных требованиях. Значения данного показателя требуют принятия мер по поддержанию необходимого качества связей между данными (ссылочной целостности);
- вектор $Z_\eta = \{z_{ll'}^\eta\}$ технологических толщин дуг графа $G^{md}(D,U)$, где $z_{ll'}^\eta$ - технологическая толщина дуги $(d_l, d_{l'})$, $z_{ll'}^\eta \in \{0,1,2...,N\}$. Данная характеристика, фактически, описывает сложность технологии совместной обработки данных элементов d_l и $d_{l'}$. Чем выше значение данного показателя, тем требуется более высокое внимание по обеспечению непротиворечивости, взаимосвязанности, доступности и достоверности данных и описывающих их метаданных ООБмД.

Таким образом, предложенные формализованные методы построения БмД обеспечивают построение структуры ООБмД РПИФ и описание информационных и поведенческих (технологических) аспектов функционирования ПБД. Анализ характеристик сформированной модели ООБмД позволяет выявить информационные объекты и элементы, а также связи (отношения) между ними, подверженные наибольшему воздействию со стороны пользователей (обращениям, обработке) и требующие особого внимания по обеспечению их качества, т.е. ранжировать патентные данные по приоритетам их обслуживания и поддержания в актуальном состоянии. Это позволяет снизить расходы на проведение работ по планированию качества данных и мероприятий по оценке и повышению качества патентных данных при создании системы повышения качества данных ПБД.

Заключение

В работе рассмотрены особенности формирования и характеристики распределенного патентного информационного фонда и патентных баз данных, играющих важную роль при проведении патентных и научных исследования. Выявлены и классифицированы причины снижения качества ПБД, возникающие на различных этапах жизненного цикла управления данными – от анализа и проектирования структур патентных данных до эксплуатации ПБД РПИФ и использования патентной информации разными категориями пользователей. Определены и сформулированы критерии и показатели качества патентной информации, определены задачи повышения качества патентных данных, решаемые на этапе предобработки данных.

Рассмотрены показатели полноты, достоверности, защищенности и доступности данных, являющиеся наиболее важными для патентных данных. Приведены аналитические выражения для их расчета и оценки.

Предложены методы формирования интегрированной объектно-ориентированной базы метаданных РПИФ, играющей важную роль при согласования, координации и стандартизации деятельности разработчиков и пользователей РПИФ при выработке критериев и показателей оценки показателей качества ПБД, обеспечения надлежащего качества патентных данных при изменениях в предметной области РПИФ. Предложенные методы позволяют построить структуру ООБмД РПИФ и определить состав описывающих ее метаданных. Анализ сформированной модели ООБмД позволяет

эффективно управлять качеством патентных данных. Рассмотренные методы и подходы использовались при построении системы управления качеством патентных данных Евразийского патентного ведомства [13].

Литература

1. *В.В. Кульба, В.О. Сиротюк* Формализованная методология повышения эффективности и качества патентных информационных фондов и опыт ее использования при формировании и развитии евразийского патентно-информационного пространства. - М.:ИПУ РАН. Монография, 2019. - 236с.
2. *Кульба В.В., Сиротюк В.О.* Модели и методы синтеза распределенной информационно-управляющей структуры патентных информационных фондов / Труды 13-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD'2020, Москва) , под общей редакцией С.Н.Васильева, А.Д.Цвиркуна, М.: ИПУ РАН, 2020. с. 1542-1551.
3. Материалы сайта ВОИС: www.wipo.int.
4. *Х.Ф. Фаязов, В.О. Сиротюк, А.В. Овчинников, А.Б. Бурцев* Формирование и развитие евразийского патентно-информационного пространства. М.: ИНИЦ «Патент», 2010.-124 с.
5. Материалы сайта Европейской патентной организации: www.epo.org.
6. *В.О. Сиротюк* Формализованные модели и методы анализа и оценки полноты патентных информационных фондов (на примере международной патентной организации) - ж. Проблемы управления, М.:ИПУ РАН, №6/2018, с.35-43.
7. *Кульба В.В., Сиротюк В.О.* Формализованная методология обеспечения полноты патентных информационных фондов. - Управление развитием крупномасштабных систем MLSD-2018. Труды 11-й международной конференции. Том 3, с.127-138, под общей редакцией С.Н.Васильева, А.Д.Цвиркуна.
8. ГОСТ Р ИСО 8000-2019. Качество данных.
9. ГОСТ Р ИСО 9000-2015. Система менеджмента качества.
10. V.V. Kulba., V.O., Sirotyuk Development of Formalized Models and Methods of Assessment of Structural, Functional and Informational Completeness of Patent Collections / Proceedings of the 11th International Conference "Management of Large-Scale System Development" (MLSD). Denvers: IEEE Catalog Number CFP18GAE-ART, 2018.: https://ieeexplore.ieee.org/document/8551922
11. *Кульба В.В., Ковалевский С.С., Косяченко С.А., Сиротюк В.О.* Теоретические основы проектирования оптимальных структур распределенных баз данных. Серия «Информатизации России на пороге XXI века». М.: СИНТЕГ, 1999.- 660 с.
12. *В.В. Кульба, В.О. Сиротюк, С.А. Косяченко* Информационная безопасность патентных ведомств: теория и практика. - М.:ИПУ РАН. Научное издание, 2017. - 166с.
13. Материалы сайта Евразийской патентной организации: www.eapo.org.