ОГЛАВЛЕНИЕ

ОТ СОСТАВИТЕЛЕЙ

Настоящий номер журнала «Электронные библиотеки» является второй частью тематического выпуска (первая часть – №5 за 2021 год) и включает статьи, подготовленные их авторами на основе материалов, представленных и доложенных на XXIII Всероссийской научной конференции «Научный сервис в сети Интернет». Конференция была проведена с 20 по 23 сентября 2021 года в режиме онлайн и традиционно была посвящена направлениям и тенденциям использования интернет-технологий в современных научных исследованиях. Организатором конференции был Институт прикладной математики им. М.В. Келдыша Российской академии наук.

М.М. Горбунов-Посадов, А.М. Елизаров

ПАМЯТИ АЛЕКСАНДРА НИКОЛАЕВИЧА ТОМИЛИНА

3 декабря 2021 г. на 89-м году ушёл из жизни выдающийся учёный, заслуженный деятель науки Российской Федерации, постоянный участник нашей конференции и всеми нами любимый Александр Николаевич Томилин.

Его научные результаты впечатляют. После окончания мехмата МГУ он выполнил ряд чрезвычайно успешных пионерских работ в области программного обеспечения противовоздушной и противоракетной обороны страны. Затем принял активнейшее участие в проектировании БЭСМ-6 – флагмана советской вычислительной техники, а также в разработке первой операционной системы для этой машины. За эти работы в 1969 году Александр Николаевич был награжден Государственной премией СССР.

Более 35 лет он преподавал на факультете ВМК МГУ и в МФТИ. Читал курсы по вычислительным системам и операционным системам ЭВМ. Много лет был председателем Государственной аттестационной комиссии ВМК МГУ.

С момента образования Российского фонда фундаментальных исследований и до последних дней Александр Николаевич был фактическим главой программистского сектора фонда, хотя обычно и не занимал там формально ведущих

административных позиций. Был хорошо знаком практически со всеми работающими в стране мало-мальски заметными программистскими коллективами, пользовался безграничным авторитетом при решении спорных вопросом, нередко возникавших при работе фонда.

Его участие в работе нашей конференции «Научный сервис в сети Интернет» неоценимо. Конечно, он не раз выступал с интереснейшими докладами и сообщениями, работал руководителем секций конференции. Но главное – он во многом формировал неизменно доброжелательную атмосферу конференции. Каждый участник получал от него свою долю внимания. Будучи неизменным тамадой на завершающих банкетах, для каждого выступающего Александр Николаевич неизменно находил точные теплые слова.

Не будет преувеличением сказать, что Александр Николаевич был душой всего научного программистского сообщества страны. Он всех нас прекрасно знал и любил, и мы всегда отвечали ему уважением и любовью.

Светлая память.

Программный комитет конференции «Научный сервис в сети Интернет»

УДК 004.65, 004.053, 005, 001.5

СЕМАНТИЧЕСКАЯ БИБЛИОТЕКА КАК СРЕДСТВО ОПРЕДЕЛЕНИЯ НАУЧНОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

О. М. Атаева[1] [0000-0003-0367-5575], В. А. Серебряков[2] [0000-0003-1423-621X]

[1,2] Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, Москва, ул. Вавилова, 40

[1] oli@ultimeta.ru

[2] serebr@ultimeta.ru

Аннотация

Рассмотрены информационная система, предназначенная для представления предметной области, связанной с наукой, и ее особенности. Выделены общие концепции для формального описания такой предметной области в базе знаний семантической библиотеки. Особенность этих областей заключается в том, что структура данных подвержена частым изменениям. Поэтому средство организации знаний, в качестве которого выступает семантическая библиотека, должно быть достаточно универсальным и не требовать глубоких технических познаний. В работе приведены описание функциональности системы и ее использования при настройке на предметную область. Для каждой области набор ресурсов может отличаться как по формату, так и по набору самих ресурсов. Набор понятий, формирующих описание контента библиотеки, должен быть настолько универсальным, чтобы мог адаптироваться под нужды конкретной области. Для представления данных использованы метаданные трех уровней.

Ключевые слова: *семантическая библиотека, онтология, представление знаний*

ВВЕДЕНИЕ

Вопросами семантической организации знаний занимались различные исследователи с древнейших времен. Библиотеки, специализированные по конкретным областям, используют обычно свои классификаторы для систематизации своих ресурсов. Такой подход обеспечивает более детальный анализ содержания

документов и соотношение смысловых понятий содержимого библиотеки с определенным направлением специализированной области знания.

Накопленные данные стали доступны широкому кругу пользователей через Сеть, функциональность цифровых библиотек становится все разнообразней, удовлетворяя *информационные потребности* пользователей.

В фокусе настоящей работы лежат предметные области, связанные с наукой, и их особенности. Выделены общие концепции для их формальных описаний в базе знаний. Особенность этих областей заключается в том, что структура данных подвержена частым изменениям [1–4]. Основной акцент сделан на представлении обобщенной модели научной предметной области и ее особенностей, реализации в поисковых системах и отличиях от классических подходов к поиску информации в научных массивах данных.

Новые проблемы и вызовы относятся также к представлению знаний в информационной среде для различных областей науки с использованием современных подходов. Для обеспечения потребления научной информации на новом уровне в первую очередь необходим переход к семантически значимому представлению научных знаний, извлекаемых из информации в цифровой среде.

Для представления данных предметной области используют метаданные трех уровней: (1) универсальные понятия без привязки к предметной области, или метаметаданные; (2) понятия для описания конкретной предметной области или метаданные, определения которых задаются в терминах первого уровня; (3) данные прикладной области как таковые, представленные в терминах метаданных второго уровня. На основе этих метаданных настраиваются интерфейсы взаимодействия с пользователями для навигации, редактирования и поиска информации.

Главной задачей создания и описания обобщенного представления научных знаний для некоторой области является помощь экспертам в организации знаний и предоставления доступа к ней [5–9]. При этом средство организации знаний должно быть достаточно универсальным и не требовать глубоких технических познаний.

Была поставлена задача создания такой информационной системы, которая могла бы учитывать все разнообразие различных типов ресурсов научной предметной области, которые могут в ней храниться, и при этом поддерживать ее терминологическое описание. Фактически такая система должна представлять собой конструктор с адаптируемой моделью контента хранимых данных для создания цифровой библиотеки любой направленности. Адаптируемая модель данных позволяет описать произвольную модель данных контента библиотеки в рамках предметной области, фиксированной в терминах тезауруса. Такая информационная система должна учитывать разнообразие типов ресурсов научной предметной области и при этом поддерживать ее терминологическое описание. Основные задачи такой системы – представление контента предметной области в виде онтологии и поддержка интеграции данных из источников. На данный момент реализован и готов к использованию дистрибутив семантической библиотеки. Ниже дано описание основных идей построения модели данных и подсистем, которые представлены в дистрибутиве информационной системы

1. О МОДЕЛИ ДАННЫХ

В информационную модель семантической библиотеки были введены понятия для описания содержимого библиотеки для некоторой предметной области [10–13]. Эти понятия позволяют сконструировать описание любых типов информационных ресурсов для этой области. При этом согласно определению информационные объекты, являющиеся непосредственно содержимым библиотеки, имеют распределенную природу, что означает, что данные могут поступать из различных источников и агрегировать информацию об информационном объекте из различных источников, непосредственно сохраняя данные в самой библиотеке или сохраняя ссылки на идентичные объекты в источниках данных.

Для описания ресурсов, составляющих контент конкретной предметной области, использованы понятия, общие для любой из них, т. е. набор понятий, формирующих описание контента библиотеки, должен быть настолько универсальным, чтобы мог адаптироваться к нуждам конкретной области.

Контент библиотеки тесно связан с тезаурусом, который поддерживает родственные связи различных типов как между самими концептами, так и между концептами и информационными объектами. Это позволяет реализовать гибкий

настраиваемый поиск, результатом которого будет сбалансированный список объектов предметной области. На основе одного и того же тезауруса определяются коллекции ресурсов самых разнообразных типов. Такой подход чрезвычайно полезен для создания раздельных пользовательских коллекций.

Фактически понятия делятся на три категории: первая включает определения понятий контента семантической библиотеки, вторая категория относится к определению понятий, необходимых для поддержки терминов в тезаурусе предметной области, и третья включает определения, необходимые для описания процессов интеграции контента этих ресурсов [14–23]. На основе этих определений описаны основные процессы, такие, как, например, интеграция данных из разных источников, категоризация/классификация, отображение разных моделей данных источников на заданную предметную область, построение классов эквивалентности и т. д.

2. АРХИТЕКТУРА

Рассмотрим формальное описание системы, определяющее ее цели, функции, внешне видимые свойства и интерфейсы. Оно включает также описание компонентов системы и их отношений наряду с принципами, управляющими ее дизайном, функционированием и возможным последующим развитием. Это описание включает программные подсистемы, визуализированные свойства этих подсистем, отношения между подсистемами и ограничения в их использовании. При этом каждая подсистема может состоять из нескольких уровней абстракции, а каждый уровень может иметь свою архитектуру. Ниже приведен список основных подсистем:

- Подсистема описания контента информационной системы;
- Подсистема управления тезаурусом;
- Подсистема автоматизированной обработки и представления данных;
- Подсистема реализации задач интеграции данных;
- Рекомендательная подсистема.

Каждая из этих подсистем отвечает за определенную функциональность и использует свое подмножество понятий из информационной модели.

3. ПОДСИСТЕМА ОПИСАНИЯ КОНТЕНТА

Рассмотрим одну из подсистем, которая определяет основные настройки системы. За универсальность определения контента системы отвечает набор понятий, составляющих информационную модель контента библиотеки Libmeta: *информационный ресурс* и *информационный объект*, которые описывают экземпляры ресурсов. *Информационный ресурс* является основной единицей описания контента библиотеки, а *информационный объект* представляет экземпляры информационных ресурсов. Каждый из них имеет собственный уникальный идентификатор. Фактически семантическое значение *информационного ресурса* является эквивалентным понятию класса *онтологии* с некоторыми ограничениями в его описании. Структура описания информационных объектов определяется понятиями *атрибут* и *набор атрибутов*, которые определяются при описании соответствующего ресурса. Атрибут является элементом описания свойств ресурса, а набор атрибутов определяется как коллекция атрибутов разных видов. Типы атрибутов следующие: *файловый, объектный, числовой, текстовый, строковый*. Помимо определения круга значений атрибута важной характеристикой являются его тип и определение количества его значений. Для описания конкретного информационного ресурса используется понятие *значение атрибута*, которое тесно связано с понятием *атрибут* и является фактически контейнером для хранения конкретных значений *информационного объекта* определенного типа.

Приведенные понятия обеспечивают структурированное описание контента и поддержку его адаптируемости. Такой подход также обеспечивает описание конкретных ресурсов и их объектов в виде RDF-троек и предоставления SPARQL точки доступа для публикации данных в машиночитаемых форматах.

В общем случае конкретная реализация модели контента библиотеки может быть основана на некоторой импортируемой онтологии, классы которой превращаются в ресурсы, свойства могут быть описаны в терминах атрибутов Libmeta, а наборы атрибутов определяют фактически домены свойств онтологий. При построении модели ресурсов библиотеки на основе этой онтологии сохраняются все URI свойств, отношений и классов выбранной онтологии. При необходимости при импортировании выбранной онтологии в систему можно изменить набор понятий, расширив или, наоборот, сократив его средствами системы.

Конечно, такой способ отображения онтологии на понятия системы LibMeta не сохраняет весь возможный перечень ограничений, накладываемых на свойства и классы онтологии изначально, но структурная ее часть сохраняется, что является достаточным для решения задач, определенных в рамках системы.

На рисунке 1 приведены основные понятия, используемые для конструирования описания предметной области в рамках этой подсистемы.

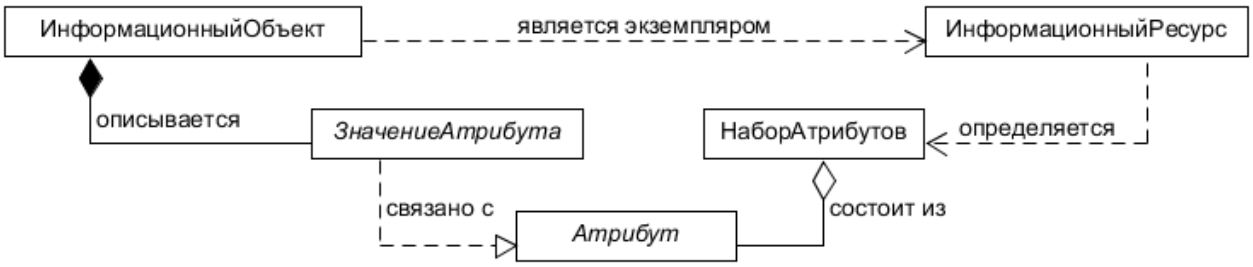


Рис. 1. Основные понятия, используемые для конструирования описания предметной области

При описании *информационных ресурсов* и определении набора их атрибутов важную роль играют *виды атрибутов*, которые формируют структурное описание ресурса. Атрибуты делятся на несколько пересекающихся видов: *поисковые, описательные, административные, идентифицирующие*. При формировании интерфейсов поиска важную роль играют именно *поисковые* атрибуты, которые используются при выполнении атрибутного поиска по типам ресурсов. Результатом такого поиска являются объекты, краткое описание которых представлено пользователю посредством описательных атрибутов.

Фактически в рамках этой подсистемы выполняется первичная настройка конфигурации контента библиотеки и ее интерфейсов под конкретную предметную область. На рисунке 2 изображена последовательность действий пользователя по настройке системы.

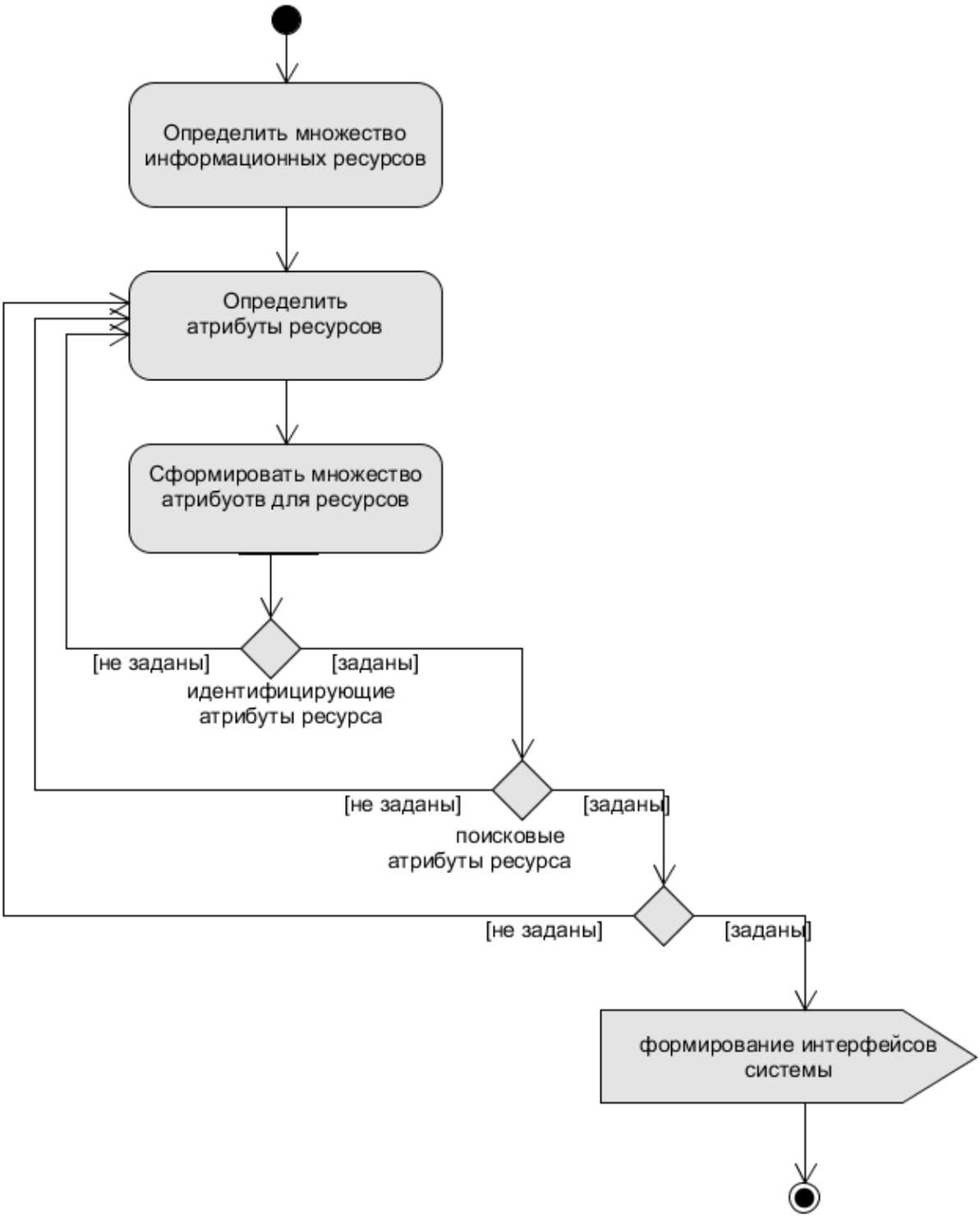


Рис. 2. Последовательность действий пользователя по настройке системы

4. ОСНОВНАЯ ФУНКЦИОНАЛЬНОСТЬ LIBMETA

Эта функциональность такова:

- создание/просмотр/редактирование информационных ресурсов и их структуры;
- создание/просмотр/редактирование информационных объектов и их структуры;
- подключение источников данных;
- загрузка данных из подключенных источников данных, в дальнейшем становящихся частью контента библиотеки;
- создание/просмотр/редактирование структуры тезауруса поддерживаемой предметной области;
- создание/просмотр/редактирование понятий тезауруса
- пакетная загрузка данных, составляющих контент библиотеки;
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным информационным объектам системы;
- атрибутный/семантический/полнотекстовый поиск по источникам данных;
- создание/просмотр/редактирование коллекций информационных объектов;
- формирование онтологии предметной области по описанию структуры информационных ресурсов и тезауруса;
- предоставление данных, составляющих контент системы в машиночитаемом формате;
- выделение связей между информационными объектами и понятиями тезауруса;
- поддержка семантических меток или фолксономии [24 – 25] для описания тематической направленности информационных объектов;
- создание/просмотр/редактирование области интересов пользователя;
- создание рекомендательной системы:
    a. на основе описания интересов пользователя;
    b. на основе рассматриваемого тезауруса предметной области;

- поддержка микротезаурусов пользователей на основе тезауруса предметной области.

Функциональность LibMeta, доступная для всех публичных пользователей:

- просмотр информационных ресурсов и их структуры;
- просмотр информационных объектов и их структуры;
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным ресурсам системы;
- атрибутный и семантический поиск по источникам данных;
- просмотр общедоступных коллекций информационных объектов.

С точки зрения авторизированного пользователя, семантическая библиотека дополнительно обеспечивает ему следующую функциональность:

- определение своего микротезауруса как расширение некоторого узла основного терминологического тезауруса, определенного в системе. Также обеспечивается поддержка создания так называемых *аннотационных онтологий* или *онтологий пользователей (фолксономии),* которые представляют собой коллективный словарь пользователей, составленный в результате процесса проставления ими семантических меток ресурсов;
- определение собственных коллекций информационных объектов;
- организация совместных тематических коллекций для групп пользователей;
- атрибутный и семантический поиск по источникам данных с возможностью сохранения результатов поиска;
- пользователь в роли администратора системы имеет доступ ко всей вышеопределенной функциональности и может воспользоваться дополнительной, доступной только ему функциональностью:
    a. может по запросу пользователей расширять описания типов ресурсов или создавать новые;
    b. может по запросу пользователей включать их объекты ресурсов в общедоступный список объектов;
    c. для групп пользователей доступны возможности редактирования определенных типов ресурсов или таксономий;

d. редактировать группы и роли пользователей и набор доступных им операций;
e. осуществлять редактирование и настройку основного терминологического тезауруса и его связей.

ЗАКЛЮЧЕНИЕ

Представлено описание информационной системы для реализации функциональности семантической библиотеки для некоторой предметной области. В результате эксперты предметной области получают возможность реализации главной задачи библиотеки – *семантического/интеллектуального* конструирования научного пространства знаний для некоторой предметной области, т. е. наделение его семантикой за счет явного выделения интеллектуально значимых связей, поддержки семантической разметки. Основным инструментов конструирования является онтология предметной области, которая позволяет осмысленно структурировать и обеспечить связность между ресурсами, которые включены в научное пространство знаний предметной области, и использование унифицированной терминологической поддержки в виде тезауруса этой предметной области. Для реализации функций открытости научного пространства знаний реализованы возможности интеграции других источников данных и связывания с их данными. Предоставление функциональности для совместной работы над развитием пространства научного знания повышает эффективность проводимых в нем исследований и расширяет возможности по его поддержке в актуальном состоянии.

СПИСОК ЛИТЕРАТУРЫ

1. *Леонова Ю.В., Федотов А.М.* Создание прототипа системы управления информационными ресурсами // Вестник Восточно-Казахстанского гос. техн. университета и журнала «Вычислительные технологии» ИВТ СО РАН. CITech-2018, Усть-Каменогорск, Казахстан. 2018. С. 47–56.
URL: http://www.ict.nsc.ru/jspui/bitstream/ICT/1879/8/Part1_46-55.pdf

2. *Кулагин М.В., Лопатенко А.С*. Научные информационные системы и электронные библиотеки. Потребность в интеграции // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Сборник докладов

Третьей Всероссийской конференции, Петрозаводск, 11–13 сентября, 2001. С. 14-19. URL: http://elib.ict.nsc.ru/ jspui/handle/ICT/1864

3. *Шокин Ю.И., Федотов А.М., Барахнин В.Б.* Проблемы поиска информации. Новосибирск: Наука, 2010. URL: https://lib.nsu.ru/xmlui/handle/ nsu/161

4. *Börner K. et al.* VIVO: A semantic approach to scholarly networking and discovery // Synthesis lectures on the Semantic Web: theory and technology. 2012. Vol. 7. No. 1. P. 1–178. https://doi.org/10.2200/ S00428ED1V01Y201207WBE002

5. *Нгок Н.Б., Тузовский А.Ф.* Обзор подходов семантического поиска // Доклады Томского государственного университета систем управления и радиоэлектроники. 2010. № 2-2 (22). С. 234–238. URL: https://journal.tusur.ru/ru/arhiv/2-2-2010/obzor-podhodov-semanticheskogo-poiska

6. *Апанович З.В., Винокуров П.С., Кислицина Т.А.* Средства визуального анализа информационного наполнения порталов, входящих в облако Linked Open Data // Труды XIII Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», г. Воронеж, 14–17 октября, 2011. С. 113–120. URL: http://ceur-ws.org/Vol-803/paper15.pdf

7. *Оробинская Е.А., Дорошенко А.Ю.* Использование онтологий для автоматической обработки текстов на естественном языке // Вестник Нац. техн. ун-та ХПИ: сб. науч. тр. Темат. вып. «Актуальные проблемы развития украинского общества». Харьков: НТУ ХПИ. 2011. № 30. С. 101–106.
URL: http://repository.kpi.kharkov.ua/handle/KhPI-Press/14950

8. *Добров Б.В., Лукашевич Н.В.* Тезаурус РуТез как ресурс для решения задач информационного поиска // Труды Всероссийской конференции Знания–Онтологии–Теории (ЗОНТ-09), Новосибирск. 2009. Т. 10. С. 250–259
URL: http://ns.math.nsc.ru/conference/zont09/reports/93Dobrov-Lukashevich.pdf

9. *Ngonga Ngomo A.C. et al.* Sorry, i don't speak SPARQL: translating SPARQL queries into natural language // Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013. P. 977–988.
URL: https://dl.acm.org/doi/10.1145/2488388.2488473

10. *Серебряков В.А., Атаева О.М.* Основные понятия формальной модели семантических библиотек и формализация процессов интеграции в ней // Программные продукты и системы. 2015. № 4. С. 180–187.

https://doi.org/10.15827/0236-235x.112.180-187

11. *Атаева О.М., Серебряков В.А.* Персональная открытая семантическая цифровая библиотека LibMeta. Конструирование контента. Интеграция с источниками LOD // Информатика и её применения. 2017. Т. 11. №2. С. 85–100. https://doi.org/10.14357/19922264170210

12. *Атаева О.М.* Информационная модель семантической библиотеки LibMeta // Программные продукты и системы. 2016. № 4. С. 36–44. http://dx.doi.org/10.15827/0236-235X.116.036-044

13. *Атаева О.М., Серебряков В.А.* Онтология цифровой семантической библиотеки LibMeta // Информатика и её применения. 2018. Т. 12. №1. С. 2–10. https://doi.org/10.14357/19922264180101

14. *Ломов П.А., Шишаев М.Г.* Интеграция онтологий с использованием тезауруса для осуществления семантического поиска // Информационные технологии и вычислительные системы. 2009. № 3. С. 49–59. URL: http://mi.mathnet.ru/itvs460

15. *Katsis Y., Papakonstantinou Y.* View-based data integration // Encyclopedia of Database Systems. 2009. P. 3332–3339. https://doi.org/10.1007/978-1-4614-8265-9_1072

16. *Xu L., Embley D.W.* Combining the Best of Global-as-View and Local-as-View for Data Integration // ISTA. 2004. Vol. 48. P. 123–136. URL: https://subs.emis.de/LNI/Proceedings/Proceedings48/GI.Band.48-9.pdf

17. *Когаловский М.Р.* Методы интеграции данных в информационных системах. 2010. URL: http://www.ipr-ras.ru/old_site/articles/kogalov10-05.pdf

18. *Карабач А.Е.* Системы интеграции информации на основе семантических технологий // Наука, техника и образование. 2014. № 2 (2). URL: https://cyberleninka.ru/article/n/sistemy-integratsii-informatsii-na-osnove-semanticheskih-tehnologiy

19. *Lenzerini M.* Data integration: A theoretical perspective // Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002. P. 233–246. http://dx.doi.org/10.1145/543613.543644

20. *Calvanese D., De Giacomo G., Lenzerini M.* Ontology of Integration and Integration of Ontologies // Description Logics. 2001. Vol. 49. No. 10-19. P. 30.

URL: http://www.diag.uniroma1.it/degiacom/papers/2001/CaDL01dl.pdf

21. *Noy N.F.* Semantic integration: a survey of ontology-based approaches // ACM Sigmod Record. 2004. Vol. 33. No. 4. P. 65–70.
http://dx.doi.org/10.1145/1041410.1041421

22. *Zhao L., Ichise R.* Ontology integration for linked data // Journal on Data Semantics. 2014. Vol. 3. No. 4. P. 237–254.
https://doi.org/10.1007/s13740-014-0041-9

23. *Ле Хоай, Тузовский А.Ф.* Разработка семантических электронных библиотек на основе онтологических моделей // Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», г. Ярославль, 14–17 октября, 2013. С. 143–151.
URL: http://ceur-ws.org/Vol-1108/paper18.pdf

24. *Noruzi A.* Folksonomies:(un) controlled vocabulary? // KO KNOWLEDGE ORGANIZATION. 2006. Vol. 33. No. 4. P. 199–203.
URL: https://core.ac.uk/download/pdf/11882508.pdf

25. *Gruber T*. Ontology of folksonomy: A mash-up of apples and oranges // International Journal on Semantic Web and Information Systems (IJSWIS). 2007. Vol. 3. No. 1. P. 1–11. URL: https://tomgruber.org/writing/ontology-of-folksonomy.htm

SEMANTIC LIBRARY AS A TOOL OF DEFINING A SCIENTIFIC SUBJECT AREA

O. M. Ataeva[1] [0000-0003-0367-5575], V. A. Serebriakov[2] [0000-0003-1423-621X]

[1, 2] Dorodnicyn Computing Center FRC CSC of RAS

[1] oli@ultimeta.ru

[2] serebr@ultimeta.ru

Abstract

The paper considers an information system designed to represent a subject area related to science and its features. Highlighted general concepts for formal descriptions of such a subject area in the knowledge base of the semantic library. The peculiarity of these areas is that the data structure is subject to frequent changes. Therefore, the means of organizing knowledge, which is a semantic library, should be sufficiently universal and not require deep technical knowledge. The paper describes the functionality of the system and its use. For each area, the set of resources can differ both in format and in the set of the resources themselves. The set of concepts that form the description of the library's content should be so universal that it can be adapted to the needs of a particular area. Three levels of metadata are used to represent the data.

Keywords: *semantic library, ontology, knowledge representation*

REFERENCES

1. *Leonova Yu.V., Fedotov A.M.* Sozdanie prototipa sistemy upravleniya informacionnymi resursami //Vestnik Vostochno-Kazahstanskogo gos. Tekhn. Universiteta i zhurnala Vychislitel'nye tekhnologii IVT SO RAN.–CITech-2018, Ust'-Kamenogorsk, Kazahstan. 2018. S. 47–56.
URL: http://www.ict.nsc.ru/ jspui/bitstream/ICT/1879/8/Part1_46-55.pdf

2. *Kulagin M.V., Lopatenko A.S.* Nauchnye informacionnye sistemy i elektronnye biblioteki. Potrebnost' v integracii // Digital Libraries: Advanced Methods and Technologies, Digital Collections, Petrozavodsk, September 11–13, 2001. S. 14–19. URL: http://elib.ict.nsc.ru/jspui/handle/ICT/1864

3. *Shokin Yu.I., Fedotov A.M., Barahnin V.B.* Problemy poiska informacii. Novosibirsk: Nauka, 2010. URL: https://lib.nsu.ru/xmlui/handle/nsu/161

4. *Börner K. et al.* VIVO: A semantic approach to scholarly networking and discovery // Synthesis lectures on the Semantic Web: theory and technology. 2012. Vol. 7. No. 1. P. 1–178. https://doi.org/10.2200/ S00428ED1V01Y201207WBE002

5. *Ngok N.B., Tuzovskij A.F.* Obzor podhodov semanticheskogo poiska // Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki. 2010. № 2-2 (22). S. 234–238.
URL: https://journal.tusur.ru/ ru/arhiv/2-2-2010/obzor-podhodov-semanticheskogo-poiska

6. *Apanovich Z.V., Vinokurov P.S., Kislicina T.A.* Sredstva vizual'nogo analiza informacionnogo napolneniya portalov, vhodyashchih v oblako Linked Open Data // Proceedings of the 13th All-Russian Scientific Conference ”Digital libraries: Advanced Methods and Technologies, Digital Collections” Voronezh, Russia, October 19–22, 2011. S. 113–120. URL: http://ceur-ws.org/Vol-803/paper15.pdf

7. *Orobinskaya E.A., Doroshenko A.Yu.* Ispol'zovanie ontologij dlya avtomaticheskoj obrabotki tekstov na estestvennom yazyke // Vestnik Nac. tekhn. un-ta HPI: sb. nauch. tr. Temat. vyp. “Aktual'nye problemy razvitiya ukrainskogo obshchestva”. Har'kov: NTU HPI. 2011. № 30. S. 101–106.
URL: http://repository.kpi.kharkov.ua/handle/KhPI-Press/14950

8. *Dobrov B.V., Lukashevich N.V.* Tezaurus RuTez kak resurs dlya resheniya zadach informacionnogo poiska // Trudy Vserossijskoj Konferencii Znaniya-Ontologii-Teorii (ZONT-09), Novosibirsk. 2009. T. 10. S. 250–259
URL: http://ns.math.nsc.ru/conference/zont09/reports/93Dobrov-Lukashevich.pdf

9. *Ngonga Ngomo A.C. et al.* Sorry, i don't speak SPARQL: translating SPARQL queries into natural language // Proceedings of the 22nd international conference on World Wide Web. ACM, 2013. P. 977–988.
URL: https://dl.acm.org/doi/10.1145/2488388.2488473

10. *Serebryakov V.A., Ataeva O.M.* Osnovnye ponyatiya formal'noj modeli semanticheskih bibliotek i formalizaciya processov integracii v nej // Programmnye produkty i sistemy. 2015. № 4. S. 180–187.
https://doi.org/10.15827/0236-235x.112.180-187

11. *Ataeva O.M., Serebryakov V.A.* Personal'naya otkrytaya semanticheskaya cifrovaya biblioteka LibMeta. Konstruirovanie kontenta. Integraciya s istochnikami LOD // Informatika i eyo primeneniya. 2017. T. 11, No. 2. S. 85–100. https://doi.org/10.14357/19922264170210

12. *Ataeva O.M.* Informacionnaya model' semanticheskoj biblioteki LibMeta // Programmnye produkty i sistemy. 2016. № 4. S. 36–44. http://dx.doi.org/10.15827/0236-235X.116.036-044

13. *Ataeva O.M., Serebryakov V.A*. Ontologiya cifrovoj semanticheskoj biblioteki LibMeta // Informatika i eyo primeneniya. 2018. T. 12, No. 1. S. 2–10. https://doi.org/10.14357/19922264180101

14. *Lomov P.A., Shishaev M.G.* Integraciya ontologij s ispol'zovaniem tezaurusa dlya osushchestvleniya semanticheskogo poiska // Informacionnye tekhnologii i vychislitel'nye sistemy. 2009. № 3. S. 49–59. URL: http://mi.mathnet.ru/itvs460

15. *Katsis Y., Papakonstantinou Y.* View-based data integration // Encyclopedia of Database Systems. 2009. P. 3332–3339. https://doi.org/10.1007/978-1-4614-8265-9_1072

16. *Xu L., Embley D.W.* Combining the Best of Global-as-View and Local-as-View for Data Integration // ISTA. 2004. Vol. 48. P. 123–136. URL: https://subs.emis.de/LNI/Proceedings/Proceedings48/GI.Band.48-9.pdf

17. *Kogalovskij M.R.* Metody integracii dannyh v informacionnyh sistemah. 2010. URL: http://www.ipr-ras.ru/old_site/articles/kogalov10-05.pdf

18. *Karabach A.E.* Sistemy integracii informacii na osnove semanticheskih tekhnologij // Nauka, tekhnika i obrazovanie. 2014. № 2 (2). URL: https://cyberleninka.ru/article/n/sistemy-integratsii-informatsii-na-osnove-semanticheskih-tehnologiy

19. *Lenzerini M.* Data integration: A theoretical perspective // Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002. P. 233–246. http://dx.doi.org/10.1145/543613.543644

20. *Calvanese D., De Giacomo G., Lenzerini M.* Ontology of Integration and Integration of Ontologies // Description Logics. 2001. Vol. 49. No. 10-19. P. 30. URL: http://www.diag.uniroma1.it/degiacom/papers/2001/CaDL01dl.pdf

21. *Noy N.F.* Semantic integration: a survey of ontology-based approaches // ACM Sigmod Record. 2004. Vol. 33. No. 4. P. 65–70.
http://dx.doi.org/10.1145/1041410.1041421

22. *Zhao L., Ichise R*. Ontology integration for linked data // Journal on Data Semantics. 2014. Vol. 3. No. 4. P. 237–254.
https://doi.org/10.1007/s13740-014-0041-9

23. *Le Hoaj, Tuzovskij A.F.* Razrabotka semanticheskih elektronnyh bibliotek na osnove ontologicheskih modelej // 15th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies, Digital Collections”. Yaroslavl, Russia, October 14–17, 2013. S. 143–151. URL: http://ceur-ws.org/Vol-1108/paper18.pdf

24. *Noruzi A.* Folksonomies:(un) controlled vocabulary? //KO KNOWLEDGE ORGANIZATION. 2006. Vol. 33. No. 4. P. 199–203.
URL: https://core.ac.uk/download/pdf/11882508.pdf

25. *Gruber T.* Ontology of folksonomy: A mash-up of apples and oranges // International Journal on Semantic Web and Information Systems (IJSWIS). 2007. Vol. 3. No. 1. P. 1–11. URL: https://tomgruber.org/writing/ontology-of-folksonomy.htm

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат техн. наук, специалист в области системного программирования и баз данных.

Olga Muratovna ATAEVA – researcher of the of Dorodnicyn computing center FRC SCS RAS, PhD

email: oli@ultimeta.ru

ORCID: 0000-0003-0367-5575

СЕРЕБРЯКОВ Владимир Алексеевич – специалист в области теории формальных языков и её приложений, доктор физ.-мат. наук, профессор, зав. Отделом Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН. Руководитель и участник разработки ряда известных программных проектов, в частности, ИСИР и ИСИР РАН, «Научный портал РАН».

Vladimir Alekseevich SEREBRIAKOV – expert in the field of theory of formal languages and its applications, doctor of sciences, professor, head of Dorodnicyn computing center FRC SCS RAS department. Head and participant in the development of a number of well-known program projects, in particular, ISIR and ISIR RAS, Scientific portal RAS.

email: serebr@ultimeta.ru

ORCID: 0000-0003-1423-621X

Материал поступил в редакцию 2 ноября 2021 года

УДК 004.65, 005, 001.5

О МОДЕЛИ ПОИСКА СИНОНИМОВ

О. М. Атаева[1] [0000-0003-0367-5575], В. А. Серебряков[2] [0000-0003-1423-621X],

Н. П. Тучкова[3] [0000-0001-5357-9640]

[1,2,3]*Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, г. Москва*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Аннотация

Рассмотрена задача нахождения наиболее релевантных документов в результате расширенного и уточненного запроса. Для ее решения предложены модель поиска и механизм предварительной обработки текста, а также совместное использование поисковой системы и модели, построенной на основе индекса с помощью алгоритмов word2vec для генерации расширенного запроса с синонимами и уточнения результатов поиска на основе подбора похожих документов в цифровой семантической библиотеке. В работе исследуется построение векторного представления документов применительно к массиву данных цифровой семантической библиотеки LibMeta. Решалась задача обогащения пользовательских запросов синонимами. При построении модели поиска совместно с алгоритмами word2vec использован подход «сначала индексация, затем обучение», что позволяет получить более точные результаты поиска. Обучение модели проводилось на базе контента библиотеки для предметной области «Математика». Приведены примеры расширенного запроса с использованием синонимов.

Ключевые слова*: модель поиска, алгоритм word2vec, синонимы, информационный запрос, расширение запроса.*

ВВЕДЕНИЕ

Задача поиска синонимов и схожих (похожих, аналогичных, «similar») документов изучается достаточно давно [1, 2]. Известен такой алгоритм, как латентный алгоритм Дирихле (LDA model) [3], основанный на статистической модели Байеса. Наибольшую популярность в свое время обрели алгоритмы векторного представления текстов серии tf-idf [3]. Схема tf-idf сокращает документы произ-

вольной длины до списков фиксированной длины и числа слов, не отражая семантическую структуру внутри документа. LDA-алгоритм использует тематическую привязку слов и тем самым способствует учету семантических связей документов и внутри документов.

Исследования, представленные в [1–3] и других известных работах, позволяют говорить о том, что неверная информация, получаемая по запросу, как правило, является результатом использования в базах данных ошибочных семантических связей, т. е. на этапе предварительной обработки данных не учитываются некоторые семантические связи терминов в тексте [4, 5]. Для научных работ, помещенных в поисковый индекс без учета семантических связей, специфичных для каждой предметной области, это означает, что они могут быть не найдены специалистами и не процитированы. В этом контексте особую роль играют предварительная обработка данных и применение современных подходов к решению задачи поиска достоверной научной информации на основе машинного обучения [6, 7]. Исходные данные, обретя определенную структуру в процессе обработки, могут использоваться уже в качестве источника достоверных знаний [8].

В настоящей работе изучается проблема поиска документов из контента семантической библиотеки, наиболее близких к информационному запросу. Для выбора релевантных документов использована процедура нахождения близких по тематике, схожих документов, которые можно получить в результате расширения запроса синонимами. Целью исследований является построение модели поиска, которая будет удовлетворять условиям наиболее полного удовлетворения *поисковой потребности* пользователя на имеющемся наборе документов сематической библиотеки.

Версию модели, построенную на поисковом индексе LibMeta [9] с помощью алгоритмов word2vec [10–12], далее будем сокращенно называть wsgMath, как принято ранее в работе [13]. Этот подход к совместному использованию индекса поисковой системы и нейросети позволяет получать релевантные модели и функции ранжирования, которые хорошо адаптируются к базовым данным.

В данной работе ставится задача связать модель поиска с предметной областью, границы которой очерчены ее тезаурусом и классификаторами. Таким об-

разом, поиск по контенту библиотеки, поиск новых терминов и новых семантических связей между терминами предметной области становится более осмысленным и точным.

