УДК 004.652.4(045) DOI: 10.17587/it.25.53-58

Ю. В. Полищук, канд. техн. наук, доц., e-mail: youra_polishuk@bk.ru,
Оренбургский государственный университет

Способ хранении электронных документов с квазиструктурированным контентом

Рассмотрен способ хранения электронных документов с квазиструктурированным контентом, который реализует минимизацию их объема хранения за счет выделения единого шаблона оформления документов, извлечения из документов фактографических данных с последующим сжатием шаблона и фактографических данных.

Ключевые слова: *квазиструктурированная информация, сжатие информации, обработка электронных документов*

Введение

В процесс работы предприятия формируется сопутствующая эксплуатационная документация, которая, как правило, представлена коллекцией документов. Среди документов единого вида можно выделить общую структуру, но оформление и порядок размещения информации в них будет различен. Последнее обусловлено тем, что документация формируется на основе внутренних стандартов предприятия или в соответствии с требованиями ГОСТ, в которых определены требования к содержимому документов.

Контент документов данного вида квазиструктурирован, т. е. представлен квазиструктурированной информацией. Под квазиструктурированной информацией понимают информацию, в которой можно выделить некую структуру, однако структура эта заранее целиком или частично не известна, либо может меняться с течением времени [1].

При длительном периоде работы предприятия скапливается большое количество электронной документации с квазиструктирированным контентом, как правило, представленной документами формата MS Word, и минимизация объема ее хранения является актуальной задачей.

Подавляющее большинство исследований в области обработки квазиструтурированной информации ориентировано на хранение непосредственно фактографических данных, а не электронных документов. В данной работе рассматривается способ хранения электронных документов с квазиструктурированным информационным наполнением. Электронный документ представляет собой квазиструктурированный информационный контент, обладающий структурой и визуальным оформлением.

Таким образом, способ хранения электронных документов с квазиструктурированным информационным наполнением должен реализовать хранение не только фактографических данных, но и их структуры и визуального оформления в документе.

Исследования в данном направлении выполнены в работе [2], а на предлагаемый способ преобразования слабоформализуемых документов для минимизации их объема при хранении получен патент [3].

К недостаткам данного способа следует отнести отсутствие математической модели информационного наполнения обрабатываемых документов, отсутствие возможности описания лексикологическим деревом фактографического контента (неунифицированной информации) для документов.

Исследования, проведенные в настоящей работе, соответствуют направлениям Федеральной целевой программы "Информационное общество" (2011—2020)", утвержденной постановлением Правительства РФ от 20.10.2010 г. № 1815-р, и критическим технологиям РФ

(технологии информационных, управляющих и навигационных систем), утвержденными Президентом РФ (Пр-899 от 7.07.2011 г.).

Постановка задачи

Для минимизации объема, требуемого для хранения электронных документов с квазиструктурированным контентом, необходимо отделить фактографические данные документа от их визуального представления в документе. Так как визуальное представление будет идентичным для всей коллекции документов, то его можно хранить в виде шаблона формы электронного документа, в котором будут храниться все визуальные особенности представления фактографической информации в контенте документов.

Таким образом, алгоритм сохранения электронных документов предложенным способом будет иметь вид, представленный на рис. 1.

Для восстановления электронного документа необходимо к фактографическому контенту выбранного документа применить шаблон формы электронного документа (рис. 2).

Практическая реализация

В качестве примера рассмотрим пример хранения электронной документации газоконденсатного месторождения. Применим предлагаемый способ для хранения коллекции документов вида "Информационная карта скважины".

Для обработки контента документов данного вида необходима разработка квазиструктурированной модели фактографического контента инфомационного наполнения документа, которая с учетом спецификации Xml Schema Definition (XSD) может быть записана следующим образом [4]:

$$S = \left\langle \begin{matrix} root, sObj, LObj, minOccurs, \\ maxOccurs, sMet, Obj\_smet \end{matrix} \right\rangle, \qquad (1)$$

где *root* — корневой объект; $root \in sObj$; *sObj* — конечное множество объектов, каждый из которых содержит фрагмент информационного наполнения документа (текст, рисунок и т. д.) или выполняет роль контейнера для одного или нескольких объектов.

