УДК 004.550

ЦИФРОВОЙ ГЕОЛОГИЧЕСКИЙ РЕПОЗИТОРИЙ И ИНФОРМАЦИЯ О СТРАТИГРАФИЧЕСКОМ ВОЗРАСТЕ (НА ПРИМЕРЕ DSPACE)

М. И. Патук[1, [0000-0003-3036-2275]], В. В. Наумова[2, [0000-0002-3001-1638]]

Федеральное государственное бюджетное учреждение науки Государственный геологический музей им. В.И. Вернадского РАН

[1]patuk@mail.ru, [2]naumova_new@mail.ru

Аннотация

Описан новый подход, связанный с извлечением терминов относительного геологического возраста из метаданных научных геологических публикаций. На основе разработанных и адаптированных подходов и технологических решений реализован комплекс макросов, реализующий функции поиска, извлечения и добавления новых метаданных к научным публикациям.

***Ключевые слова**: информационные технологии, науки о Земле, репозиторий, научные публикации, стратиграфический возраст*

ВВЕДЕНИЕ

В рамках разработки Информационно-аналитической геологической среды "GeologyScience.ru", которая обеспечивает единую точку доступа к геологическим данным и системам их обработки на территории России [1, 2], был создан и поддерживается блок управления научными геологическими публикациями – http://repository.geologyscience.ru [3].

Для оперативного и полноценного доступа к информации необходимо как можно полнее описать каждую научную публикацию, снабдив ее, по возможности, достаточным перечнем поисковых признаков в виде метаданных. Стандартным подходом в таких случаях является классификация источников с помощью словарей или тезаурусов [4].

Названный репозиторий создан на основе свободно распространяемого ПО DSpace v.6.3. По умолчанию в системе настроены для поиска следующие метаданные: дата публикации, авторы, наименование публикации, тематика. Тематика содержит ключевые слова, которые авторы указали для своей работы.

Но не все публикации, особенно русскоязычные, содержат ключевые слова. Кроме того, тематика публикации зачастую богаче того тематического перечня, что указан авторами.

Как уже отмечалось [3], данный репозиторий является составной частью информационно-аналитической среды "GeologyScience.ru". Но для целей полноценной интеграции с другими компонентами системы имеющийся тематический перечень оказался недостаточным. Нами уже был добавлен дополнительный параметр для поиска – УДК. Но данный параметр доступен только для русскоязычных публикаций и то далеко не для всех.

Следующим шагом в расширении поисковых возможностей репозитория мы решили сделать добавление относительного геологического возраста. За основу были взяты международная хроностратиграфическая шкала [5] и ее российская адаптация с сайта ВСЕГЕИ [6]. Сразу стоит оговорится, что строго придерживаться данной шкалы оказалось достаточно трудно, а в некоторых случаях просто невозможно. Довольно часто в публикациях содержатся устаревшие указания возраста, например, третичный период (отменен в середине XX века), которые зачастую невозможно привести к современной хроностратиграфической шкале. Кроме того, российская шкала и ее временные интервалы в некоторых случаях не совпадают с международной хроностратиграфической шкалой – например, рифей и венд.

В принципе в литературе уже достаточно подробно описано извлечение дополнительных данных из текстовых документов на примере географических названий [7]. Хотя методика, предложенная для извлечения географических наименований, для нас является избыточной, у нас не стоит задача сопрягать извлеченные данные с геоинформационными системами. Но основа изложенного подхода вполне применима для извлечения относительного геологического возраста.

В качестве СУБД в нашей версии DSpace используется PostgreSQL v.11.1. Как отмечено [8], в PostgreSQL реализован встроенный механизм полнотекстового поиска (FTS), который реализует обработку текстовой информации с возможностью расширения базовых функций за счет дополнительных словарей. Кроме того, до-

ступно создание дополнительных модулей на различных языках программирования. Данный механизм подробно описан [9]. Качество работы полнотекстового поиска в PostgreSQL напрямую зависит от полноты и качества используемых справочников. Справочником в нашем случае является иерархический словарь, созданный на основе хроностратиграфической шкалы.

Для англоязычных текстов проблем идентификации возраста, с использованием этого словаря, практически не возникает, особенно для англоязычных авторов. Редким исключением являются синонимичные написания терминов, например, Paleo – Palaeo.