Структура работы следующая: в первом разделе изложены принципы построения модели поиска; во второй части описано построение расширенных запросов на основе векторного представления текстов; далее приведены примеры, заключение и список цитирования.

1. ОСОБЕННОСТИ МОДЕЛИ ПОИСКА

Необходимо отметить, что существуют подходы с использованием моделей, построенных с помощью алгоритмов, обученных на общедоступных наборах данных. Как правило, эти наборы не включают специальные предметные области и не учитывают их терминологическую специфику. Расширение поисковых запросов [14–16] синонимами также требует предварительно составленных словарей синонимов. Можно использовать такие ресурсы, как WordNet[1] или RuWordNet[2], но основная проблема заключается в том, что синонимы из предварительно составленных словарей не привязаны к индексируемым данным, и их использование не улучшает результаты. Поэтому была выбрана модель использования поискового индекса совместно с векторной моделью индексируемых данных из предметной области «Математика», построенная с помощью алгоритма word2vec и обученная на математической предметной области.

Для реализации такого подхода выбрана последовательная схема работы с данными, а именно:

- определяется предметная область;
- определяется словарь, соответствующий предметной области;
- на основе связей между терминами словаря выявляются связи между документами, статьями, авторами и т. д.

Поиск и отслеживание связей выполняются следующим образом:

- производится предварительная обработка текстов;
- применяются алгоритмы машинного обучения для обработки и анализа текстов;

[1] https://wordnet.princeton.edu/
[2] https://ruwordnet.ru/ru

- применяются векторные представления документов и запросов для *ранжирования* результатов поиска.

Такой подход увеличивает вероятность того, что система будет более *точно реагировать* на информационную потребность пользователя и выдавать более *релевантные* ответы.

В процессе исследований была определена архитектура подсистемы поиска семантической библиотеки, которая состоит из:

- компонента предварительной обработки текста для представления документов в формате, пригодном для поиска, эффективной загрузки и хранения данных и обеспечения быстрого доступа к ним;
- компонента формирования полнотекстового индекса документов;
- компонента построения векторной модели на основе индекса с помощью алгоритмов word2vec;
- компонента обработки запросов и представления их в формате, удобном для выражения информационных потребностей пользователя на естественном языке, обогащенных синонимами из предметной области;
- компонента формирования результатов на основе оценок соответствия документа запросу, с использованием контента библиотеки.

Особенность этого подхода – в гибком сочетании всех инструментов библиотеки, таких как тезаурусы, классификаторы и энциклопедия для поиска синонимов и схожих документов, а также оценки результатов на их основе.

На рис. 1 представлены основные шаги формирования поисковой выдачи в библиотеке LibMeta. Строка запроса, поступающая из интерфейса полнотекстового поиска, проходит через блок *Анализатор*. В нем строка разбивается на слова, затем проводятся их анализ и преобразование. Из модели wsgMath извлекаются и фильтруются синонимы к словам, что позволяет сформировать расширенный запрос, с помощью которого из полнотекстового индекса извлекаются соответствующие документы.

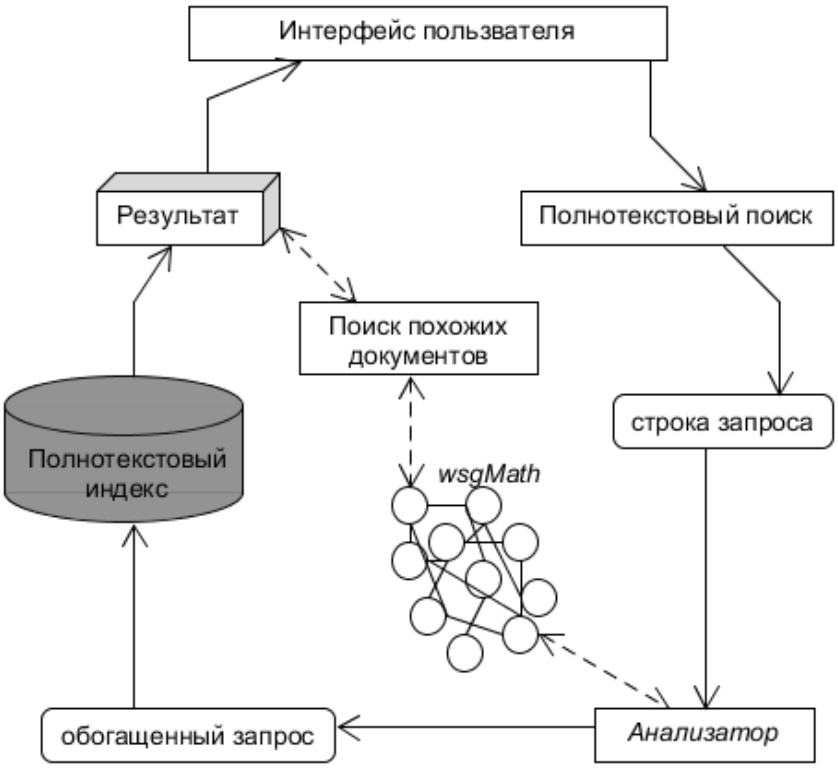


Рис. 1. Схема использования поисковой системы и нейросетевой модели.

Применение расширенной версии word2vec (doc2vec или paragraph2vec, в разных источниках) позволяет ввести дополнительный элемент, такой как *Метка фрагмента текста* или всего документа, и, основываясь на векторах этих меток, подбирать похожие документы не только по точному совпадению ключевых слов или терминов, но и, *основываясь на контексте* отдельных фрагментов или всего документа.

Замечание 1. Метка фрагмента текста используется для выдачи близких по смыслу документов, которые не попадают в поисковую выдачу, но могут представлять интерес для пользователя.

2. ПОСТРОЕНИЕ И ОБУЧЕНИЕ МОДЕЛИ ВЕКТОРНОГО ПРОСТРАНСТВА ПОИСКОВОЙ СИСТЕМЫ

2.1. Предобработка статей

Одним из необходимых этапов подготовки данных к их загрузке в определенных текстовых форматах в уже подготовленную инфраструктуру данных являются *предобработка* и *очистка* этих данных.

В нашем случае данные предоставлялись файлами в формате TeX, оформленными с разными стилями и метакомандами, т. е. сначала было необходимо заменить все авторские тэги на стандартные, очистить документы от специальных символов и неизвестных тэгов. При этом совсем избежать ручной обработки не удалось, но удалось свести ее к минимуму.

На рис. 2 приведен пример просмотра терминов в системе LiMeta, связи которых сформировались на этапах *предобработки* и *очистки* данных.

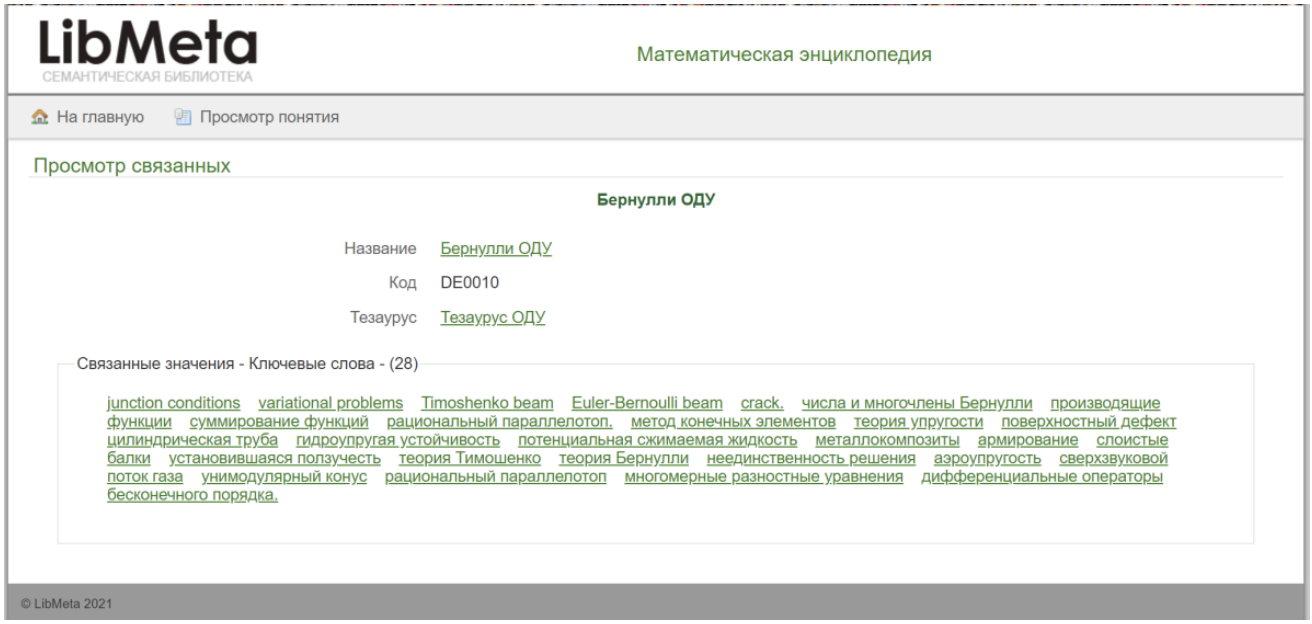


Рис. 2. Пример реализации модуля предобработки в LibMeta.

Модуль предварительной обработки выполнен на языке программирования Python вместе с интеграцией open-source библиотеки TexSoup версии 2015 года и разбит на следующие блоки:

- очистка документа;
- преобразование статьи в древовидное представление;

- обработка всех узлов дерева, запись исправленного документа.

На рис. 3 представлены основные этапы предварительной обработки текстов.

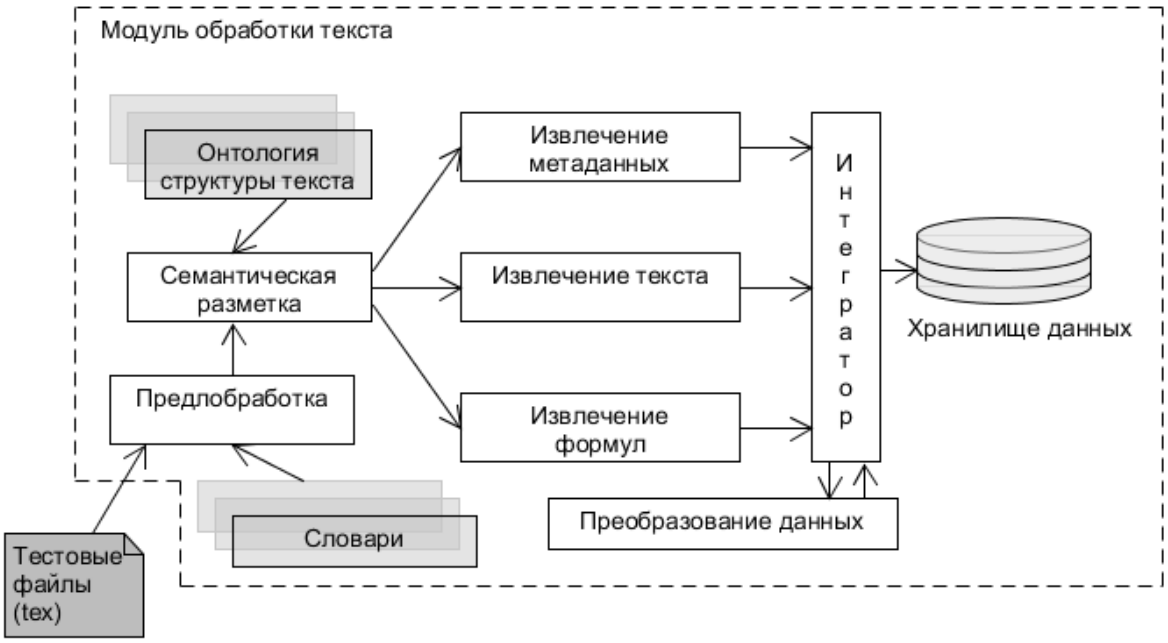


Рис. 3. Схема предварительной обработки текстов.

2.2. Построение индекса и обучение word2vec

Целью обучения модели wsgMath было получение синонимов, которыми можно было бы расширить поисковый запрос и получить неучтенные ранее семантические связи для дальнейшего извлечения информации, релевантной запросу.

Используемая в работе модель поиска реализует интеграцию модели, построенной на основе нейронной сети word2vec, и полнотекстового индекса. Интеграция нейронной сети и индекса может выполняться следующими способами:

- сначала обучение на корпусе текстов, затем индексация текстов и совместное использование обученной модели и индекса при поиске;
- сначала индексация, затем обучение на индексированных данных и совместное использование при поиске;
- сначала обучение, затем извлечение/создание полезных ресурсов обученной сетью, а потом индексация всех ресурсов, и новых, и исходных.

В библиотеке LibMeta использован подход «сначала индексация, затем обучение». Также изучены проблемы, как предоставить больше точных результатов на основе расширенных запросов [14–16] и как дать пользователям более «умные рекомендации» для дальнейшего поиска на основе найденных документов из предметной области в LibMeta.

На основе массива предварительно обработанных статей был построен *индекс полных текстов* на базе библиотеки поиска с открытым исходным кодом Apache Lucene[3], написанный на Java[4]. Этот индекс используется подсистемой полнотекстового поиска библиотеки, и он же использовался для обучения алгоритма и извлечения контекстов.

Слова в контексте, близкие к рассматриваемому, трактуются как синонимы (контекстно-зависимые синонимы, в данном случае) и анализируются. Осуществляется их лексико-семантический анализ, т. е. определяются части речи, словоформы и собственные связи, в том числе со словарями и тезаурусами предметной области. На основе модели wsgMath численно оценивается близость контекстно-зависимых синонимов. С помощью этих оценок выбираются кандидаты, а затем из них – наилучшие с наибольшими оценками. Для дальнейшего сравнения могут быть использованы коды классификаторов, если с ними связаны выбранные слова.

В Таблице 1 приведены примеры слов со связями между словами (в первой строке стоит главное слово, в столбцах ниже – выявленные).

[3] https://lucene.apache.org/

[4] https://www.java.com/ru/

Таблица 1. Все связи слова

простран-ство	***краевой***	***задача***	***краевой***	***напряжение***	***оста-точный***
оператор	граничный	решение	интегральный	деформация	концен-тратор
множество	интегральный	уравнение	дифференциаль-ный	упрочнение	упрочне-ние
функция		условие	уравнение	пластический	усталость
		система			
		функция			

Таблица 1 содержит примеры синонимов, определенных по частям речи и коэффициенту близости слов контента библиотеки на основе модели wsgMath.

3. ПРИМЕРЫ (ЗАПРОС С СИНОНИМАМИ)

3.1. Расширение синонимами

Рассмотрим термин «*краевая задача*», состоящий из двух слов – «*задача» и «краевая*», каждое из которых имеет собственные синонимы, представленные в Таблице 1.

В контекст термина как одной единицы попадают такие синонимы, как *[решение, уравнение, условие, система, тип, функция, область, работа].* При этом сам термин «*краевая задача»* имеет следующие *синонимы: «граничное уравнение», «граничное условие», «граничная функция», «интегральная функция»,* которые были определены на основе высоких оценок близости следующих пар синонимов и в соответствии с паттерном «прилагательное + *термин(сущ)»* на модели wsgMath*:*

sim (задача, решение) = 0.91

sim (задача, уравнение) = 0.86

sim (задача, условие) = 0.82

sim (задача, система) = 0.79

sim (задача, функция) = 0.73

Замечание 2. При построении синонимичных терминов не используются синонимы слов, определенных как именованная сущность на основе словаря, который включает в себя список персон, встречающихся в математической энциклопедии. Но при этом отметим, что в множество синонимов {*коши* (*Коши*)} попало слово *риман* (*Риман*), а в множество синонимов {*лаплас* (*Лаплас*)} попало слово *фурье* (*Фурье*).

Таким образом, были выбраны работы с высокими оценками близости синонимов.

1. О положительном радиально-симметрическом решении задачи дирихле для одного нелинейного уравнения и численном методе его получения

score = 0.90484273

2. О корректности краевой задачи на прямой для трех аналитических функций

score = 0.902505

3. Проекционные процедуры нелокального улучшения линейно управляемых процессов

score = 0.8816618

4. Краевая задача для частного вида уравнения эйлера-дарбу с интегральными условиями и специальными условиями сопряжения на характеристике

score = 0.846388

5. Теорема валле-пуссена для одного класса функционально-дифференциальных уравнений

score = 0.84127665

3.2. Поиск похожих документов

Рассмотрим пример использования элемента *Метка фрагмента текста* для процесса *ранжирования* документов на основе модели wsgMath при поиске схожих документов.

Когда документ поступает в систему, то извлекается его текущее векторное представление, выполняется поиск и возвращаются *метки ближайших документов, косинусное расстояние которых превышает некоторый порог, определенный экспериментально как 0,6.*

Далее можно также использовать для сравнения коды классификаторов как один из вариантов оценки похожих документов. При этом возможны различные варианты, связанные с наличием или отсутствием кодов классификации MSC[5] и УДК[6] у исходных документов:

- Документы, поступающие в систему, размечены кодами классификаторов MSC и УДК. В этом случае при выявлении документов, косинусное расстояние у которых превысило заданный порог, можно сравнивать коды классификаторов и устанавливать соответствие MSC и УДК. Если коды УДК отличаются у схожих документов, то можно их указать как смежные предметные области (приложения результатов, междисциплинарные исследования и пр.).
- Документы не снабжены кодами, но ключевые слова соответствуют предметной области, и в словаре (тезаурусе, энциклопедии) есть коды классификаторов. В этом случае сравниваются коды ключевых слов, и документам приписываются соответствующие коды.

На рис. 4 приведен пример соответствия кодов классификаторов, полученных на основе контента LibMeta и процедуры выявления синонимов. В данном случае было выявлено, что коду УДК 515.128 соответствуют такие коды MSC, как 54E20, 54E40, 54D65 и т. д.

[5] https://cran.r-project.org/web/classifications/MSC.html

[6] https://teacode.com/online/udc/

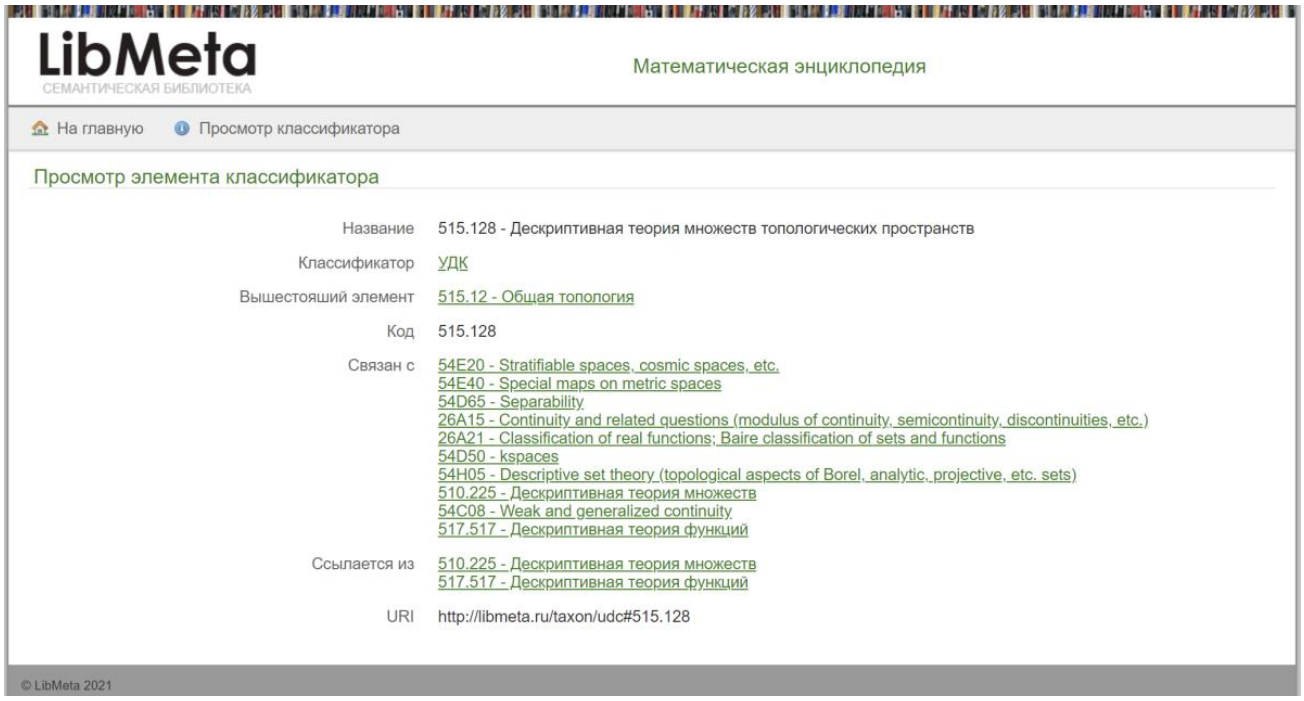


Рис. 4. Пример соответствия кодов классификаторов MSC и УДК.

ЗАКЛЮЧЕНИЕ

В представленном исследовании получены следующие основные результаты.

Показано, что предварительная обработка входных массивов данных (текстов научных статей) позволяет учесть в дальнейшем дополнительные семантические связи и улучшить качество поиска.

Использование механизма интеграции нейронной сети и индекса позволяет реализовать варианты поисковой модели для получения релевантных документов с заданной точностью.

Совместное использование индекса поисковой системы и нейросети позволяет получать релевантные модели и функции ранжирования, которые хорошо адаптируются к базовым данным.

Предложенная модель поиска позволяет также устанавливать соответствие кодов классификаторов для близких документов, находить синонимы при контекстном сравнении и ранжировать документы на основе метки фрагмента.

Выявлены проблемы для дальнейшего изучения – развитие механизма оценки качества поиска с использованием различных метрик, использование английских и русских синонимов для обогащения запроса и улучшения качества поиска, оценки скорости обучения модели. Решение этих проблем вытекает из проделанных исследований, которые позволяют сформулировать конкретные задачи для улучшения качества поиска. Это составление словарей синонимов предметной области, связанных с классификаторами и эталонных документов, связанных с терминами тезауруса предметной области. Такие ресурсы могут позволить в дальнейшем улучшить качество поиска на основе алгоритмов машинного обучения.

Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований, проект № 20-07-00324, и в рамках темы Министерства науки и высшего образования РФ «Математические методы анализа данных и прогнозирования».

СПИСОК ЛИТЕРАТУРЫ

1. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval. ACM Press, New York, 1999. 518 p.

2. *Salton G.* Introduction to Modern Information Retrieval. McGraw-Hill, 1983, 513 p.

3. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. V. 3. P. 993–1022.

4. *Furnas G.W., Landauer T.K., Gomez L.M., Dumais S.T.* The vocabulary problem in human-system communication // Commun. ACM. 1987. V. 30, No. 11 P. 964–971.

5. *Biswas G., Bezdek J., Oakman R.L.* A knowledge-based approach to online document retrieval system design. In Proc. ACM SIGART Int. Symp. Methodol. Intell. Syst. 1986. P. 112–120.

6. *Мак-Каллок У.С., Питтс В.* Логическое исчисление идей, относящихся к нервной активности // Автоматы. Под ред. К. Э. Шеннона и Дж. Маккарти. М.: Изд-во иностр. лит., 1956. С. 363–384 (Перевод англоязычной статьи 1943 г.).

7. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. URL: http://www.machinelearning.ru/ (доступно 26.10.2021)

8. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.

9. *Атаева О.М., Серебряков В.А.* Онтология цифровой семантической библиотеки LibMeta // Информатика и её применения. 2018. Т. 12. № 1. С. 2–10.

10. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.

11. *Mikolov T., Yih W.T., Zweig C.* Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.

12. *Le Q., Mikolov T.* Distributed Representations of Sentences and Document // International Conference on Machine Learning. 2014. P. 1188–1196.

13. *Ataeva O.M., Sererbryakov V.A., Tuchkova N.P.* Using Applied Ontology to Saturate Semantic Relations // Lobachevskii Journal of Mathematics. 2021. V. 42. No. 8. P. 1776–1785.

14. *Voorhees E.M.* Query expansion using lexical-semantic relations. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Dublin, Ireland, 1994.

15. *Buckley C., Salton G., Allan J., Singhal A.* Automatic query expansion using SMART: TREC 3, presented at the 3rd Text Retr. Conf. (TREC), 1995.

16. *Efthimiadis E.N.* Query expansion // Annu. Rev. Inf. Sci. Technol. 1996. V. 31. No. 5. P. 121–187.

ON THE SYNONYM SEARCH MODEL

O. M. Ataeva[1] [0000-0003-0367-5575], V. A. Serebriakov[2] [0000-0003-1423-621X],

N. P. Tuchkova[3] [0000-0001-5357-9640]

[1,2,3]*Dorodnicyn Computing Centre FRC CSC RAS, Moscow*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Abstract

The problem of finding the most relevant documents as a result of an extended and refined query is considered. For this, a search model and a text preprocessing mechanism are proposed, as well as the joint use of a search engine and a neural network model built on the basis of an index using word2vec algorithms to generate an

extended query with synonyms and refine search results based on a selection of similar documents in a digital semantic library. The paper investigates the construction of a vector representation of documents based on paragraphs in relation to the data array of the digital semantic library LibMeta. Each piece of text is labeled. Both the whole document and its separate parts can be marked. The problem of enriching user queries with synonyms was solved, then when building a search model together with word2vec algorithms, an approach of "indexing first, then training" was used to cover more information and give more accurate search results. The model was trained on the basis of the library's mathematical content. Examples of training, extended query and search quality assessment using training and synonyms are given.

Keywords: *search model, word2vec algorithm, synonyms, information query, query extension*

REFERENCES

1. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval. ACM Press, New York, 1999. 518 p.

2. *Salton G.* Introduction to Modern Information Retrieval. McGraw-Hill, 1983, 513 p.

3. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. V. 3. P. 993–1022.

4. *Furnas G.W., Landauer T.K., Gomez L.M., Dumais S.T*. The vocabulary problem in human-system communication // Commun. ACM. 1987. V. 30. No. 11. P. 964–971.

5. *Biswas G., Bezdek J., Oakman R.L.* A knowledge-based approach to online document retrieval system design. In Proc. ACM SIGART Int. Symp. Methodol. Intell. Syst. 1986. P. 112–120.

6. *Mak Kallok U.S., Pitts V.* Logicheskoe ischislenie idej otnosyashchihsya k nervnoj aktivnosti Avtomaty Ed. Shennon i Dzh Makkarti M: Izd-vo inostr. Lit. 1956. S. 363–384. Perevod anglijskoj stati 1943 g.

7. Professionalnyj informacionno analiticheskij resurs posvyashchennyj mashin nomu obucheniyu raspoznavaniyu obrazov i intellektualnomu analizu dannyh URL: http://www.machinelearning.ru/ (access 26.10.2021)

8. *Gavrilova T.A., Horoshevskij V.F.* Bazy znanij intellektualnyh sistem SPb: Piter. 2000, 384 s.

9. *Ataeva O.M., Serebryakov V.A.* Ontologiya cifrovoj semanticheskoj biblioteki LibMeta //Informatika i eyo primeneniya. 2018. V. 12. No. 1. S. 2–10.

10. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.

11. *Mikolov T., Yih W.T., Zweig C.* Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.

12. *Le Q., Mikolov T.* Distributed Representations of Sentences and Document // International Conference on Machine Learning. 2014. P. 1188–1196.

13. *Ataeva O.M., Sererbryakov V.A., Tuchkova N.P.* Using Applied Ontology to Saturate Semantic Relations // Lobachevskij Journal of Mathematics. 2021. V. 42. No. 8. P. 1776–1785.

14. *Voorhees E.M.* Query expansion using lexical-semantic relations. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Dublin, Ireland, 1994.

15. *Buckley C., Salton G., Allan J., Singhal A.* Automatic query expansion using SMART: TREC 3, presented at the 3rd Text Retr. Conf. (TREC), 1995.

16. *Efthimiadis E.N.* Query expansion // Annu. Rev. Inf. Sci. Technol. 1996. V. 31. No. 5. P. 121–187.

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат техн. наук, специалист в области системного программирования и баз данных.

Olga Muratovna ATAEVA – researcher of the of Dorodnicyn computing center FRC SCS RAS, PhD, expert in the field of system programming and databases.

email: oli@ultimeta.ru
ORCID: 0000-0003-0367-5575

СЕРЕБРЯКОВ Владимир Алексеевич – специалист в области теории формальных языков и её приложений, доктор физ.-мат. наук, профессор, зав. отделом Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН. Руководитель и участник разработки ряда известных программных проектов, в частности, ИСИР и ИСИР РАН, «Научный портал РАН».

Vladimir Alekseevich SEREBRIAKOV – expert in the field of theory of formal languages and its applications, doctor of sciences, professor, head of Dorodnicyn computing center FRC SCS RAS department. Head and participant in the development of a number of well-known program projects, in particular, ISIR and ISIR RAS, Scientific portal RAS.

email: serebr@ultimeta.ru
ORCID: 0000-0003-1423-621X

ТУЧКОВА Наталия Павловна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-мат. наук, окончила ВМиК МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – senior researcher of Dorodnicyn computing center FRC SCS RAS, PhD in physics with a math degree, graduated from CS Faculty of Lomonosov MSU. The expert in the field of algorithmic languages and information technologies.

email: natalia_tuchkova@mail.ru
ORCID: 0000-0001-5357-9640

Материал поступил в редакцию 05 ноября 2021 года

УДК 004.4

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ WIKIDATA ДЛЯ ФОРМИРОВАНИЯ МЕТАДАННЫХ ДОКУМЕНТОВ ЭЛЕКТРОННЫХ МАТЕМАТИЧЕСКИХ КОЛЛЕКЦИЙ

П. О. Гафурова[1] [0000-0002-1544-155X], **А. М. Елизаров**[2] [0000-0003-2546-6897], **Е. К. Липачёв**[3] [0000-0001-7789-2332]

[1–3] *Институт информационных технологий и интеллектуальных систем Казанского федерального университета, ул. Кремлевская, 35, г. Казань, 420008*

[1]pogafurova@gmail.com, [2]amelizarov@gmail.com, [3]elipachev@gmail.com

Аннотация

Представлены методы создания цифровых математических коллекций, включающих неструктурированные наборы документов. Эти наборы содержат материалы сборников научных конференций, а также статьи из архивов математических журналов «доцифрового» периода.

Формирование обязательного набора метаданных названных документов произведено с помощью программных инструментов фабрики метаданных цифровой математической библиотеки Lobachevskii DML. Для уточнения и пополнения наборов метаданных документов цифровых коллекций использованы методы извлечения знаний из Wikidata.

Разработана система SPARQL-запросов для поиска в Wikidata информации о документах электронных коллекций и их авторах. Обозначен набор сущностей Wikidata, определяющих признаки поиска, а также последующую фильтрацию полученных результатов.

Предложены методы уточнения и дополнения библиографических ссылок, приведенных в статьях. При формировании метаданных документов ретро-коллекций произведен поиск в Wikidata сведений о годах жизни авторов статей, а также URL веб-страниц с информацией о статьях и их авторах. Приведены результаты формирования нескольких новых электронных коллекций цифровой библиотеки Lobachevskii-DML.

***Ключевые слова**: Wikidata, метаданные, фабрика метаданных, цифровая математическая коллекция, цифровая математическая ретро коллекция, цифровая математическая библиотека, Lobachevskii-DML.*

ВВЕДЕНИЕ

В настоящее время происходящие изменения в инфраструктуре научных коммуникаций поставили целый ряд новых задач по управлению знаниями, а каждый этап жизненного цикла научной публикации предполагает его сопровождение специализированными программными инструментами (например, [1–4]).

Сегодня необходимым элементом научного исследования стало описание связей новых научных результатов с полученными ранее, что может быть выполнено в современных условиях только при наличии в Сети научного контента за весь период исследования рассматриваемых научных проблем. Такие связи устанавливаются все более активно во всех научных дисциплинах, поэтому можно утверждать, что в настоящее время формируется общее пространство научных знаний (см., например, [5]). В частности, основные направления интеграции математического знания определены в проектах “The Global Digital Mathematics Library” (GDML) и “World Digital Mathematics Library” (WDML) [6–8].

Метаданные являются основой коммуникации в Сети и используются на всех этапах жизненного цикла научной публикации (например, [9]). В настоящее время все научные документы оформляются для публикации с помощью программных инструментов – в англоязычной научной литературе для обозначения этого процесса используется термин “born-digital” (см. [10]). Современные правила подготовки научных публикаций, как в специализированных научных журналах, так и в сетевых изданиях, содержат требования по включению в соответствующие документы предметных классификаторов, ключевых слов, ORCID авторов и некоторой другой информации (например, [11, 12]). Именно на основании этой информации формируется набор метаданных научного документа.

При создании электронных коллекций научных документов, изданных в «доцифровой» период, возникают определенные проблемы с формированием обязательного набора метаданных документа (см., например, [13]). В такой ситуации с помощью методов анализа структуры и стилевых особенностей документа

можно сформировать основной набор его метаданных, включающий название этого документа, список его авторов и библиографию [14–17].

Ключевые слова и предметные классификаторы, такие как УДК [18] и Mathematics Subject Classification (MSC) [19], являются обязательными атрибутами современной научной публикации. Для создания или расширения списка ключевых слов используются методы текстового анализа. Подбор предметных классификаторов для математических статей производится методами автоматической классификации и категоризации (например, [20–23]). Но этих методов недостаточно для получения полного набора метаданных. Например, при формировании научных ретро-коллекций возникают проблемы даже с получением полной информации об авторах документов. Отметим, что новые методы формирования метаданных математических документов разрабатываются в проектах создания цифровых математических библиотек (см., например, [24, 25]).

На протяжении последних лет новые электронные математические коллекции формируются нами в рамках проекта создания цифровой библиотеки Lobachevskii Digital Mathematical Library (Lobachevskii-DML, https://lobachevskii-dml.ru/). Основной целью этого проекта является построение системы взаимосвязанных программных сервисов, обеспечивающих создание, обработку, хранение и управление объектами цифровых библиотек, а также интеграцию создаваемых электронных коллекций в агрегирующие цифровые математические библиотеки [26–28].

В настоящей работе представлен метод создания обязательного набора метаданных документов ретро-коллекций цифровой математической библиотеки. Термин «обязательный набор метаданных» понимается нами в соответствии со схемой метаданных EuDML [29]. В качестве источника пополнения метаданных использованы открытые ресурсы Веба. С помощью программных инструментов фабрики метаданных цифровой математической библиотеки Lobachevskii-DML реализованы основные процессы текстового анализа документов электронных ретро-коллекций, в частности, выделение именованных сущностей. С помощью разработанной системы запросов произведен поиск в Сети информации, необходимой для получения метаданных, с последующей экстракцией информационных объектов. После автоматизированного проведения фильтрации и нормали-

зации полученная информация включается в набор метаданных. Как один из основных результатов, представлен процесс формирования обязательного набора метаданных ретро-коллекции статей журнала «Известия физико-математического общества при Казанском университете» – одной из электронных коллекций цифровой библиотеки Lobachevskii-DML.

В разделе 1 выделены основные процессы создания математических электронных коллекций для их включения в цифровые библиотеки. Отмечены особенности формирования метаданных документов таких коллекций в соответствии со схемами агрегирующих цифровых библиотек.

В разделе 2 выделены наиболее важные проблемы формирования метаданных документов электронных математических ретро-коллекций.

Третий раздел посвящен методам получения информации из Wikidata с целью пополнения и уточнения метаданных документов электронных коллекций.

В четвертом разделе приведены алгоритмы пополнения метаданных документов ретро-коллекций цифровой библиотеки Lobachevskii-DML информацией, полученной из Wikidata.

1. ЦИФРОВЫЕ МАТЕМАТИЧЕСКИЕ БИБЛИОТЕКИ КАК ЧАСТЬ СПЕЦИАЛИЗИРОВАННОЙ НАУЧНОЙ ИНФРАСТРУКТУРЫ

Как отмечено в [8, 24], цифровым математическим библиотекам отводится роль основного интегратора математического знания, представленного в научных документах, опубликованных когда-либо. Обзор наиболее значительных цифровых математических библиотек приведен в [25, 30].

Проблемы интеграции знаний, полученных в области математики за весь «печатный» период развития этой науки, рассматривались в целом ряде проектов. Даже если эти проекты носили локальный характер, методы и инструменты, разрабатываемые в ходе их выполнения, были ориентированы на всеобъемлющую интеграцию математических знаний (см., например, [24]), а достигнутый уровень развития позволил поставить вопрос создания Всемирной цифровой математической библиотеки WDML.

Цифровая библиотека Lobachevskii DML включает в себя ряд электронных коллекций, при создании которых было необходимо выполнить полный цикл их

формирования: от оцифровки бумажных документов до загрузки цифровых документов и их метаданных в библиотеку. К числу таких коллекций относятся, например, «Труды Математического центра им. Н.И. Лобачевского» (далее – «Труды …» [31], а также ретро-коллекция «Известия физико-математического общества при Казанском университете» (“Bulletin de la Société Physico-Mathémaique de Kasan”) (далее – «Известия …» [32]. «Труды …» издаются с 1998 года, а до 2015 года большая часть их выпусков была только на бумажных носителях. Архивы журнала «Известия …» хранятся в Научной библиотеке Казанского университета только в бумажном виде и, как правило, в единичных экземплярах.

Создание электронной коллекции математических документов состоит из следующих основных этапов:

- Сканирование и оптическое распознавание документов;
- Разбиение архива документов на группы на основании сходства структуры и стилевого оформления документов;
- Определение начальной и завершающей страниц статей в файлах каждой группы документов;
- Разделение файлов на отдельные статьи на основании данных, полученных на предыдущем этапе;
- Поиск и выделение из документов основных метаданных;
- Уточнение метаданных;
- Поиск и пополнение метаданных информацией из Wikidata;
- Формирование метаданных статей по xml-схемам цифровой библиотеки;
- Интеграция электронной коллекции в соответствующую цифровую математическую библиотеку;
- Нормализация метаданных по xml-схемам агрегирующих цифровых библиотек.

Метаданные документов электронных коллекций, представленных в настоящей статье, были сформированы программными сервисами фабрики метаданных цифровой библиотеки Lobachevskii-DML [28, 33]. Эти сервисы реализуют методы, основанные на анализе структуры документов и особенностях их стилевого оформления [14, 16]. В основе реализации этих инструментов лежат методы ана-

лиза структуры документов (см., например, [14–17, 34, 35]). Также при формировании метаданных были применены стандартные алгоритмы текстового анализа (см., например, [36–38]).

Особенностью электронной коллекции «Труды …», как и многих других сборников материалов конференций, является отсутствие единых изначально сформулированных требований к структуре научных документов, включенных в эти издания. Это обстоятельство усложняет процесс извлечения метаданных методами, основанными на анализе структуры документа и его стилевых признаков. Так, например, ключевые слова, аннотации и предметные классификаторы присутствуют в статьях лишь в незначительном количестве сборников, в то время как эта информация необходима для формирования наборов метаданных по схемам агрегаторов математических документов [29, 39].

Далее, с использованием инструментов фабрики метаданных цифровой библиотеки Lobachevskii-DML нами была проведена процедура нормализации метаданных в соответствии с DTD-правилами и XML-схемами Journal Archiving and Interchange Tag Suite (NISO JATS) [40]. Для этого был сформирован набор метаданных в виде item-структуры, включающей как содержание метаданных, так и информацию о языке их представления. Это позволило включить в набор метаданных не только фамилии и имена авторов, приведенные в статье, но также варианты их написания на других языках. В результате работы соответствующего программного приложения был сгенерирован набор файлов в формате JATS, которые описывают каждую статью из обрабатываемого источника [33, 41].

Одной из структурных особенностей формата метаданных JATS является необходимость выбора основного языка представления статьи, а остальные языки объявляются альтернативными. Это создает сложности при формировании мультиязычных коллекций. Поэтому выбор основного языка представления – один из вопросов, которые приходится решать при создании xml-представления документов. Один из вариантов решения этой проблемы – использование языка, на котором написаны статьи, однако это не всегда позволяет организовать адекватный поиск в коллекциях цифровой библиотеки Lobachevskii-DML, потому что электронные коллекции этой библиотеки содержат в основном статьи на русском языке, а большая часть материалов ретро-коллекций – документы на дореформенном русском языке. При обработке таких документов возникают сложности в

написании названий статей и имен авторов, а также дополнительной информации, необходимой для формирования метаданных.