Для объектов-контейнеров доступны следующие метасвойства: $smet_c$ — определяет объект в качестве контейнера; *mixed* — разре-

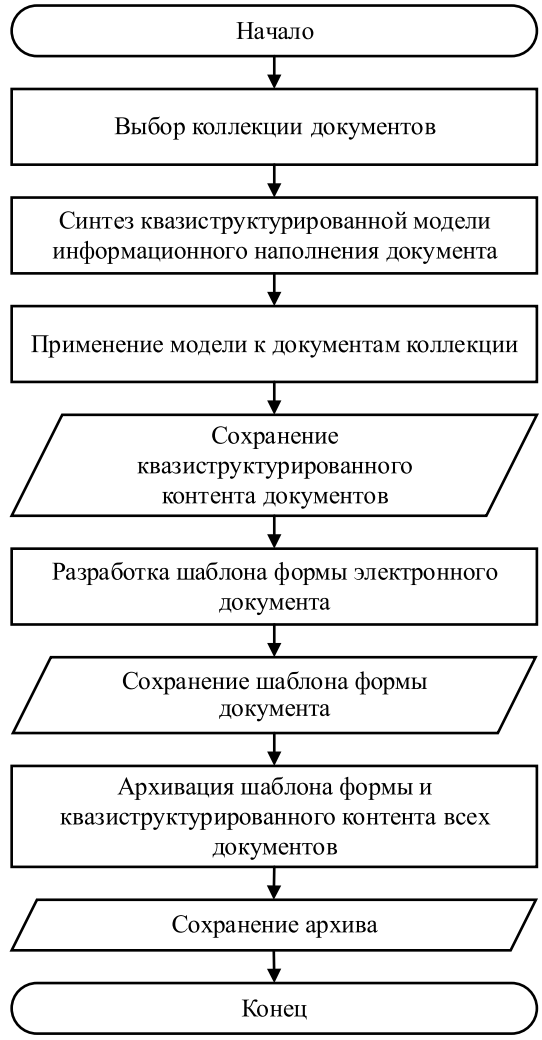

Рис. 1. Алгоритм сохранения электронных документов

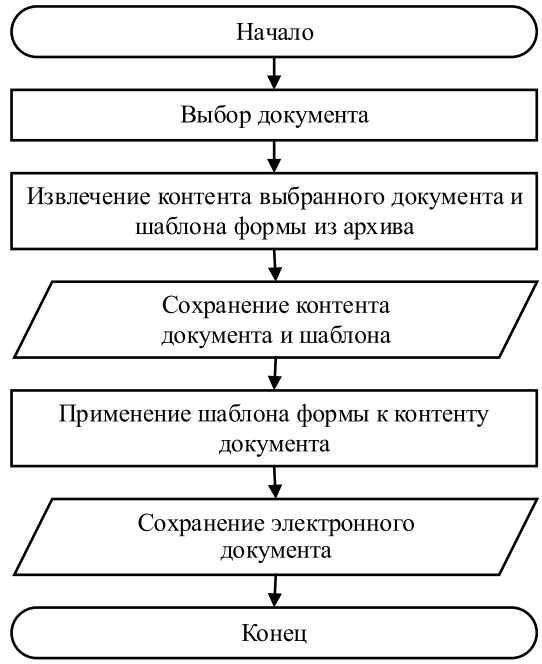

Рис. 2. Алгоритм восстановления электронного документа

шает использование объектов-потомков в произвольном порядке. *LObj* — отображение, определенное на множестве *sObj*, такое что $sObj \xrightarrow{LObj} \{obj_1, \ldots, obj_n\}$, где $obj_i \in sObj$ — дочерний объект; *n* — число дочерних объектов; *Obj_met* — отображение, определенное на множестве *sObj*, такое что $sObj \xrightarrow{Obj\_smet} \rightarrow \{smet_c | smet_c, mixed | smet_1, \ldots, smet_k\}$, где $smet_i \in sMet$ — метасвойство ограничения на содержимое объекта; *minOccurs* — функция, определяющая минимально возможное число раз использования объекта в модели; *maxOccurs* — функция, определяющая максимально возможное число раз использования объекта в модели.