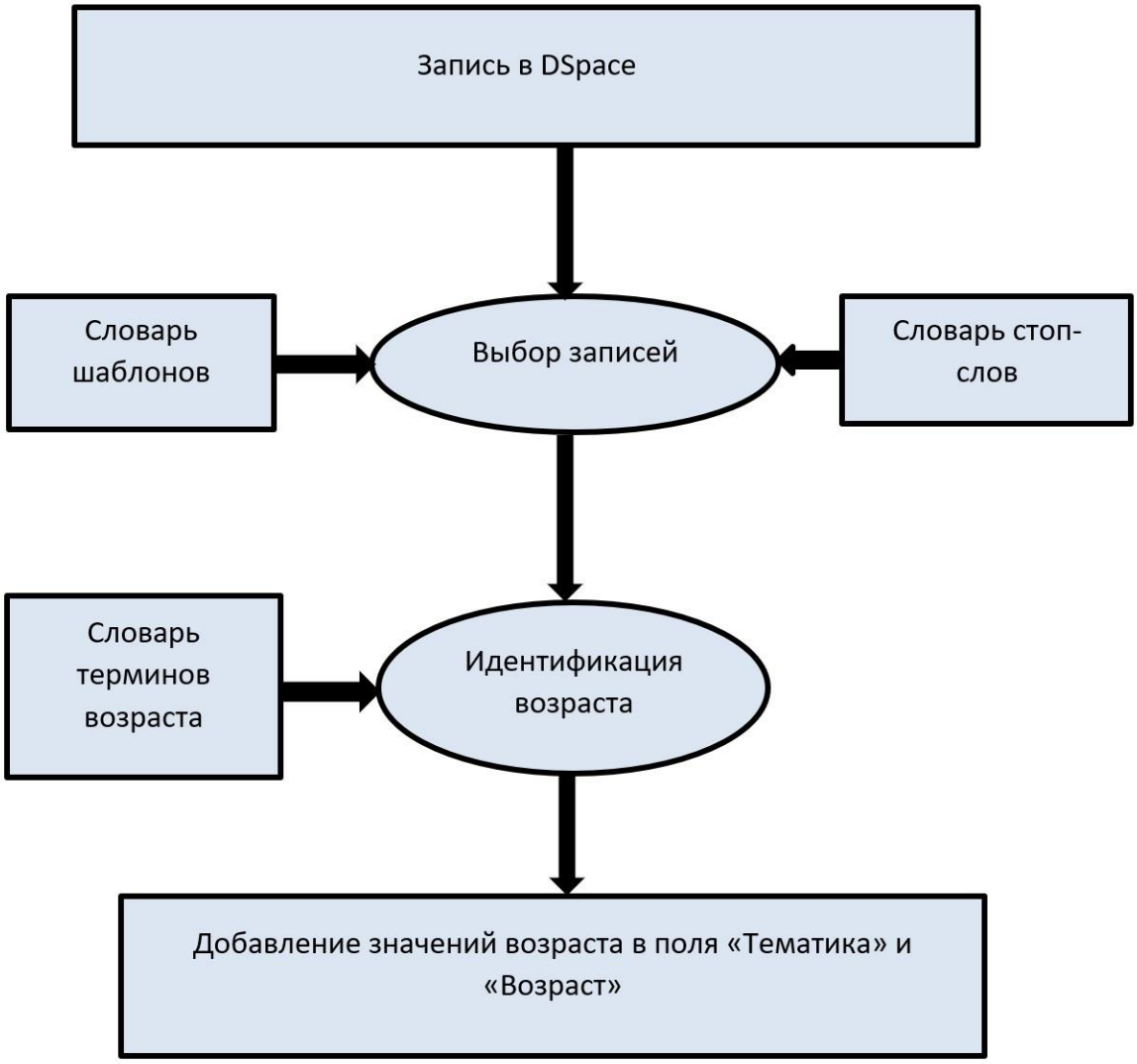

Рис. 1. Алгоритм идентификации относительного геологического возраста.

Немного сложнее происходит обработка англоязычных текстов русскоязычных авторов. В тексты вносятся русскоязычные кальки принятых в русскоязычных статьях сокращений и жаргонных терминов, например, carbon вместо carboniferous (каменноугольный период) или meso- вместо mesozoic (мезозойская эра). Основные проблемы при извлечении терминов возраста связаны с русскоязычными статьями. Кроме уже упомянутых сокращений и жаргонных терминов, присутствуют различные словоформы с использованием кратких прилагательных – нижне-, средне-, ранне- и т. п. Для корректной работы алгоритма извлечения приходится создавать полноценный тезаурус (словарь замен), с множеством словоформ исходного понятия. Дополнительно необходимо использовать словарь стоп-слов, чтобы не происходило ложной идентификации, например, Пермского края в качестве Пермского периода Палеозойской эры.