Сложности метаописания документов ретро-коллекций в формате JATS возникают и со статьями, опубликованными частями в различных номерах журнала, а также со статьями, которые имеют продолжения, причем об этом, как правило, говорится только в тексте статьи-продолжения, имеющей то же самое название.

2. АРХИВНЫЕ МАТЕМАТИЧЕСКИЕ ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ: ПРОБЛЕМЫ ФОРМИРОВАНИЯ МЕТАДАННЫХ

К архивным (ретро-коллекциям) мы относим электронные коллекции, которые содержат документы (статьи из периодических изданий (журналов), книги, препринты, сборники докладов конференций), напечатанные в период до широкого внедрения информационных технологий не только в процесс создания документа его авторами, но и в последующие этапы жизненного цикла этой публикации (см., например, [1, 3, 4]). Распространение научных знаний в этот период (обычно обозначаемый как «доцифровой») осуществлялось исключительно посредством печатных изданий. Как следствие, документы этих изданий, как правило, не содержат атрибутов, обязательных для изданий, распространяемых в Сети, таких, как предметные классификаторы, ключевые слова, аннотации, аффилиации авторов.

Отдельную категорию образуют исторические научные коллекции (ретро-коллекции), к которым можно отнести научные документы, опубликованные в периодических изданиях до начала XX века, а русскоязычные научные издания – до орфографической реформы 1918 года.

В работах [42–44] приведены результаты применения сервисов фабрики метаданных к документам ретро-коллекций цифровой библиотеки Lobachevskii-DML. Нормализация метаданных была проведена в них в соответствии с xml-схемами обязательного набора EuDML [29, 45].

Отметим наиболее важные проблемы формирования метаданных документов ретро-коллекций:

- разнообразие типов научных документов, размещенных в одном выпуске журнала, – статьи, доклады, письма, протоколы, объявления с отличающейся структурой оформления;

- отсутствие в статьях предметных классификаторов;
- отсутствие в статьях ключевых слов, характеризующих область исследования;
- отсутствие аннотаций к статьям;
- проблема с поиском в документе авторов статей – авторы могут быть указаны как в начале статьи, так и на последней ее странице;
- проблемы с полным описанием авторов: у автора статьи могут быть указаны только фамилия и начальная буква имени; встречаются сокращения фамилий авторов до инициалов (например, встречающееся сокращение «А. В.» соответствует статьям главного редактора А. Васильева (например, Известия. физ.-мат. общества, 1894 год, том 4));
- фамилия автора может снабжаться титулом (например, «Свящ. И. Максимовъ» (Известия физ.-мат. общества, 1915 год, том 11)), что требует использования отдельных шаблонов в методах экстракции метаданных;
- в статьях не указаны места работы авторов;
- в названиях теорем, а также в ссылках на статьи фамилии авторов приведены на языке оригинала;
- ссылки на литературу часто приведены в сносках либо непосредственно в тексте без полного библиографического описания;
- русскоязычные электронные ретро-коллекции содержат документы, использующие орфографию до реформы русского языка 1918 года.

3. КАКИЕ ЗНАНИЯ МОЖНО ИЗВЛЕЧЬ ИЗ WIKIDATA

Wikidata является базой знаний Википедии и центральной платформой управления данными для Википедии, а также родственных ей проектов (sister project) (см., например, [46, 47]). С момента запуска Wikidata в 2012 году на сайте этого проекта при участии более 5 млн. зарегистрированных пользователей собраны данные о 96633609 записях (информация на ноябрь 2021 года) (текущую статистику можно получить в [48]). Значительный профессиональный интерес к этому проекту объясняется тем, что Wikidata охватывает широкий спектр общих и специализированных знаний, актуальных во многих областях применения. Большая часть утверждений Wikidata снабжена сведениями об их происхождении, а

также дополнительными контекстными данными, такими как временная достоверность. Кроме того, приводимые данные связаны с внешними наборами данных во многих областях знаний, и информация продублирована на различных языках.

Объекты реального мира представлены в Wikidata элементами (items). Каждому элементу назначен числовой идентификатор с префиксом “Q”. Элементам соответствуют Wikipage in the Wikidata main namespace. Wikipage каждого элемента организована в виде свойств (properties) и утверждений (statements). Экземпляры свойств и утверждений также называют сущностями (entity), они имеют свои идентификаторы (с префиксом “Q” для утверждений и с префиксом “P” для свойств), которые служат важным источником метаданных item’a [49]. И у элементов, и у свойств имеются метка, описание и (многоязычные) псевдонимы.

Модель данных Wikidata приведена в [50]. Особенности работы с именованными сущностями в Wikidata выделены в работе [51]. Формулы присутствуют во всех математических статьях. Методы представления математических формул в Wikidata описаны в [52].

На страницах Wikidata, приведенных на рис. 1, представлена информация, которая использовалась при пополнении метаданных к статье А. Маркова «Распространение закона больших чисел на величины, зависящие друг от друга», опубликованной в номере 4 за 1906 год «Известий …». Рабочие процессы создания ретро-коллекции журнала «Известия …» и особенности формирования метаданных ее документов описаны в [43, 44]. Наиболее существенной является проблема идентификации авторов статей при фильтрации результатов запросов. Авторы статей этой коллекции указаны в выпусках журнала только фамилией и инициалами, иногда даже с одним инициалом (например, «А. Марковъ»). Поэтому при обработке результатов запросов потребовались фильтрация по нескольким признаков и дальнейшая проверка с привлечением экспертов.

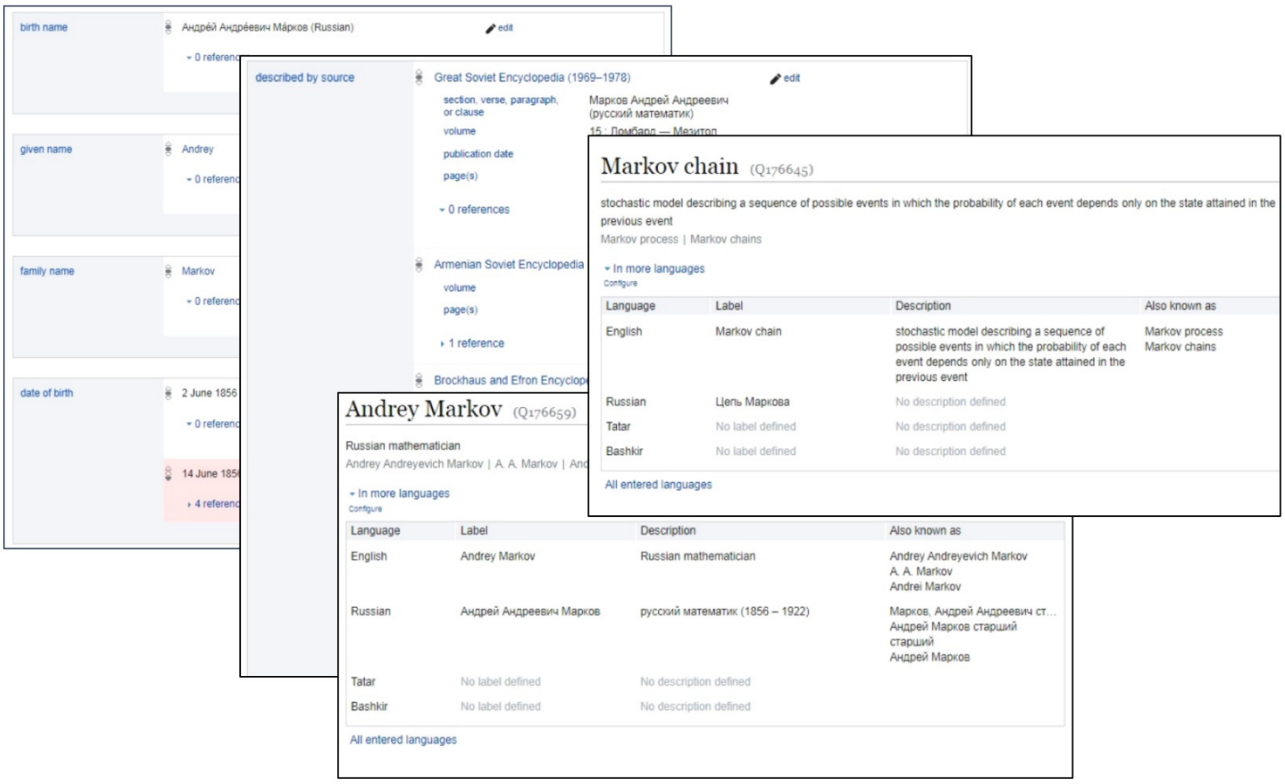


Рис. 1. Страницы Wikidata по запросам «Андрей Марков» и «Цепь Маркова».

Укажем основные свойства, значения которых были включены в состав метаданных: birth name (“Андре́й Андре́евич Ма́рков (Russian)”), given name (“Andrey”), family name (“Markov”), date of birth (“2 June 1856[Julian]”, “14 June 1856[Gregorian]”), date of death (“20 July 1922[Gregorian]”), occupation (“mathematician”, “statistician”, “university teacher”), field of work (“probability theory”, “mathematical analysis”, “number theory”), employer (“Saint Petersburg Academy of Sciences”, “Saint Petersburg State University”). Идентификаторы ID (“Q176659” и “Q176645”) также сохранены в метаданных – в дальнейшем с их помощью можно отслеживать обновление информации на соответствующих страницах Wikidata.

4. АЛГОРИТМЫ ПОПОЛНЕНИЯ МЕТАДАННЫХ ИНФОРМАЦИЕЙ ИЗ WIKIDATA

В данном разделе приведены алгоритмы формирования фундаментального набора метаданных документов ретро-коллекций цифровой библиотеки Lobachevskii-DML. Информация, которая по ряду причин оказалась недоступной для извлечения методами текстового и структурного анализа, была пополнена из открытых научных ресурсов Сети через систему поисковых запросов.

С помощью инструментов фабрики метаданных цифровой библиотеки Lobachevskii-DML была произведена обработка оцифрованных документов ретроколлекций. В результате удалось выделить из текстов следующие метаданные (в скобках приведены названия групп тегов в соответствии с xml-схемами JATS):

- название издания (“journal-title-group” и “trans-title-group”),
- время и место публикации (“publisher”, “pub-date”),
- название статьи (“title-group”) на одном из языков (дореформенном русском, английском, немецком, французском),
- фамилия автора с инициалами или только с одним инициалом (“contrib-group”),
- том и номер выпуска периодического издания (“volume”, “issue”),
- номера первой и последней страниц публикации (“star-page”, “end-page”).

Обработка сносок, часто присутствующих в текстах статей, позволяет получить информацию о названии и авторах (или только об одном из возможных авторов) тех статей, на которые указывают ссылки.

Для получения дополнительной информации, в частности, формирования фундаментального набора метаданных по схеме EuDML, предложен следующий алгоритм.

Имеющиеся метаданные преобразуются в csv-формат. Фамилии авторов и названия статей дополняются вариантами их написания на современном русском языке (в случае использования в документе дореформенной орфографии), а также производится транслитерация. Все названное необходимо для формирования шаблонов поисковых запросов.

На следующем этапе формируются поисковые запросы, включающие шаблоны, полученные на предыдущем этапе.

Далее выполняются стандартные операции по обработке полученных данных (см., например, [53]).

С помощью сформированных поисковых запросов можно уточнить (или дополнить) информацию об именах и отчествах авторов, месте работы, годах их публикационной активности, а также добавить URL сайтов, содержащих биографии и другую информацию об авторах.

Отметим, что описанный подход оказался результативным только в тех случаях, когда авторы документов электронных коллекций отражены в сетевом научном пространстве.

Для уточнения, а также пополнения уже сформированных метаданных были использованы открытые научные ресурсы, в частности, Wikipedia, Wikidata, DBPedia и Freebase [13, 54]. Для поиска и извлечения информации из Сети был применен инструментарий пакетов wikipedia и pywikibot (см., например, [55–57]). Также была разработана система SPARQL-запросов к ресурсам Wikidata.

Приведем теперь алгоритм извлечения метаданных из открытых научных ресурсов Сети, с обработкой полученных данных (см. Алгоритм 1).

```
<article-meta>
  <article-id>14_1_1_0</article-id>
  <title-group>
    <article-title xmllang="fr">Rapport sur les travaux de M. Pieri</article-title>
  </title-group>
  <contrib-group>
    <contrib contrib-type="author">
      <name-alternatives>
        <name>
          <surname xmllang="fr">Peano</surname>
          <string-name xmllang="fr">G. Peano</string-name>
        </name>
      </name-alternatives>
    </contrib>
  </contrib-group>
  <volume>14</volume>
  <volume-series>2</volume-series>
  <pub-date>
    <year>1904</year>
  </pub-date>
  <issue>1</issue>
  <issue-part>1</issue-part>
</article-meta>
```

Рис. 2. Фрагмент набора метаданных статьи G. Peano “Rapport sur les travaux de M. Pieri”, опубликованной в журнале «Известия …» (серия 2, том 14, номер 1 за 1904 год). Метаданные сформированы по схеме NISO JATS с помощью инструментов фабрики метаданных цифровой библиотеки Lobachevskii-DML. Из текста документа экстрагированы: название статьи на французском языке, фамилия и инициал автора, номера начальной и финальной страниц документа.

На вход алгоритма подается набор

$$M=\{d\_1.xml,d\_2.xml,\ldots,d\_m.xml\},$$

состоящий из файлов с метаданными документов ретро-коллекции. На рис. 2

приведен фрагмент такого файла. Из рассматриваемой статьи удалось извлечь только ее название (“Rapport sur le travaux de M. Pieri”) и фамилию автора (G. Peano). Отметим, что автор статьи в приведенном примере – известный математик Джузеппе Пеано (1858–1932); в настоящее время в Сети имеется информация об этом математике, как и о многих других авторах формируемой электронной ретро-коллекции.

Как уже было сказано, полученные наборы метаданных являются неполными, так как в соответствии со схемами интегрирующих цифровых математических библиотек требуется существенно больший объем метаинформации о научных документах. В частности, полученных метаданных недостаточно для формирования фундаментального набора по xml-схемам цифровой математической библиотеки EuDML [29].

Алгоритм 1: Извлечение метаданных из открытых научных источников Сети

```
load M=[d1.xml, d2.xml,..., dm.xml]
for d in M:
  # Разбор XML-дерева:
  md=parse(d).getroot()
 # Найти группу тегов с данными об авторах
 # (схема NISO JATS):
 for authors in md.findall(“./conrib_group/
              [@content_type=’authors’]”):
   # Выбрать группу тегов для каждого автора:
   for author in md.findall(“./contrib/
                    [@contrib_type=’author’]”):
     # Найти тег идентифицирующий имя автора и его инициалы:
  name_author_in_paper=author.find(‘string-name’)
     # Перевести и транслитерировать имя и фамилию автора:
•    if language(name_author_in_paper) !=’ru’:
•     name_author_ru=translate_ru()
•    if language(name_author_in_paper)==’ru’:
•     name_author_en=transliterate()
•    if language(name_author_in_paper)==’ru-old’:
•     name_author_ru=translate_ru_old()
•     name_author_en=transliterate()
•      # Сформировать список шаблонов поисковых запросов:
•    patterns=pattern_list()
•      # Выбрать и соединиться с точкой доступа
• #    (Wikipedia, Wikidata, DBPedia):
•    point=point_connect()
```

```
    results=[]
      # Поиск с каждым из шаблонов:
    for p in patterns:
     result=search(p)
          # Обработка результатов:
     extracting(result)
     cleaning(result)
     similarity(result)
     results.append(result)
    end for
      # Запись новых метаданных в соответствии с XML-схемой:
    normalization(results)
  end for
end for
```

В качестве источника пополнения метаданных нами были использованы открытые ресурсы Сети. Программные инструменты фабрики метаданных цифровой математической библиотеки Lobachevskii-DML на основе текстового анализа документов электронных ретро-коллекций позволяют извлекать такие метаданные, как название статьи, библиографические ссылки, диапазоны страниц, фамилии авторов на языке оригинала (русском, дореформенном русском, немецком, французском или английском). В настоящее время в Сети об авторах большинства статей формируемой электронной коллекции имеются сведения, которые отсутствовали в самих статьях. Это сделало возможным извлечение из сетевых ресурсов недостающей информации об авторах статей, в частности, о вариантах написания на различных языках их фамилий, имен и отчеств, мест работы в момент написания статьи (см. рис. 3 и 4).

Отметим, что имеется ряд сервисов связывания данных, содержащих объекты математического знания. Большинство из них имеет точку подключения (SPARQL endpoint) [58].

```
<article-meta>
  <article-id>14_1_1_0</article-id>
  <title-group>
    <article-title xmllang="fr">Rapport sur les
    travaux de M. Pieri</article-title>
  </title-group>
  <contrib-group>
    <contrib contrib-type="author">
      <name-alternatives>
        <name>
          <surname xmllang="fr">Peano</surname>
          <surname xmllang="ru">Пеано</surname>
          <given-names xmllang="fr">G.</given-names>
          <given-names xmllang="it">Giuseppe
          </given-names>
          <given-names xmllang="ru">Джузеппе
          </given-names>
        </name>
        <string-name xmllang="fr">G. Peano
        </string-name>
        <string-name xmllang="it">Giuseppe Peano
        </string-name>
        <string-name xmllang="ru">Джузеппе Пеано
        </string-name>
      </name-alternatives>
      <aff xmllang="fr">University of Turin</aff>
    </contrib>
  </contrib-group>
  <volume>14</volume>
  <volume-series>2</volume-series>
  <pub-date>
    <year>1904</year>
  </pub-date>
  <issue>1</issue>
  <issue-part>1</issue-part>
</article-meta>
```

Рис. 3. Фрагмент фундаментального набора метаданных, сформированных по Алгоритму 1. Ранее сформированный набор (см. рис. 2) был дополнен информацией об авторе статьи.

E	F	G	H	I	J	K	L
автор	исходное	название	исходное название статьи	WikidataURI	MathN	ZbMATHAuthorID	OpenLibraryID
H. Poincare		Rapport sur les travaux de M. Hilbert		Q81082		poincare.henri	OL7476098A
P. Mansion		Rapport sur les travaux de M. Barbarin		null		mansion.paul	OL3775794A
C. A. Laisant		Rapport sur les travaux de M. Lemoine		Q25318		laisant.ch-a	OL2426857A
G. Peano		Rapport sur les travaux de M. Pieri		Q191029		peano.giuseppe	OL32329A

Рис. 4. Дополнительные метаданные, экстрагированные из сетевых источников. В частности, получены URL страниц, содержащих упоминание о рассматриваемом документе.

На рис. 5 приведен пример обработки результата, полученного по запросу

в Wikidata и в ходе последующей фильтрации по ряду признаков, ограничивающих результаты запроса принадлежностью к научной деятельности. По запросу, содержащему фамилию автора статьи, в Wikidata было обнаружено 229 элементов. После проведения процедуры уточнения выделен элемент с меткой Q16648192, содержащий информацию об авторе статьи.

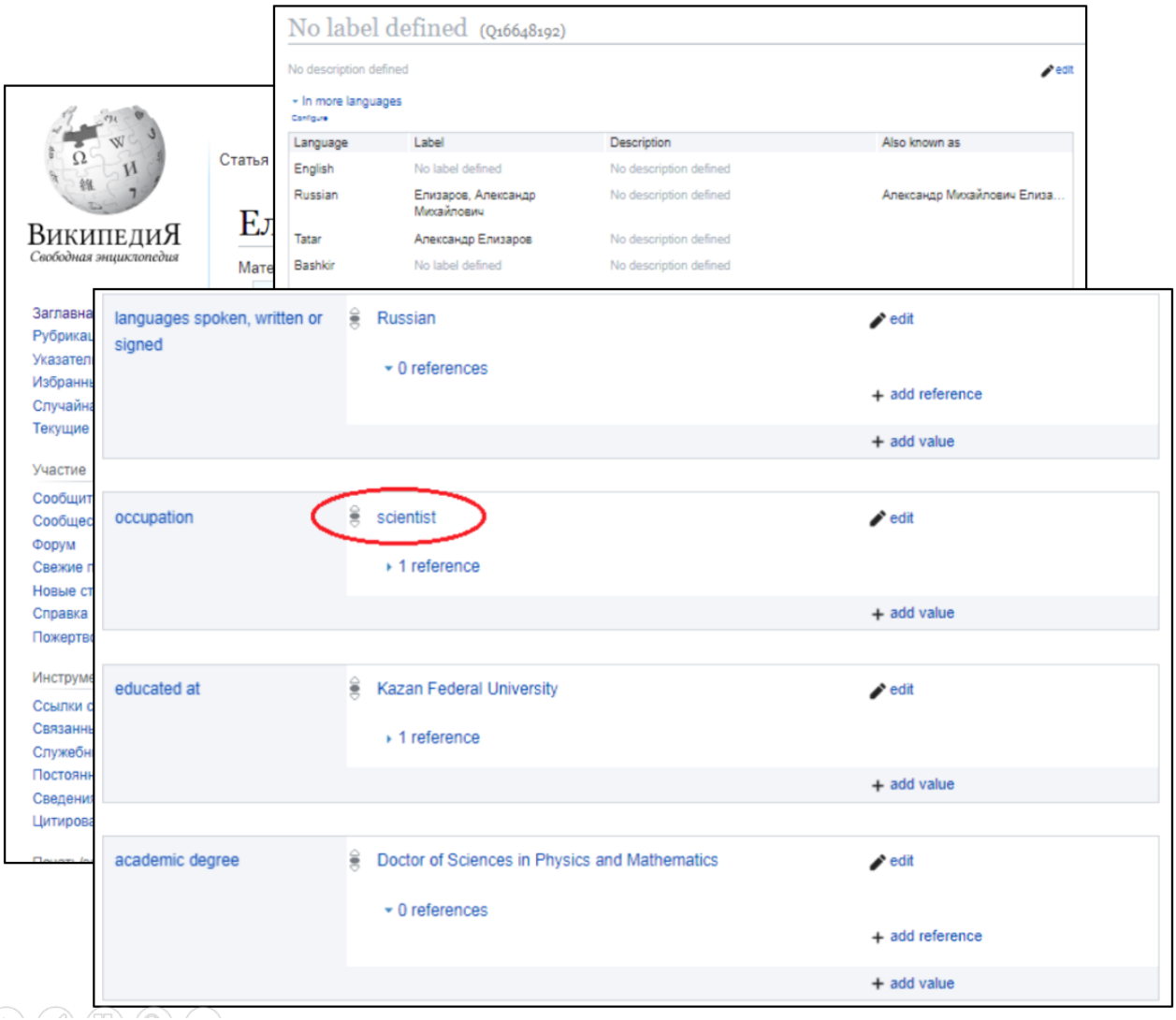


Рис. 5. Страница персоны в Wikidata. Отметим, что по данной персоне в Wikipedia имеется только русскоязычная страница, поэтому на странице Wikidata отсутствует label.

В таблице 1 приведены основные свойства (properties), информация из которых была использована при формировании дополнительных метаданных.

Таблица 1. Основные свойства Wikidata, используемые в алгоритме пополнения метаданных (на примере элемента с ID Q570859)

Property	ID	Пример	Jats_Tag
family name	P734	Chebotaryov	<surname>
given name	P735	Nikolai	<given-names>
name in native language	P1559	Николай Григорьевич Чеботарёв (Russian)	<string-name>
birth name	P1477	Николай Григорьевич Чеботарёв (Russian)	<string-name>
date of birth	P569	3 June 1894*Julian*, 15 June 1894*Gregorian*, 1894	<def-list>
date of death	P570	2 July 1947, 1947	<def-list>
occupation	P106	Mathematician(Q170790), university teacher (Q1622272)	<def-list>
employer	P108	Kazan Federal University (Q113788)	<aff>
member of	P463	Academy of Sciences of the USSR (Q2370801)	<aff-alterna-tives>
academic degree	P512	Doctor of Sciences in Physics and Mathematics (Q17281097)	<degrees>
field of work	P101	number theory (Q12479), algebra (Q3968), function theory (Q4455174)	<def-list>
notable work	P800	Chebotarev's density theorem (Q1425529), Chebotarev theorem on roots of unity (Q17007435)	<def-list>

Важным свойством, используемым в SPARQL-запросах к Wikidata, является свойство occupation (P106). С его помощью можно произвести фильтрацию результатов запроса, оставив только страницы с информацией о персонах, связанных с научной деятельностью (см. таблицу 2).

Таблица 2. Основные критерии отбора item по названию и количество соответствующих страниц

occupation (P106)	ID	Количество результатов
scientist	Q901	444354
mathematician	Q170790	35182
researcher	Q1650915	148544
university teacher	Q1622272	164512

Отметим, что при создании запросов к Wikidata учитываются синонимичные property, предоставляющие различные способы получения одних и тех же данных, например, свойства “name in native language” и “birth name” имеют одинаковые значения.

Ниже приведен Алгоритм 2 пополнения метаданных документов электронных коллекций с помощью запросов к Wikidata. На Рис. 6 приведены фрагменты класса Table и структуры item, используемые в алгоритме при формировании информации об авторе статьи.

```
class Table
{
    public List<item> familyname
    public List<item> initials
    public List<string> uri
    public List<item> Props
    public string type;
…..
}
```

```
struct item
{
    public string Property;
    public string ID;
    public string Jats_Tag;
    public string lang;
}
```

Рис. 6. Фрагменты класса и структуры, используемых для описания автора в Алгоритме 2.

Алгоритм 2: Пополнение метаданных документа цифровой коллекции

```
read metadata_set
List<string> authors_result = selected content from tag <authors>
#Список метаданных в xml формате
List <XElement> metadata
foreach authors_str in authors_results
        List <string> authors = Split(authors_str)
        foreach author in authors
                # поиск автора в Wikidata Wikidata,
                form SPARQL requests for Wikidata by family name (P734)
                # SPARQL запросы (Рис. 7, Рис. 8)
                get list Request_list from request
                filter by initials (from birth name or name in native language),
                filter by occupation set in Table 2
                if Request_list.Length>1 then необходима ручная экспертиза
                else
                List<Table> Props = new List<Table>
```

```
14:             fill in the attributes ID, Jats_Tag, Property for each class instance
15:             foreach Prop in Props
16:                 form SPARQL requests for Wikidata: property is Prop.ID
17:                 get content for Prop.Content
                #Формирование metadata set
18:             form a metadata_set using list Props
19:         metadata.Add(metadata_set)
20: form new metadata_set
21: save new metadata_set
```

Поиск происходит при помощи службы MediaWiki API [59]. Она позволяет вызывать MediaWiki API из SPARQL и получать результаты из запроса SPARQL. Ниже представлены некоторые запросы, которые используются в этом алгоритме.

Стандартный запрос поиска в Wikidata дополнительной информации по автору статьи (соответствует шагу 7 Алгоритма 2; осуществляет поиск в Wikidata страниц документов с фамилией автора статьи) представлен на рис. 7.

```
select ?item where {
?item rdfs:label "Елизаров"@ru.
?item wdt:P31 wd:Q101352.
}
```

Рис. 7. Запрос со свойством “instance of” (P31) с явным указанием сущности “family name” (Q101352).

Далее производится поиск по сущностям, полученным на шаге 7 Алгоритма 2. С помощью фильтрации по принадлежности к профессии (“scientist” или другого значения из Таблицы 2) результаты сужаются, в большинстве случаев до ссылок на страницы статей искомого автора (Рис. 8).

```
SELECT DISTINCT ?item ?itemLabel WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language "ru". }
  {
    SELECT DISTINCT ?item WHERE {
      ?item p:P734 ?statement0.
      ?statement0 (ps:P734/(wdt:P279*)) wd:Q21507140.
      ?item p:P106 ?statement1.
      ?statement1 (ps:P106/(wdt:P279*)) wd:Q901.
    }
    LIMIT 100
  }
}
```

Рис. 8. Поисковый запрос по сущности, полученной на предыдущем шаге алгоритма, с фильтрацией по значению “scientist” (Q901) свойства “occupation” (P106).

Запрос по получению всех метаданных, указанных в Таблице 1, представлен на рис. 9. Результат включает не только ссылку на сущность (entity), но и значение этой сущности.

```
select * where {
     wd:Q570859 wdt:P734 ?family_name_id.
  ?family_name_id rdfs:label ?family_name filter(lang(?family_name)
= 'ru')
  wd:Q570859 wdt:P735 ?given_name_id.
  ?given_name_id rdfs:label ?given_name filter(lang(?given_name) =
'ru')
  wd:Q570859 wdt:P1559 ?name_in_native_language.
  wd:Q570859 wdt:P1477 ?birth_name.
   wd:Q570859 wdt:P569 ?date_of_birth.
   wd:Q570859 wdt:P570 ?date_of_death.
   wd:Q570859 wdt:P106 ?occupation_id.
  ?occupation_id rdfs:label ?occupation filter(lang(?occupation) =
'ru')
   wd:Q570859 wdt:P108 ?employer_id.
  ?employer_id rdfs:label ?employer filter(lang(?employer) = 'ru')
  wd:Q570859 wdt:P463 ?member_of_id.
  ?member_of_id rdfs:label ?member_of filter(lang(?member_of) =
'ru')
   wd:Q570859 wdt:P512 ?academic_degree_id.
  ?academic_degree_id rdfs:label ?academic_degree filter(lang(?aca-
demic_degree) = 'ru')
```

```
   wd:Q570859 wdt:P101 ?field_of_work_id.
  ?field_of_work_id rdfs:label ?field_of_work fil-
ter(lang(?field_of_work) = 'ru')
   wd:Q570859 wdt:P800 ?notable_work_id.
  ?notable_work_id rdfs:label ?notable_work filter(lang(?nota-
ble_work) = 'ru')
}
```

Рис. 9. Запрос по получению всех метаданных, указанных в Таблице 1.

```
<contrib-group>
    <contrib contrib-type="author">
        <name-alternatives>
            <name>
                <surname id="Q21493235" xml:lang="ru">Чеботарёв</surname>
                <surname id="Q21493235" xml:lang="en">Chebotaryov</surname>
                <given-names id="Q5486169" xml:lang="ru">Николай</given-names>
                <given-names id="Q5486169" xml:lang="en">Nikolai</given-names>
                <string-name id="P1559" xml:lang="ru">Николай Григорьевич Чеботарёв</string-name
                <string-name id="P1477" xml:lang="ru">Николай Григорьевич Чеботарёв</string-name
            </name>
        </name-alternatives>
        <bio>
            <def-list id="P569">
                <def-item>3 June 1894 Julian</def-item>
                <def-item>15 June 1894 Gregorian, </def-item>
                <def-item>1894</def-item>
            <def-list>
            <def-list id="P570">
                <def-item>2 July 1947</def-item>
                <def-item>1947</def-item>
            <def-list>
            <def-list id="P106">
                <def-item id="Q170790">mathematician</def-item>
                <def-item id="Q1622272">university teacher</def-item>
            <def-list>
            <def-list id="P101">
                <def-item id="Q12479">number theory</def-item>
                <def-item id="Q3968">algebra</def-item>
                <def-item id="Q4455174">function theory</def-item>
            <def-list>
            <def-list id="P800">
                <def-item id="Q1425529">Chebotarev's density theorem</def-item>
                <def-item id="Q17007435">Chebotarev theorem on roots of unity</def-item>
            <def-list>
        </bio>
        <aff id="Q113788">Kazan Federal University</aff>
        <aff-alternatives id="Q2370801">Academy of Sciences of the USSR </aff-alternatives>
        <degrees id="Q17281097">Doctor of Sciences in Physics and Mathematics </degrees>
    </contrib>
</contrib-group>
```

Рис. 10. Фрагмент Jats-представления документа с метаописанием автора статьи по информации, полученной из Wikidata.

Далее производится обработка результатов SPARQL-запросов, включающая проведение трансформации в набор метаданных в формате JATS. Фрагмент полученных метаданных приведен на рис. 10.

Отметим, что при внутреннем представлении документов электронной коллекции используются id сущности Wikidata.

ЗАКЛЮЧЕНИЕ

Представлен метод создания обязательного набора метаданных документов электронных ретро-коллекций цифровой математической библиотеки Lobachevskii-DML. Названный набор соответствует известной схеме метаданных EuDML, широко используемой в настоящее время в цифровых математических библиотеках. При формировании такого набора метаданных документов, опубликованных в «доцифровой» период и включаемых в ретро-коллекции, возникает ряд проблем, связанных в первую очередь с недостаточностью имеющейся информации, необходимой для создания метаданных. Поэтому в качестве источника пополнения такой информации предложено использовать открытые ресурсы Веба, в частности, Wikidata.

С помощью программных инструментов созданной ранее фабрики метаданных цифровой математической библиотеки Lobachevskii-DML реализованы основные процессы текстового анализа документов электронных ретро-коллекций и выделены именованные сущности. Разработана система запросов для организации поиска в Сети информации, необходимой для получения метаданных, с последующей экстракцией информационных объектов. После автоматизированного проведения фильтрации и нормализации полученная информация включается в набор метаданных. Приведены алгоритмы пополнения метаданных документов ретро-коллекций информацией, полученной из Wikidata.

Одними из основных результатов проведенного исследования стали формирование обязательного набора метаданных ретро-коллекции статей журнала «Известия физико-математического общества при Казанском университете» и ее включение в состав цифровой математической библиотеки Lobachevskii-DML.

Благодарности

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 21-11-00105).

СПИСОК ЛИТЕРАТУРЫ

1. *Bartling S., Friesike S*. Towards Another Scientific Revolution // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing. 2014. P. 3–15 (2014). https://doi.org/10.1007/978-3-319-00026-8_1.

2. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the One-Brain Barrier: The Tetrapod Model of Mathematical Knowledge // Math. Intelligencer. 2021. Vol. 43. P. 78–87 (2021). https://doi.org/10.1007/s00283-020-10006-0.

3. *Елизаров А.М., Зуев Д.С., Липачёв Е.К.* Управление жизненным циклом электронных публикаций в информационной системе научного журнала // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2014. № 4. С. 81–88.

4. *Binfield P.* Novel Scholarly Journal Concepts // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. P. 155–163. https://doi.org/10.1007/978-3-319-00026-8_10.

5. *Ataeva O., Kalenov N., Serebriakov V., Sotnikov A*. Informational Infrastructure of the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 1–10. URL: http://ceur-ws.org/Vol-2990/rpaper1.pdf, last accessed 2021/11/07.

6. *Ion P.D.F.* Mathematics and the World Wide Web // In: Carette J., Aspinall D., Lange C., Sojka P., Windsteiger W. (Eds.) Intelligent Computer Mathematics. CICM 2013. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2013. Vol. 7961. https://doi.org/10.1007/978-3-642-39320-4_15.

7. *Ion P.D.F., Watt S.M.* The Global Digital Mathematics Library and the International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics, 2017. Lecture Notes in Artificial Intelligence. 2017. Vol. 10383. P. 56–69. https://doi.org/10.1007/978-3-319-62075-6_5.

8. Developing a 21st Century Global Library for Mathematics Research.

Washington: The National Academies Press, 2014. 142 p. https://doi.org/10.17226/18619.

9. *Xie I., Matusiak K.* Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016.

10. Born-digital. URL: https://en.wikipedia.org/wiki/Born-digital, last accessed 2021/11/07.

11. Author Guide – ScholarOne Manuscripts. Clarivate Analytics. 2019. P. 1–70. URL: https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/dlm_uploads/2019/10/ScholarOne-Manuscripts-Author-Guide.pdf, last accessed 2021/11/07.

12. Author tutorials. Writing a journal manuscript. Springer Nature Switzerland AG, 2021. URL: https://www.springernature.com/gp/authors/campaigns/writing-a-manuscript, last accessed 2021/11/07.

13. *Gafurova P., Elizarov A., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings.2021. Vol. 2990. P. 39–49. URL: http://ceur-ws.org/Vol-2990/rpaper4.pdf, last accessed 2021/11/07.

14. *Биряльцев Е.В., Елизаров А.М., Жильцов Н.Г., Липачёв Е.К., Невзорова О.А., Соловьев В.Д.* Методы анализа семантических данных математических электронных коллекций // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2014. № 4. С. 12–17.

15. *Tkaczyk D., Tarnawski B., Bolikowski Ł.* Structured Affiliations Extraction from Scientific Literature // D-Lib Magazine. 2015. Vol. 21. No. 11/12. https://doi.org/10.1045/november2015-tkaczyk.

16. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Automated system of services for processing of large collections of scientific documents // CEUR Workshop Proceedings. 2016. Vol. 1752. P. 58–64.

17. *Tkaczyk D.* New Methods for Metadata Extraction from Scientific Literature. arXiv:1710.10201v1. 2017. URL: https://arxiv.org/pdf/1710.10201v1.pdf, last accessed 2021/09/09.

18. Universal Decimal Classification. URL: https://udcc.org/index.php, last accessed 2021/09/09.

19. MSC2020–Mathematics Subject Classification System.
URL: https://mathscinet.ams.org/msnhtml/msc2020.pdf, last accessed 2021/09/09.

20. *Řehůřek R., Sojka P.* Automated Classification and Categorization of Mathematical Knowledge // In: Autexier S., Campbell J., Rubio J., Sorge V., Suzuki M., Wiedijk F. (Eds.) Intelligent Computer Mathematics. CICM 2008. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2008. Vol. 5144. P. 543–557. https://doi.org/10.1007/978-3-540-85110-3_44.

21. *Хайдаров Ш.М., Ямалутдинова Г.Ш.* Рекомендательная система классификации физико-математических документов // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. С. 480–486.
URL: https://doi.org/ 10.20948/abrau-2018-57. http://keldysh.ru/abrau/2018/theses/57.pdf.

22. *Schubotz M., Scharpf P., Teschke O., Kühnemund A., Breitinger C., Gipp B*. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // In: Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020. arXiV:2005.12099v1. 25 May 2020.

23. *Nevzorova O., Almukhametov D.* Towards a Recommender System for the Choice of UDC Code for Mathematical Articles // CEUR Workshop Proceedings. 2021. Vol. 3036. P. 54–62.
URL: http://ceur-ws.org/Vol-3036/paper04.pdf, last accessed 2021/11/07.

24. *Rocha E.M., Rodrigues J.F.* Disseminating and preserving mathematical knowledge // In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 3–21.

25. *Elizarov A.M., Lipachev E.K., Zuev D.S*. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 317–325.

26. *Elizarov A.M., Lipachev E.K*. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 326–333. URL: http://ceur-ws.org/Vol-2022/paper50.pdf, last accessed 2021/11/07.

27. *Elizarov A.M., Lipachev E.K.* Big Math Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. Vol. 2523. P. 59–72.

URL: http://ceur-ws.org/Vol-2523/invited08.pdf, last accessed 2021/11/21.

28. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Базовые сервисы фабрики метаданных цифровой математической библиотеки Lobachevskii-DML // Электронные библиотеки. 2020. Т. 23. №3. С. 336–381. https://doi.org/10.26907/1562-5419-2020-23-3-336-381.

29. EuDML metadata schema specification (v2.0–final). https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2021/11/11.

30. *Elizarov A., Lipachev E.* Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 25–38. URL: http://ceur-ws.org/Vol-2990/rpaper3.pdf, last accessed 2021/11/07.

31. Электронная коллекция: Труды математического центра им. Н. И. Лобачевского. URL: https://lobachevskii-dml.ru/journal/tmt, last accessed 2021/11/07.

32. Электронная коллекция: «Известия физико-математического общества при Казанском университете». URL: https://lobachevskii-dml.ru/journal/izfmo2, https://lobachevskii-dml.ru/journal/izfmo3, last accessed 2021/11/07.

33. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // CEUR Workshop Proceedings. 2021. Vol. 2813. P. 13–21. URL: http://ceur-ws.org/Vol-2813/rpaper01.pdf, last accessed 2021/11/07.

34. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.K.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // In: Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE. 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064.

35. *Elizarov A., Lipachev E.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360. URL: http://ceur-ws.org/Vol-2543/spaper05.pdf, last accessed 2021/11/07.

36. *Lane H., Hapke H., Howard C.* Natural Language Processing in Action: Under-standing, analyzing, and generating text with Python. Manning Publications, 2019.

37. Natasha. URL: https://github.com/natasha/natasha, last accessed

2021/11/07.

38. Проект Natasha. Набор качественных открытых инструментов для обработки естественного русского языка (NLP).
URL: https://habr.com/ru/post/516098/, last accessed 2021/11/07.