Для разработки квазиструктурированной модели фактографического контента инфомационного наполнения документа "Информационная карта скважины" используем "Способ формирования квазиструктурированных моделей фактографического информационного наполнения документов" [5].

Структура модели документа "Информационная карта скважины" описывает основную информацию о газовых скважинах, такую как подключение, конструкция, информация об исследованиях, рабочие дебиты и т. д.

Модель "Информационная карта скважины" представлена базовым сегментом, рассмотрим его структуру подробнее (рис. 3).

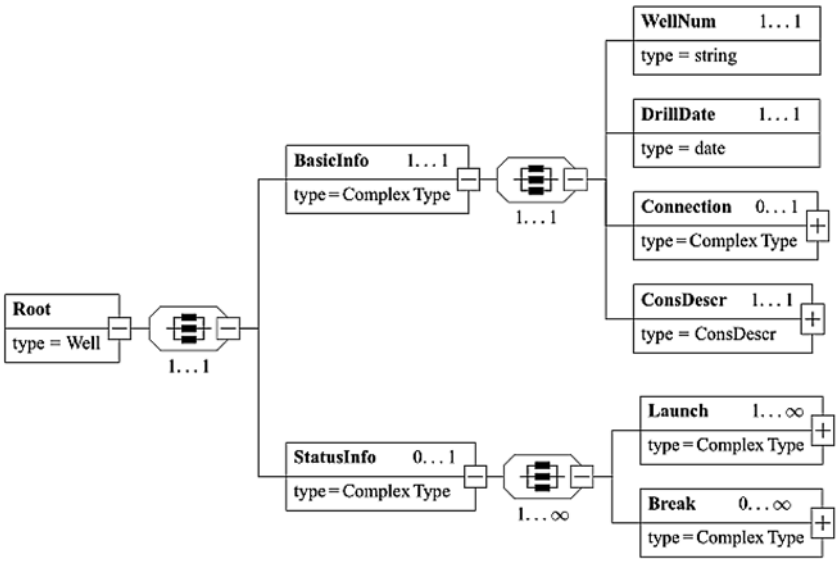

Рис. 3. Структура базового сегмента модели "Информационная карта скважины"

Базовый сегмент состоит из двух сегментов: обязательного **BasicInfo** и необязательного **StatusInfo**. Сегмент **BasicInfo** хранит базовую информацию о пробуренной скважине. Он состоит из двух объектов **WellNum** — номер скважины и **DrillDate** — дата бурения, а также двух сегментов: **Connection** (рис. 4) и **ConsDescr** (рис. 5).

Обязательный сегмент **ConsDescr** используется для описания конструкции скважины и состоит из двух необязательных объектов: **PerfDiap** — информация о перфорации и **TrunkDepth** — информация об открытом стволе. Оба рассмотренных объекта могут быть использованы неограниченное число раз.

Сегмент **StatusInfo** хранит информацию о пусках и остановках скважины. Он состоит из двух сегментов: обязательного сегмента **Launch** (рис. 6), который может быть использован в контенте неограниченное число раз, и необязательного сегмента **Break** (рис. 7), который также может быть использован неограниченное число раз.

Данный сегмент состоит из трех объектов: **FieldTitle** — название месторождения, **Train** — номер шлейфа, к которому подключена скважина и **UKPG** — номер установки комплексной подготовки газа (УКПГ), к которой поступает продукция от скважины. Рассмотренные объекты необязательны к использованию в документе.