Схема извлечения терминов возраста выглядит следующим образом. Анализируются метаданные публикаций в следующем порядке: наименование статьи, краткое содержание, ключевые слова. Для улучшения быстродействия алгоритма создается выборка записей на основе словаря шаблонов – сокращенных словоформ терминов возраста. Если встречается термин возраста в наименовании или в кратком содержании, то происходит поиск этого термина в ключевых словах. Если искомый термин там находится, то происходит добавление этого термина в поле «Возраст». Если не находится, то дополнительно термин добавляется в ключевые слова. Блок схема алгоритма поиска и добавления терминов возраста представлена на рис. 1.

В DSpace наряду с обычным поиском встроен так называемый фасетный поиск. В фасетах по умолчанию выводится несколько терминов с наибольшей частотой встречаемости. Также в фасетах есть возможность вывода иерархически устроенных терминов. Поскольку шкала хроностратиграфических возрастов является иерархической, мы использовали эту возможность при добавлении терминов в поле «Возраст». Значения возраста при этом выглядят следующим образом: Эра::Период::Эпоха::Ярус (например, если был найден термин – «Неоплейстоцен», то в поле «Возраст» будет внесено следующее значение – Кайнозой::Четвертичная::Плейстоцен::Неоплейстоцен). При таком способе указания возраста появляется возможность поиска по любой из его составных частей, т. е. или по

эре, или по периоду, или по эпохе, или по ярусу. В окне фасетного поиска для возраста при таком способе организации термина происходит подсчет числа вхождений по каждой из составных частей (рис. 2). В поле «Тематика» значение возраста вносится в сокращенном варианте, т. е. только извлеченный термин.

Рис. 2. Страница иерархического фасетного поиска по возрасту.

ЗАКЛЮЧЕНИЕ

Разработан и адаптирован новый подход, связанный с извлечением понятий относительного геологического возраста из метаданных публикаций, на основе международной хроностратиграфической шкалы.

Предложены новые технические решения: обработка записей репозитория с фильтрацией по специализированному словарю терминов, выделением новых понятий и добавлением их в список метаданных.

На основе разработанных и адаптированных подходов и технологических решений реализован комплекс макросов, реализующий функции поиска, извлечения и добавления новых метаданных к научным публикациям.

Благодарности

Научные исследования выполняются в рамках Государственного задания ФГБУН Государственного геологического музея им. В.И. Вернадского РАН по теме № 0140-2019-0005 «Разработка информационной среды интеграции данных естественнонаучных музеев и сервисов их обработки для наук о Земле».

СПИСОК ЛИТЕРАТУРЫ

1. *Наумова В.В., Платонов К.А., Еременко В.С., Патук М.И., Дьяков С.Е.* Информационно-аналитическая среда для поддержки научных исследований в геологии: текущее состояние и перспективы развития // Труды XVII Международной конференции «Распределенные информационно-вычислительные ресурсы. Цифровые двойники и большие данные (DICR-2019)», 2019. С. 139–147.

2. *Naumova V.V., Eremenko V.S., Platonov K.A., Dyakov S.V., Patuk M.I., Eremenko A.S.* Development of geographically distributed information-analytical geological environment // Russian Journal of Earth Sciences. 2019. Vol. 19, No. 6., ES6012, URL: https://doi.org/DOI:10.2205/2019ES000696.

3. *Патук М.И., Наумова В.В., Ерёменко В.С.* Цифровой репозиторий "geologyscience.ru": открытый доступ к научным публикациям по геологии России // *Электронные библиотеки. 2020. Т. 23*, № 6. С. 1324–1338. URL: https://doi.org/10.26907/1562-5419-2020-23-6-1324-1338.

4. *Федотов А.М., Идрисова И.А., Самбетбаева М.А., Федотова О.А.* Использование тезауруса в научно-образовательной информационной системе // Вестник НГУ. Серия: Информационные технологии. 2015. Т. 13, вып. 2 С. 86–102.

5. Interactive international chronostratigraphic chart, URL: https://stratigraphy.org/timescale/

6. Стратиграфическая основа ГК-200 и ГК-1000, URL: https://vsegei.ru/ru/info/stratigraphy/

7. *Жижимов О.Л., Леонова Ю.В.* О географической привязке контента текстовых документов // CEUR Workshop Proceedings. 2020. SDM 2019 – Proceedings of the All-Russian Conference "Spatial Data Processing for Monitoring of Natural and Anthropogenic Processes", 2019. Vol. 2534, P. 241–247.