39. *Bouche T., Rákosník J.* Report on the EuDML External Cooperation Model // in: Kaiser K., Krantz S.G., Wegner B. (Eds.) Topics and Issues in Electronic Publishing, JMM, Special Session. San Diego, 2013. P. 99–10.
URL: https://www.emis.de/proceedings/TIEP2013/07bouche_rakosnik.pdf, last accessed 2021/11/11.

40. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html, last accessed 2021/01/05.

41. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M*. Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2543. P. 136–148.
URL: http://ceur-ws.org/Vol-2543/rpaper13.pdf, last accessed 2021/11/07.

42. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Lobachevskii-DML: формирование архивных математических коллекций // Научный сервис в сети Интернет: труды XXII Всероссийской научной конференции. М.: ИПМ им. М.В. Келдыша, 2020. С. 171–183. https://doi.org/10.20948/abrau-2020-23.

43. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Metadata Extraction Methods for Organizing a Retro-Collection in the Lobachevskii Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2784. P. 62–71.
URL: http://ceur-ws.org/Vol-2784/rpaper06.pdf, last accessed 2021/11/07.

44. *Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Алгоритмы формирования метаданных математических ретро-коллекций на основе анализа структурных особенностей документов // Электронные библиотеки. 2021. Т. 24, №2. С. 238–271. https://doi.org/10.26907/1562-5419-2021-24-2-238-270.

45. *Jost M., Bouche T., Goutorbe C., Jorda J.P.* D3.2: The EuDML metadata schema. Revision: 1.6 as of 15th December 2010.
URL: http://www.mathdoc.fr/publis/d3.2-v1.6.pdf, last accessed 2021/11/11.

46. *Vrandečić D., Krötzsch M.* Wikidata: a free collaborative knowledgebase // Communications of the ACM. October 2014. Vol. 57. Issue 10. P. 78–85.
https://doi.org/10.1145/2629489.

47. Wikipedia: Wikidata. URL: https://en.wikipedia.org/wiki/Wikidata, last accessed 2021/11/07.

48. Statistics – Wikidata.
URL: https://www.Wikidata.org/wiki/Special:Statistics, last accessed 2021/11/07.

49. Wikidata: Glossary.
URL: https://www.Wikidata.org/wiki/Wikidata:Glossary, last accessed 2021/11/07.

50. *Erxleben F., Günther M., Krötzsch M., Mendez J., Vrandečić D*. Introducing Wikidata to the Linked Data Web // In: Mika P. et al. (Eds.) The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science. Springer, Cham. 2014. Vol. 8796. P. 50–65. https://doi.org/10.1007/978-3-319-11964-9_4.

51. *Geiß J., Spitz A., Gertz M.* NECKAr: A Named Entity Classifier for Wikidata // In: Rehm G., Declerck T. (Eds.) Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science. Springer, Cham. 2018. Vol. 10713. P. 115–129. https://doi.org/10.1007/978-3-319-73706-5_10.

52. *Scharpf Ph., Schubotz M., Gipp B.* Mathematics in Wikidata // CEUR Workshop Proceedings. 2021. Vol. 2982. P. 1–14.
URL: http://ceur-ws.org/Vol-2982/paper-1.pdf, last accessed 2021/11/07.

53. *Knoblock C.A., Szekely P.* A scalable architecture for extracting, aligning, link-ing, and visualizing multi-Int data // Proc. SPIE 9499, Next-Generation Analyst III, 949907 (15 May 2015). https://doi.org/10.1117/12.2177119.

54. *Андреичев М.Д., Гафурова П.О., Елизаров А.М., Липачёв Е.К.* Пополнение метаданных документов математических цифровых ретро-коллекций методом семантических сетей // Научный сервис в сети Интернет: труды XXIII Всероссийской научной конференции (20–23 сентября 2021 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2021. С. 22–33. https://doi.org/10.20948/abrau-2021-22. URL: https://keldysh.ru/abrau/2021/theses/22.pdf, last accessed 2021/11/07.

55. *Ayers P., Matthews C., Yates B.* How Wikipedia Works: And How You Can Be a Part of It. No Starch Press, San Francisco, CA, 2008.

56. Wikipedia Documentation.
URL: https://wikipedia.readthedocs.io/en/latest/code.html, last accessed 2021/11/07.

57. Pywikibot Documentation.
URL: https://doc.wikimedia.org/pywikibot/master/index.html, last accessed 2021/11/07.

58. SPARQL Query Language for RDF/W3C.
URL: https://www.w3.org/TR/rdf-sparql-query/. last accessed 2021/11/07.

59. MediaWiki is a collaboration and documentation platform brought to you by a vibrant community. URL: https://www.mediawiki.org/wiki/MediaWiki, last accessed 2021/11/07.

EXTRACTION OF WIKIDATA KNOWLEDGE FOR THE METADATA FORMATION FOR DOCUMENTS OF ELECTRONIC MATHEMATICAL COLLECTIONS

P. O. Gafurova[1] [0000-0002-1544-155X], A. M. Elizarov[2] [0000-0003-2546-6897],
E. K. Lipachev[3] [0000-0001-7789-2332]

[1–3] *Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

[1]pogafurova@gmail.com, [2]amelizarov@gmail.com, [3]elipachev@gmail.com

Abstract

Methods for creating digital mathematical collections that include unstructured sets of documents are presented. These sets contain materials from scientific conferences, as well as articles from the archives of mathematical journals of the "pre-digital" period.

Using the software tools of the metadata factory of the digital mathematical library Lobachevskii DML, a mandatory set of metadata for digital collection documents was formed. To refine and replenish the metadata sets, knowledge extraction methods from Wikidata were used.

To search Wikidata for information about digital collection documents and their authors, a system of SPARQL queries has been developed. A set of Wikidata entities is defined, which determine the features of the search, as well as the subsequent filtering of the results.

Methods for clarifying and supplementing the bibliographic references given in the articles are proposed. When forming the metadata of documents of retrocollec-

tions, a search was made in Wikidata for information about the years of life of the authors of articles, as well as URLs of web pages with information about articles and their authors. The results of the formation of several new digital collections of the Lobachevskii-DML digital library are presented.

***Keywords**: Wikidata, metadata, metadata factory, digital mathematical collection, retrodigitized mathematical collection, Digital Mathematical Libraries, Lobachevskii-DML.*

REFERENCES

1. *Bartling S., Friesike S*. Towards Another Scientific Revolution // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing. 2014. P. 3–15. https://doi.org/10.1007/978-3-319-00026-8_1.

2. *Carette J., Farmer W.M., Kohlhase M., Rabe F.* Big Math and the One-Brain Barrier: The Tetrapod Model of Mathematical Knowledge // Math. Intelligencer. 2021. Vol. 43. P. 78–87. https://doi.org/10.1007/s00283-020-10006-0.

3. *Elizarov A.M., Zuev D.S., Lipachev E.K*. Lifecycle Management of Electronic Publications in Information Systems Scientific Journal // Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies. 2014. No. 4. P. 81-88.

4. *Binfield P.* Novel Scholarly Journal Concepts // In: Bartling S., Friesike S. (Eds.) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, 2014. P. 155–163. https://doi.org/10.1007/978-3-319-00026-8_10.

5. *Ataeva O., Kalenov N., Serebriakov V., Sotnikov A*. Informational Infrastructure of the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings 2990 (2021) 1–10. URL: http://ceur-ws.org/Vol-2990/rpaper1.pdf, last accessed 2021/11/07.

6. *Ion P.D.F.* Mathematics and the World Wide Web // In: Carette J., Aspinall D., Lange C., Sojka P., Windsteiger W. (Eds.) Intelligent Computer Mathematics. CICM 2013. Lecture Notes in Computer Science. 2013. Vol. 7961. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39320-4_15.

7. *Ion P.D.F., Watt S.M.* The Global Digital Mathematics Library and the

International Mathematical Knowledge Trust // ICM 2017: Intelligent Computer Mathematics. Lecture Notes in Artificial Intelligence. 2017. Vol. 10383. P. 56–69. https://doi.org/10.1007/978-3-319-62075-6_5.

8. Developing a 21st Century Global Library for Mathematics Research. Washington: The National Academies Press, 2014. 142 p. https://doi.org/10.17226/18619.

9. *Xie I., Matusiak K.* Discover Digital Libraries: Theory and Practice. Elsevier Inc., 2016.

10. Born-digital. URL: https://en.wikipedia.org/wiki/Born-digital, last accessed 2021/11/07.

11. Author Guide – ScholarOne Manuscripts. Clarivate Analytics. 2019. P. 1–70. URL: https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/dlm_uploads/2019/10/ScholarOne-Manuscripts-Author-Guide.pdf, last accessed 2021/11/07.

12. Author tutorials. Writing a journal manuscript. Springer Nature Switzerland AG, 2021. URL: https://www.springernature.com/gp/authors/campaigns/writing-a-manuscript, last accessed 2021/11/07.

13. *Gafurova P., Elizarov A., Lipachev E.* Algorithms for Integration of Unstructured Mathematical Documents into the Common Digital Space of Scientific Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 39–49. URL: http://ceur-ws.org/Vol-2990/rpaper4.pdf, last accessed 2021/11/07.

14. *Biryal'tsev E., Elizarov A., Zhil'tsov N., Lipachev E., Nevzorova O., Solov'ev V.* Methods for Analyzing Semantic Data of Electronic Collections in Mathematics // Automatic Documentation and Mathematical Linguistics. 2014. Vol. 48, No. 2. P. 81–85. https://doi.org/10.3103/S000510551402006X.

15. *Tkaczyk D., Tarnawski B., Bolikowski Ł.* Structured Affiliations Extraction from Scientific Literature // D-Lib Magazine. 2015. Vol. 21, No. 11/12. https://doi.org/10.1045/november2015-tkaczyk.

16. *Elizarov A.M., Lipachev E.K., Khaydarov S.M.* Automated system of services for processing of large collections of scientific documents // CEUR Workshop Proceedings. 2016. Vol. 1752. P. 58–64.

17. *Tkaczyk D.* New Methods for Metadata Extraction from Scientific

Literature. arXiv:1710.10201v1. 2017. URL: https://arxiv.org/pdf/1710.10201v1.pdf, last accessed 2021/09/09.

18. Universal Decimal Classification. URL: https://udcc.org/index.php, last accessed 2021/09/09.

19. MSC2020–Mathematics Subject Classification System.
URL: https://mathscinet.ams.org/msnhtml/msc2020.pdf, last accessed 2021/09/09.

20. *Řehůřek R., Sojka P.* Automated Classification and Categorization of Mathematical Knowledge // In: Autexier S., Campbell J., Rubio J., Sorge V., Suzuki M., Wiedijk F. (Eds.) Intelligent Computer Mathematics. CICM 2008. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2008. Vol. 5144. P. 543–557.
https://doi.org/10.1007/978-3-540-85110-3_44.

21. *Khaydarov S.M., Yamalutdinova G.S.* Recommender System of Physical and Mathematical Documents Classification. CEUR Workshop Proceedings. 2018. Vol. 2260. P. 480–486. URL: http://ceur-ws.org/Vol-2260/57_480-486.pdf, last accessed 2021/11/07.

22. *Schubotz M., Scharpf P., Teschke O., Kühnemund A., Breitinger C., Gipp B*. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // In: Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020. arXiV:2005.12099v1. 25 May 2020.

23. *Nevzorova O., Almukhametov D.* Towards a Recommender System for the Choice of UDC Code for Mathematical Articles // CEUR Workshop Proceedings. 2021. Vol. 3036. P. 54–62.
URL: http://ceur-ws.org/Vol-3036/paper04.pdf, last accessed 2021/11/07.

24. *Rocha E.M., Rodrigues J.F.* Disseminating and preserving mathematical knowledge // In: Borwein J.M., Rocha E.M., Rodrigues J.F. (Eds.). Communicating Mathematics in the Digital Era. A K Peters, Ltd., 2008. P. 3–21.

25. *Elizarov A.M., Lipachev E.K., Zuev D.S*. Digital Mathematical Libraries: Overview of Implementations and Content Management Services // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 317–325.

26. *Elizarov A.M., Lipachev E.K*. Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University // CEUR Workshop Proceedings. 2017. Vol. 2022. P. 326–333. URL: http://ceur-ws.org/Vol-2022/paper50.pdf, last accessed 2021/11/07.

27. *Elizarov A.M., Lipachev E.K.* Big Math Methods in Lobachevskii-DML Digital Library // CEUR Workshop Proceedings. 2019. Vol. 2523. P. 59–72.
URL: http://ceur-ws.org/Vol-2523/invited08.pdf, last accessed 2021/11/21.

28. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Basic Services of Factory Metadata Digital Mathematical Library Lobachevskii-DML // Russian Digital Libraries Journal. 2020. V. 23, No. 3. P. 336–381.
https://doi.org/10.26907/1562-5419-2020-23-3-336-381.

29. EuDML metadata schema specification (v2.0–final).
https://initiative.eudml.org/eudml-metadata-schema-specification-v20-final, last accessed 2021/11/11.

30. *Elizarov A., Lipachev E.* Digital Libraries and the Common Digital Space of Mathematical Knowledge // CEUR Workshop Proceedings. 2021. Vol. 2990. P. 25–38.
URL: http://ceur-ws.org/Vol-2990/rpaper3.pdf, last accessed 2021/11/07.

31. Digital Collection: Proceedings of Lobachevskii mathematical center.
URL: https://lobachevskii-dml.ru/journal/tmt, last accessed 2021/11/07

32. Digital Collection: “Izvestia of the Physics and Mathematics Society at Kazan University” (“Bulletin de la Société Physico-Mathémaique de Kasan”).
URL: https://lobachevskii-dml.ru/journal/izfmo2,
https://lobachevskii-dml.ru/journal/izfmo3, last accessed 2021/11/07.

33. *Elizarov A., Lipachev E.* Digital Library Metadata Factories // CEUR Workshop Proceedings. 2021. Vol. 2813. P. 13–21.
URL: http://ceur-ws.org/Vol-2813/rpaper01.pdf, last accessed 2021/11/07.

34. *Elizarov A.M., Khaydarov Sh.M., Lipachev E.K.* Scientific Documents Ontologies for Semantic Representation of Digital Libraries // In: Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE. 2017. P. 1–5. https://doi.org/10.1109/RPC.2017.8168064.

35. *Elizarov A., Lipachev E.* Methods of Processing Large Collections of Scientific Documents and the Formation of Digital Mathematical Library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.
URL: http://ceur-ws.org/Vol-2543/spaper05.pdf, last accessed 2021/11/07.

36. *Lane H., Hapke H., Howard C.* Natural Language Processing in Action: Under-standing, analyzing, and generating text with Python. Manning Publications, 2019.

37. Natasha. URL: https://github.com/natasha/natasha, last accessed 2021/11/07.

38. Proekt Natasha. Nabor kachestvennyh otkrytyh instrumentov dlya obrabotki estestvennogo russkogo yazyka (NLP).
URL: https://habr.com/ru/post/516098/, last accessed 2021/11/07.

39. *Bouche T., Rákosník J.* Report on the EuDML External Cooperation Model // in: Kaiser K., Krantz S.G., Wegner B. (Eds.) Topics and Issues in Electronic Publishing, JMM, Special Session. San Diego, 2013. P. 99–10.
URL: https://www.emis.de/proceedings/TIEP2013/07bouche_rakosnik.pdf, last accessed 2021/11/11.

40. Journal Article Tag Suite. URL: https://jats.nlm.nih.gov/about.html, last accessed 2021/01/05.

41. *Gafurova P.O., Elizarov A.M., Lipachev E.K., Khammatova D.M*. Metadata Normalization Methods in the Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2543. P. 136–148.
URL: http://ceur-ws.org/Vol-2543/rpaper13.pdf, last accessed 2021/11/07.

42. *Гафурова П.О., Gafurova P. O., Elizarov A. M., Lipachev E. K*. Lobachevskii-DML: Formation of Archival Mathematical Collections // Nauchnyj servis v seti Internet: trudy XXII Vserossijskoj nauchnoj konferencii. M.: IPM im. M.V. Keldysha, 2020. S. 171–183. https://doi.org/10.20948/abrau-2020-23.

43. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Metadata Extraction Methods for Organizing a Retro-Collection in the Lobachevskii Digital Mathematical Library // CEUR Workshop Proceedings. 2020. Vol. 2784. P. 62–71.
URL: http://ceur-ws.org/Vol-2784/rpaper06.pdf, last accessed 2021/11/07.

44. *Gafurova P.O., Elizarov A.M., Lipachev E.K.* Algorithms for Formation of Metadata Mathematical Retro Collections Based on Analysis of Structural Features of Documents // Russian Digital Libraries Journal. 2021. Vol. 24. No. 2. P. 238–271.
https://doi.org/10.26907/1562-5419-2021-24-2-238-270.

45. *Jost M., Bouche T., Goutorbe C., Jorda J.P.* D3.2: The EuDML metadata schema. Revision: 1.6 as of 15th December 2010.
URL: http://www.mathdoc.fr/publis/d3.2-v1.6.pdf, last accessed 2021/11/11.

46. *Vrandečić D., Krötzsch M.* Wikidata: a free collaborative knowledgebase // Communications of the ACM. October 2014. Vol. 57. Issue 10. P. 78–85.

https://doi.org/10.1145/2629489.

47. Wikipedia: Wikidata. URL: https://en.wikipedia.org/wiki/Wikidata, last accessed 2021/11/07.

48. Statistics – Wikidata.
URL: https://www.Wikidata.org/wiki/Special:Statistics, last accessed 2021/11/07.

49. Wikidata: Glossary.
URL: https://www.Wikidata.org/wiki/Wikidata:Glossary, last accessed 2021/11/07.

50. *Erxleben F., Günther M., Krötzsch M., Mendez J., Vrandečić D*. Introducing Wikidata to the Linked Data Web // In: Mika P. et al. (Eds.) The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science. Springer, Cham. 2014. Vol. 8796. P. 50–65. https://doi.org/10.1007/978-3-319-11964-9_4.

51. *Geiß J., Spitz A., Gertz M.* NECKAr: A Named Entity Classifier for Wikidata // In: Rehm G., Declerck T. (Eds.) Language Technologies for the Challenges of the Digital Age. GSCL 2017. Lecture Notes in Computer Science. Springer, Cham. 2018. Vol. 10713. P. 115–129. https://doi.org/10.1007/978-3-319-73706-5_10.

52. *Scharpf Ph., Schubotz M., Gipp B.* Mathematics in Wikidata // CEUR Workshop Proceedings. 2021. Vol. 2982. P. 1–14.
URL: http://ceur-ws.org/Vol-2982/paper-1.pdf, last accessed 2021/11/07.

53. *Knoblock C.A., Szekely P.* A scalable architecture for extracting, aligning, link-ing, and visualizing multi-Int data // Proc. SPIE 9499, Next-Generation Analyst III, 949907 (15 May 2015). https://doi.org/10.1117/12.2177119.

54. *Andreichev M.D., Gafurova P.O., Elizarov A.M., Lipachev E.K.* Replenishment of Documents of Mathematical Digital Retro-collections by Searching in Semantic Web. Nauchnyj servis v seti Internet: trudy XXIII Vserossijskoj nauchnoj konferencii (20–23 sentyabrya 2021 g., onlajn). M.: IPM im. M.V. Keldysha, 2021. S. 22–33. https://doi.org/10.20948/abrau-2021-22.
URL: https://keldysh.ru/abrau/2021/theses/22.pdf, last accessed 2021/11/07.

55. *Ayers P., Matthews C., Yates B.* How Wikipedia Works: And How You Can Be a Part of It. No Starch Press, San Francisco, CA, 2008.

56. Wikipedia Documentation.
URL: https://wikipedia.readthedocs.io/en/latest/code.html, last accessed 2021/11/07.

57. Pywikibot Documentation.
URL: https://doc.wikimedia.org/pywikibot/master/index.html,

last accessed 2021/11/07.

58. SPARQL Query Language for RDF/W3C.
URL: https://www.w3.org/TR/rdf-sparql-query/. last accessed 2021/11/07.

59. MediaWiki is a collaboration and documentation platform brought to you by a vibrant community. URL: https://www.mediawiki.org/wiki/MediaWiki, last accessed 2021/11/07.

СВЕДЕНИЯ ОБ АВТОРАХ

ГАФУРОВА Полина Олеговна – магистр математики, аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Polina GAFUROVA – Magister of Mathematics, Kazan (Volga Region) Federal University.

Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: pogafurova@gmail.com;

ORCID: 0000-0002-1544-155X

ЕЛИЗАРОВ Александр Михайлович – доктор физико-математических наук, профессор, заслуженный деятель науки Республики Татарстан, профессор кафедры программной инженерии Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Alexander Michailovich ELIZAROV – Doctor of Physics and Mathematics, Professor, Honoured Worker of Science of the Republic of Tatarstan, Kazan Federal University.

Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

email: amelizarov@gmail.com;

ORCID: 0000-0003-2546-6897

ЛИПАЧЁВ Евгений Константинович – кандидат физико-математических наук, доцент, доцент кафедры Интеллектуальных технологий поиска Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Evgeny Konstantinovich LIPACHEV – Candidate of Physics and Mathematics, Associate Professor, Kazan Federal University.

Current scientific interests: data mining, recommender systems, cloud computing, knowledge extraction technologies.

URL: https://kpfu.ru/Evgeny.Lipachev.

email: elipachev@gmail.com;

ORCID: 0000-0001-7789-2332

Материал поступил в редакцию 11 ноября 2021 года

УДК 01, 002.53

РЕЙТИНГ ЖУРНАЛА В БИБЛИОГРАФИЧЕСКОЙ БАЗЕ

М. М. Горбунов-Посадов[1] [0000-0002-7044-8287], Т. А. Полилова[2] [0000-0003-4628-3205]

[1, 2]Институт прикладной математики им. М.В. Келдыша Российской академии наук, Миусская пл., 4, Москва, 125047

[1]gorbunov@keldysh.ru, [2]polilova@keldysh.ru

Аннотация

Инструмент построения рейтингов научных журналов является одним из востребованных сервисов библиографических баз. Задача построения рейтинга обычно делится на две основные подзадачи: определение референтной группы журналов и вычисление показателя рейтинга для журналов этой группы. Практика показывает, что для корректного сопоставления журналов необходимым условием является ограничение референтной группы исключительно журналами определенной тематики. В случае методических ошибок, допущенных на этапе выделения референтной группы, значения показателя журналов в рейтинге могут сильно отличаться от ожидаемых.

Например, в рейтинге журналов в Российском индексе научного цитирования (РИНЦ) по двухлетнему импакт-фактору в тематическом направлении «Математика» классические фундаментальные математические журналы вопреки ожиданиям не выходят на первые позиции рейтинга. Первые позиции заняли журналы, для которых математика не является доминирующей профильной дисциплиной. Анализ статистических данных о тематике публикуемых статей и цитирований в журналах, занимающих лидирующие позиции рейтинга РИНЦ, показывает, что на показатели рейтинга существенно повлияла мультидисциплинарность этих журналов.

Отмеченное недоразумение подводит к мысли о том, что в подсчет рейтинга в данном случае следовало вовлекать не все статьи журнала, а только относящиеся к данному тематическому направлению. Вместе с тем вопросы вызывает и сложившаяся схема тематической классификации направлений. Более перспективной представляется набирающая популярность классификация «снизу вверх», работающая на представительном массиве статей. Здесь тематиче-

ские кластеры вычленяются на основе понятия близости статей, трактуемого как близость их библиографических связей. И далее тематическая принадлежность статьи не назначается волевым решением автора или редакции, а строго формально вычисляется на основе ее библиографического списка.

***Ключевые слова**: научная публикация, цитирование, рейтинг журналов, тематическая классификация, импакт-фактор, мультидисциплинарность, библиографическая ссылка, со-цитирование, классификация снизу вверх, тематическая кластеризация, Citation Topics.*

ВВЕДЕНИЕ

Современные библиографические базы накопили огромный массив информации о научных публикациях. Научные издательства регулярно поставляют в библиографические базы широкий набор метаданных издаваемой научной продукции. В библиографические базы попадают данные о журналах, названия и аннотации статей, ключевые слова, коды тематических направлений, данные об авторах статей и организациях, аффилированных с авторами. В базы загружаются библиографические списки литературы, используемой в статье, которые преобразуются в формальные структуры. Эти структуры допускают машинную обработку и обеспечивают отождествление элемента библиографического списка со статьей. В последние годы благодаря международным кодам DOI, ORCID, ROR качество обработки библиографических списков и процедур отождествления библиографических ссылок со статьями заметно улучшилось.

Развитые библиографические базы предоставляют своим многочисленным пользователям интернет-доступ к различным инструментам для поиска журналов, статей и авторов, интересующих читателя. На странице автора, найденной через поисковый сервис базы, любой посетитель может увидеть все публикации данного автора, размещенные в этой базе. Автору становится доступной информация о числе цитирований каждой его статьи. Доступны также ссылки на статьи, цитирующие публикации автора. Пользователь может осуществить поиск интересующего журнала или группы журналов по указанным поисковым атрибутам. Перейдя на страницу журнала, посетитель получает более подробные сведения о журнале и осуществляемой им редакционной политике.

Карточка журнала содержит информацию о тематике журнала — обычно приводятся тематические коды из используемого рубрикатора.

Данные, хранящиеся в библиографической базе, предоставляют пользователю статистические сведения и наукометрические показатели журнала, в частности: число опубликованных статей и цитирований статей журнала, позиции журнала в тематических рейтингах, распределение публикаций по авторам или организациям, по тематике публикуемых или цитирующих статей, распределение цитирующих публикаций по журналам или организациям и т. д.

Библиографическая база становится мощным инструментом, позволяющим пользователю, не имеющему специальной подготовки в области наукометрии, проводить несложные аналитические исследования. Как отмечено в [1], появилась новая целевая аудитория, которая не занимается наукометрическими исследованиями на профессиональном уровне, но проявляет большой интерес к наукометрическим показателям и рейтингам научных журналов. К этой аудитории относятся сотрудники организаций и фондов, финансирующих научные исследования, а также научные работники. Обладая общей информационной культурой, зная методологические принципы проведения научных исследований, имея опыт анализа и обработки данных, специалисты из разных научных областей активно интерпретируют результаты наукометрических исследований, получаемые профессионалами.

Следует отметить, что и в среде профессиональных специалистов в области наукометрии до сих пор не утихают споры по фундаментальным методологическим вопросам: идет поиск новых индикаторов, способствующих более адекватному сравнению показателей журналов, новых подходов, обеспечивающих аккуратную нормализацию цитируемости по разным предметным областям на уровне журналов и на уровне отдельных статей [2]. Приходит понимание, что наукометрические исследования должны уделять больше внимания географическим, социальным и языковым измерениям, что приведет к расширению набора показателей и аналитических приемов, используемых для оценки эффективности исследований [3].

Одним из наиболее востребованных сервисов библиографических баз является инструмент построения рейтингов журналов по выбранным показателям,

например, по известному широкой научной аудитории показателю импакт-фактора. Ученые, являясь заинтересованными потребителями результатов ранжирования журналов, в состоянии оценить адекватность рейтингов в своей научной области, построенных профессиональными наукометристами. Ученые имеют личный опыт подготовки и издания научных статей в журналах по своей тематике и, безусловно, знают не понаслышке авторитетные журналы своего научного направления.

Задача построения тематического рейтинга обычно делится на два этапа:

- выбор референтной группы журналов в исследуемом тематическом направлении,
- вычисление показателей импакт-фактора для журналов референтной группы на выбранном поле цитирующих журналов в определенном временном интервале.

Результаты рейтинга зависят от методических приемов и алгоритмических решений, которые принимаются на каждом из перечисленных этапов. Анализ методик построения рейтингов научных журналов в некоторых случаях позволяет выявить источники перекосов, ошибок или неточностей, обнаружить групповые интересы или предвзятости составителей рейтинга. Поэтому важными основополагающими принципами, сформулированными в Лейденском Манифесте наукометрии [4], объявлены открытость процедур сбора и анализа данных для проведения наукометрических оценок, а также возможность самим ученым-исследователям проверять правильность определения результатов наукометрического анализа.

Известно, что если поведение построенной математической модели не соответствует физическим реалиям, то теоретические положения, ограничения или допущения в математической модели требуют дополнительного тщательного анализа. Аналогично, если результаты тематического рейтинга научных журналов не отвечают ожиданиям специалистов данной научной области, следует проанализировать не только формулу вычисления рейтинга (она может быть весьма простой), но и методику отбора референтной группы, принятые ограничения (расширения) множества цитирующих журналов, способы учета особенностей цитируемости в различных научных областях.

1. РЕЙТИНГИ ЖУРНАЛОВ В РОССИЙСКОМ ИНДЕКСЕ НАУЧНОГО ЦИТИРОВАНИЯ

Рассмотрим инструменты, которые предоставляет Российский индекс научного цитирования (РИНЦ) [5], функционирующий на платформе библиографической базы eLibrary.ru [6]. В настоящее время в РИНЦ индексируется более 5700 научных журналов.

РИНЦ формирует рейтинги журналов по таким показателям как:

- число цитирований,
- двухлетний и пятилетний импакт-факторы с возможными ограничениями по типам научных изданий и области цитирования,
- индекс Херфиндаля по цитирующим журналам или организациям авторов,
- индекс Хирша

и другим наукометрическим показателям.

С помощью экранной формы РИНЦ можно построить общий рейтинг журналов по перечисленным показателям, а также тематические рейтинги для журналов определенного научного направления.

Построим рейтинг РИНЦ «Двухлетний импакт-фактор РИНЦ» по тематическому направлению «Математика» за 2019 год (рис. 1). В этом рейтинге при расчете показателей импакт-фактора учитываются цитирования из всех журналов, включенных в РИНЦ: первые 10 позиций занимают журналы, представленные в Таблице 1.

Результаты рейтинга, представленные в таблице, не вполне соответствуют ожиданиям ученых-математиков. Существующий математический портал Math-Net.ru [7] дает более адекватную картину на поле математических журналов. Наиболее известными и авторитетными в среде математиков являются журналы математической тематики, издаваемые Российской академией наук (РАН), Математическим институтом им. В.А. Стеклова РАН, Российской академией наук, Отделением математических наук и другими авторитетными академическими структурами.

В приведенном выше рейтинге РИНЦ обращает на себя внимание лидирующая позиция журнала «Геометрия и графика», который не относится к топовым

математическим журналам. Чтобы понять, в чем причина столь высоких показателей этого журнала в направлении «Математика», рассмотрим более внимательно анкету журнала и статистические отчеты о тематике издаваемых статей и тематике цитирующих публикаций.

Таблица 1

Рейтинг по двухлетнему импакт-фактору РИНЦ для журналов в тематическом направлении «Математика» за 2019 год (позиции рейтинга 1–10)

№ в рейтинге ИФ-2 РИНЦ	Название журнала	Значение показателя в рейтинге 2-ИФ РИНЦ
1.	Геометрия и графика	1,899
2.	Информатика и автоматизация	1,684
3.	Вестник Самарского государственного технического университета. Серия: Физико-математические науки	1,024
4.	Известия Российской академии наук. Серия математическая	0,978
5.	Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления	0,957
6.	Известия Иркутского государственного университета. Серия: Математика	0,932
7.	Вычислительная механика сплошных сред	0,760
8.	Математические заметки	0,757
9.	Экономика и математические методы	0,750
10.	Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications)	0,744

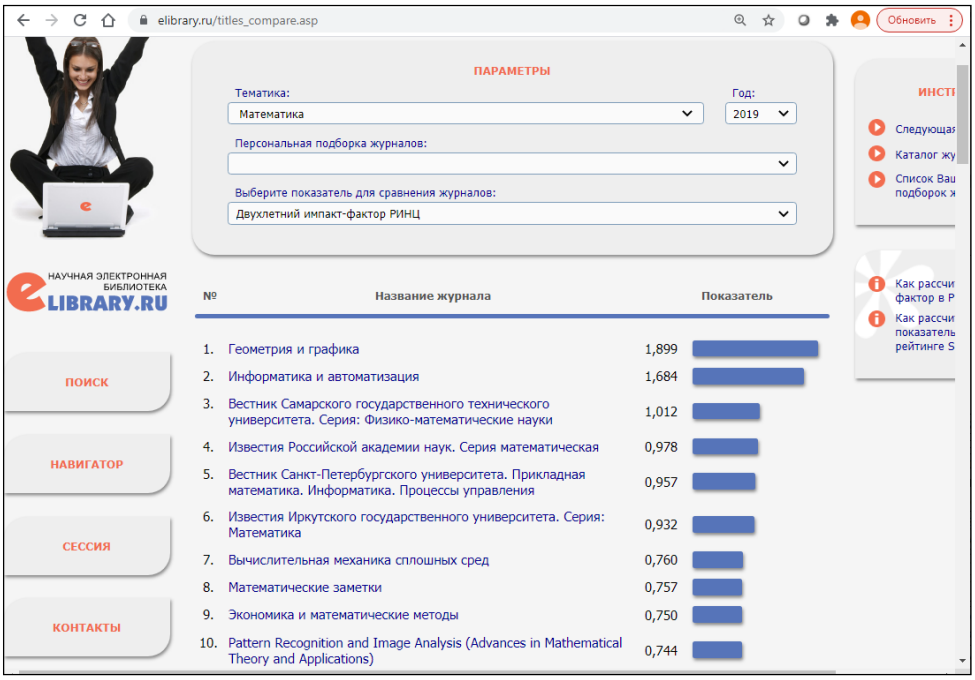


Рис. 1. Рейтинг по показателю «Двухлетний импакт-фактор РИНЦ» по тематическому направлению «Математика» за 2019 год.

Журнал «Геометрия и графика»

Учредителем журнала является частное лицо — В.И. Вышнепольский.

В анкете журнала указано, что этот научный журнал посвящен проблемам геометрии, черчения, компьютерной графики, преподаванию графических дисциплин и других тем, связанных с геометрией и графикой. Важным направлением является исследование отраслевых особенностей применения геометрии и компьютерной графики в строительстве, машиностроении, разработке программного обеспечения и т. д.

Журнал указал следующие тематические рубрики.

РУБРИКИ ГРНТИ:

143500. Высшее профессиональное образование. Педагогика высшей профессиональной школы

272100. Геометрия

281700. Теория моделирования

РУБРИКИ OECD:

101. Mathematics

503. Educational sciences

СПЕЦИАЛЬНОСТИ ВАК:

050100. Инженерная геометрия и компьютерная графика

Рассмотрим распределение статей журнала по тематике (рис. 2).

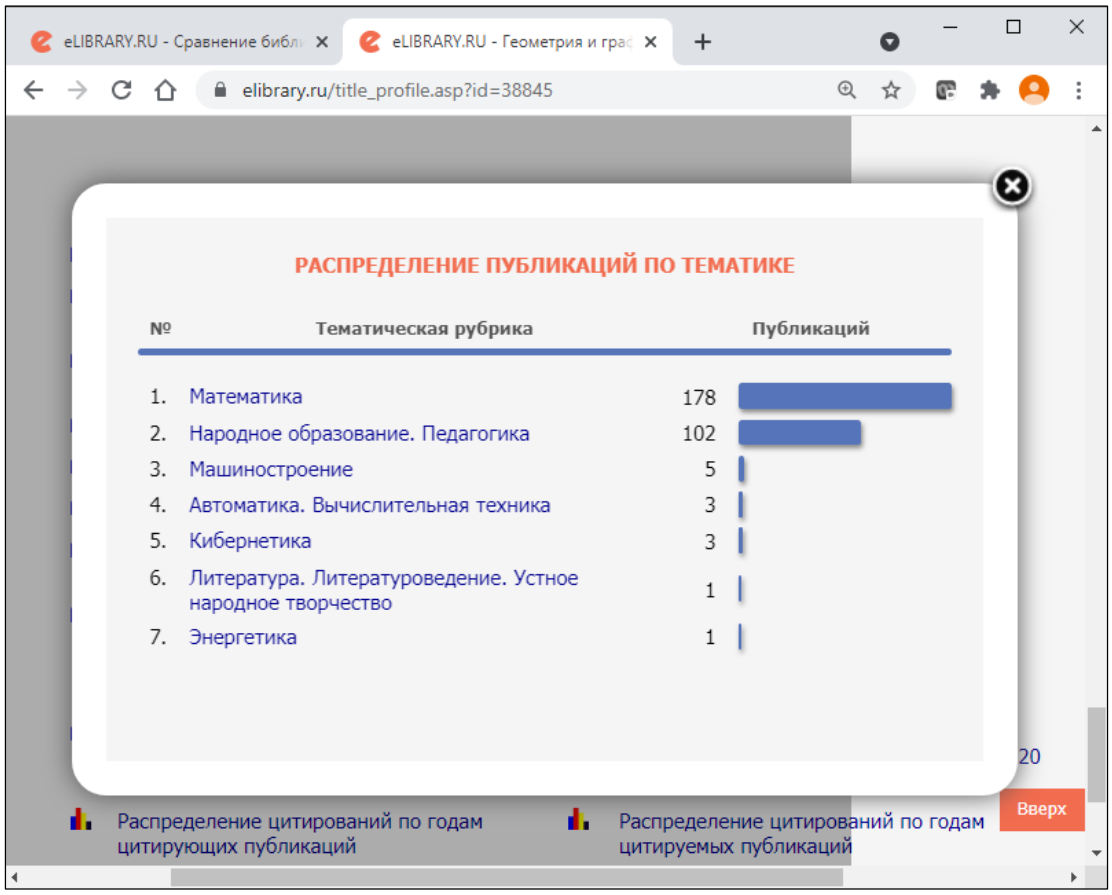


Рис. 2. Статистический отчет РИНЦ «Распределение публикаций по тематике» журнала «Геометрия и графика».

Будем считать, что в этом статистическом отчете, как и в аналогичных последующих, каждая статья заявлена по одной (главной) тематике. В то же время статья, вообще говоря, может быть приписана к нескольким тематическим направлениям. Как в этом случае будет выглядеть отчет? Поскольку разработчики не дают точного описания содержания отчета, предположение о множественном присутствии одной статьи в отчете приходится не принимать во внимание. Однако этой неопределенностью можно пренебречь, поскольку нас в первую очередь будут интересовать не столько точные подсчеты, сколько качественные оценки.

Всего статей, опубликованных в журнале «Геометрия и графика» — 293. По направлению «Математика» в отчете присутствуют 178 статей, т. е. 61%. всех статей. Однако, если обратиться к распределению цитирующих публикаций по тематике для журнала «Геометрия и графика» (рис. 3), обнаружатся иные пропорции.

Всего в отчете зафиксировано 832 цитирующих статьи. Цитирующих статей по теме «Математика» — 290 (35%). Как показано в работе [8], высокие показатели цитируемости журналу обеспечили статьи, относящиеся к тематическому направлению «Народное образование. Педагогика» (около 49% всех цитирующих статей). В РИНЦ предусмотрен прицельный отчет «Распределение цитирований по тематике цитирующих публикаций», который дал бы более полную картину по тематике цитирующих публикаций, но, к сожалению, этот отчет оказался недоступным.

Журнал «Геометрия и графика» в рейтинге по двухлетнему импакт-фактору РИНЦ 2019 года в направлении «Народное образование. Педагогика» занимает 10-е место. В этом направлении показатели цитируемости журналов значительно выше, чем в направлении «Математика». Так, лидер рейтинга журнал «Вестник Мининского университета» (Нижний Новгород) имеет показатель двухлетнего импакт-фактора 5,336. Показатели ведущих математических журналов значительно скромнее. Таким образом, высокий показатель журнала «Геометрия и графика» в рейтинге по импакт-фактору в направлении «Математика», как и следовало ожидать, во многом обеспечили цитирования из статей, не относящихся к математике.