Сегмент **Launch** (см. рис. 6) используется для хранения информации о работе скважины. Данный сегмент состоит из обязательного сегмента **Debit**, который может быть использо-

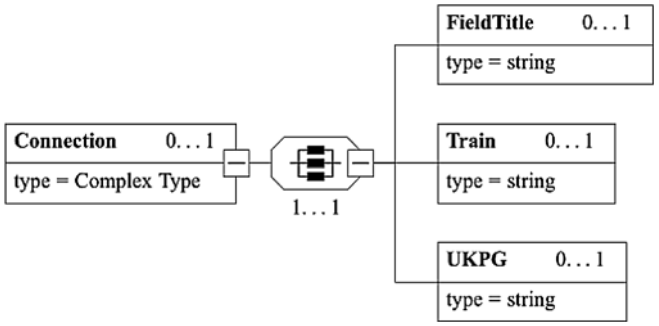

Рис. 4. Структура сегмента Connection

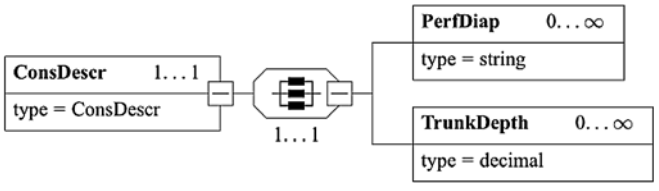

Рис. 5. Структура сегмента ConsDescr

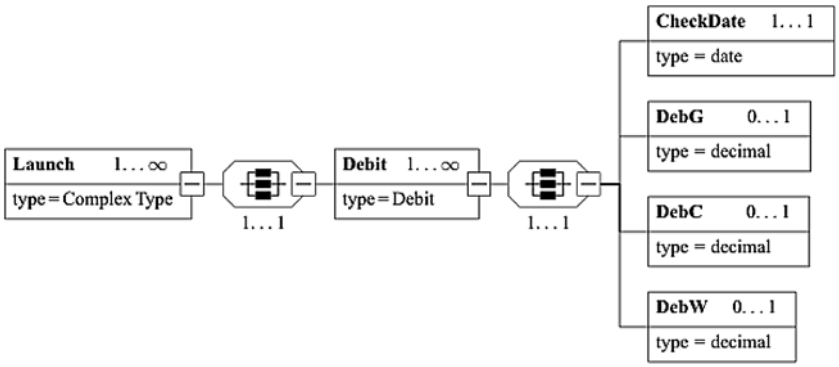

Рис. 6. Структура сегмента Launch

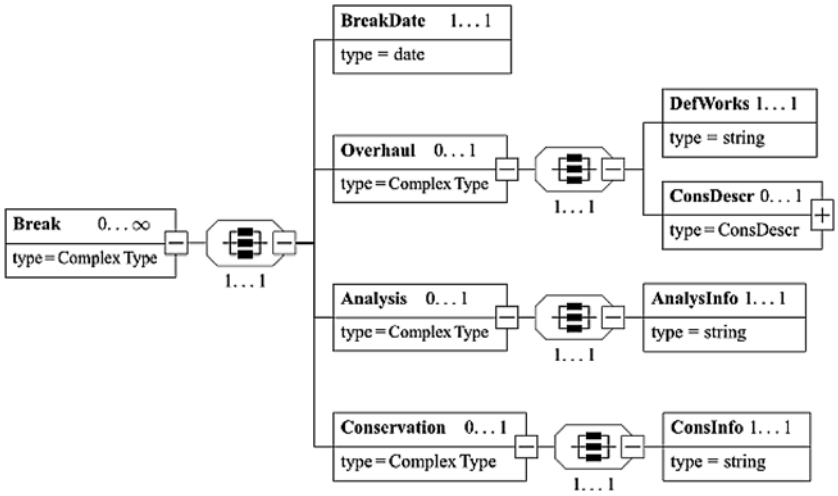

Рис. 7. Структура сегмента Break

ван неограниченное число раз. Сегмент **Debit**, в свою очередь, состоит из обязательного объекта **CheckDate** — дата отчетного месяца, и необязательных объектов **DebG** — месячный дебит газа, **DebC** — месячный дебит конденсата, **DebW** — месячный дебит воды.