8. *Жижимов О.Л.* Технология извлечения географических названий из текстовых документов на основе инструментария PostgreSQL // Вестник восточно-казахстанского государственного технического университета им. Д. Серикбаева, 2018. № 3-1, С. 195–203.

9. *Бартунов О., Сигаев Ф.* Введение в полнотекстовый поиск в PostgreSQL, URL: http://citforum.ru/database/postgres/fts/bib.shtml.

DIGITAL GEOLOGICAL REPOSITORY AND INFORMATION ON STRATIGRAPHIC AGE (ON THE EXAMPLE OF DSPACE)

M. I. Patuk[1] [0000-0003-3036-2275], V. V. Naumova[2, [0000-0002-3001-1638]

Vernadsky State Geological Museum RAS, Moscow (Russia)

[1]patuk@mail.ru, [2]naumova_new@mail.ru

Abstract

A new approach related to the extraction of terms of relative geological age from the metadata of scientific geological publications is described. Based on the developed and adapted approaches and technological solutions, a set of macros is implemented that implements the functions of searching, extracting and adding new metadata to scientific publications.

Keywords: *information technologies, Earth sciences, repository, scientific publications, stratigraphic age*

REFERENCES

1. *Naumova V.V., Platonov K.A., Eremenko V.S., Patuk M.I., Dyakov S.V.* Informacionno-analiticheskaja sreda dlja podderzhki nauchnyh issledovanij v geologii: tekushhee sostojanie i perspektivy razvitija // Trudy XVII Mezhdunarodnoj konferencii «Raspredelennye informacionno-vychislitel'nye resursy. Cifrovye dvojniki i bol'shie dannye (DICR-2019)», 2019. P. 139–147.

2. *Naumova V.V., Eremenko V.S., Platonov K.A., Dyakov S.V., Patuk M.I., Eremenko A.S.* Development of geographically distributed information-analytical geological environment // Russian Journal of Earth Sciences. 2019. V. 19, No. 6, ES6012, URL: https://doi.org/10.2205/2019ES000696.

3. *Patuk M.I., Naumova V.V., Eryomenko V.S.* Cifrovoj repozitorij "geology-science.ru": otkrytyj dostup k nauchnym publikaciyam po geologii Rossii // Elektronnye biblioteki. 2020. V. 23, №6. P. 1324–1338. URL: https://doi.org/10.26907/1562-5419-2020-23-6-1324-1338.

4. *Fedotov A.M., Idrisova I.A., Sambetbaeva M.A., Fedotova O.A.* Ispol'zovanie tezaurusa v nauchno-obrazovatel'noj informacionnoj sisteme // Vestnik NGU, Seriya: Informacionnye tekhnologii. 2015. V. 13, vyp. 2. S. 86–102.

5. Interactive international chronostratigraphic chart, URL: https://stratigraphy.org/timescale/

6. Stratigraficheskaya osnova GK-200 i GK-1000, URL: https://vsegei.ru/ru/info/stratigraphy/

7. *Zhizhimov O.L., Leonova Yu.V.* O geograficheskoj privyazke kontenta tekstovyh dokumentov // CEUR Workshop Proceedings. 2020. SDM 2019 – Proceedings of the All-Russian Conference "Spatial Data Processing for Monitoring of Natural and Anthropogenic Processes", 2019. V. 2534, P. 241–247.

8. *Zhizhimov O.L.* Tekhnologiya izvlecheniya geograficheskih nazvanij iz tekstovyh dokumentov na osnove instrumentariya PostgreSQL // Bulletin of D. Serikbayev EKTU, 2018, № 3-1, P. 195–203.

9. *Bartunov O., Sigaev F.* Vvedenie v polnotekstovyj poisk v PostgreSQL, URL: http://citforum.ru/database/postgres/fts/bib.shtml.

СВЕДЕНИЯ ОБ АВТОРАХ

ПАТУК Михаил Иванович – к. г.-м. н., и. о. н. с., научный отдел Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Michail Ivanovich PATUK – PhD, scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: patuk@mail.ru

ORCID: 0000-0003-3036-2275

НАУМОВА Вера Викторовна – д. г.-м. н., г. н. с., зав. Научным отделом Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия

Vera Viktorovna NAUMOVA – Prof., head of SGM scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: Naumova_new@mail.ru

ORCID: 0000-0002-3001-1638

Материал поступил в редакцию 15 мая 2021 года