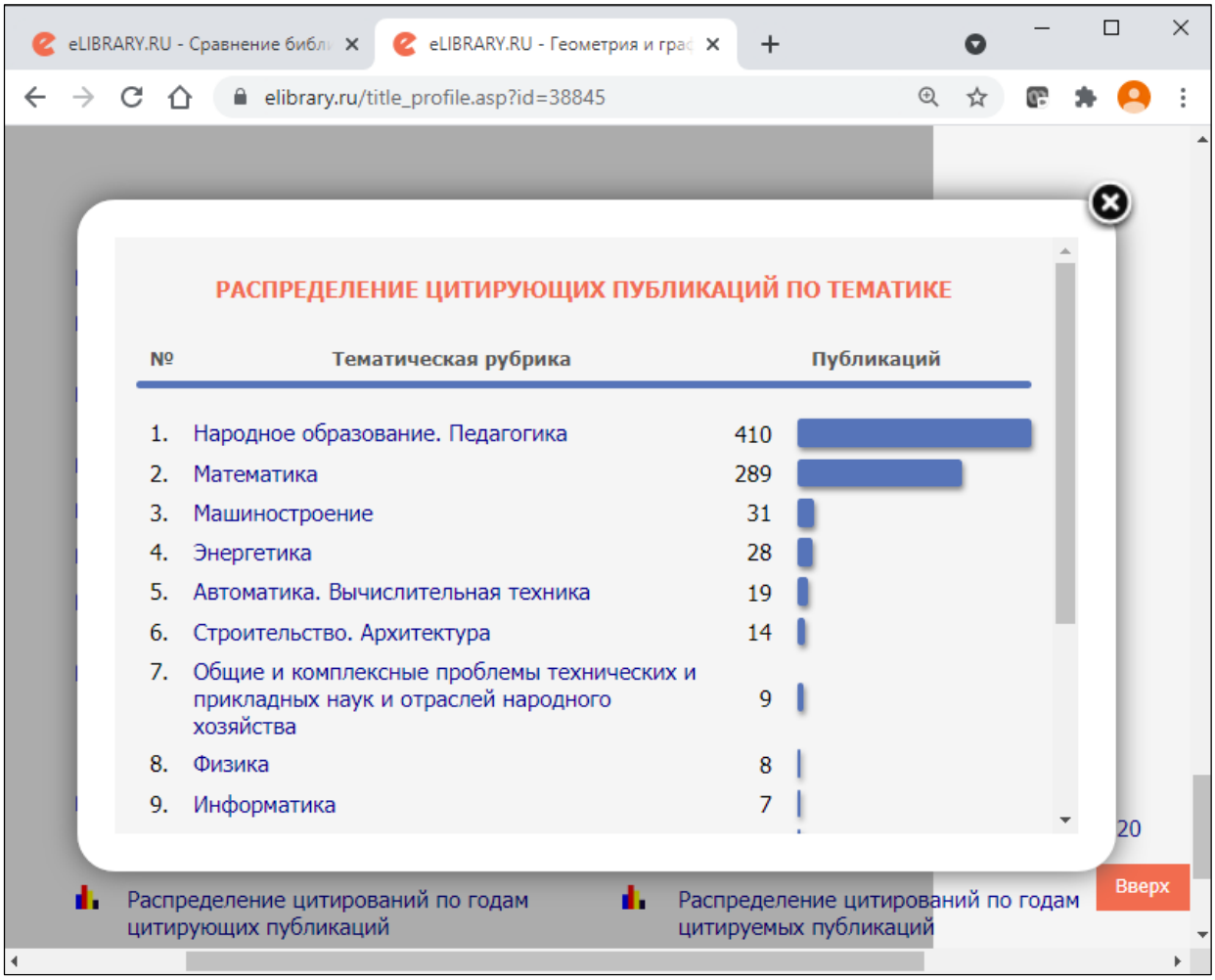


Рис. 3. Статистический отчет РИНЦ «Распределение цитирующих публикаций по тематике» журнала «Геометрия и графика».

В описании методики построения рейтингов РИНЦ по импакт-фактору нет каких-либо сведений об учете специфики цитирования в разных тематических направлениях. Также не сообщается, что методика предусматривает соответствующее нормирование показателей цитирования на уровне статей.

В результате рассмотрения статистических данных о тематике цитирующих статей можно сделать заключение, что на импакт-фактор журнала «Геометрия и графика» существенно повлияла мультидисциплинарность журнала. Журнал достаточно точно определил свои тематические рубрики, например, по классификатору OECD (“Mathematics”, “Educational sciences”). Однако на незаслуженно высокую позицию в рейтинге по импакт-фактору в направлении «Математика»

журнал выводят не относящиеся к делу ссылки из статей по тематике «Народное образование. Педагогика».

Возможен и другой подход к построению рейтинга. В работе [9] выдвинут тезис о том, что оценку влияния *автора* публикаций в некоторой научной области следует проводить на основе показателей цитируемости публикаций, относящихся к этой научной области. Например, авторитет автора в научном направлении «Физика» определяется на основе подсчета цитирований его публикаций по физике, и при этом не следует включать в расчет цитирования публикаций этого автора по другим научным направлениям. Эту идею можно распространить и на оценку влияния *журнала* в разных тематических направлениях. Если бы в подсчете рейтинга в разделе «Математика» учитывались массив статей только по теме «Математика» и цитирования только этих статей, то показатели у журнала «Геометрия и графика», по-видимому, были бы скромнее. И такая методика подсчета точнее определяла бы место журнала «Геометрия и графика» в рейтинге по математическому направлению.

Журнал «Информатика и автоматизация»

Издателем журнала является Санкт-Петербургский Федеральный исследовательский центр РАН. Анкета журнала декларирует, что он является научным, научно-образовательным, междисциплинарным журналом с базовой специализацией в области информатики, автоматизации и прикладной математики.

Журнал отнесен к следующим тематическим направлениям.

РУБРИКИ ГРНТИ:

500000. Автоматика. Вычислительная техника

270000. Математика

280000. Кибернетика

282300. Искусственный интеллект

РУБРИКИ OECD:

101. Mathematics

102. Computer and information sciences

202. Electrical engineering, electronic engineering

СПЕЦИАЛЬНОСТИ ВАК:

010102. Дифференциальные уравнения, динамические системы и оптимальное управление

010105. Теория вероятностей и математическая статистика

010109. Дискретная математика и математическая кибернетика

051301. Системный анализ, управление и обработка информации (по отраслям)

051311. Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

051315. Вычислительные машины, комплексы и компьютерные сети

051317. Теоретические основы информатики

051319. Методы и системы защиты информации, информационная безопасность

Журнал занимает 1-е место в интегральном рейтинге SCIENCE INDEX за 2020 год по тематике «Автоматика. Вычислительная техника» и 1-е место по тематике «Кибернетика». В интегральном рейтинге SCIENCE INDEX по направлению «Математика» журнал занимает 2-е место.

Рассмотрим статистический отчет журнала по тематике публикуемых статей (рис. 3). Данные отчета показывают, что доминирующим тематическим направлением журнала является направление Computer Science, включающее тематики «Автоматика. Вычислительная техника», «Кибернетика», «Информатика». Именно по этим трем тематикам опубликовано больше всего статей — 694. По теме «Математика» опубликовано 270 статей.

На рис. 4 представлен статистический отчет о распределении цитирующих статей по тематике.

Доля статей по теме «Математика» от числа статей по темам Computer Science (трем рубрикам «Автоматика. Вычислительная техника», «Кибернетика» и «Информатика») составляет 39%. Доля цитирующих статей по теме «Математика» по отношению к числу цитирующих статей по темам Computer Science составляет только 22%. Таким образом, в журнале «Информатика и автоматизация» математические статьи цитируются хуже, чем статьи по темам Computer Science. Этот вывод совпадает с общепринятыми представлениями о более скромном цитировании статей по математике. Следовательно, можно утвер-

ждать, что учет статей по темам Computer Science и их цитирований заметно повышает показатель журнала «Информатика и автоматизация» в рейтинге по импакт-фактору в направлении «Математика».

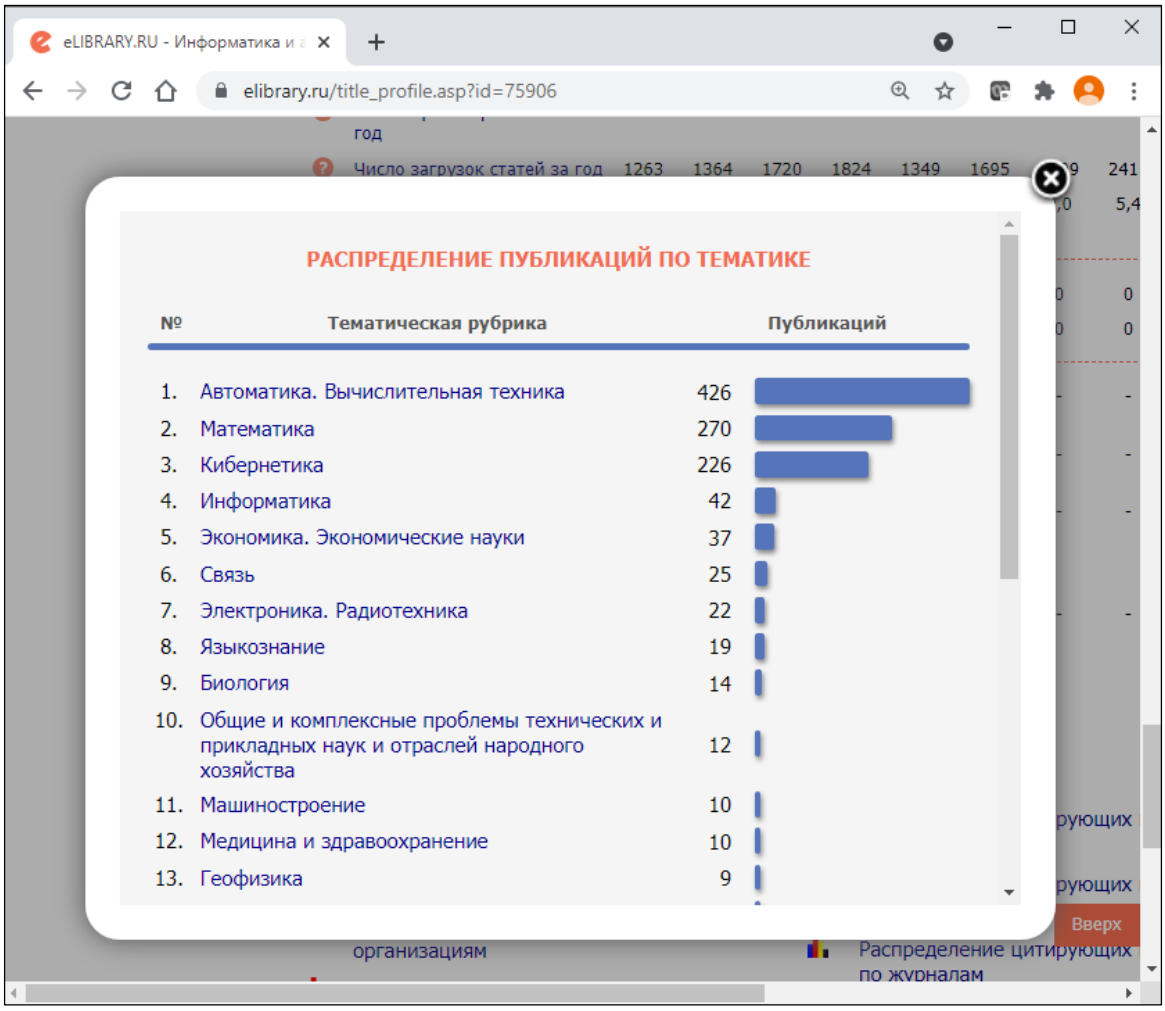


Рис. 3. Статистический отчет РИНЦ «Распределение публикаций по тематике» журнала «Информатика и автоматизация».

Если бы при расчете рейтинга журналов по импакт-фактору в разделе «Математика» учитывались только статьи по теме «Математика» и их цитирования и не учитывались статьи и цитирования по другим темам (в частности, по темам Computer Science), то показатель импакт-фактора по теме «Математика» журнала «Информатика и автоматизация» был бы ниже.

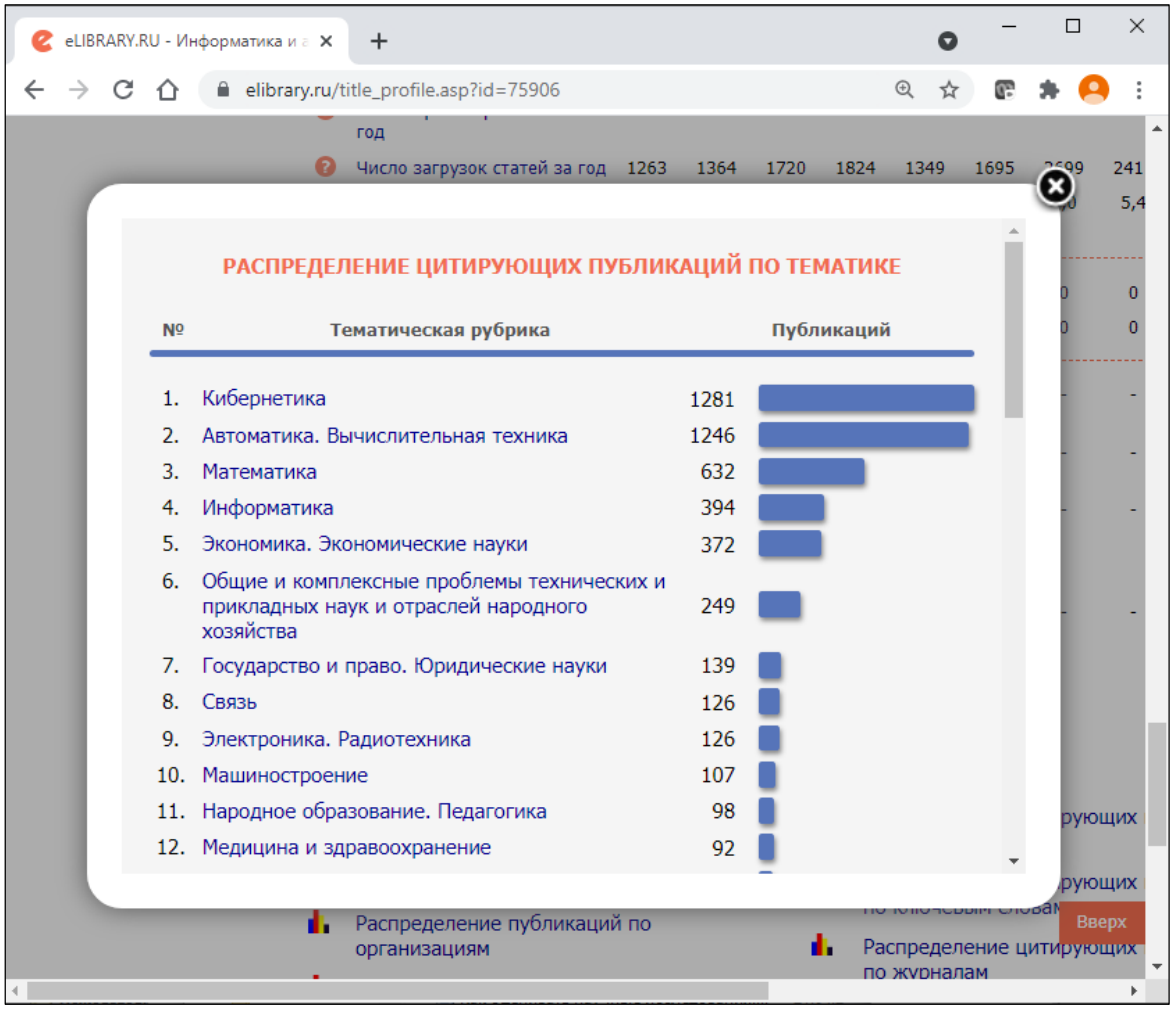


Рис. 4. Статистический отчет РИНЦ «Распределение цитирующих публикаций по тематике» журнала «Информатика и автоматизация».

Журнал «Вестник Самарского государственного технического университета. Серия: Физико-математические науки»

Журнал публикует оригинальные статьи и заказные обзоры по направлениям «Дифференциальные уравнения и математическая физика», «Механика деформируемого твёрдого тела», «Математическое моделирование, численные методы и комплексы программ».

РУБРИКИ ГРНТИ:

272900. Обыкновенные дифференциальные уравнения

273100. Дифференциальные уравнения с частными производными

273300. Интегральные уравнения

273500. Математические модели естественных наук и технических наук. Уравнения математической физики

274100. Вычислительная математика

301900. Механика деформируемого твердого тела

305100. Комплексные и специальные разделы механики

301700. Механика жидкости и газа

281500. Теория систем автоматического управления

281700. Теория моделирования

РУБРИКИ OECD:

101. Mathematics

203. Mechanical engineering

205. Materials engineering

СПЕЦИАЛЬНОСТИ ВАК:

010102. Дифференциальные уравнения, динамические системы и оптимальное управление

010204. Механика деформируемого твердого тела

051318. Математическое моделирование, численные методы и комплексы программ

В анкете журнала в РИНЦ отмечено, что Вестник не является мультидисциплинарным журналом, но, по-видимому, эта декларация нужна для того, чтобы не относить Вестник к группе мультидисциплинарных журналов РИНЦ, выделенных в отдельную категорию. Формальных критериев отнесения журнала к мультидисциплинарной группе РИНЦ не формулирует. Можно построить в РИНЦ рейтинг для мультидисциплинарных журналов, выделенных в отдельную группу. Обращает на себя внимание тот факт, что в этот рейтинг будут включены журналы, публикующие статьи в широком диапазоне естественно-научных, технических, гуманитарных и общественных направлений. Вестник, вообще говоря, не вписывается в такую группу мультидисциплинарных журналов.

Рассмотрим статистический отчет журнала по тематике публикуемых статей (рис. 5). В журнале доминируют статьи по трем темам: «Математика» «Механика» и «Физика». Больше всего статей в журнале опубликовано по теме «Ма-

тематика» — 639. Число статей по темам «Механика» и «Физика» в сумме немного больше, чем число статей по математике — 656.

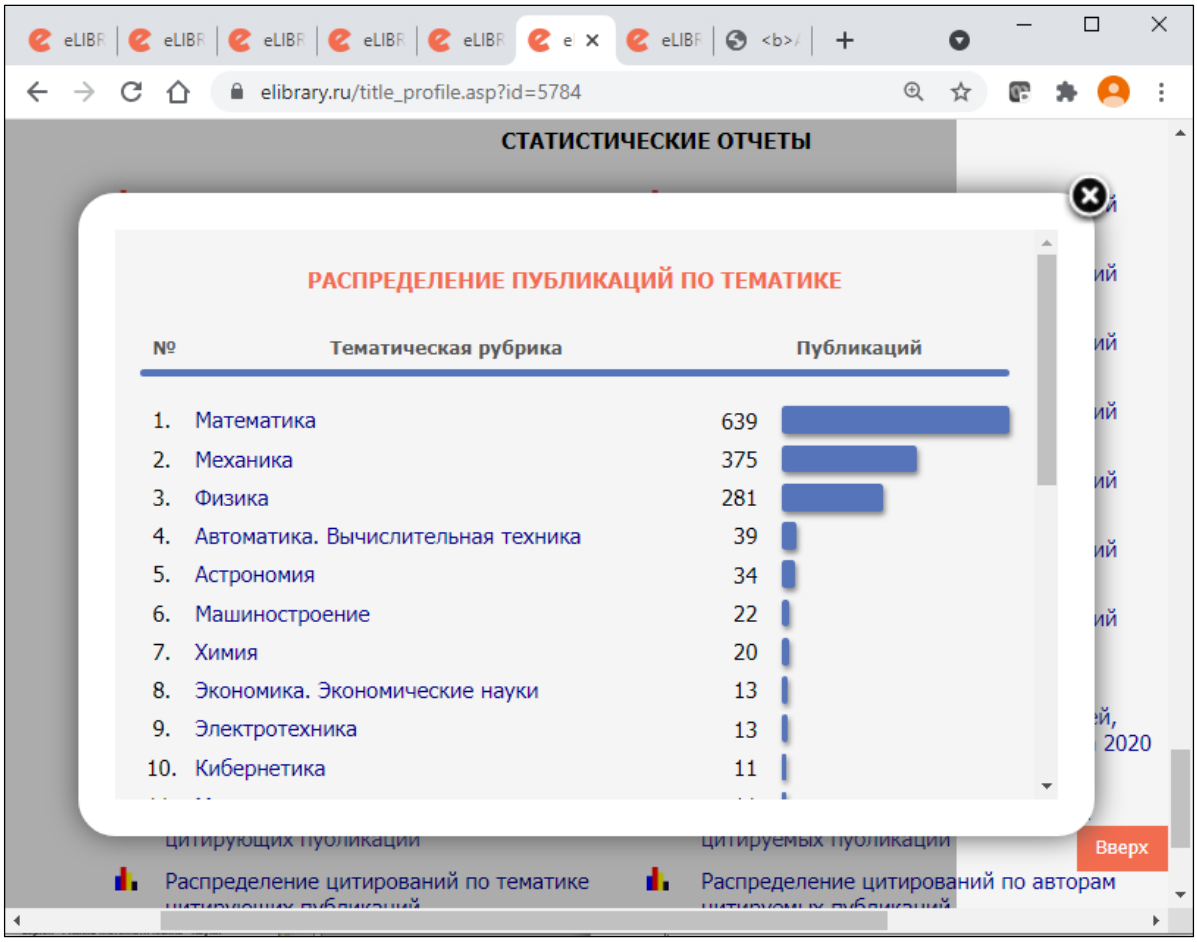


Рис. 5. Статистический отчет РИНЦ «Распределение публикаций по тематике» журнала «Вестник Самарского государственного технического университета. Серия: Физико-математические науки».

Рассмотрим статистический отчет по тематике цитирующих статей (рис. 6).

Доминирующими ожидаемо стали следующие три тематических направления: «Математика», «Механика» и «Физика». Тематическое направление «Физика» не прозвучало в декларациях редакции на сайте журнала, но этому факту есть простое объяснение. Математика и физика в научной среде считаются близкими направлениями. Физики активно используют математический аппарат для получения фундаментальных и прикладных результатов. Математики развивают математические методы для решения физических задач.

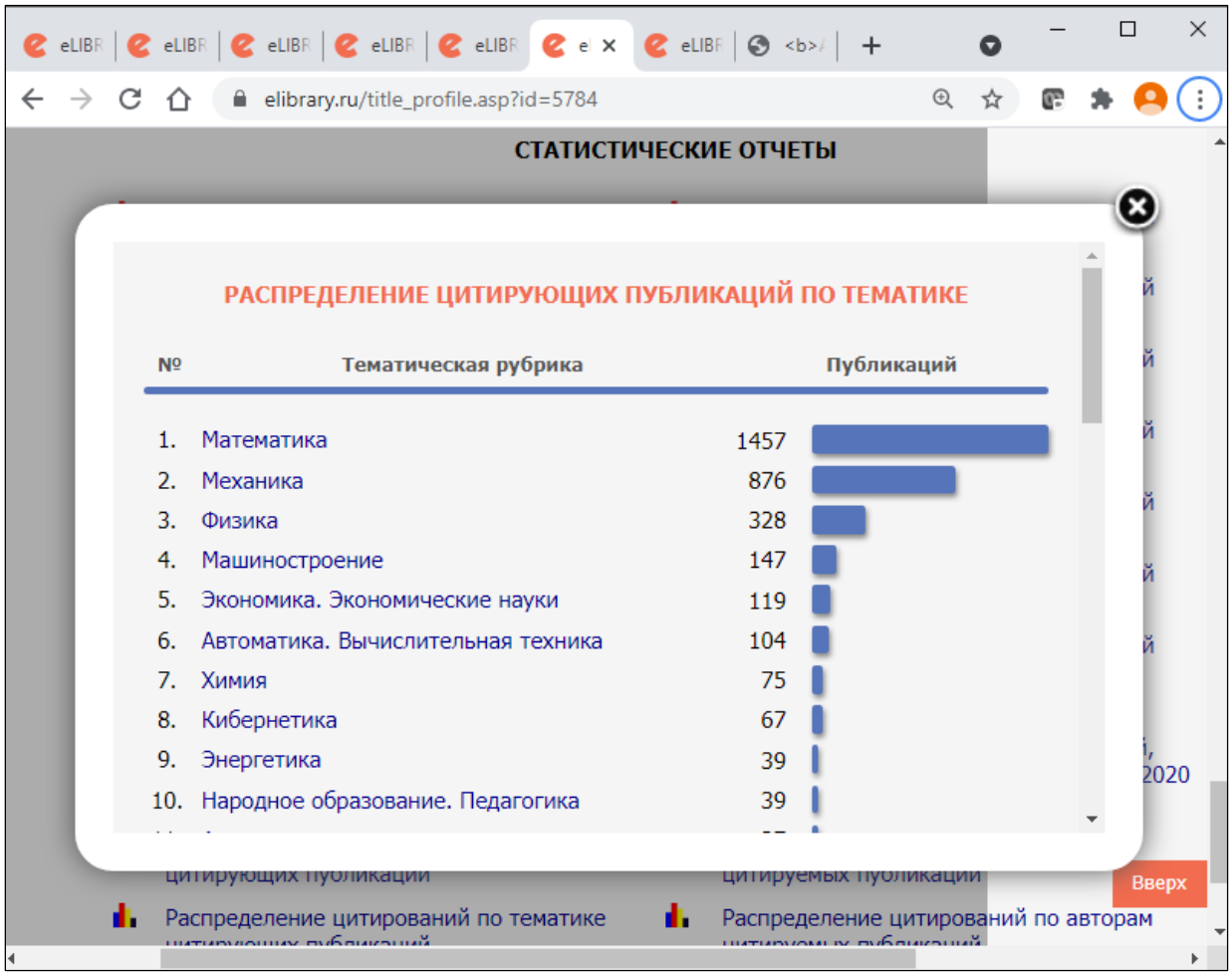


Рис. 6. Статистический отчет РИНЦ «Распределение цитирующих публикаций по тематике» журнала «Вестник Самарского государственного технического университета. Серия: Физико-математические науки».

Если перевести фокус на проблемы тематических рейтингов, то можно отметить более высокую цитируемость статей по направлению «Физика» по сравнению с цитируемостью статей по направлению «Математика». Но, скорее всего, цитируемость статей журнала в направлении «Физика» сильно зависит от конкретной темы исследований.

Появление Вестника на третьей позиции тематического рейтинга в направлении «Математика» не вызывает удивления или резкого отторжения. Однако и здесь более адекватные показатели рейтинга в направлении «Математика» могут быть получены в случае, если при расчете показателя импакт-фактора учитывать только статьи по теме «Математика» и цитирования этих статей.

Журнал «Известия Российской академии наук. Серия математическая»

Издателями журнала являются Математический институт им. В.А. Стеклова РАН (Москва) и Российская академия наук. Журнал публикует статьи по всем разделам современной математики. Особое внимание уделяется алгебре, математической логике, теории чисел, математическому анализу, геометрии, топологии, дифференциальным уравнениям. Редакция заявляет раздел «Математика» по трем классификаторам: ГРНТИ (категория 270000), OECD (категория 101. Mathematics), специальности ВАК (категория 010100.)

На рис. 7 представлен статистический отчет РИНЦ о тематике статей, публикуемых в журнале, на основе собранных данных за весь период его размещения в eLibrary.

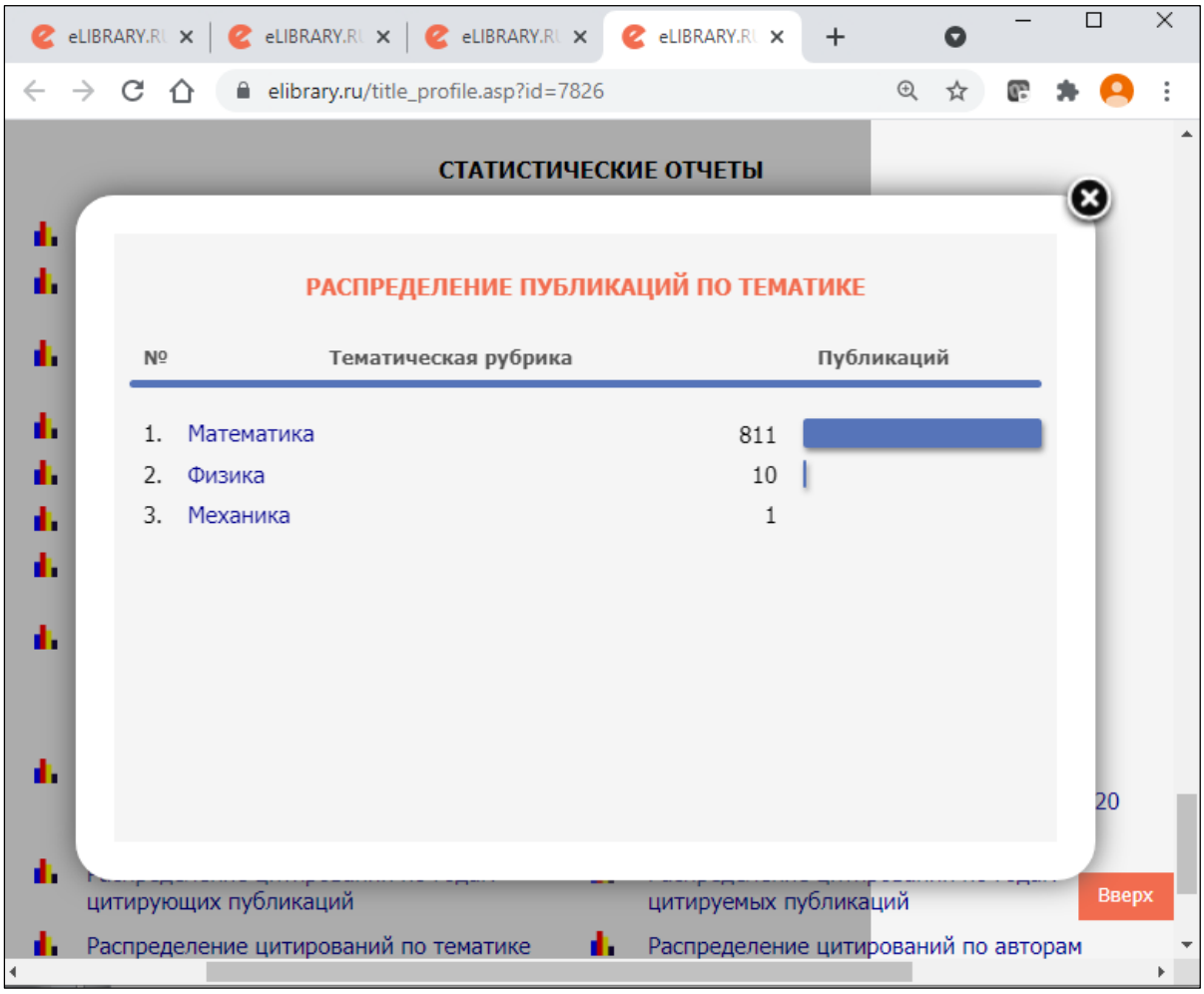


Рис. 7. Статистический отчет РИНЦ «Распределение публикаций по тематике» журнала «Известия Российской академии наук. Серия математическая».

Из представленных статистических данных следует, что практически все статьи, публикуемые в журнале, относятся к теме «Математика»: по этой теме опубликовано 811 статей, по теме «Физика» — 10, и по теме «Механика — 1.

Если посмотреть статистический отчет по цитирующим статьям, то вновь можно убедиться в том, что по теме «Математика» зафиксировано подавляющее большинство статей — 18734. На второй позиции отчета указаны 3483 статьи по теме «Физика», по теме «Автоматика. Вычислительная техника» — 241, теме «Химия» — 210, по теме «Механика» — 182. По остальным тематикам, присутствующим в отчете, зафиксировано меньше сотни цитирующих статей.

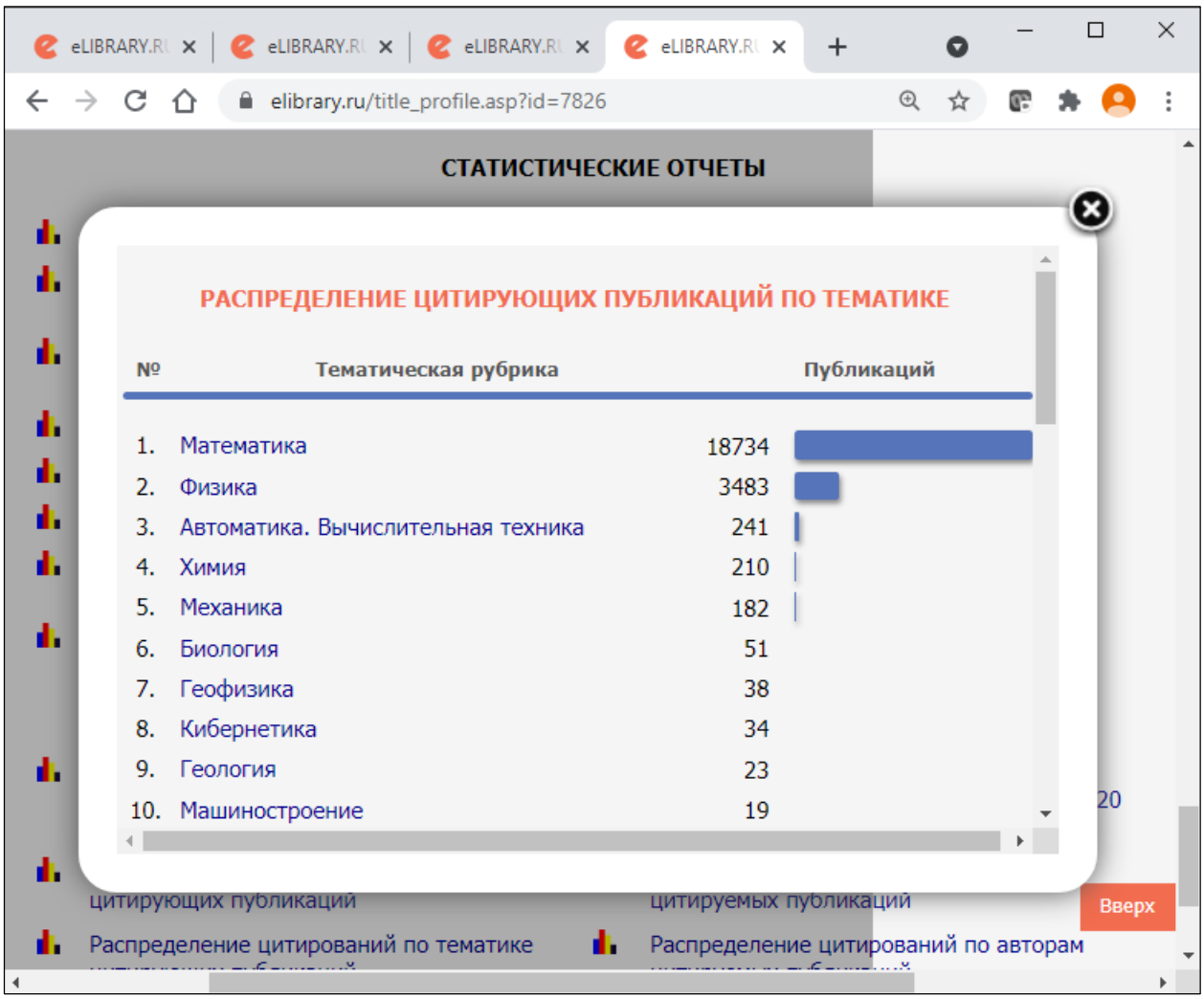


Рис. 8. Статистический отчет РИНЦ «Распределение цитирующих публикаций по тематике» журнала «Известия Российской академии наук. Серия математическая».

Подведем итог рассмотрения статистических данных по тематике публикуемых и цитирующих статей для четырех журналов, занявших лидирующие позиции в рейтинге РИНЦ по двухлетнему импакт-фактору в направлении «Математика».

Первые две позиции в рейтинге по двухлетнему импакт-фактору, занятые журналами «Геометрия и графика» и «Информатика и автоматизация», очевидно, не соответствуют ожиданиям специалистов из области «Математика». Причина столь высоких показателей этих журналов состоит в том, что при расчете показателей рейтинга учитывались статьи, не относящихся к направлению «Математика», и цитирования этих статей. Рассмотренные два журнала обладают выраженной мультидисциплинарностью в том смысле, что помимо направления «Математика» в журналах дополнительно присутствуют другие, хорошо цитируемые научные дисциплины. В обоих журналах математика не является доминирующей дисциплиной.

Журнал «Вестник Самарского государственного технического университета. Серия: Физико-математические науки» также является мультидисциплинарным: в журнале были опубликованы в большом количестве статьи по двум другим дисциплинам — «Физика», «Механика». Но доминирующим тематическим направлением является «Математика». Эффект влияния на показатели импакт-фактора статей и цитирований, относящихся к двум дисциплинам («Физика», «Механика»), не столь очевиден. Более точно определить влияние этих двух направлений можно только при получении доступа к данным о цитированиях по каждому тематическому направлению.

Журнал «Известия Российской академии наук. Серия математическая» можно назвать классическим математическим журналом: число статей по второй тематике (физике) в этом журнале незначительно. Журнал хорошо известен в кругу профессиональных математиков. Этот журнал вполне мог бы претендовать на лидирующую позицию в тематическом рейтинге по направлению «Математика». Но в рассмотренном рейтинге журналу приходится соревноваться с мультидисциплинарными журналами, включающими статьи по тематикам, которые традиционно имеют более высокие показатели цитируемости.

2. ПРОБЛЕМА ВЫБОРА РЕФЕРЕНТНОЙ ГРУППЫ ЖУРНАЛОВ

На наш взгляд, причина выхода непрофильных журналов на первые позиции в рейтинге по направлению «Математика» кроется в методике формирования референтной группы журналов. Референтная группа — это множество «похожих» журналов, относящихся к общей тематике, определяемой тематическим классификатором базы. Как поступать в том случае, когда журнал содержит статьи по другим тематикам, не относящимся к тематике рейтинга?

Известно, что в РИНЦ ведена категория «Мультидисциплинарные журналы». По-видимому, к этой категории отнесены журналы, которые декларируют мультидисциплинарность в своей редакционной политике и инициативно объявляют себя «мультидисциплинарными». Такие журналы, как правило, ориентированы на широкий круг научных направлений, относящихся к тематическим категориям верхнего уровня классификатора (такую мультидисциплинарность условно можно назвать «сильной»).

Рассмотренные нами журналы имеют разную степень мультидисциплинарности. Журнал «Геометрия и графика» можно отнести к журналам с «сильной» мультидисциплинарностью, поскольку он содержит две доминирующие тематики из разных категорий верхнего уровня классификатора. Журналы «Информатика и автоматизация» и «Вестник Самарского технического университета» можно отнести к журналам, имеющим «среднюю» мультидисциплинарность. Классический математический журнал «Известия Российской академии наук. Серия математическая» относится в предложенной шкале скорее к журналам со «слабо выраженной» мультидисциплинарностью. В рейтинге РИНЦ по двухлетнему импакт-фактору первые три позиции заняли журналы с высокой и средней степенями мультидисциплинарности.

Мультидисциплинарность зависит от выбранной системы тематической классификации. На рассмотренных примерах мы видим, что разработчики РИНЦ опираются на несколько систем классификации:

- ГРНТИ — российский Государственный рубрикатор научно-технической информации, представляющий собой универсальную иерархическую классификацию областей знания, принятую для систематизации научно-технической информации в России и государствах СНГ,

- OECD — международный классификатор OECD (Organization for Economic Cooperation and Development),
- номенклатуру специальностей Высшей аттестационной комиссии (ВАК),

а также на свой собственный рубрикатор, сочетающий в себе подходы перечисленных классификаторов, адаптированные для решения конкретных потребностей информационных сервисов. Вероятно, в РИНЦ немалую роль играют экспертные оценки и особые решения при отнесении журнала к той или иной тематической категории. К сожалению, РИНЦ не предоставляет пользователю достаточных справочных материалов с обоснованием выбора или модификации конкретного классификатора и описанием применяемых методик формирования референтных групп.

Существует проблема, связанная со структурой используемого классификатора. Приведем пример. При расчете показателей журналов в некоторых типах рейтингов РИНЦ в одну рубрику попадают два тематических класса: “Mathematics” и “Computer and information sciences”. В классификаторе OECD эти направления выделены в два самостоятельных направления. Известно, что статьи по теме “Computer and information sciences” традиционно имеют более высокие показатели цитирования по сравнению с математическими статьями. Объединение этих двух направлений в РИНЦ могло привести, в частности, к тому, что в тематическом рейтинге по «Математике» журнал «Информатика и автоматизация» имеет лучшие показатели по сравнению с показателями классических математических журналов. Таким образом, от выбора классификатора зависят показатели эффективности журналов (позиции в тематических рейтингах).

В справочных материалах РИНЦ дана информация о применении методов нормализации показателей цитирования, учитывающих специфику тематических категорий на уровне журналов, при построении отдельных видов рейтингов журналов. Однако нормализация показателей цитирования на уровне отдельных статей, судя по справочным материалам РИНЦ, не применяется. При рассмотрении результатов рейтинга РИНЦ по импакт-фактору в направлении «Математика» было показано, что именно «непрофильные» статьи и цитирования могут незаслуженно вывести «непрофильные» журналы на первые позиции чужого рейтинга.