Необязательный сегмент **Break** (рис. 7) предназначен для хранения информации об остановках и причинах этих остановок в работе скважины. Он состоит из обязательного объекта **BreakDate** — дата остановки, а также трех необязательных сегментов: **Overhaul** — используется для хранения информации о капитальном ремонте скважины, **Analysis** — используется для хранения информации о результатах гидродинамических исследованиях скважины и **Conservation** — используется для хранения информации о консервации скважины.

В состав сегмента **Overhaul** входят обязательный объект **DefWorks** — комплекс выполненных ремонтных работ и необязательный сегмент **ConsDescr**, хранящий информацию о текущем состоянии конструкции скважины.

Разметив содержимое электронных документов с помощью описанной модели, извлечем фактографические данные из документов. Полученные данные будут представлены в формате XML. Процесс разметки содержимого электронного документа с помощью модели может быть реализован как в ручном, так и в автоматизированном режиме [6].

Используя модель документа и формат Office Open XML, формируем шаблон формы электронного документа [7].

В результате проделанных операций для коллекции документов будет получено их фактографическое содержимое и шаблон формы электронного документа данного вида.

Таким образом, объем хранимой информации был существенно снижен без потери фактографических данных коллекции документов. Объем результирующий информации может быть дополнительно снижен за счет применения современных технологий архивирования данных.

Восстановление электронного документа до первоначального вида осуществляется с помощью алгоритма восстановления электронного документа, рассмотренного ранее, и может быть проиллюстрировано с помощью схемы, показанной на рис. 8.

Для оценки эффективности описанного способа хранения электронных документов с квазиструктурированным контентом были подготовлены наборы коллекций, состоящие из 1 000 и 2 000 документов "Информационная карта скважины". Результаты сравнения представлены в таблице.

Сравнение эффективности способа хранения электронных документов

Вариант хранения электронного документа	1000 документов	2000 документов
Исходный размер	29,2 Мбайт	58,5 Мбайт
Сжатие архивацией (7-Zip)	23,9 Мбайт	47,8 Мбайт
Сжатие алгоритмом	6,83 Мбайт	13,6 Мбайт
Сжатие алгоритмом + архивация (7-Zip)	1,06 Мбайт	2,11 Мбайт