3. О КЛАССИФИКАТОРАХ

Как отмечают многие специалисты в области наукометрии, не существует идеальных классификаторов. В практической работе и интерпретации результатов работы сервисов библиографической базы приходится учитывать конкретные особенности выбранных систем тематической классификации.

Так, многие библиографические базы используют систему классификации Web of Science. Однако эта система имеет свои особенности, в частности, она не позволяет однозначно классифицировать содержание междисциплинарных журналов, таких как Science и Nature: эти и подобные журналы сложно отнести к одной или нескольким категориям. Следует иметь в виду, что тематические области в библиографической базе, как и категории Web of Science, как правило, предназначены для организации поиска информации, и их использование для иной цели может иметь нежелательные последствия.

Тема применения классификаторов является весьма чувствительной для разработчиков библиографических баз. В настоящее время существуют десятки и сотни классификаторов, привязанных к крупным библиотекам, известным библиографическим базам. Многие страны разрабатывают свои собственные национальные классификаторы научных областей. Существует проблема гармонизации различных систем тематической классификации.

В недавнем отчете компании Clarivate [10] представлен альтернативный подход к формированию референтных групп, который можно характеризовать как подход «снизу вверх». Предлагаемый подход альтернативен принятому сейчас подходу, основанному на использовании тематических классификаторов научных дисциплин (этот способ классификации можно назвать подходом «сверху вниз»).

Подход «снизу вверх» предлагает использовать в качестве меры тематической близости статей их библиографические связи: прямое цитирование, пересечения пристатейных библиографических списков (со-цитирование). В единый тематический кластер объединяются наиболее близкие статьи. Тематический кластер постоянно увеличивается по мере появления новых цитирующих статей. Есть и другой эффект: с появлением новых цитирований статья может оказаться включенной дополнительно в другой тематический кластер. Тематические кла-

стеры нижнего уровня объединяются в более крупные кластеры посредством предъявления более слабых требований к близости публикаций. Таким образом формируются кластеры более высокого уровня, отражающие более общие тематические категории.

Кластеры разных уровней получают названия (метки). Для кластеров нижнего уровня метка может формироваться автоматически, например, на основе наиболее значимых ключевых слов из статей кластера. Для кластеров среднего и высшего уровней метки (категории) могут назначаться экспертами в предметной области, привязываясь к привычным или удобным названиям, отражающим тематику статей, сгруппированных в кластере. Эксперты при этом могут опираться на характерные ключевые слова и привычные для ученых категории, в частности, категории Web of Science. Строящиеся таким образом тематические кластеры получили название “Citation Topics” («Темы цитирования») [10].

В декабре 2020 года такой новой системой классификации Citation Topics, разработанной совместно с ведущей научной группой из Центра научно-технологических исследований (CWTS) Лейденского университета (Нидерланды), был дополнен аналитический продукт InCites компании Clarivate. InCites — система, позволяющая получать разнообразные библиометрические показатели авторов, организаций, журналов, а также проводить многоаспектные наукометрические исследования. InCites предоставлял пользователю возможность отбора статей по классификаторам, построенным по принципу «сверху вниз». Теперь наряду с этим система включает классификатор Citation Topics [11], построенный по принципу формирования категорий «снизу вверх».

В данный момент алгоритмический классификатор Citation Topics включает 10 категорий высокого уровня, 326 категорий среднего уровня и 2444 категории (кластера) нижнего уровня. Алгоритм классификатора обработал более 60 млн статей, и более 50 млн были отнесены к тому или иному тематическому кластеру. Несмотря на принципиальное различие между методологиями классификации «снизу вверх» (кластеризация цитирований) и классификации «сверху вниз» (категории журналов), группы в этих структурах во многом совпадают.

Создание тематических классификаторов в описанной технологии «снизу-вверх» снимает некоторые проблемы, свойственные классификаторам, постро-

енным по принципу «сверху-вниз». В частности, технология тематической классификации «снизу вверх», на наш взгляд, дает более надежную объективную основу для формирования референтных групп с целью дальнейшего ранжирования или сравнения библиометрических показателей журналов.

В работе [12] представлены результаты исследования графа цитирования для журналов физико-математического направления, входящих в библиографическую базу Math-Net.Ru. Исследование проведено для 120 журналов, общий объем обработанных библиографических ссылок — около 96 тысяч. Построенный граф цитирования, идеологически близкий кластеризации Citation Topics, показал достаточно высокую плотность, свидетельствующую об активном взаимном цитировании статей, опубликованных в журналах. Граф цитирований распался на пять модулей (кластеров), которые, как оказалось, могут быть интерпретированы по направлениям исследований: фундаментальная математика, математическое моделирование, экспериментальная и теоретическая физика, дискретная математика и прикладная математика и компьютерные науки. Для графа цитирования был построен рейтинг журналов на основе аналога показателя PageRank [13]. Оказалось, что полученный рейтинг вполне согласуется с представлениями российских математиков об авторитете и популярности математических журналов, участвующих в рейтинге.

ЗАКЛЮЧЕНИЕ

Одна из основных проблем продуктивного построения рейтинга журналов в тематическом направлении связана с адекватным выделением референтной группы журналов. Тематика журналов такой группы определяется системой классификации научных направлений, используемой в библиографической базе. Недостаток существующих систем классификации, приводящий к появлению неубедительных рейтингов, заключается в их декларативном характере, основанном на классификации «сверху вниз». Декларируемое журналом тематическое направление нередко исходит из субъективных оценок и не отражает тематики публикуемых статей. Технология формирования тематических кластеров по принципу «снизу вверх» существенно точнее и объективнее структурирует множество тематически разнообразных журнальных статей. Тематическая близость статей определяется здесь посредством анализа близости библиографических

списков, выявления прямых ссылок и со-цитирований. Вычленение референтных групп журналов на основе подхода «снизу вверх» является более надежным фундаментом для построения тематических рейтингов, позволяя конструктивно учитывать мультидисциплинарность на уровне статей.

СПИСОК ЛИТЕРАТУРЫ

1. *Leydesdorff L., Wouters P., Bornmann L.* Professional and Citizen Bibliometrics: Complementarities and Ambivalences in the Development and Use of Indicators – A State-of-the-Art Report // Scientometrics. 2016. Vol. 109. No. 3. P. 2129–2150. https://doi.org/10.1007/s11192-016-2150-8

2. *Москалева О.В.* Научные публикации как средство коммуникации. В книге: Руководство по наукометрии: индикаторы развития науки и технологии. Акоев М.А., Маркусова В.А., Москалева О.В., Писляков В.В. Екатеринбург, 2021. С. 140–176. https://doi.org/10.15826/B978-5-7996-3154-3

3. *Akoev M., Markusova V., Moskaleva O., Pislyakov V.* Handbook on Scientometrics: Science and Technology Development Indicators, Second edition. 2021. 358 p. https://doi.org/10.15826/B978-5-7996-3154-3

4. *Hicks D., Wouters P., Waltman L. et al.* Bibliometrics: The Leiden Manifesto for research metrics. Nature. Vol. 520, 429–431 (2015). https://doi.org/10.1038/520429a

5. Российский индекс научного цитирования. URL: https://www.elibrary.ru/project_risc.asp

6. Научная электронная библиотека eLibrary.ru. URL: https://www.elibrary.ru/defaultx.asp

7. *Общероссийский портал Math-Net.Ru*. URL: http://www.mathnet.ru/

8. *Полилова Т.А.* Рейтинги журналов в РИНЦ как инструменты анализа и влияния // Препринты ИПМ им. М.В. Келдыша. 2021. № 40. 35 с. https://doi.org/10.20948/prepr-2021-40

9. *Vinkler P.* Evaluation of publications by the part-set method // Scientometrics. 2021. V. 126. P. 2737–2757. https://doi.org/10.1007/s11192-020-03841-7

10. *Шомшор М. (Martin Szomszor), Адамс Д. (Jonathan Adams), Пендлбери Д. (David A. Pendlebury), Роджерс Г. (Gordon Rogers)*. Классификация

данных: как делать осознанный выбор, ведущий к желаемым результатам. Отчет о международном исследовании Института научной информации (ISI).
URL: https://img06.en25.com/Web/ClarivateAnalytics/%7B3f970f80-b453-451f-80a2-8d1792213e83%7D_Clarivate_ISI_Data_categorization_report_RU.pdf

11. *Potter I*. Introducing Citation Topics in InCites – Clarivate. Dec 3, 2020. URL: https://clarivate.com/blog/introducing-citation-topics/

12. *Печников А.А., Чебуков Д.Е.* Структура графа цитирования журналов Math-Net.Ru // Научный сервис в сети Интернет: труды XXIII Всероссийской научной конференции (20–24 сентября 2021 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2021. https://doi.org/10.20948/abrau-2021-2

13. *PageRank.* Wikipedia, the free encyclopedia.
URL: https://en.wikipedia.org/wiki/PageRank

THE RATING OF THE JOURNAL IN THE BIBLIOGRAPHIC DATABASE

M. M. Gorbunov-Posadov [1] [0000-0002-7044-8287], T. A. Polilova[2] [0000-0003-4628-3205]

[1, 2]Keldysh Institute of Applied Mathematics, Miusskaya sq., 4, Moscow, 125047, Russia

[1]gorbunov@keldysh.ru, [2]polilova@keldysh.ru

Abstract

The tool for building ratings of scientific journals is one of the popular services of bibliographic databases. The task of building a rating is usually divided into two main subtasks: determining the reference group of journals and calculating the rating indicator for journals of this reference group. Practice shows that for the correct comparison of journals, a necessary condition is to limit the reference group to exclusively journals of a certain subject. In the case of methodological errors made at the stage of selecting a reference group, the values of the journal index in the rating may differ greatly from the expected ones.

For example, in the ranking of journals in the Russian Science Citation Index (RSCI) according to the two-year impact factor in the thematic area “Mathematics”, classical fundamental mathematical journals, contrary to expectations, do not reach

the first positions of the rating. The first positions were taken by journals for which mathematics is not the dominant profile discipline. Analysis of statistical data on the subject of published articles and citations in journals that occupy leading positions in the RSCI rating shows that the multidisciplinary nature of these journals significantly influenced the rating indicators.

The noted misunderstanding leads to the idea that in this case, not all the articles of the journal should have been involved in the calculation of the rating, but only those related to this thematic area. At the same time, the existing scheme of thematic classification of directions also raises questions. The "bottom-up" classification, which is gaining popularity and works on a representative array of articles, seems to be more promising. Here thematic clusters are isolated on the basis of the concept of proximity of articles, interpreted as the proximity of their bibliographic links. And further, the thematic affiliation of the article is not assigned by the volitional decision of the author or the editorial board, but is strictly formally calculated on the basis of its bibliographic list.

Keywords*: scientific publication, citation, rating of journals, thematic classification, impact factor, multidisciplinary, bibliographic reference, co-citation, bottom-up classification, thematic clustering, Citation Topics.*

REFERENCES

1. *Leydesdorff L., Wouters P., Bornmann L.* Professional and Citizen Bibliometrics: Complementarities and Ambivalences in the Development and Use of Indicators – A State-of-the-Art Report // Scientometrics. 2016. Vol. 109, No. 3. P. 2129–2150. https://doi.org/10.1007/s11192-016-2150-8

2. *Moskaleva O.V.* Nauchnye publikatsii kak sredstvo kommunikatsii. V knige: Rukovodstvo po naukometrii: indikatory razvitiia nauki i tekhnologii. Akoev M.A., Markusova V.A., Moskaleva O.V., Pisliakov V.V. Ekaterinburg, 2021. S. 140–176. https://doi.org/10.15826/B978-5-7996-3154-3.

3. *Akoev M., Markusova V., Moskaleva O., Pislyakov V*. Handbook on Scientometrics: Science and Technology Development Indicators, Second edition. 2021. 358 p. https://doi.org/10.15826/B978-5-7996-3154-3.

4. *Hicks D., Wouters P., Waltman L. et al.* Bibliometrics: The Leiden Manifesto for research metrics // Nature. 2015. V. 520. P. 429–431. https://doi.org/10.1038/520429a

5. Rossiiskii indeks nauchnogo tsitirovaniia. URL: https://www.elibrary.ru/project_risc.asp

6. Nauchnaia elektronnaia biblioteka eLibrary.ru. URL: https://www.elibrary.ru/defaultx.asp

7. Obshcherossiiskii portal Math-Net.Ru. URL: http://www.mathnet.ru/

8. *Polilova T.A.* Reitingi zhurnalov v RINTs kak instrumenty analiza i vliianiia // Preprinty IPM im. M.V. Keldysha. 2021. № 40. 35 s. https://doi.org/10.20948/prepr-2021-40

9. *Vinkler P.* Evaluation of publications by the part-set method // Scientometrics. 2021. Vol. 126. P. 2737–2757. https://doi.org/10.1007/s11192-020-03841-7

10. *Shomshor M. (Martin Szomszor), Adams D. (Jonathan Adams), Pendlberi D. (David A. Pendlebury), Rodzhers G. (Gordon Rogers)*. Klassifikatsiia dannykh: kak delat osoznannyi vybor, vedushchii k zhelaemym rezultatam. Otchet o mezhdunarodnom issledovanii Instituta nauchnoi informatsii (ISI). URL: https://img06.en25.com/Web/ClarivateAnalytics/%7B3f970f80-b453-451f-80a2-8d1792213e83%7D_Clarivate_ISI_Data_categorization_report_RU.pdf

11. *Potter I.* Introducing Citation Topics in InCites – Clarivate. Dec 3, 2020. URL: https://clarivate.com/blog/introducing-citation-topics/

12. *Pechnikov A.A., Chebukov D.E.* Struktura grafa tsitirovaniia zhurnalov Math-Net.Ru // Nauchnyi servis v seti Internet: trudy XXIII Vserossiiskoi nauchnoi konferentsii (20–24 sentiabria 2021 g., onlain). M.: IPM im. M.V. Keldysha, 2021. https://doi.org/10.20948/abrau-2021-2

13. *PageRank*. Wikipedia, the free encyclopedia. URL: https://en.wikipedia.org/wiki/PageRank

СВЕДЕНИЯ ОБ АВТОРАХ

ГОРБУНОВ-ПОСАДОВ Михаил Михайлович – главный научный сотрудник Института прикладной математики им. М.В. Келдыша РАН, доктор физико-математических наук;

Mikhail Mikhailovich GORBUNOV-POSADOV – Keldysh Institute of Applied Mathematics, chief researcher

email: gorbunov@keldysh.ru

ORCID: 0000-0002-7044-8287

ПОЛИЛОВА Татьяна Алексеевна – старший научный сотрудник Института прикладной математики им. М.В. Келдыша РАН, доктор физико-математических наук, лауреат Премии Президента РФ в области образования;

Tatyana Alekseevna POLILOVA – senior researcher of the Keldysh Institute of Applied Mathematics Russian Academy of Sciences.

email: polilova@keldysh.ru.

ORCID: 0000-0003-4628-3205

Материал поступил в редакцию 21 октября 2021 года

УДК 004.43 БК 22.18

ПЕРСПЕКТИВЫ ФУНКЦИОНАЛЬНОГО ПРОГРАММИРОВАНИЯ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛЕНИЙ

Л. В. Городняя[0000-0002-4639-9032]

Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирский государственный университет, г. Новосибирск

lidvas@gmail.com

Аннотация

Статья посвящена результатам анализа современных тенденций функционального программирования, рассматриваемого как метапарадигма решения проблем организации параллельных вычислений и многопоточных программ для многопроцессорных комплексов и распределённых систем. С учетом мультипарадигмальности параллельного программирования использован парадигмальный анализ языков и систем функционального программирования. Такой анализ позволяет снижать сложность решаемых задач методами декомпозиции программ на автономно развиваемые компоненты, оценивать их сходство и различия. Учёт парадигмальных особенностей необходим при прогнозировании хода процессов применения программ, а также при планировании их изучения и разработки. Есть основания рассчитывать, что функциональное программирование помогает повышать производительность программ. Показано разнообразие парадигмальных характеристик, присущих подготовке и отладке долгоживущих программ параллельных вычислений.

Ключевые слова: *функциональное программирование, парадигмальная декомпозиция, параллельные вычисления, система программирования, мультипарадигмальность.*

ВВЕДЕНИЕ

Рассматривая идеи функционального программирования (ФП) и практику их применения при анализе проблем, средств и методов организации параллельных вычислений, можно обратить внимание, что уже создано значительное количество языков и систем параллельного программирования, решающих многие проблемы подготовки многопоточных программ для многопроцессорных комплексов и распределённых систем. Новые работы по ФП существенно нацелены на поиск более производительных решений проблем параллельного программирования [1].

Изложение начинается с обсуждения особенностей термина «функциональное программирование», основных принципов и ограничений, обнаруживающихся при переносе техники ФП на параллельные вычисления (ПВ). Затем рассматривается ряд парадигм ПВ, поддержанных в известных языках программирования (ЯП). Анализируется проблема учёта уровня новизны в задачах ПВ и перехода к производственному ФП.

1. О ТЕРМИНЕ «ФУНКЦИОНАЛЬНОЕ ПРОГРАММИРОВАНИЕ»

Понятие «функциональное программирование» в настоящее время допускает два смысла. Исторически в начале 1960-х годов Дж. Маккарти провозгласил, что все понятия программирования могут трактоваться как функции или результат их применения. В середине 1970-х годов Дж. Бекус привлек внимание к ФП, призывая преодолеть узость так называемого «бутылочного горлышка» оператора присваивания. Собственно определение термина не было дано, оно использовалось интуитивно без особых разночтений и постепенно стало восприниматься как отдельная парадигма, противостоящая императивному программированию, в которой осознано, что *правильность важнее эффективности*. ФП даёт приоритет организации вычислений и сложных структур данных, а вопросы эффективной обработки памяти и управления процессами уходят на второй план.

Примерно с середины 1990-х годов кристаллизовалась идея «чистого функционального программирования» как раздела дискретной математики, исследующего прямые однозначные функции в дискретном пространстве над значени-

ями типа целых чисел, символьных строк или графов. Чистое ФП можно рассматривать как математическую основу более общего производственного ФП. Постепенно это «чистое ФП» стали называть просто «ФП». Такое разделение смыслов примерно соответствует различию в стандартах на академические и производственные ЯП. Теперь общее определение ФП начинают с утверждения, что характеристическая его особенность заключается в том, что *одинаковые формулы в одинаковом контексте имеют одинаковое значение.* Из этого вытекает исключение присваиваний, глобальных переменных, побочных эффектов, работы с внешними устройствами, передач управления. Эта изящная формула уязвима из-за отсутствия единого, точного определения термина «контекст».

Понятие «контекст» в программистской литературе обладает некоторой двойственностью. Это одновременно фрагмент программы, иногда называемый областью существования или видимости, и контекстная таблица соответствия символов и их значений, используемых в этом фрагменте. Семантика ЯП может по-разному ограничивать правила воздействия на контекстную таблицу. В одном ЯП может одновременно существовать две и более контекстных таблицы для одного фрагмента – статика и динамика, а кроме того, глобальный и локальный контексты. Системы программирования (СП) для одного ЯП могут по-разному решать вопрос о порядке перебора контекстных таблиц для определения значения формулы, более того, такой порядок может быть изменён по опциям из программы или задания на её компиляцию.

Можно обратить внимание, что при оптимизирующей компиляции нередко контекст – это линейный участок между двумя соседними присваиваниями. В таком случае можно считать, что на внутреннем языке компилятора происходит приведение программы к функциональной форме, удобной для обнаружения невычисляемых, константных или дублирующихся формул. Компилятор заменяет константную формулу на константу, вычисленную при компиляции, а дублирующиеся формулы — на переменную, значение которой будет вычислено при выполнении программы. Не исключено, что возможность при решении сложных задач безопасно выполнять подобные оптимизирующие преобразования является сильным мотивом использования чистого ФП. Эта же причина позволяет ФП быть полезным дополнением для любой парадигмы параллельных вычислений.

Если контекстом считать линейный участок между двумя соседними присваиваниями, то приведённая выше характеристика не отличает парадигму ФП от других парадигм. В процессе компиляции программ практикуется именно такая декомпозиция программ на линейные участки с целью решения проблем распределения памяти. При переходе к ФП каждый такой интервал можно представить как отдельный контекст, в котором каждое значение связано со своей локальной переменной.

2. СЕМАНТИЧЕСКИЕ ПРИНЦИПЫ

Обычно ФП подразумевает поддержку ряда семантических и прагматических принципов, удобных для создания функциональных моделей на этапе исследовательского компьютерного эксперимента, естественного при решении новых и сложных задач**.** Программы обработки любых данных могут быть заданы символьными формами, в которых выделен один элемент – представление функции, остальные – ее аргументы. Границы применимости таких форм определены интерпретатором или компилятором, позволяющим вычислять значения символьных форм. Функциональное программирование поддерживает *внешние* семантические принципы представления алгоритмов, такие как универсальность, параметризация, самоприменимость, и системно-реализационные *внутренние* прагматические принципы поддержки информационной обработки, такие как интегральность ограничений, неизменяемость данных, строгость результата. Семантическим принципам программист следует при подготовке программы. Прагматические принципы обеспечивает система программирования, освобождая программиста от непринципиальных решений, не зависящих от природы задачи.

Универсальность. Понятия «функция» и «значение» представляются теми же средствами, как и любые данные для компьютерной обработки. Исторически близкое понятие – «принцип хранимой программы».

Этот принцип позволяет строить представления функций из их частей и вычислять части по мере поступления и обработки данных. Нет принципиальных ограничений на манипулирование средствами языка, функциями из определения семантики языка, конструкциями реализации языка в СП и выражениями в программе. Всё, что понадобилось при реализации ЯП, может пригодиться при его

применении. Равноправие разноуровневых средств обуславливает открытый характер систем функционального программирования. Строго говоря, программирование, в отличие от математики, вообще не имеет дела ни со значениями, ни с функциями. Программирование работает с данными, которые *могут представлять* значения или функции.

Идея хранимой программы впервые сформулирована в описании аналитической машины Чарльза Беббиджа, через сто лет она реализована в компьютерах Конрада Цусе и определении машины Алана Тьюринга, позднее провозглашена в архитектуре Джона фон Неймана [2]. Исторически подтверждена возможность такого символьного представления информации, при котором нет принципиального различия в природе данных для изображения значений и функций. Следовательно, нет и препятствий для обработки представлений функций теми же средствами, какими обрабатываются данные. Поэтому представления функций можно строить из их частей – символов. Их даже можно формировать по ходу процесса вычислений, поступления и обработки информации о них. Именно так компиляторы конструируют программы.

При компиляции программ выполняется распределение памяти для функций, переменных и констант. Эффективность такого распределения зависит от учёта особенностей базовых средств обработки машинных кодов, что обычно формулируется как тип данных, удобных для обработки компьютером, но несколько противоречит принципу универсальности. Вообще, тип данных – это множество объектов с соответствующим ему набором допустимых операций. Обычно тип данных задают в тексте программы или выводят его при статическом анализе, чтобы экономно распределять память и обнаруживать некоторые ошибки несоответствия выбора операций при обработке кодов данных. Динамическое управление вычислениями и конструированием программ может по существу требовать представления и более детального анализа типов данных в процессе вычислений.

Параметризация. Представление любой выделенной формулы можно рассматривать как параметр некоторой функции. Это значит, что части представления функций можно вычислять в зависимости от промежуточных результатов и конструировать функции, учитывающие условия их применения, в частности, расположение их определений и вызовов на разных уровнях иерархии представления

программы. Так работают интерпретаторы и отладчики программ. Реализация языков ФП обычно содержит механизмы, приспособленные к дополнению определений отдельных элементов программируемой системы. Любая символьная форма в определении функции может быть выделена из него как параметр и, наоборот, подставлена в него. Функции-переменные допустимы равноправно с обычными функциями-константами и могут быть значениями аргументов или выработаны в качестве результатов других функций.

Процесс вычисления результатов по заданным аргументам методов выполнения определённого алгоритма часто рассматривается как исполнение неизменяемой программы, заранее определённой конструкции – константы. В большинстве ЯП реализация функций подразумевает процесс вычисления результатов по заданным аргументам методов выполнения определённого алгоритма. Функция – это соответствия между аргументами и результатами; и то, и другое, и сама функция могут быть значениями переменных. Отсутствие навыков работы с функциональными переменными говорит лишь о том, что надо осваивать такую возможность, потенциал которой может превзойти ожидания теперь, когда программирование становится все более компонентно-ориентированным.

Обеспечение многократного использования данных выполняется с помощью именования. Переменная – именованная часть памяти, предназначенная для многократного доступа к изменяющимся данным, а константа – к неизменным данным. Переменные отличаются от именованных констант частотой изменения связи между именем и соответствующим ему данным. Поэтому можно их реализацию делать одним и тем же способом. Именования данных обычно используются при их описании, предназначенном для многократного использования описанных конструкций – значений или функций.

Самоприменимость. Представления рекурсивных функций прямо или косвенно используют сами себя, что позволяет строить ясные лаконичные символьные формы.

Примеры самоприменимости дают многие математические функции, особенно рекурсивные, такие как факториал, числа Фибоначчи, суммирование рядов и многие другие, определение которых использует математическую индукцию. В технологии программирования некоторым сходством обладает метод раскрутки

программ. Этот метод сводит организацию процесса программирования к ряду шагов, каждый из которых даёт или работоспособную часть программы, или инструмент для выполнения очередных шагов раскрутки. Первые реализации языка Lisp были выполнены методом раскрутки, причем в составе системы сразу были предусмотрены и интерпретатор, и компилятор функций. Оба эти инструмента были весьма точно описаны на самом языке Lisp, причем основной объем описаний не превосходил пару страниц. Похожая история произошла и при разработке языка Си, ядро которого по трудоёмкости оценивалось как один человеко-месяц.

3. ПРАГМАТИЧЕСКИЕ ПРИНЦИПЫ

Прагматические принципы поддерживают система программирования, точнее, её разработчики.

Интегральность ограничений. Оперативный пересмотр распределения памяти или её освобождения поддержан для профилактики необоснованных простоев памяти.

Интегральность ограничений на пространственные характеристики позволяет исключать необоснованные простои памяти. Бывает, что не хватает памяти принципиально не на всю задачу, а лишь на отдельные блоки данных, возможно мало существенные для ее решения. Такая проблема в системах ФП решается принципом интегральности ограничений на пространственные характеристики. Многие СП поддерживают чётко фиксированное распределение памяти на части для хранения собственно программы, используемых в ней переменных, констант, стека вызовов, динамически размещаемых данных – «куча». Возникают ситуации, когда одни из таких частей исчерпаны, в то время как в других остаётся недоиспользованное пространство. В системах ФП эту проблему решает специальная функция – «мусорщик» (garbage collector), пытающаяся при недостатке любой области памяти автоматизировать перераспределение или освобождение памяти.

Представление данных в памяти может быть скрыто специальными функциями, освобождающими программиста от непринципиальных проблем для решения более важных задач, отдавая приоритет алгоритму и структурам данных. Управление обработкой значений и организация доступа к памяти актуальны после выбора и отладки принципиальных решений, без преждевременного отвлечения на проблему повторного использования памяти при её дефиците. Новые

реализации механизма «мусорщик» рационально учитывают преимущества восходящих процессов на больших объемах памяти.

*Неизменяемость данных***.** Представление каждого результата применения функции размещается в новой части свободной памяти без искажения аргументов этой функции, которые могут быть полезны для других функций.

Таким образом существенно упрощается отладка программ и обеспечивается обратимость любых действий. Можно быть уверенным, что все промежуточные результаты сохранены, их можно анализировать и снова использовать в любой момент. Если определение функции – это статическая конструкция, то процесс можно рассматривать как композицию из функций, разворачиваемую по этой конструкции в динамике. Таким образом можно поддерживать уточнение или улучшение программируемых решений по ходу вычислений. Реализация языка ФП может содержать списки свойств символов, приспособленные к внешнему доопределению отдельных элементов поведения программируемой системы. Такие списки использовались в ранних работах по представлению знаний.

В этом плане вычисление – это процесс решения задачи, сводимой к обработке данных, – чисел, кодов или символов, рассматриваемых как модели реальных объектов, представленных с помощью обозначений, смысл которых может изменяться по ходу обработки данных и появления внешних данных с сохранением прежних смыслов. Отдельный аспект связан с переходом от целых чисел к вещественным, возможно допускающим изменение точности представления по ходу вычислений. Логически они остаются константами, а реализационно обрабатываются как переменные. Кроме того, возникают некоторые отклонения от принципа неизменяемости данных в пользу восстановимости данных, не так уж влияющие на подготовку и отладку программ, если это поддерживается на системном уровне или возникает на заключительных этапах отладки.

*Строгость результата***.** Любое число результатов функции может быть представлено как одна символьная форма, из которой при необходимости можно выбрать нужный результат. Такой принцип удобен для описания интерпретатора программ. Всегда ясна граница между аргументами и результатами, размещаемыми в стеке, – результат последней вычисленной функции расположен на вер-

шине стека. Нередко этот принцип трактуется как требование однозначности математических функций, что приводит к сомнениям в правомерности функций целочисленного деления, извлечения корня, обратных тригонометрических функций и многих других категорий математических функций.

Способы определения функций были достаточно различными ещё задолго до появления компьютеров. Отличаются и подходы к сохранению результатов функций, в частности, в виде таблиц или специальных приборов типа логарифмической линейки. Могут отличаться и методы решения одних и тех же задач. Например, существует более двадцати методов сортировки, результаты которых будут одинаковы. Различия при выборе метода зависят от условий применения программы, особенностей сортируемых данных и критериев эффективности. Выбор техники реализации функции обычно зависит и от способов определения правила и методов получения результата функции по заданному правилу. При одних и тех же аргументах способы могут быть различны, например:

- алгоритм (поиск наибольшего общего делителя);
- таблица (сложение или умножение для целых чисел);
- процесс (непосредственное измерение);
- устройство (вольтметр, термометр, часы);
- формализованный текст (процедура, подпрограмма, макрос и т. п.).

Использование чисел и кодов нередко обусловлено поддержкой эффективности, надёжности и безопасности, смягчением роли человеческого фактора. Тем не менее, во многих современных информационных сервисах можно видеть решения, существенно снижающие и надёжность, и безопасность. Работа с паролями теперь часто имеет кнопочку для показа его текста. Нередко диалог с «личным кабинетом» содержит простенькую процедуру смены пароля. Идентификация пользователя на сайтах, работающих с деньгами и документами, происходит по IP-адресу, без учёта того, что один компьютер может быть во владении разных пользователей. Многие банковские инструменты на неожиданные ситуации вместо диагностики практикуют отказ в обслуживании.

4. СЛЕДСТВИЯ ПРИНЦИПОВ

Представление алгоритмов в виде функциональных программ даёт практически значимые следствия: конструктивность, факторизация и доказуемость вытекают из семантических принципов, а из прагматических принципов следуют интуитивные скрытые модели продолжаемости процессов, обратимости действий, параллелизма, дающие основу для интуитивного выстраивания функциональных моделей, допускающих непосредственный компьютерный эксперимент.

Конструктивность является следствием принципа универсальности, позволяющего представления программ обрабатывать так же, как любые данные. Это даёт поддержку мета-компиляция, включая синтаксически управляемые методы генерации и анализа программ, а также однородно-гомогенные представления программ, внешне сохраняющих аналогию или подобие обрабатываемым данным или прототипам, в том числе смешанные и частичные вычисления, оптимизирующие преобразования, макрогенерацию и многое другое, необходимое для создания операционных систем и систем программирования.

Факторизация непосредственно вытекает из принципа параметризации с учетом принципа универсальности. Любой помеченный фрагмент программы может быть вынесен из её представления и ассоциирован с определённым именем, допуская возможность восстановления исходного представления. Можно обратить внимание, что параметры при вызове функции вычисляются на одном и том же уровне иерархии, в общем контексте согласно принципу неизменяемости данных. Поэтому порядок вычисления параметров не имеет значения, может быть произвольным. Это позволяет декомпозировать программу на автономно развиваемые модули и накапливать правильность, а также представлять параллельные потоки, ленивые (отложенные) или опережающие вычисления. Можно сказать, что программа приводится в факторизованную форму по тем или иным параметрам в зависимости от цели её преобразования. Благодаря обратимости действий, т. е. неизменяемости данных, процесс отладки обретает сходимость.

Доказуемость основана на связи принципа самоопределимости с методами рекурсии, математической индукции и логики. Появляется возможность логически выводить отдельные свойства программ и благодаря этому обнаруживать некоторые трудно уловимые ошибки, что повышает надёжность и безопасность

программ, хотя не позволяет решить проблему правильности в полном объёме. Подобным образом формируется подход к разработке программ по методике раскрутки с выделением минимального ядра, с последующими шагами его расширения до полного практичного решения задачи.

Следствия принципов системно-реализационной прагматической поддержки информационной обработки в системах ФП выражаются в использовании интуитивных моделей, таких как продолжаемость процессов (бесконечность), обратимость действий и параллелизм.

Продолжаемость процессов интуитивно вытекает из прагматической поддержки принципа интегральности ограничений, позволяющего значительную часть работы выполнять на основе модели неограниченной памяти без особой заботы о её границах и разнообразии характеристик скорости доступа к разным структурам данных. Прагматика таких решений нацелена на освобождение внимания программиста от не самых важных проблем в пользу существа решаемой задачи. Во многих языках ФП поддержана имитация работы с бесконечными структурами данных.

Обратимость действий базируется на иллюзии неизменяемости данных, механизмы которой скрыты в СП, их применение почти не требует заботы при подготовке программы и основного объёма отладки. Это позволяет поддерживать механизм мемоизации функций на ранее обработанных аргументах. Фактически необходимые изменения данных, такие как повторное использование памяти, просто автоматизированы, программист может позволить себе не вмешиваться в реализацию таких средств до тех пор, пока не возникнут проблемы с производительностью.

Параллелизм основан на принципе строгого результата, позволяющего при необходимости любое число представленных результатов рассматривать как общую структуру из них. Дополнением является принцип параметризации, гарантирующий одинаковость контекста при вычислении параметров функции одного уровня. Поскольку результаты часто являются аргументами объемлющих функций, логично возникает и сопутствующий *принцип единого аргумента*. Функция любого числа аргументов может быть преобразована в функцию одного аргумента. Появляется возможность представлять независимые потоки и объединять

их в многопоточные или многопроцессорные программы, в общий проблемно-ориентированный комплекс. Это делает ФП удобным для работы с программами, нацеленными на организацию параллельных процессов. Кроме того, возможность перехода от списка параметров или результатов к строгому результату или единому аргументу позволяет отойти от привычной схемы операций, отображающей два операнда в один результат, к операциям, отображающим один ряд операндов в другой ряд результатов, что может соответствовать структуре некоторых аппаратных узлов и тем самым допускать представление более эффективных решений.

Такой комплект следствий из принципов ФП позволяет уточнять и улучшать запрограммированные решения при отладке программ решения новых задач, допускает *множественность* определений функций при исследовании свойств решаемой задачи на уровне редактируемых константных символьных форм. Множественные определения символов в рамках настраиваемой интерпретации обеспечивают, кроме общеизвестного полиморфизма, более управляемые схемы построений, отладки и конструирования программ.

5. ПАРАДИГМАЛЬНАЯ ХАРАКТЕРИСТИКА

Используя систематизацию парадигм программирования на основе различий в приоритетах принятия программируемых решений, можно сделать вывод, что одной из причин сложности разработки программ параллельных вычислений является их скрытая мультипарадигмальность [3]. Для разработки параллельной программы многое требуется продумывать по-разному одновременно в различных парадигмах, удобных для решения отдельных подзадач без возможности решения полного комплекса подзадач в единой обстановке. Наиболее очевидно парадигмальные различия видны при решении задач масштабирования вычислений на различные многопроцессорные комплексы, синхронизации взаимодействий над локальной и общей памятью в многопоточных программах, представлении естественного асинхронного параллелизма уровня постановок задач и достижения высокой производительности программ с учётом критерия полноты и равномерности загрузки доступных многопроцессорных комплексов или распределённых систем. Создано заметное число языков и систем программирования

(ЯиСП), позволяющих решать отдельные из этих задач в рамках парадигм, поддержка которых представлена в разных языках программирования (Таблица 1).

Таблица 1.

№	*Проблема*	*Парадигма*	*ЯП и API*
1	Масштабирование	Многопроцессорное программирование	VHDL, XC, СИГМА, bash, Occam, mpC, Эль-76, Limbo, Kotlin, MPI
2	Синхронизация потоков	Синхронное (синхронизующее) программирование	APL, VAL, Sisal, Alef, E, X10, LuNA, Charm, Kotlin, Go, Java, Scala, Rust, Пифагор, OpenMP
3	Постановки задач	Асинхронное программирование	БАРС, Haskell, Erlang, JavaScript, Python, C#
4	Производительность программ	Высокопроизводительное программирование	Setl, HPF, G, Sparkel, mpC, Sanscript, D, Rest, F#

Таким образом, для каждой из трудно решаемых задач параллельных вычислений уже сформирована отдельная удобная парадигма её решения и создан ряд ЯП, поддерживающих такую парадигму. Любая из таких парадигм может быть дополнена моделями и методами ФП. Различие между парадигмами проявляется в упорядочении важности средств и методов, используемых при решении отдельных задач, другие задачи требуют иного упорядочения. В каждый момент разработки программы обычно используется одна парадигма. Соответственно и в разных ЯП выделяется одна ведущая парадигма. Ряд таких ЯП (Kotlin, APL, VAL, Sisal, Go, Haskell, Erlang, F#) изначально относят к функциональным, почти все содержат подъязыки, позволяющие представлять программы в функциональном стиле.

Требования к решению достаточно сложных задач ПВ связано с целым рядом затруднений, что влечёт необходимость использования разных парадигм на разных этапах их создания и фазах их жизни. При переходе к технологии параллельного программирования важна гарантия получения практичного результатов в заданные сроки, что требует поддержки полного спектра парадигм, используемых на разных этапах разработки программ. Трудоёмкость использования разных парадигм при решении одной задачи обычно минимизируется созданием многоязыковых систем, допускающих по мере необходимости возможность перехода

от одной парадигмы к другой без затрат на освоение разных интерфейсов. Это показывает целесообразность создания мультипарадигмального языка ПВ, поддерживающего одновременно все основные парадигмы параллелизма, дополненные ФП.

6. ПАРАДИГМАЛЬНАЯ ДЕКОМПОЗИЦИЯ

Как показывает опыт применения языков БАРС и Haskell, в таких случаях удобно выделять в определении ЯП отдельные подъязыки, поддерживающие основные парадигмы или монады, нацеленные на конкретные модели подготовки и представления программ так, чтобы на каждом этапе разработки программы локализовать использование одной парадигмы, характеризуемой сравнительно небольшим набором средств и методов в рамках одного способа мышления. Каждая парадигма имеет свои наполнение категорий семантических систем и упорядочение их роли в процессе программирования [3].

Средства многопроцессорного программирования обычно опираются на данные и характеристики доступной архитектуры, включая основной многопроцессорный комплекс и взаимосвязи между его элементами. Оперирование комплексом позволяет инициировать процессы функционирования отдельных процессоров, их блокировку, возобновление и отмену процессов. По ходу процессов возможны обмены данными по определённым протоколам, результаты которых можно рассматривать как цель программы. Принятие решений начинается с определения пространства возможных многопроцессорных комплексов, что можно рассматривать как особую разновидность памяти со своей дисциплиной функционирования и взаимодействия элементов. Далее происходит выбор подходящих конфигураций и структурирования пространства итерирования процессов, предназначенных для выполнения на отдельных процессорах. Затем полученная схема управления процессами наполняется собственно действиями, выполняющими вычисления. Обычно предпочитаются процедуры без рекурсии, освобожденные

от сложностей управления вычислениями и побочных эффектов над общей памятью. Строится многопроцессорная программа, допускающая в динамике реконфигурацию многопроцессорного комплекса[1].

При синхронном (синхронизующем) многопоточном программировании выделяются достаточно чёткие схемы управления вычислениями, которые разделяются на регулярные участки и типичные модели программы, удобные для распараллеливания. Обычно выделяются фрагменты, свободные от побочных эффектов в памяти, и допускается неимперативное управление временем исполнения потоков программы с учётом иерархии схемы управления программой и некоторых временных отношений. Принятие решений начинается с выбора стандартных схем управления, используется понятие «пространство итерирования», которое можно структурировать в зависимости от распределения данных и методов их хранения, что может влиять на эффективность и дисциплину обработки многоуровневой памяти. Управление итерированием может использовать произвольные предикаты. Схема управления над пространством итерирования потоков наполняется фрагментами, сравнительно простыми для отладки, возможно отлаженными заранее или программируемыми автономно. Для вывода результатов программы выделяется специальные средства формирования результатов вычислений, полученных на равноправных потоках многопоточной программы. Таким образом получается многопоточная программа с динамически изменяемым пространством потоков над локальной памятью с возможностью эпизодической синхронизации их отдельных фрагментов[2].

Асинхронное программирование нацелено на предельное выражение независимых элементов программы, обусловленных природой решаемой задачи, что может быть базой для максимального распараллеливания при условии представления специальных схем организации вычислений с учётом специфики доступного оборудования. Начинается принятие решений с выбора схем управления действиями и представления условий их срабатывания. Действия могут использо-

1 Курс «Параллельное программирование с использованием OpenMP и MPI». URL: https://mooc.tsu.ru/mooc-openedu/mpi/

2 Курс «Основы MPI». URL: https://habr.com/ru/post/121925/

вать ту или иную дисциплину обработки памяти. Поддерживается неявное и программируемое разнообразие дисциплин доступа к памяти, включая иерархию неоднородный памяти, и схем вычислений – фрагментов, наполняющих общую схему программы процедурами или библиотечными модулями. Получается схема программируемых синхросетей, асинхронно управляющая выполнением действий в зависимости от условий их готовности к выполнению.

Для высокопроизводительного программирования необходим переход от отдельного прогона программы к учёту перспектив её многократного применения и улучшения. Появляется возможность использовать недогруженные мощности многопроцессорных комплексов выполнением фрагментов программы в расчёте на предстоящие в будущем прогоны при данных, возможно востребованных в предстоящих прецедентах её отладки и применения. Примерно так организуют программы, предназначенные для сверх быстрого реагирования на опасные события, например, при прогнозировании цунами. Принятие решений начинается со схем и моделей вычислений, возможно над общей памятью, но с приоритетом локальной памяти. Управление вычислениями учитывает особенности многократного выполнения программы при её отладке и применении, включая возможность наследования результатов между сеансами и сравнения измеримых характеристик производительности версий программы. Получается ряд улучшаемых версий программы, выбор одной из которых может учитывать особенности или изменение текущих условий применения, включая конфигурацию многопроцессорного комплекса и требования к измеримым характеристикам производительности программ.

Долгоживущие и учебные ЯП, как и новые ЯП нашего века, обычно мультипарадигмальны. Есть основания для заключения, что успешная практика параллельного программирования требует мультипарадигмальной поддержки полного спектра парадигм параллельных вычислений, допускающей их развитие, пополнение и применение по мере необходимости с возможностью перехода к очередной парадигме без изменения общеязыковой и системной обстановки.

7. ТРУДОЁМКОСТЬ

Жаргон современного практического программирования использует понятие «язык программирования» как «входной язык или расширенное подмножество ЯП типовой СП, функционирующей на базе определённой конфигурации оборудования». Различие заключается в том, что СП обычно сопровождает реализацию ЯП расширяемым комплектом библиотечных модулей и может не поддерживать отдельные сложности семантики ЯП. В результате происходит сглаживание видимых на практике различий между разными ЯиСП. Кроме того, прямые измерения трудоёмкости программирования и производительности программ почти не отражают зависимости результата от принятых программистом решений и выбора конструкций ЯП. Хотя программируемые решения представляются в терминах ЯП, их влияние растворяется в весьма сложном комплексе, наследующем производительность СП и оборудования с большим доминированием характеристик элементной базы и аппаратуры. Таким образом, существует проблема создания методики, позволяющей выявлять такие зависимости совмещением прямых измерений с результатами экспертных оценок особенностей ЯиСП, возможно отличающихся от оценок ЯП [4].

Есть основания при прогнозировании трудоёмкости параллельного программирования учитывать не только степень изученности решаемых задач, но ещё и уровень квалификации и способностей разработчиков программы решения задачи, умеющих преодолевать понятийную сложность реализуемых и используемых программируемых и программных средств, особенности функционирования которых могут выходить за пределы привычных представлений [5]. Для параллельных вычислений степень изученности часто наследует результат ранее созданной последовательной программы решения задачи, обладающей математически точной постановкой. Возникает соблазн, тормозящий осознание или создание более эффективного параллельного алгоритма. Кроме того, уровень квалификации специалистов опирается на опыт императивно-процедурного программирования, препятствующего восприятию более сложных зависимостей в параллельных процессах. Понятийная сложность решаемых задач выходит за пределы обычного образования, и сами понятия обретают более широкое толкование, отчасти противоречащее привычному интуитивному пониманию. Представление

результатов оценки понятийной сложности, позволяющее структурировать пространство таких параметров, рассмотрено в статье [3].

8. СТЕПЕНЬ ИЗУЧЕННОСТИ

Практическое функциональное программирование существенно зависит от выбора постановок задач, решения которых представляются системами функций, сравнительно простых, не слишком трудоёмких и удобных при отладке. По степени изученности существенно различаются следующие категории постановок задач, влияющие на выбор методов решения задач и трудоёмкость их программирования:

– новые;
– исследовательские;
– практичные;
– точные.

Для новых постановок задач характерны отсутствие доступного прецедента практичного решения задачи, новизна используемых средств или недостаток опыта исполнителей. Задачи параллельных вычислений поставлены ещё в докомпьютерную эпоху, и постановки многих таких задач обладают математической точностью. Тем не менее, часть задач параллельного программирования их решений приходится рассматривать как новые из-за стремительного обновления ИТ, элементной базы и нерешённых образовательных проблем программирования в целом. Любая проблема, не получившая хорошего решения, остаётся в статусе новой задачи независимо от времени её постановки. Исследовательские постановки задач параллельного программирования в настоящее время следуют тенденции функционального подхода и изучения схем структурирования пространства итерирования процессов, позволяющего оптимизировать обработку памяти. Функциональный подход можно рассматривать как методику сведения решений сложных задач к композициям из планарных проекций, достаточно удобных для понимания, анализа и обработки. Практичные постановки математических задач ПВ, нацеленные на актуальность и удобство применения, преимущественно используют мощность доступных ИТ для воспроизведения математических моделей, ранее ограниченных низкой эффективностью оборудования, а

теперь получивших перспективу стать новой информационной революцией. Точные постановки большинства задач параллельного программирования сложились на базе последовательных алгоритмов и включают в себя испытание возможностей используемых средств, связанных с мерой организованности ранее созданной императивной программы. Некоторые сложности связаны с использованием однопроцессорных конфигураций как исходной модели параллельных многопоточных программ. Не исключено, что предельным случаем удобнее рассматривать двухпроцессорные конфигурации. Появление собственно параллельных алгоритмов встречается не так уж часто, хотя в середине 1990-х годов проводились конференции такого направления.

9. ОБРАЗОВАНИЕ И КВАЛИФИКАЦИЯ

Обучение параллельным вычислениям входит в образовательные программы многих университетов, что достаточно для понимания их сложности и постановки задач их исследования. Проблемой является переход к практике реализации высокопроизводительных программ, удовлетворяющих особо сложным, трудно удостоверяемым критериям надёжности и безопасности. Ещё ряд проблем связан с формированием интуитивной грамматики деятельности, достаточной для практичного программирования. Возможное решение этих проблем предлагается в проекте языка Синхро, предназначенного для учебного применения [6]

Характерной чертой системного или функционального подхода как ведущего метода программирования является переход к классам задач при содержательном анализе постановок задач. Границы этого класса устанавливаются выбором процесса решения задач. Переход к экспериментам на суперкомпьютерах показал, что именно системные решения могут дать весомый вклад в производительность ПВ, причём такой вклад может превышать теоретические прогнозы. Это можно рассматривать как обоснование необходимости более фундаментального подхода к программированию, особенно к системному программированию и его математическим основам, естественно представимым в чисто функциональном программировании [7].

При создании, формировании и исследовании математических моделей как фундаментального базиса для решения особо трудных проблем эффективности, надёжности и безопасности программного обеспечения важную роль играет развитие моделей, связанных со временем и ресурсами и слабо представленных в курсах по классической математике, тем не менее, доступных в рамках ФП.

Экстенсивное развитие ИТ заметно опережает возможности человека оперативно осваивать новые возможности аппаратуры и системных средств ИТ, выходящие за пределы пользовательского уровня, поддерживаемого поставщиками инструментария и программных продуктов. Миссия программирования – создание инструментария, позволяющего повышать качество информационных систем, включая поиск новых решений по обеспечению надежности и безопасности ИТ [7].

10. ПЕРЕНОС ПРИНЦИПОВ НА ПАРАЛЛЕЛЬНЫЕ ВЫЧИСЛЕНИЯ

При переходе к многократно используемым программам и параллельным вычислениям *становятся более важными успешный опыт применения и производительность программ, чем их формальная правильность и эффективность*. Принцип универсальности имеет два аспекта — равноправие программ и данных и полная диагностичность определений функций. При решении задач ПВ универсальность сохраняет аспект равноправия программ и данных, традиционно востребованный в задачах операционных систем. Полнота диагностичности функций, удобная для сборки программ из отлаженных функций, может создавать проблемы из-за нарастания числа потоков в многопроцессорных программах. Иногда эта проблема преодолевается подбором выражений, не требующих ветвлений. Объём необходимой диагностики может быть отчасти сокращён средствами статического анализа и контроля типов данных.

Роль параметризации возрастает, она даёт решения проблем по реорганизации потоков при настройке на разные конфигурации многопроцессорных комплексов, требующей декомпозиции фрагментов программы. Факторизация программ на схемы и фрагменты позволяет разделять компоненты по уровню сложности отладки и наследовать правильность ранее отлаженных составляющих программы. Используются функции, не требующие предварительного вычисления параметров, подобно макротехнике.

Самоопределимость в виде рекурсивных функций обычно рассматривается как усложнение, влекущее опасное разрастание стека. Здесь следует обратить внимание, что многие системы ФП предлагают ряд решений, таких как отложенные действия, мемоизация, восходящая рекурсия, методы динамического программирования и оптимизация рекурсий сведением к циклам, во многих случаях позволяющие практически исключить чрезмерное разбухание стека. Да и поддержка работы со стеком в рамках принципа интегральности ограничений может быть поддержана более эффективно, чем в большинстве ЯиСП.

Несколько сложнее с прагматическими принципами, требующими пересмотра системных решений на уровне разработки систем программирования. Интегральность ограничений обычно рассматривается по отношению к памяти. Естественно расширить её на временные границы, подобно тому, как в языке mpC происходит перераспределение объёма вычислений при обнаружении неравномерной загрузки процессоров [8].

Неизменяемость данных сохраняется на уровне локальных потоков, но вызывает проблемы при переходе к общей памяти. Не исключено, что механизмы общей памяти в большей мере требуют восстановимости данных, не считая математических аспектов работы с разноуровневой памятью, копиями, репликами и т. п., что похоже на динамическое редактирование сложных конструкций, пока проиллюстрированного на задачах работы с DSL-языками [1].

Строгость результатов убедительно поддержана в языке Sisal как концепция пространств итерирования, строящегося над перечислимыми множествами с помощью операций скалярного и декартового произведений [9].

11. ПРАКТИЧНЫЕ КОМПРОМИССЫ

Чистое ФП можно рассматривать как методику функционального моделирования при создании программ решения сложных задач. Более общая парадигма ФП позволяет переходить от таких функциональных моделей к производственным структурам данных и принимать при необходимости практичные решения по их обработке в зависимости от реальных условий. Кроме принципов и следствий из них в реальных СП производственная парадигма ФП допускает уравновешивающие механизмы, внешне в ЯП выглядящие как специальные функции.

Например, в языках Lisp 1.5, Clisp, Cmucl и других представителях лисповского семейства обычно предоставлены такие компромиссные функции:

- универсальность смягчается функциями статического и динамического контроля типов данных;
- параметризация дополняется функциями доступа к общей и внешней памяти;
- самоопределимость преодолевается функциями, имитирующими привычные схемы циклов;
- интегральности ограничений противодействуют функции для программируемого распределения памяти;
- неизменяемости данных противостоят деструктивные функции, позволяющие исключать избыточный расход памяти;
- строгость результатов расширяется функциями ввода-вывода данных, работы с файлами, протоколами и многими другими, дающими связь программы с окружающим миром.

ЗАКЛЮЧЕНИЕ

В феврале 2021 года состоялась 22-я конференция, посвященная современным тенденциям функционального программирования [1]. Представленные доклады убедительно показали нацеленность ФП на решение многих проблем организации ПВ.

В рамках ФП возможен подход, позволяющий учитывать особенности решаемых задач и методов программирования, необходимых для решения проблем ПВ, влияющих на их решения в зависимости от приоритетов в выборе языковых средств и реализационных конструкций. Можно наметить линию, позволяющую сравнивать языки, выделяя сопоставимые примеры программ, и анализировать результаты прямых измерений производительности программ, выделяя особенности базовых средств и реализационных решений в СП, нацеленных на улучшение создаваемых программных продуктов. Программирование давно уже стало массовой профессией, требующей объективных метрик для оценки качества программируемых решений. Парадигмальные ошибки, обнаруживаемые при эксплу-

атации привычных СП на современном многопроцессорном оборудовании, показывают, что часть из них были просто незаметны до появления сетей, мобильных устройств и суперкомпьютеров.

Многие вопросы пока не получили практичного ответа. Появление новых парадигм можно связать с проявлением круга новых задач, решение которых пока вызывает трудности. Не ясно, насколько целесообразно взаимодействие парадигм и каким должен быть механизм их взаимодействия. Остаются в стороне образовательные проблемы овладения новыми парадигмами. Возможно, в этом деле помогут методы визуализации программ [5]. Кроме того, особенности парадигм лишь отчасти выражаются на уровне представления программы, часть являются требованиями к прагматике системно-реализационной поддержки в ЯиСП, другие вообще имеют отношение к скрытой интуитивной грамматике деятельности. Граница между семантикой программируемых решений и прагматикой их системной поддержки у каждой парадигмы своя.

В целом следует отметить, что следствия из семантических и прагматических принципов ФП и высокая моделирующая сила аппарата функций, расширенные специальными функциями практичных компромиссов, позволяют полезно дополнять основные парадигмы ПВ и практиковать работы по повышению производительности программ. В этом плане существенный вклад дают мемоизация, отложенные и опережающие вычисления, возможность синтаксического конструирования и декомпозиции программ на автономно развиваемые модули, а на более общем уровне — мета-программирование, позволяющее строить специализированные проблемно ориентированные DSL-языки программирования. Мемоизация позволяет радикально снижать сложность многократно повторяемых вычислений. Отложенные и опережающие вычисления дают возможность перераспределения нагрузки. Синтаксическое конструирование – механизм надёжного использования прототипов и структур обрабатываемых данных. Декомпозиция программ приводит к выводу автономно развиваемых модулей, изменение которых не требует повторного программирования смежных модулей. Мета-программирование может быть применено для создания подъязыков, учитывающих индивидуальные особенности процессоров, а также для обустройства переходов от одного языка к другому.

СПИСОК ЛИТЕРАТУРЫ

1. *Koopman P., Michels S., Plasmeijer R.* Dynamic Editors for Well-Typed Expressions // Trends in Functional programming/ 22nd International Symposium, TFP 2021, February 17–19, 2021. Springer, LNCS 12834. P. 44–66.

2. *Городняя Л.В., Кирпотина И.А.* О проблеме достоверности доступной в Интернете исторической фактографии // Сборник трудов SoRuCom-2017. Четвертая Международная конференция «Развитие вычислительной техники в России и странах бывшего СССР: история и перспективы». Зеленоград, 3–5 октября 1917 г. Под редакцией д. ф.-м. н. А.Н. Томилина. М.: ФГБОУ ВО «РЭУ им. Г.В. Плеханова», 2017. С. 40–49.

3. *Городняя Л.В.* О представлении результатов анализа языков и систем программирования // Научный сервис в сети Интернет: труды XX Всероссийской научной конференции (17–22 сентября 2018 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2018. С. 262–277. https://doi.org/10.20948/abrau-2019-03

4. *Городняя Л.В.* Подход к оценке трудоёмкости программирования **//** Научный сервис в сети Интернет: труды XXII Всероссийской научной конференции (21–25 сентября 2020 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2020. С. 192–209. https://doi.org/10.20948/abrau-2020-3
https://keldysh.ru/abrau/2020/theses/3.pdf

5. *Авербух В.Л.* Визуализация программного обеспечения. Екатеринбург: ИММ УрО РАН, 1995. 168 с.

6. *Городняя Л.В.* Учебный язык параллельного программирования СИНХРО // Языки программирования и компиляторы—2017. Труды конференции. Южный федеральный университет; под ред. Д.В. Дуброва. Ростов-на-Дону: Изд-во Южного федерального университета, 2017. С. 92–97.
URL: http://plc.sfedu.ru/files/PLC-2017-proceedings.pdf

7. *Городняя Л.В.* Перспективно стратегические парадигмы программирования Академика Андрея Петровича Ершова. 5-я международная конференция «Развитие вычислительной техники в России, странах бывшего СССР и СЭВ (SORUCOM 2020)». Москва, 6–8 октября 2020 г. С. 83–97.

8. mpC: A Multi-Paradigm Programming Language for Massively Parallel Computers // ACM SIGPLAN Notices. 1996. Vol. 31. No. 2. P. 13–20.

9. *Kasyanov V.N.* Sisal 3.2: functional language for scientific parallel programming. Enterprise Information // Systems. 2013. Vol. 7. No. 2. P. 227–236.

PERSPECTIVES OF FUNCTIONAL PROGRAMMING OF PARALLEL COMPUTATIONS

L. V. Gorodnyaya [0000-0002-4639-9032]

A.P. Ershov Institute of Informatics Systems (IIS)

Abstract

The article is devoted to the results of the analysis of modern trends in functional programming, considered as a metaparadigm for solving the problems of organizing parallel computations and multithreaded programs for multiprocessor complexes and distributed systems. Taking into account the multi-paradigm nature of parallel programming, the paradigm analysis of languages and functional programming systems is used. This makes it possible to reduce the complexity of the problems being solved by methods of decomposition of programs into autonomously developed components, to evaluate their similarities and differences. Consideration of such features is necessary when predicting the course of application processes, as well as when planning the study and organizing the development of programs. There is reason to believe that functional programming has the ability to improve programs performance. A variety of paradigmatic characteristics inherent in the preparation and debugging of long-lived parallel computing programs are shown.

Keywords*: functional programming, paradigm decomposition, parallel computing, multi-paradigm programming languages.*

.

REFERENCES

1. *Koopman P., Michels S., Plasmeijer R.* Dynamic Editors for Well-Typed Expressions // Trends in Functional programming/ 22nd International Symposium, TFP 2021, February 17–19, 2021. Springer, LNCS 12834. P. 44–66.

2. *Gorodnyaya L.V., Kirpotina I.A.* On the Problem of Reliability of Historical Factography Available on the Internet // Proceedings of the SoRuCom-2017. Forth International Conference «Computer Technology in Russia and in the Former Soviet Union». Zelenograd, the city of Moscow, October 3–5. Prof. A.N. Tomilin, Ed. Moscow, 2017. P. 40-49.

3. *Gorodnyaya L.V*. O predstavlenii rezul'tatov analiza yazykov i sistem programmirovaniya // Nauchnyj servis v seti Internet: trudy XX Vserossijskoj nauchnoj konferencii (17-22 sentyabrya 2018 g., g. Novorossijsk). M.: IPM im. M.V. Keldysha, 2018. S. 262–277. https://doi.org/10.20948/abrau-2019-03

4. *Gorodnyaya L.V.* Podhod k ocenke trudoyomkosti programmirovaniya // Nauchnyj servis v seti Internet: trudy XXII Vserossijskoj nauchnoj konferen-cii (21–25 sentyabrya 2020 g., onlain). M.: IPM im. M.V. Keldysha, 2020. S. 192–209.
https://doi.org/10.20948/abrau-2020-3
https://keldysh.ru/abrau/2020/theses/3.pdf

5. *Averbuh V.L.* Vizualizaciya programmnogo obespecheniya. Ekaterin-burg: IMM UrO RAN, 1995. 168 s.

6. *Gorodnyaya L.V.* Uchebnyj yazyk parallel'nogo programmirovaniya SINHRO // Yazyki programmirovaniya i kompilyatory—2017. Trudy konferen-cii. Yuzhnyj federal'nyj universitet; pod red. D.V. Dubrova. Rostov-na-Donu: Izd-vo Yuzhnogo federal'nogo universiteta, 2017. S. 92–97.
URL: http://plc.sfedu.ru/files/PLC-2017-proceedings.pdf

7. *Gorodnyaya L.V.* Perspektivno strategicheskie paradigmy programmirovaniya Akademika Andreya Petrovicha Ershova. 5-ya mezhdunarodnaya konferenciya «Razvitie vychislitel'noj tekhniki v Rossii, stranah byvshego SSSR i SEV (SORUCOM 2020)». Moskva, 6–8 oktyabrya 2020 g. S. 83–97.

8. mpC: A Multi-Paradigm Programming Language for Massively Parallel Computers // ACM SIGPLAN Notices. 1996. Vol. 31. No. 2. P. 13–20.

9. *Kasyanov V.N.* Sisal 3.2: functional language for scientific parallel programming. Enterprise Information // Systems. 2013. Vol. 7. No. 2. P. 227–236.

СВЕДЕНИЯ ОБ АВТОРЕ

ГОРОДНЯЯ Лидия Васильевна – старший научный сотрудник Института систем информатики имени акад. А.П. Ершова СО РАН, доцент Новосибирского государственного университета, специалист в области системного программирования и образовательной информатики.

Lidia Vasiljevna GORODNYAYA – Senior Researcher of A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Associate Professor of Novosibirsk State University, a specialist in system programming and educational informatics.

email: lidvas@gmail.com
ORCID: 0000-0002-4639-9032

Материал поступил в редакцию 5 ноября 2021 года

УДК 004

ПОВЫШЕНИЕ КАЧЕСТВА МЕТАДАННЫХ НАУЧНЫХ ПУБЛИКАЦИЙ С ПОМОЩЬЮ ОТЧЕТОВ CROSSREF

А. В. Ермаков[0000-0002-6054-0813]

Институт прикладной математики им. М.В. Келдыша Российской академии наук, Миусская пл., 4, Москва, 125047

Ermakov@Keldysh.ru

Аннотация

Рассмотрены вопросы, связанные с повышением качества метаданных научных публикаций, размещаемых в библиографической базе данных Crossref. Всю информацию, содержащуюся в метаданных, полученных от издателей научных публикаций, Crossref анализирует и отображает в различных отчетах. Отчеты дают издателям представление о полноте и корректности представленных библиографических данных. Качество метаданных прямо или косвенно влияет на количество просмотров и ссылок на публикацию, соответственно, на рейтинги научных изданий, авторов и организаций.

Ключевые слова*: метаданные публикаций, отчеты Crossref, цитирование, рейтинги научных изданий.*

ВВЕДЕНИЕ

Подготовка научной публикации неразрывно связана с тематикой и направлением исследований, в которых работают авторы. Как правило, публикации являются продолжением работ научного коллектива в данном направлении и опираются на предыдущие результаты исследований, выполненных ранее либо самими авторами, либо их научными руководителями или коллегами. Эта опора в публикациях на предшествующие результаты научных исследований сопровождается цитированием соответствующих научных материалов. Корректность цитирования очень важна не только для авторов новой научной статьи, но и для авторов цитируемых статей, а также для их научных изданий, так как прямо или косвенно влияет на количество просмотров и ссылок на публикацию, соответственно, на рейтинги научных изданий, авторов и организаций.

Метаданные публикаций необходимы, чтобы сделать библиографические данные о самой статье, авторах и т. п. идентифицируемыми, открываемыми и доступными для других.

Всем публикуемым научным материалам присваивается цифровой идентификатор DOI (Digital Object Identifier), обеспечивающий ссылку (URL) на постоянное местонахождение объекта или информацию о нём (метаданные) в интернете.

Подавляющее число авторов имеют ORCID – это уникальный код (ID), который автор научных работ получает для идентификации полученных им результатов и написанных трудов. Главная задача ORCID – однозначное определение исследователя во всех библиографических базах данных.

Ассоциация Crossref [1], членом которой с 2016 г. является Институт прикладной математики (ИПМ) им. М.В. Келдыша, поддерживает совместную всемирную службу взаимной цитируемости, функционирующую как своеобразный шлюз между электронными платформами издателей. Эта служба не хранит полные тексты научных публикаций, но заносит в свою базу данных информацию о связи публикаций с помощью технологии DOI [2], а также метаданные опубликованных научных материалов.

Разрабатываемые инструменты Crossref (и некоторых других организаций, таких как Google Scholar, Scopus и Web of Science, которые используют различные источники для своих данных цитирования) облегчают как автору публикации, так и читателям поиск, цитирование, оценку, повторное использование результатов научных исследований. Кроме того, статистика «считывания» (переходов на публикацию), а также обратных цитирований [3] (ссылок на публикации, ссылающиеся на данную статью) в той или иной мере оказывают влияние на рейтинг научных сотрудников и всего Института в целом. Поэтому ИПМ и как издателю научных материалов, и как научно-исследовательскому центру важно отслеживать корректность метаданных выпускаемых научных публикаций.

Для проверки качества метаданных Crossref разработал достаточно большой набор инструментов, которые помогают издателю оценить и улучшить свои метаданные.

В данной работе рассмотрены отчеты, к которым любой издатель или заинтересованный автор может получить доступ на сайте Crossref. Анализ инфор-

мации, представленной в этих отчетах, позволяет качественно оценить полноту отгружаемого контента, связанного с издаваемыми научными материалами, и наметить пути улучшения инфраструктуры научных публикаций и инструментов взаимодействия с Crossref.

1. СПИСОК ИЗДАНИЙ С ДОСТУПНЫМИ ДЛЯ ПРОСМОТРА МЕТАДАННЫМИ

Подробно расскажем об отчетах Crossref на примере некоторых научных изданий ИПМ и Казанского федерального университета (КФУ). На примере изданий ИПМ покажем, как издательство может с помощью отчетов Crossref находить ошибки и/или неточности в метаданных публикаций и вносить необходимые исправления. На примере изданий КФУ будет показано, как заинтересованный автор может обнаружить неточности или неполноту метаданных интересующего его издания. Далее автор может обратиться к издателю, указав на неточности, и попросить их исправить.

Любое научное издательство может таким же образом изучить отчеты Crossref, фокусируя внимание на своих научных изданиях.

Всем организациям, сотрудничающим с Crossref, предоставляется возможность не только увидеть информацию о своих изданиях, но и внести необходимые корректировки.

Информация представляется в виде трех списков, упорядоченных по алфавиту:

- Журналы
  https://www.crossref.org/06members/51depositor.html
- Сборники конференций
  https://www.crossref.org/06members/51depositorCP.html
- Книги
  https://www.crossref.org/06members/51depositorB.html

На рис. 1 представлен фрагмент алфавитного списка мировых издателей журналов, среди которых ИПМ (Keldysh Institute of Applied Mathematics) представлен тремя журналами:

- Mathematica Montisnigri (Математика Черногории);
- Mathematicheskoe modelirovanie (Математическое моделирование);
- Keldysh Institute Preprints (Препринты ИПМ).

Рис. 1. ИПМ в списке издателей журналов.

На рис. 2 представлен фрагмент алфавитного списка мировых издателей сборников материалов конференций, среди которых ИПМ (Keldysh Institute of Applied Mathematics) представлен материалами двух конференций [4]:

• Futurity designing. Digital reality problems (Проектирование будущего. Проблемы цифровой реальности) – 3 выпуска – 2018, 2019, 2020 гг.

• Scientific Services & Internet (Научный сервис в сети Интернет) – 5 выпусков – 2016 и 2017 гг., а затем объединенные в серию 2018, 2019 и 2020 гг.

На рис. 3 представлен фрагмент алфавитного списка мировых издателей журналов, среди которых КФУ (Kazan Federal University) представлен 11 журналами. Далее, в разделе 6, мы более подробно остановимся на одном из них – журнале «Электронные библиотеки» (Russian Digital Libraries Journal) [5].

В различных инструментах Crossref (при загрузке контента, определении ссылки и т. д.) в зависимости от требований простоты-сложности и оперативности-точности используется тот или иной алгоритм. Но в любом случае автору или сотруднику издательства рекомендуется самостоятельно проверить корректность результата распознавания библиографических записей и точность полученной ссылки.

Рис. 2. ИПМ в списке издателей сборников материалов конференций.

2. ОТЧЕТ ВКЛАДЧИКА

Возможно, чтобы подчеркнуть тот факт, что издатель, загружая метаданные в Crossref, получает определенную выгоду и некоторые бесплатные сервисы, авторы проекта стали использовать определенные банковские термины: размещение контента назвали депонированием, самого издателя – вкладчиком и т. д.

Отчеты вкладчика для каждого издателя используются для проверки основной информации о регистрациях DOI. Отчеты «привязываются» к трем ключевым спискам издателей, поддерживаемым Crossref, – списку издателей журналов, списку издателей сборников конференций и списку издателей монографий (книг) (см. разд. 1). В настоящее время нет отчетов вкладчиков для других типов контента, например, фотографий, картинок, видео и аудио-файлов.

Индексная страница обновляется еженедельно. Отчеты на уровне заголов-

ков обновляются по мере обновления метаданных. Можно получить и проанализировать отчет по всем изданиям ИПМ (журналам, сборникам конференций, монографиям), но мы для примера рассмотрим один из журналов.

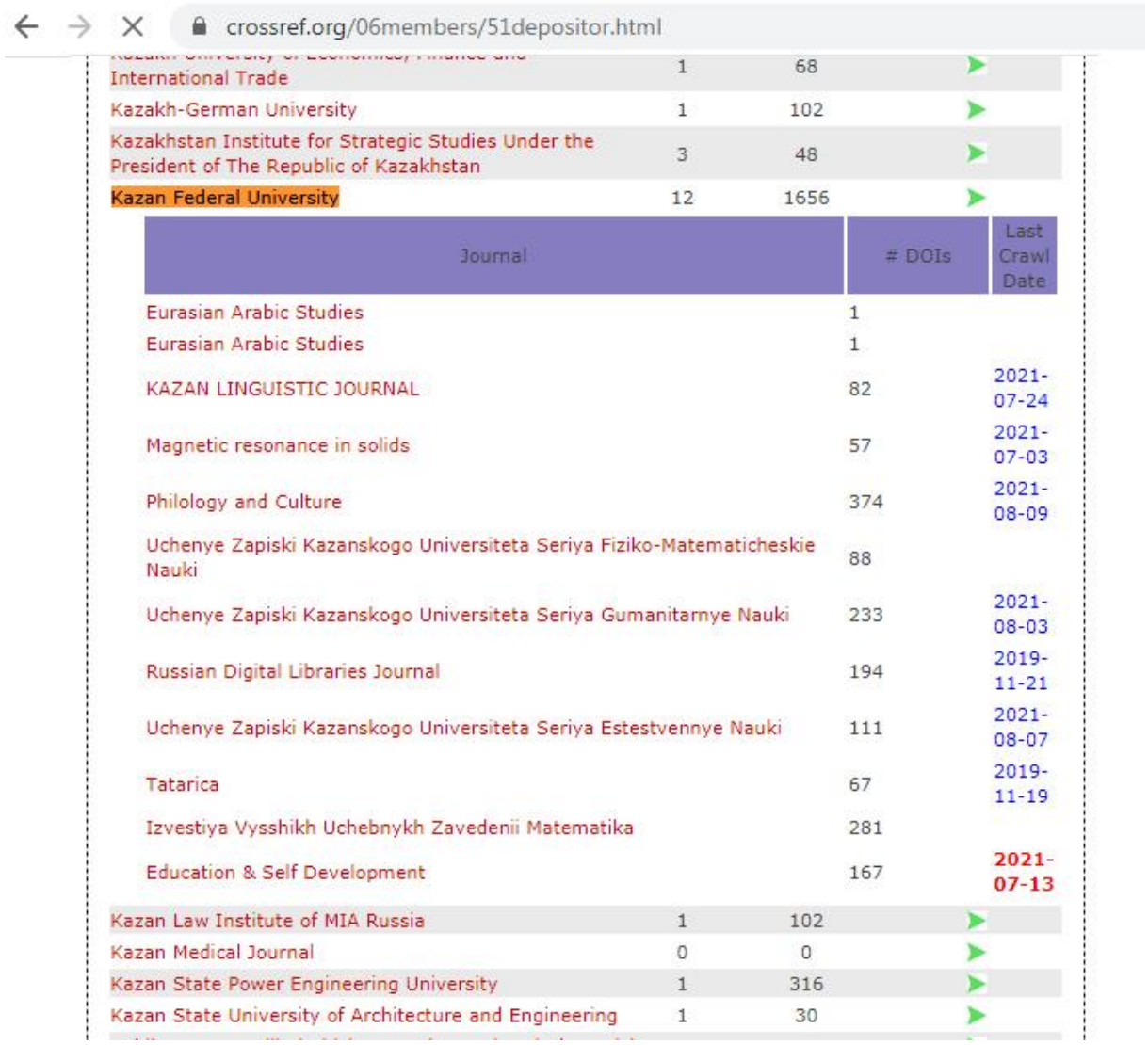


Рис. 3. КФУ в списке издателей журналов.

Выбрав в списке журналов (см. рис. 1) издательство Keldysh Institute of Applied Mathematics, а затем, выбрав, например, Mathematicheskoe modelirovanie, получим подробный отчет, где для каждого DOI указаны префикс владельца, временная метка, дата последнего обновления записи и количество цитирований (по данным Crossref) данной публикации (см. рис. 4).

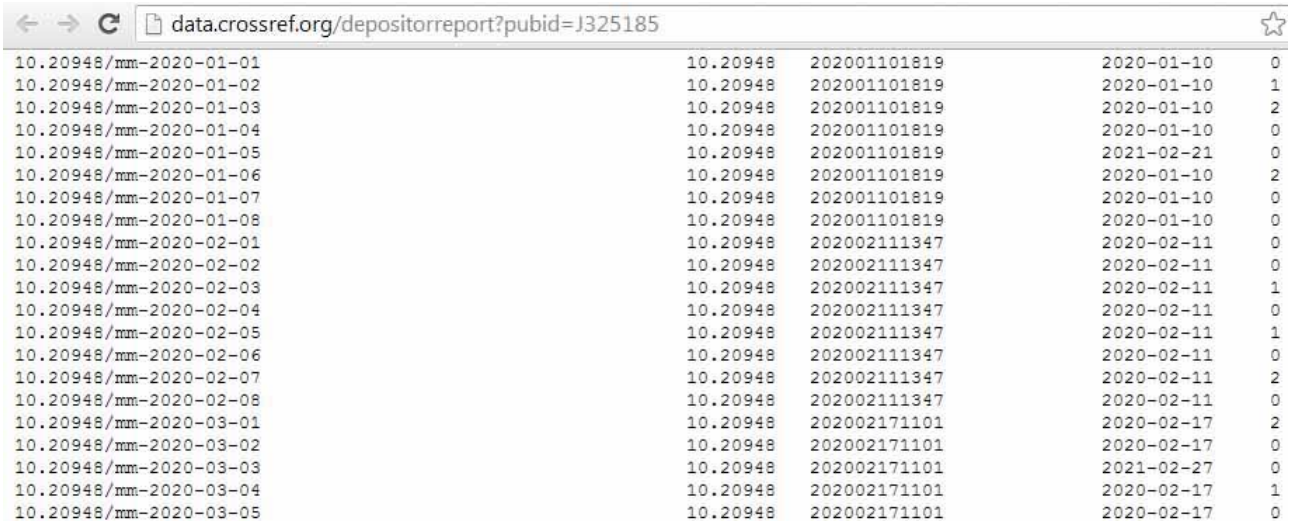


Рис. 4. Фрагмент Отчета вкладчика для журнала «Математическое моделирование».

3. ОТЧЕТ О КОНФЛИКТЕ

Как известно, DOI — это уникальный идентификатор, поэтому для каждого элемента контента всегда должен быть только один DOI. И издатель получит отчет о конфликтах, если у него есть хотя бы один конфликт с DOI.

Важно исправить эти конфликты как можно скорее, потому что они могут привести к проблемам в будущем. Наличие двух разных DOI для одного и того же контента означает, что исследователь не будет знать, какой из них цитировать, рискуя тем самым исказить количество цитирований. Кроме того, издатель может забыть, что у него есть два DOI на один объект, и обновить только один из них при изменении контента. Это означает, что любой, кто воспользуется другим DOI, который не обновляли, перейдет по неработающей ссылке. Поэтому плохие метаданные следует быстро устранить и тем самым решить проблему.

Отчет о конфликте показывает, где два (или более) DOI были отправлены с одинаковыми метаданными, или указывает на то, что у издателя научной публикации при отправке в Crossref метаданных могли быть повторяющиеся DOI.

Все конфликты DOI, связанные со статьями в журналах, сборниках конференций или монографиях, отмечаются в едином Отчете о конфликте на вебсайте Crossref (рис. 5). Если у нас есть активные конфликты, мы будем ежемесячно получать напоминание по электронной почте.

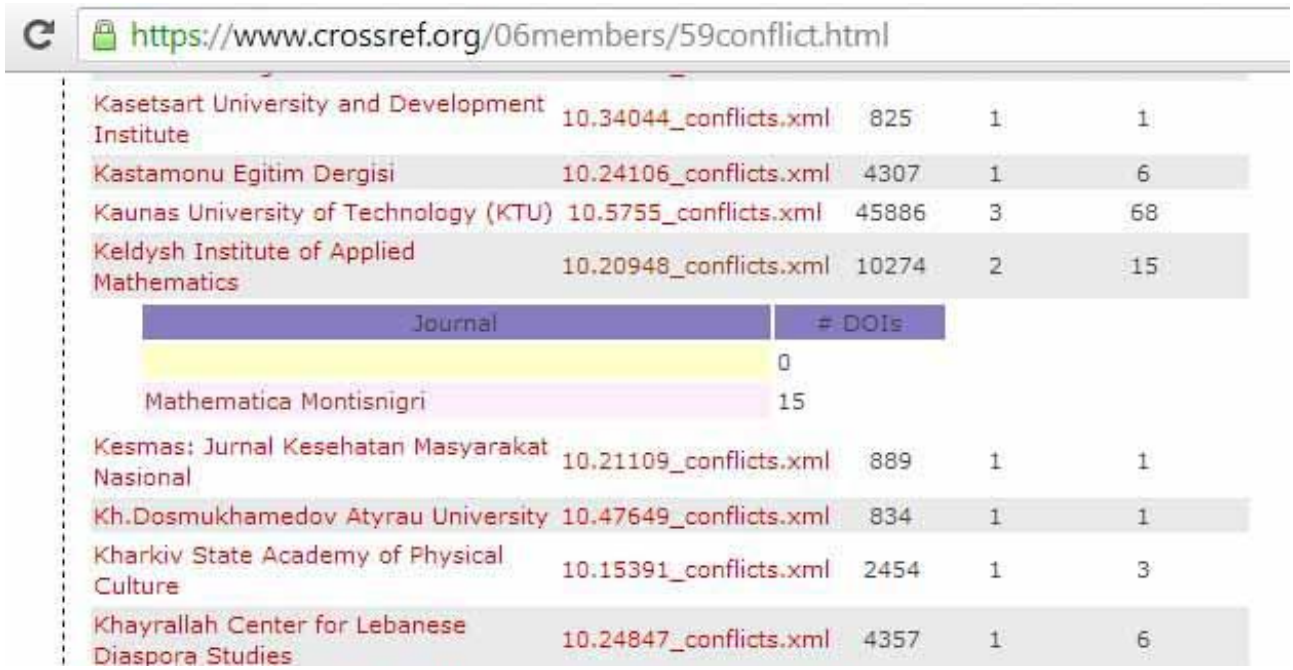


Рис. 5. Журнал «Математика Черногории» в списке конфликтов DOI.

В отчете о конфликте DOI, представленном на рис. 5, говорится о наличии 15 конфликтных ситуаций, возникших при размещении в Crossref метаданных журнала «Математика Черногории». Первая строчка в этом отчете с неуказанным названием журнала и нулевым количеством конфликтов – это небольшой глюк программистов Crossref, о чем мы им сообщили и с чем они полностью согласились.