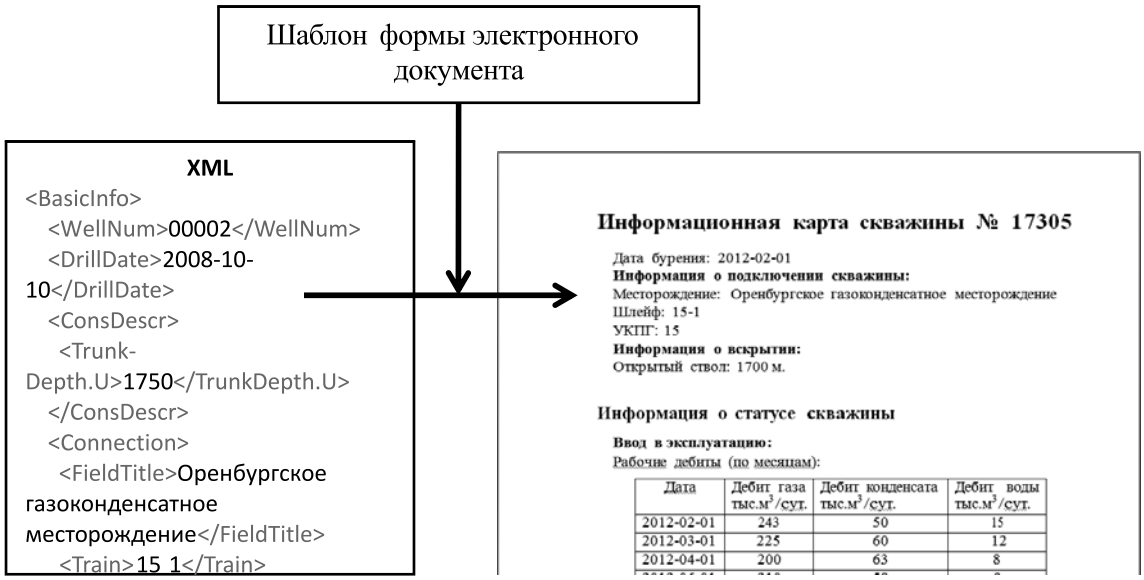

Рис. 8. Восстановление электронного документа

Заключение

Рассмотренный способ хранения электронных документов с квазиструктурированным контентом минимизирует их объем при хранении за счет выделения единого шаблона оформлении документов, извлечения из документов фактографических данных с последующим сжатием шаблона и фактографических данных, что обеспечивает сжатие электронных документов до 22 раз эффективнее стандартных архиваторов.

На описанный в работе способ хранения электронных документов с квазиструктурированным контентом получен патент [8].

Список литературы

1. **Палей Д.** Моделирование квазиструктурированных данных // Открытые системы. 2002. № 9. С. 57—64.

2. **Черников Б. В.** Технология хранения слабоформализуемых документов на основе лексикологического синтеза // Информатика и ее применение. 2009. Т. 3, Вып. 4. С. 64—75.

3. **Черников Б. В.** Способ преобразования слабоформализуемых документов для минимизации их объема при хранении. Патент на изобретение, рег. № 2413985 от 10.03.2011. М.: Роспатент, 2011.

4. **Полищук Ю. В., Черных Т. А.** Синтез квазиструктурированных моделей информационного наполнения электронных // Вестник компьютерных и информационных технологий. 2012. № 6. С. 20—27.

5. **Полищук Ю. В.** Способ формирования квазиструктурированных моделей фактографического информационного наполнения документов. Патент на изобретение, рег. № 2517428 от 28.03.2014. М.: Роспатент, 2014.

6. **Полищук Ю. В., Ларин А. В.** Автоматизация процесса разметки управленческого контента электронных документов с квазиструктурированным информационным наполнением на основе паттернов // Вестник компьютерных и информационных технологий. 2013. № 3. С. 55—60.

7. **Воутер В. В.** Open XML — Кратко и доступно. Open XML Technical Evangelist. Microsoft, 2007. 101 с.

8. **Полищук Ю. В., Полищук П. В.** Способ преобразования документов для минимизации их объема при хранении электронных документов с квазиструктурированным информационным наполнением. Патент на изобретение, рег. № 2625611 от 17.07.2017. М.: Роспатент, 2017.

Yu. V. Polishuk, PhD, Associate Professor
of Computer Security mathematical software and information systems, Orenburg State University

The Method of Storing Electronic Documents with Semistructured Content

The accompanying operational documentation is formed in the work process of the enterprise, which, as a rule, is represented by a collection of documents, among which for documents of a single type it is possible to single out a general structure, but the design and procedure for placing information in them will be different. The latter is because the documentation is formed on the basis of internal standards of the enterprise or in accordance with the requirements of GOST in which the requirements for the content of documents are defined. The content of documents of this type is semistructured, i.e. is represented by semistructured

information. The semistructured information is understood as information in which a certain structure can be identified, but this structure is completely unknown in advance or may change with time. With a long period of the enterprise's work, a large number of electronic documentation with semistructured content is accumulated, as a rule, represented by MS Word documents and minimization of its storage volume is an actual task. The described method of storing electronic documents is to minimize the amount of storage of electronic documents with semistructured content by allocating a single template for processing documents, extracting factual data from documents, and then compressing the template and factographic data. The restoring the document is done by extracting from the archive the factual data of the form template and specified content of the document, and applying the form template to the newly received document content. The method provides compression of electronic documents with semistructured content up to 22 times more efficient than standard archivers.

Keywords: *semistructured data; information compression; electronic documents processing*

DOI: 10.17587/it.25.53-58

References

1. **Palej D.** *Modelirovanie kvazistrukturirovannyh dannyh* (Simulation of quasi-structured data), *Otkrytye Sistemy*, 2002, no. 9, pp. 57—64 (in Russian).

2. **Chernikov B. V.** Tehnologija hranenija slaboformalizuemyh dokumentov na osnove leksikologicheskogo sinteza (Storage technology of poorly formalized documents based on lexicological synthesis), *Informatika i ee Primenenie*, 2009, vol. 3, iss. 4, pp. 64—75 (in Russian).

3. **Chernikov B. V.** *Sposob preobrazovanija slaboformalizuemyh dokumentov dlja mini-mizacii ih ob#ema pri hranenii* (The method of converting poorly formalized documents to minimize their volume during storage), Patent na izobretenie, reg. № 2413985 ot 10.03.2011, Moscow, Rospatent, 2011 (in Russian).

4. **Polishhuk Ju. V., Chernyh T. A.** *Sintez kvazistrukturirovannyh modelej informacionnogo napol-nenija jelektronnyh dokumentov* (Synthesis of quasi-structured models of informational filling of electronic), *Vestnik Komp'ju-Ternyh I Informacionnyh Tehnologij*, 2012, no. 6, pp. 20—27 (in Russian).

5. **Polishhuk Ju. V.** *Sposob formirovanija kvazistrukturirovannyh modelej faktogra-ficheskogo informacionnogo napolnenija dokumentov* (The method of forming quasi-structured models of factual content of documents), Patent na izobretenie, reg. № 2517428 ot 28.03.2014, Moscow, Rospatent, 2014 (in Russian).

6. **Polishhuk Ju. V., Larin A. V.** *Avtomatizacija processa razmetki upravlencheskogo kontenta jelek-tronnyh dokumentov s kvazistrukturirovannym informacionnym napolneniem na osnove patternov* (Automating the process of marking management content of electronic documents with quasi-structured content based on patterns), *Vestnik komp'juternyh i infor-macionnyh tehnologij*, 2013, no. 3, pp. 55—60 (in Russian).

7. **Vouter V. V.** *Open XML — Kratko i dostupno* (Open XML — Brief and Available), Open XML Technical Evangelist, Microsoft, 2007, 101 p. (in Russian).

8. **Polishhuk Ju. V., Polishhuk P. V.** *Sposob preobrazovanija dokumentov dlja minimizacii ih ob#joma pri hranenii jelektronnyh dokumentov s kvazistrukturirovannym informacionnym napolneniem* (The method of converting documents to minimize their volume when storing electronic documents with quasi-structured information content), Patent na izobretenie, reg. № 2625611 ot 17.07.2017, Moscow, Rospatent, 2017 (in Russian).

УДК 004.421 DOI: 10.17587/it.25.58-63

Т. О. Перемитина, канд. техн. наук, доц., e-mail: peremitinat@mail.ru,
И. Г. Ященко, канд. геол.-минер. наук, e-mail: sric@ipc.tsc.ru,
Федеральное государственное бюджетное учреждение науки Институт химии нефти СО РАН, Томск

Алгоритм комплексного анализа многомерных данных об экологическом состоянии окружающей среды

Рассматриваются методические вопросы реализации и практического применения комплексного подхода к анализу многомерных данных об экологическом состоянии пространственно-распределенных объектов исследования. Подход основан на сочетании метода главных компонент и метода пространственного анализа с применением геоинформационных систем. Метод главных компонент применяется для статистической обработки и анализа данных, метод пространственного анализа используется для учета пространственных свойств исследуемых объектов. В статье приведены результаты анализа многомерных данных о состоянии природной среды и здоровья населения территорий Сибирского и Дальневосточного федеральных округов.

Ключевые слова: *многомерные данные, метод главных компонент, геоинформационные системы, окружающая среда, экология*

Введение

Мониторинг состояния окружающей среды сопряжен с обработкой постоянно увеличивающегося массива разрозненной совокупной информации об объектах исследования. Аналитические процессы предполагают обработку