А нам нужно разобраться в причине конфликта и постараться его исправить. Кликнув название журнала, получаем полный отчет обо всех 15 конфликтных ситуациях. Пример одной из них:

Created: 2020-03-19 14:23:13.0

ConfID: 5583368

CauseID: 1465359938

OtherID: 1462438643

JT: Mathematica Montisnigri

MD: Jokanović, 46 ,null,5,2019,A breaf survey on Armendariz and central Armendariz rings

DOI: 10.20948/mathmon-2019-46-1(Journal) (5583368-N)

DOI: 10.20948/mathmontis-2019-46-1(Journal)

Как выяснилось, предпринятая редакцией журнала попытка перейти на новые идентификаторы DOI одновременно с использованием старых идентификаторов недопустима с точки зрения идеологии DOI и технологии Crossref.

Как видно из представленного примера, первая статья 46-го тома была загружена дважды с различными идентификаторами (DOI). Хотя для каждого элемента контента всегда должен быть только один DOI, так как наличие двух разных DOI для одного и того же контента может ввести в заблуждение читателей, знакомящихся с материалами данного издания. Да и сами издатели могут запутаться при изменении метаданных и повторной загрузке контента.

Crossref предлагает 3 сценария исправления конфликтов DOI:

Сценарий 1. Если назначили два DOI разным элементам контента, но случайно отправили одни и те же метаданные для них обоих. В этом случае один из DOI имеет неверные метаданные. Если повторно загрузить исправленные метаданные этого DOI, конфликт будет разрешен.

Сценарий 2. Если назначили два DOI одному и тому же элементу контента. В этом случае вы можете разрешить конфликт, назначив один из DOI в качестве основного, а другой — в качестве псевдонима. Псевдоним DOI будет автоматически перенаправляться на основной DOI, поэтому достаточно будет поддерживать только основной.

Сценарий 3. Если два DOI относятся к разным элементам контента, но их метаданные настолько похожи, что был отмечен конфликт. Это происходит, когда в элементы включено очень мало метаданных. Лучше всего зарегистрировать дополнительные метаданные, чтобы устранить конфликт. Или же можно принять конфликт, удалив статус конфликта и установив для него статус «разрешено». Это не повлияет на записи метаданных или DOI, но устранит конфликты из отчета о конфликтах.

Для нас самым простым решением было удалить «неверный» DOI. Но удалять DOI нельзя – это фундаментальный принцип DOI. Поэтому было принято решение идти по второму пути, предложенному Crossref, – «неверный» DOI назвать псевдонимом «правильного» с помощью подсистемы администрирования DOI (рис. 6).

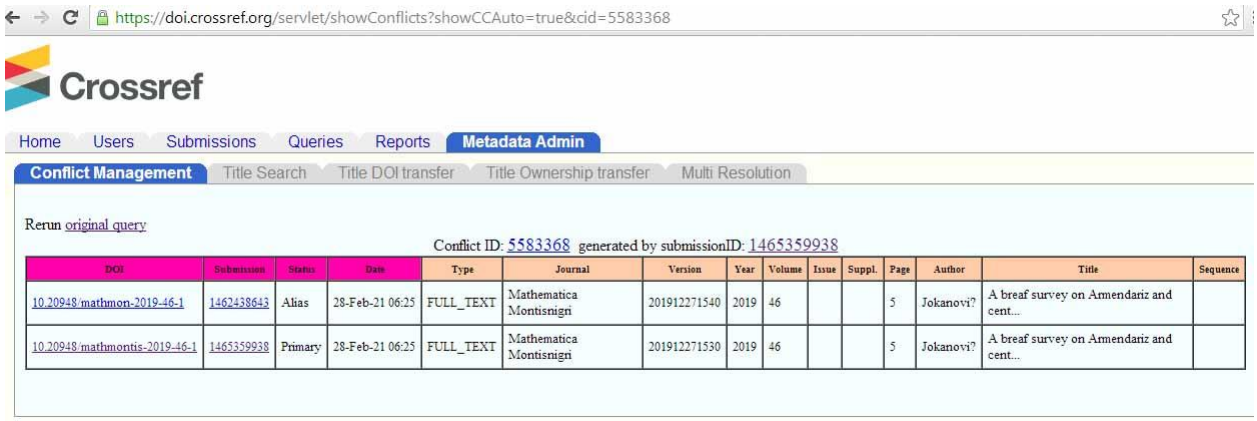


Рис. 6. Изменение статуса DOI для разрешения конфликта.

После изменения статуса «неверного» DOI любое обращение к нему вызовет автоматический переход по ссылке, указанной в первичном («правильном») DOI.

Проведя таким образом обновление контента всех «неверных» DOI и изменив их статус на «псевдоним», мы исключили издания ИПМ из списка конфликтов DOI.

4. ОТЧЕТ О ПОЛЯХ ИЛИ ОТСУТСТВУЮЩИХ МЕТАДАННЫХ

Отчет о полях или отсутствующих метаданных содержит подробную информацию о полноте метаданных. Он так же, как и «Отчет вкладчика», «привязывается» к одному из трех ключевых списков издателей, поддерживаемых Crossref, – списку издателей журналов, списку издателей сборников конференций и списку издателей монографий (книг) (см. разд. 1).

К данному отчету можно получить доступ, выбрав значок (зеленая стрелка вправо) рядом с именем издателя научных материалов в одном из вышеуказанных списков.

Следует отметить, что наборы метаданных, передаваемых для статей журналов, научных публикаций в сборниках конференций и монографий, несколько различаются. Для любой научной публикации передаются ее название на русском и английском языках, список авторов (для каждого автора имя (First name) и фамилия (Surname) на русском и английском языках, а также ORCID), год издания, количество или диапазон страниц. Для журналов и сборников добавляются номер тома и/или номер выпуска. Кроме того, последние 2 года мы стали за-

гружать в Crossref аннотации научных публикаций и списки литературы, что очень важно для отслеживания взаимного цитирования [3].

Рассмотрим информацию по нашим журналам, представленную в отчетах о полях или отсутствующих метаданных (рис. 7).

REPORT DATE:

Publication Title	Ignore Fields	V	I	P	A	S	V/I	V/I/P	V/I/P/A	P/A	T	N	F	missing iParadigms URLs	Total DOIs[4]
Keldysh Institute Preprints	none CHANGE	878	0	0	0	275	0	0	0	0	0	0	0	na	878
Matematicheskoe modelirovanie	none CHANGE	25	0	0	0	29	0	0	0	0	0	0	25	na	162
Mathematica Montisnigri	none CHANGE	0	89	0	2	21	0	0	0	0	0	2	1	na	89

Рис. 7. Отчет о полях или отсутствующих метаданных в журналах ИПМ.

Заголовки столбцов таблицы на рис. 7 (красным отмечены поля, с точки зрения Crossref неполные или некорректно заполненные):

- Participation Title – название издания;
- Ignore Fields – указание, отсутствие каких полей следует игнорировать;
- V=volume –номер тома;
- I=issue – номер выпуска;
- P=page – количество или диапазон страниц;
- A=author – автор;
- S=single-author – единственный автор;
- T=articletitle – название статьи;
- N=no-first-name – не задано имя автора;
- F=first name initial only – в качестве имени заданы только инициалы автора;
- U=missing iParadigms Url– Crossref сотрудничает (с 2008 г.) с iParadigms LLC, предлагая своим членам — ведущим научным и профессиональным издателям — возможность проверки оригинальности работ с помощью служб CrossCheck и iThenticate. База данных CrossCheck включает полнотекстовые журналы ведущих академических издателей и достаточно быстро растет по мере того, как издатели-участники Crossref подписываются на эту услугу.

Как мы видим, с помощью Отчета о полях или отсутствующих метаданных Crossref подчеркивает – несмотря на то, что некоторые библиографические ме-

таданные являются необязательными для целей регистрации контента, настоятельно рекомендуется всем издателям регистрировать максимально полные метаданные для каждого зарегистрированного элемента. И даже выделяет красным цветом в отчете поля, для которых, с точки зрения Crossref, издатель грубо не выполняет указанные рекомендации.

Анализируя информацию, представленную на рис. 7, можно отметить, что для всех выпусков Препринтов ИПМ не заданы номера тома, для всех томов журнала «Математика Черногории» не заданы номера выпусков. Для журнала «Математическое моделирование» замечания касаются не всех выпусков, а только нескольких томов начального периода присвоения DOI научным публикациям, когда нашему журналу присваивали DOI какие-то внешние организации, даже некорректно зарегистрированные в Crossref.

Следует отметить, что всем статьям журнала «Математика Черногории» DOI присваивали мы. И, соответственно, можем вносить необходимые исправления и/или дополнения в наборы метаданных, загруженные в Crossref. С журналом «Математическое моделирование» ситуация более сложная – до 2020 года в результате конкурсов издание переходило из рук в руки и получало совершенно различные DOI. При этом отгружались метаданные научных статьей журнала, абсолютно различные по полноте и корректности. И теперь ИПМ, как издатель журнала, не имеет ни прав, ни возможностей что-либо исправить или дополнить в информации о выпусках 2016–2019 гг.

Анализируя далее замечания Crossref по регистрации метаданных публикаций, мы видим отметки в столбце задания единственного автора (S=single-author) и некоторые другие.

Похожие замечания имеются по метаданным сборников конференций и монографий. Мы с этими (и другими) замечаниями Crossref внимательно разбираемся и по возможности стараемся устранить, внося необходимую правку в описание метаданных.

Однако издателю конкретного журнала можно отказаться от подобной навязчивости Crossref с помощью переключателя Change (второй столбец, рис. 7 – IgnoreFields), указав, отсутствие каких полей следует игнорировать (рис. 8).

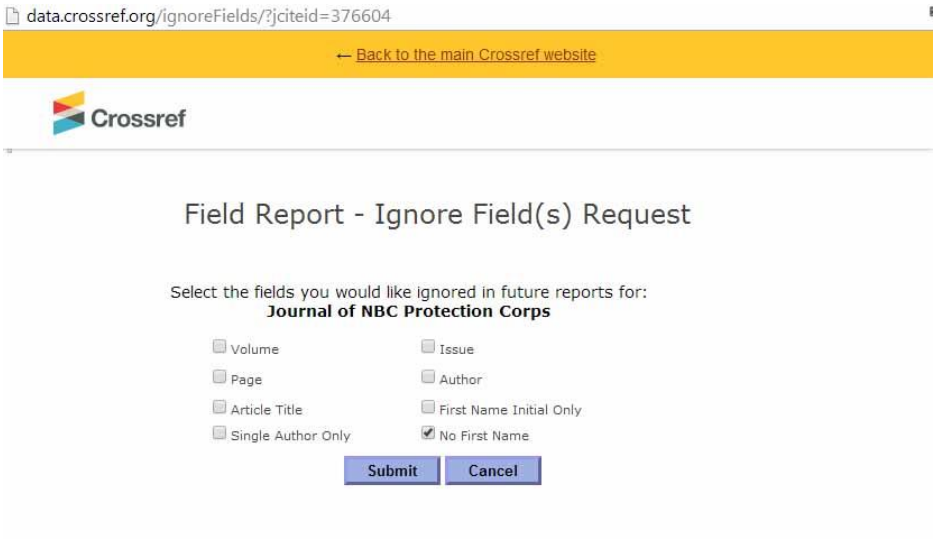


Рис. 8. Указание, отсутствие каких полей следует игнорировать.

5. ОТЧЕТ СКАНЕРА DOI

Отчет сканера DOI выполняется только для журналов и, соответственно, «привязывается» к списку издателей журналов.

Сканер DOI делает выборку статей для каждого журнала конкретного издателя, чтобы убедиться, что заданные DOI переводятся на соответствующую страницу. Для каждого просматриваемого журнала выбираются DOI, количество которых равно примерно 5% от общего числа DOI для журнала (максимум до 50 DOI). На рис. 9 перечислены все журналы, издаваемые ИПМ, и для каждого указаны общее количество DOI и дата последнего сканирования.

Рис. 9. Дата последнего сканирования журналов ИПМ.

Получить доступ к деталям работы поискового робота для данного журнала можно, выбрав дату в столбце «Дата последнего сканирования».

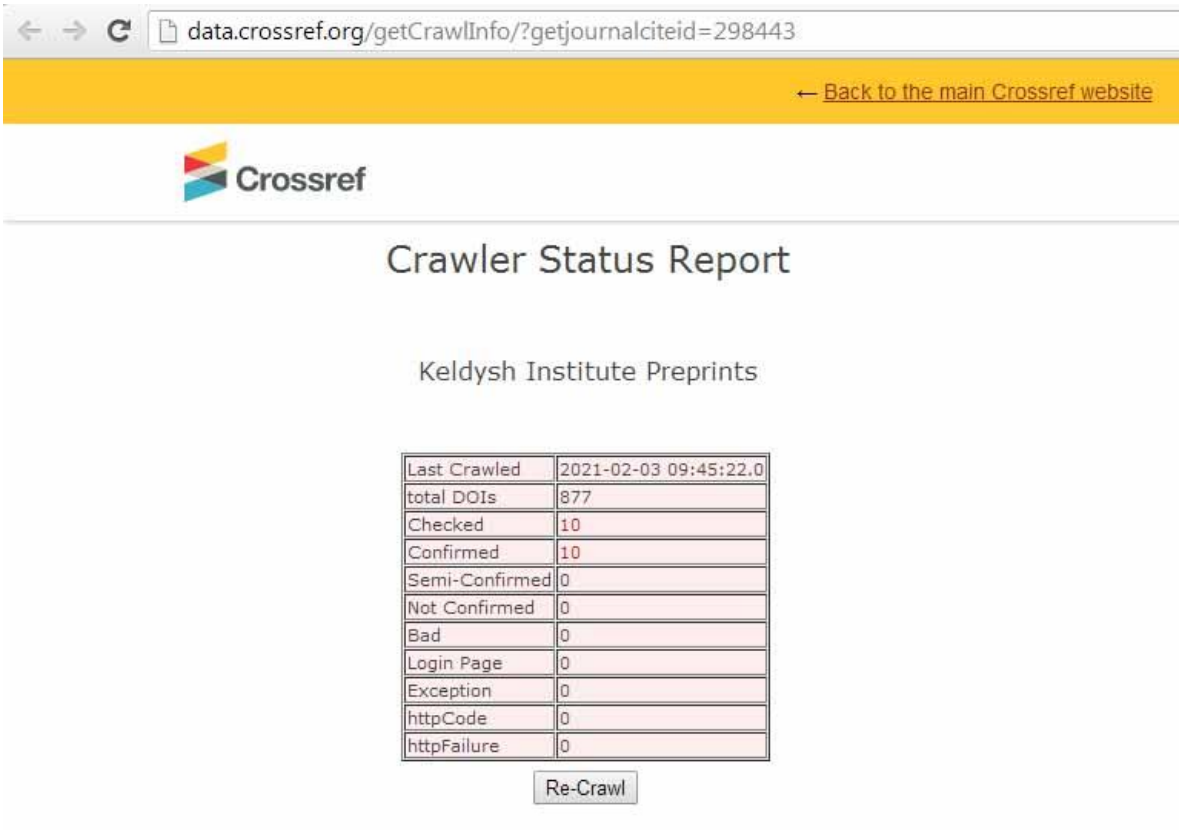


Рис. 10. Отчет о работе сканера DOI для препринтов ИПМ.

Никаких ошибок, как следует из отчета сканера на рис. 10, у Препринтов ИПМ нет. При этом просканировано (и подтверждена корректность) 10 DOI. Эти поля «кликабельны», поэтому можно посмотреть более подробно результаты процедуры сканирования (рис. 11, 12).

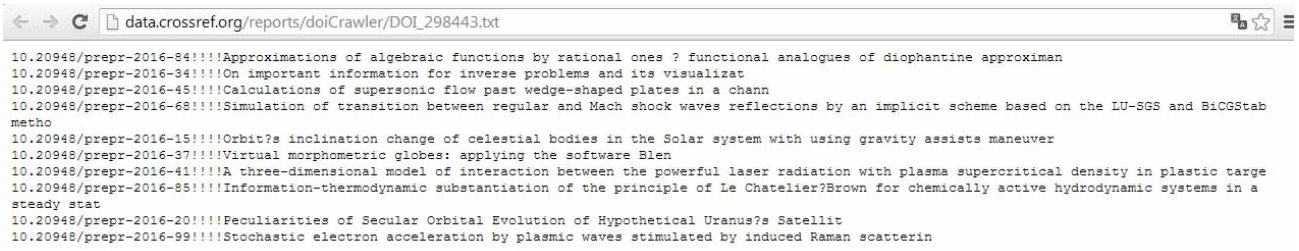


Рис. 11. Список выбранных для сканирования DOI.

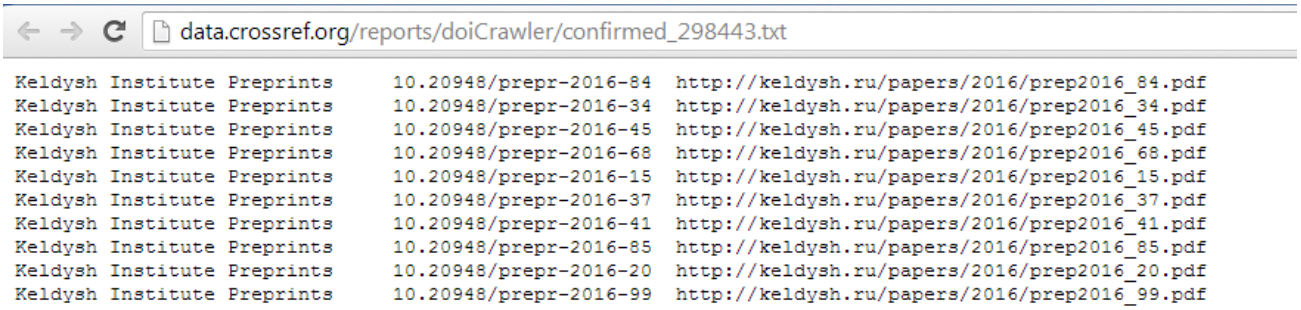


Рис. 12. Проверенные и «подтвержденные» DOI.

Эта функция (сканирование DOI), на наш взгляд, достаточно интересная и полезная, особенно для тех изданий, у которых изменялись сервер базы данных или администрация, так как такого рода изменения могли внести серьезные ошибки в процедуру депонирования контента и сделать недоступными материалы изданий.

Однако реализация этой функции программистами Crossref представляется немного странной и вызывает несколько вопросов:

- Почему, например, для Препринтов ИПМ, насчитывающих 877 выпусков с начала присвоения DOI, проверяется только 10 элементов, хотя в описании алгоритма говорится о 5% от общего числа?
- Почему при повторном сканировании (Re-Crawl) используется та же выборка, что и несколько лет назад, при предыдущей проверке? Ведь издателю интересно отслеживать корректность не только старых материалов, но и всех публикаций исследуемого журнала.

Пока что мы не получили ответа на эти вопросы от Службы поддержки Crossref, но продолжаем активное взаимодействие с ней.

Еще одно важное дополнение связано с тем, что Crossref – это один из регистраторов DOI. А так как регистраторы DOI не обмениваются метаданными, то издатели, связанные с другими регистрационными сервисами, не смогут воспользоваться инструментами Crossref.

6. ОТЧЕТ ОБ УЧАСТИИ

Авторы проекта и разработчики Crossref призывают издателей не просто размещать метаданные научных публикаций, а делать их максимально полными. Кроме того, Crossref призывает научное и издательское сообщество активно

пользоваться предлагаемыми сервисами, разработанными для анализа полноты и корректности загружаемой информации, тем самым как бы участвуя в развитии и расширении набора этих услуг.

Для каждого издателя, сотрудничающего с Crossref, существует отдельный Отчет об участии (Participation report), который показывает, какой процент их депонированных данных зарегистрирован для каждого из десяти ключевых элементов метаданных. Отчеты об участии наглядно показывают, где есть пробелы и что можно улучшить в плане полноты метаданных.

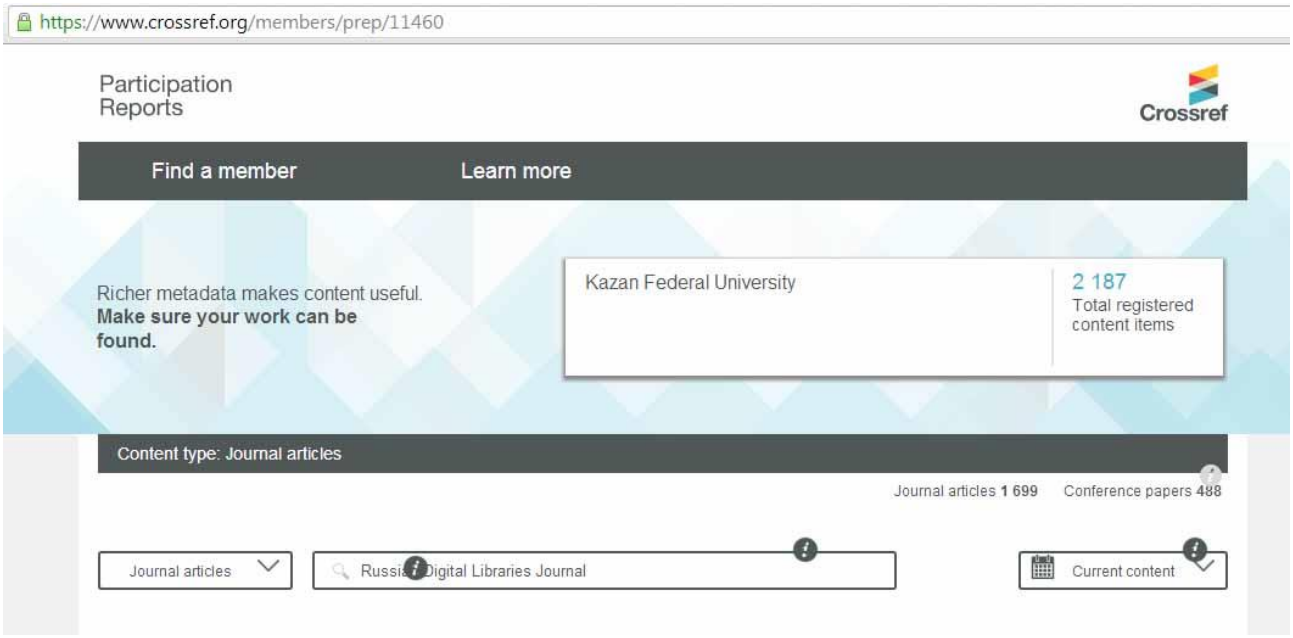


Рис. 13. Отчет об участии для научных публикаций КФУ.

На рис. 13 представлен заголовок Отчета об участии издателя Kazan Federal University – общее число элементов контента 2187, в том числе журнальных статей 1699, статей в сборниках материалов конференций 488. В нижней части заголовка (рис. 13) есть 2 меню – выбор типа научной публикации (слева) и выбор анализируемого периода (все время депонирования, текущий период – последние 2 года, «старые» материалы – данные, загруженные более 2 лет назад). Центральное поле заголовка Отчета об участии позволяет ввести название журнала, сборника или даже название публикации и проанализировать полноту соответственно загруженных метаданных.

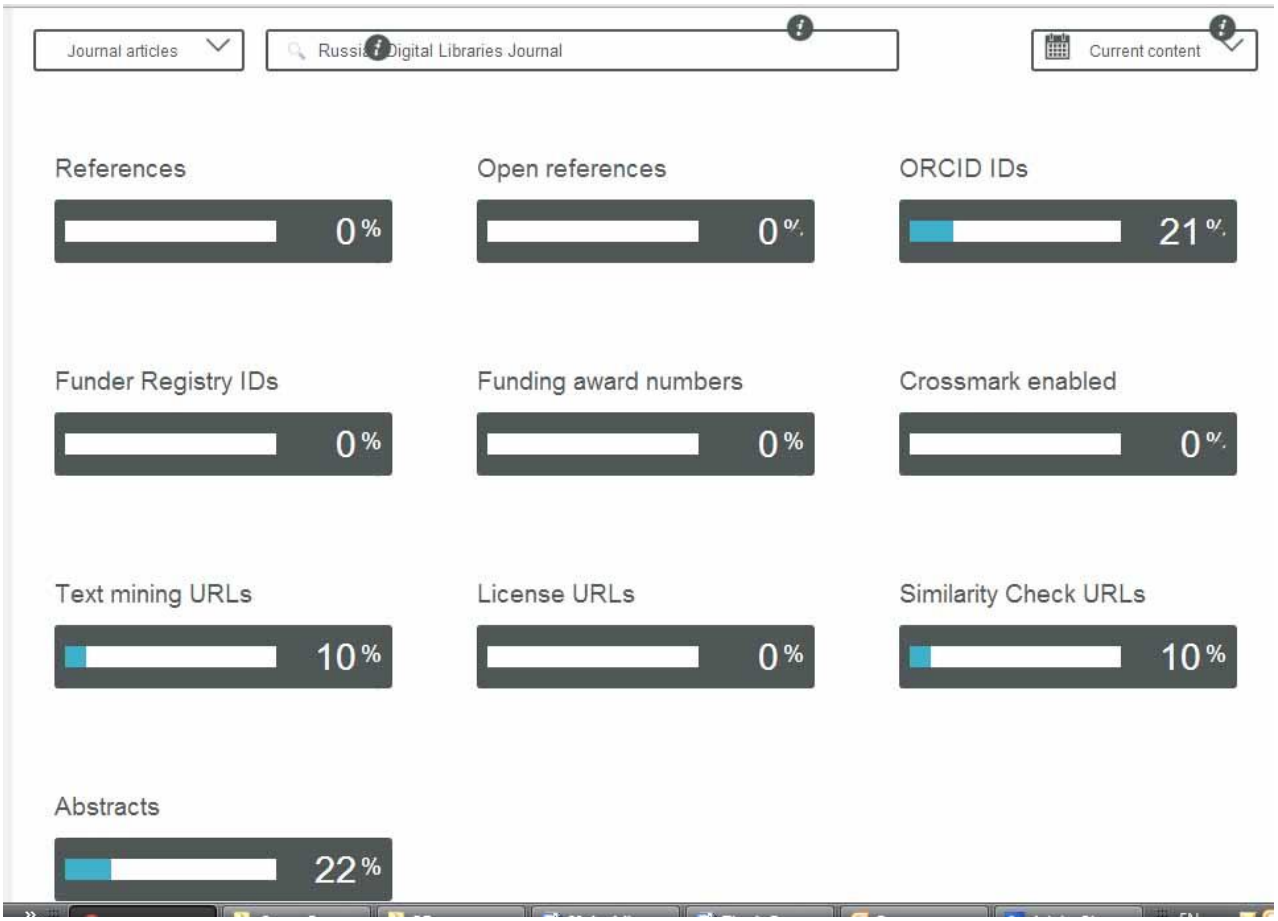


Рис. 14. Процентные показатели Отчета об участии журнала «Электронные библиотеки» (“Russian Digital Libraries Journal”).

Основная часть Отчета об участии для журнала «Электронные библиотеки» показана на рис. 14 (задан текущий период (current time) – последние 2 года):

- для 0 % публикаций загружен список литературы (References);
- 0 % ссылок открыты (Open References) – доступны всем пользователям всех сервисов Crossref (нет доступных ссылок, так как нет списков литературы);
- для 21 % авторов указан ORCID;
- для 0 % публикаций указаны имя и идентификатор (Funder Registry IDs) спонсора – хотя бы одной из организаций, финансировавших исследование;
- для 0 % публикаций указан номер гранта финансирования (Funding award numbers);
- доля контента (в данном случае 0 %), использующего службу Crossmark (Crossmark-enabled), которая дает читателям быстрый и легкий доступ

к текущему статусу элемента контента (в рамках политики издателя в отношении исправлений, опровержений, отзыва и других обновлений);

- процент зарегистрированного контента (в данном случае 10 %), содержащего URL-адреса для интеллектуального анализа текста и данных (Text-mining URLs) научной публикации – автоматического анализа и извлечения информации из большого количества документов. В настоящий момент большинство научных организаций мира (и КФУ, как нам кажется, в том числе) не заинтересованы в задании специального набора инструкций, с помощью которых кто-то зачем-то будет исследовать их научные материалы;
- процент метаданных публикаций (в данном случае 0 %), содержащих URL-адреса, указывающие на лицензию (License URLs), определяющую условия, на которых читатели могут получить доступ к контенту;
- процент метаданных публикаций (в данном случае 10 %), которые включает URL-адреса для проверки схожести (Similarity Check URLs), для изданий, сотрудничающих с CrossCheck и iThenticate;
- 22 % метаданных публикаций включают аннотации (Abstracts), что дает более глубокое понимание содержания работы.

На наш взгляд, не стоит гнаться за 100 % показателями, но при этом должно быть понятно, что более полное и аккуратное заполнение метаданных публикации в той или иной мере влияет [6] на рейтинги изданий, авторов и организаций. А указание грантов и фондов поддержки научной деятельности положительно влияет на взаимоотношения с этими фондами.

СПИСОК ЛИТЕРАТУРЫ

1. Ассоциация Crossref. URL: https://www.crossref.org/about/
2. International DOI Foundation (IDF). URL: https://www.doi.org/
3. *Ермаков А.В.* Библиографическая ссылка как инструмент автора и читателя // Научный сервис в сети Интернет: труды XXII Всероссийской научной конференции (21–25 сентября 2020 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2020. С. 268–275. https://doi.org/10.20948/abrau-2020-55
URL: https://keldysh.ru/abrau/2020/theses/55.pdf
4. *Слепенков М.И.* Материалы конференций в онлайновой библиотеке ИПМ им. М.В. Келдыша // Препринты ИПМ им. М.В. Келдыша. 2020. № 18. 16 с.

http://doi.org/10.20948/prepr-2020-18
URL: http://library.keldysh.ru/preprint.asp?id=2020-18

5. Журнал «Электронные библиотеки».URL: https://elbib.ru
6. *Полилова Т.А.* Инфраструктура научных публикаций // Препринты ИПМ им. М.В. Келдыша. 2009. № 15. 30 с.
URL: http://library.keldysh.ru/preprint.asp?id=2009-15

IMPROVING THE QUALITY OF METADATA SCIENTIFIC PUBLICATIONS WITH CROSSREF REPORTS

Alexey V. Ermakov[0000-0002-6054-0813]

Ermakov@Keldysh.ru

Keldysh Institute of Applied Mathematics, Miusskaya sq., 4, Moscow, 125047, Russia

Abstract

Issues related to improving the quality of metadata of scientific publications placed in the Crossref bibliographic database are considered. All information contained in metadata obtained from publishers of scientific publications, Crossref analyzes and displays in various reports. The reports give publishers an idea of the completeness and correctness of the bibliographic data provided. The quality of metadata directly or indirectly affects the number of views and links to a publication, respectively, on the ratings of scientific publications, authors and organizations.

Keywords*: metadata of publications, Crossref reports, citations, ratings of scientific publications.*

REFERENCES

1. Crossref Association. URL: https://www.crossref.org/about/
2. International DOI Foundation (IDF). URL: https://www.doi.org/
3. *Ermakov A.V.* Bibliographic reference as a tool for the author and reader // Conference material: “Scientific service & Internet: proceedings of the 22nd All-Russian Scientific Conference (September 21-25, 2020, online)”. M.: KIAM
URL: https://keldysh.ru/abrau/2020/theses/55.pdf

4. *Slepenkov M.I.* Conference proceedings in the on-line library of Keldysh Institute // KIAM Preprint № 18, Moscow, 2020. 16 p.
http://doi.org/10.20948/prepr-2020-18
URL: http://library.keldysh.ru/preprint.asp?id=2020-18

5. Russian Digital Libraries Journal. URL: https://elbib.ru

6. *Polilova T.A.* Infrastructure of scientific publications // KIAM Preprint № 15, Moscow, 2009. 30 p.
URL: http://library.keldysh.ru/preprint.asp?id=2009-15

СВЕДЕНИЯ ОБ АВТОРЕ

Ермаков Алексей Викторович – старший научный сотрудник Института прикладной математики им. М.В. Келдыша Российской академии наук

Alexey V. Ermakov – Senior Researcher, Keldysh Institute of Applied Mathematics.

email: Ermakov@Keldysh.ru,
ORCID: 0000-0002-6054-0813

Материал поступил в редакцию 24 октября 2021 года

УДК 004.6, 004.9

ЭЛЕКТРОННЫЕ АРХИВЫ ДЛИТЕЛЬНОГО СРОКА ЖИЗНИ: МОДЕРНИЗАЦИЯ И ИНТЕГРАЦИЯ

А. Г. Марчук[1] [0000-0001-8455-725X], С. Н. Трошков[2] [0000-0003-2952-9509], И. А. Крайнева[3] [0000-0002-0601-9795]

[1-3]Институт систем информатики им. А.П. Ершова СО РАН

[1]mag@iis.nsk, [2]kamronis@xtech.ru, [3]cora@iis.nsk

Аннотация

За период около двадцати лет в Институте систем информатики им. А.П. Ершова Сибирского отделения РАН (ИСИ СО РАН) были созданы информационные системы исторической направленности: Электронный архив академика А.П. Ершова, Фотоархив Сибирского отделения РАН, Архив газеты «Наука в Сибири», Открытый архив СО РАН и др. У каждого из этих ресурсов есть своя специфика, но в целом их контент базируется на общей социальной и территориальной основе научной и общественной деятельности СО АН СССР/РАН и Новосибирского Академгородка. В статье рассмотрены некоторые проблемы интеграции/дезинтеграции разрозненных электронных ресурсов на общую платформу на базе имеющихся и создаваемых инструментов.

Ключевые слова*: интеграция электронных ресурсов, качественная информация, открытые архивы, междисциплинарность, история науки, Сибирское отделение РАН, проприетарное ПО, Semantic Web, Drupal.*

ВВЕДЕНИЕ

С середины 1990-х гг. интернет и электронные ресурсы стремительно захватывают пространство деловой, общественной и научной активности России. Центры использования информационных технологий в историко-культурных, научно-теоретических, историко-биографических и прочих исследованиях помимо Москвы (чл.-корр. Л.И. Бородкин, д. и. н. И.М. Гарскова, д. и. н. Ю.Ю. Юмашева), действуют в Новосибирске (к. и. н. Ю.П. Холюшкин, д. ф.-м. н. А.Г. Марчук), Барнауле (д. и. н. В.Н. Владимиров), Томске (д. и. н. С.А. Некрылов), Ижевске (д. филолог. н. В.А. Баранов), Перми (д. и. н. С.П. Корниенко), Краснояр-

ске (И.А. Кижнер) и др. Государственные архивы, которые, как правило, являются основным местом паломничества исследователей, проводят оцифровку научно-справочного аппарата [1].

В работе пойдет речь о цикле жизни, возможностях и перспективах подобных ресурсов.

1. ИНФОРМАЦИОННЫЕ СИСТЕМЫ ДЛЯ ИССЛЕДОВАНИЙ В ГУМАНИТАРНЫХ НАУКАХ

Под информационными системами (ИС) мы понимаем совокупность технического, программного, организационного и финансового обеспечения, а также персонала, способного обеспечивать работоспособность этого комплекса и выполнение проекта. Минимальное количество персонала в таком проекте по опыту ИСИ СО РАН – около 10 человек: программисты, историки, информационные специалисты (операторы), переводчики. Специалистами в области применения информационных технологий в гуманитарной сфере из Пермского университета предложена спецификация историко-ориентированных систем как «особого класса систем, предназначенных для хранения, организации исторической информации, обеспечения доступа к ней и ее аналитической обработки в соответствии с потребностями исторических исследований и/или образования» [2]. Особый интерес вызывают системы, содержащие помимо исторической информации исследовательский инструментарий (поисковый, аналитический, распознавание текста, внутритекстовые гиперссылки и др.). Авторы выделяют два подхода к созданию ИС: источнико-ориентированный, когда основой системы является массив одного источника, а его структура становится моделью системы, и проблемно-ориентированный, когда модель строится на основе рассматриваемой предметной области.

В соответствии с данной классификацией ИС, созданные в ИСИ СО РАН, являются источнико-ориентированными: Фотоархив СО РАН [3] вмещает на своей платформе два разных источника – фотодокументы и газету «Наука в Сибири», материалы, связанные тематически и органически, поскольку многие фото выполнены фотокорреспондентами – сотрудниками редакции газеты. Электронный архив А.П. Ершова и Открытый архив СО РАН [4] помимо сканированных документов содержат сканы фотографий и печатных научных трудов. Мы считаем

наши ИС источнико-ориентированными и по другой причине: по наличию образов подлинных документов (полнотекстовые ИС), транскрипция которых является дополнением, позволяющим знакомиться с трудночитаемыми текстами. Кроме того, наши ИС являются интернет-ориентированными: то, что системы названы электронными, не дает представления о степени их открытости. И в завершение данной темы отметим, что наши ИС – это системы, позволяющие осуществлять описание документов с удаленных рабочих мест – распределенные. Итак, наши ИС – источнико-ориентированные, поскольку позволяют знакомиться с образами подлинных документов (полнотекстовые), интернет-ориентированные распределенные системы. Источнико-ориентированный подход демонстрируют создатели таких ресурсов, как «Берестяные грамоты» [5], «Подвиг народа» [6], Архив академика В.И. Вернадского [7] и др.

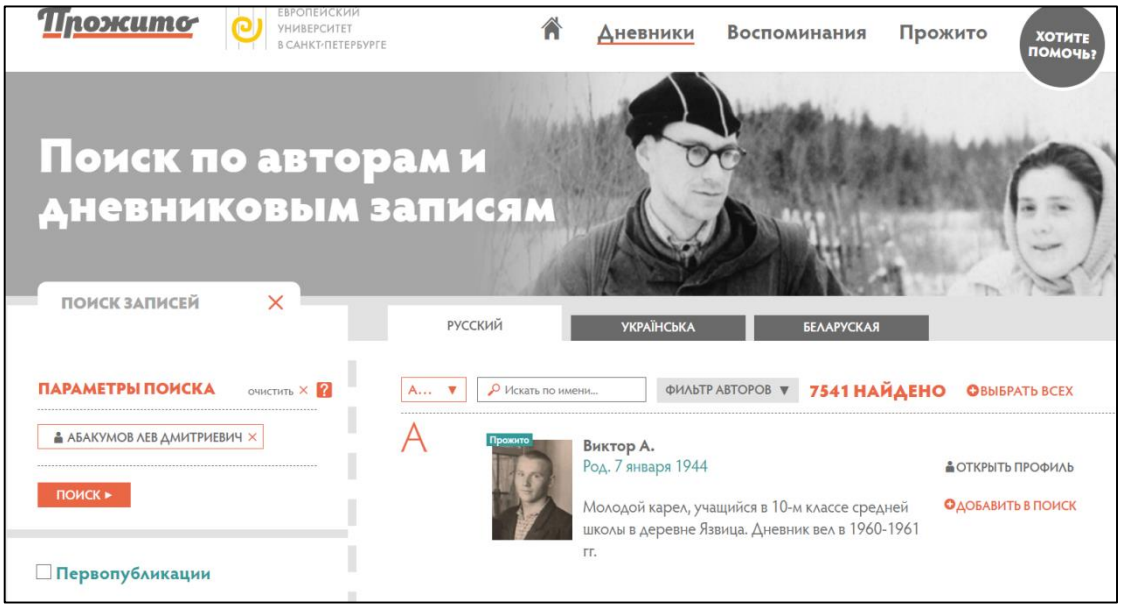


Рис. 1. Страница корпуса «Дневники» на сайте «Прожито»

Смешанный подход (проблемно-ориентированный и источнико-ориентированный), на наш взгляд, демонстрируют те историко-ориентированные ИС, которые, помимо образов документов, таких как фотографии, карты-схемы, содержат преимущественно транскрипции рукописных документов. К таковым можно отнести, например, проекты «Бессмертный барак» [8], «Истмат» [9]. На «Истмате» для транскрибированных документов при-

ведены ссылки на архивные источники. Сайт «Прожито», на котором собраны личные дневники и воспоминания, содержит не только сканы артефактов, но и транскрипции уже опубликованных документов, а также ссылки на аналогичные источники в других ИС, в частности, в музейном Госкаталоге (Рис. 1) [10]. При консультационной поддержке ИСИ СО РАН выполнен проект «Электронный архив Ольги Михайловны Фрейденберг (1890–1955) [11]. О.М. Фрейденберг – филолог-классик, философ культуры, двоюродная сестра поэта Бориса Пастернака. В контент архива включено исследование, выполненное на его материалах.

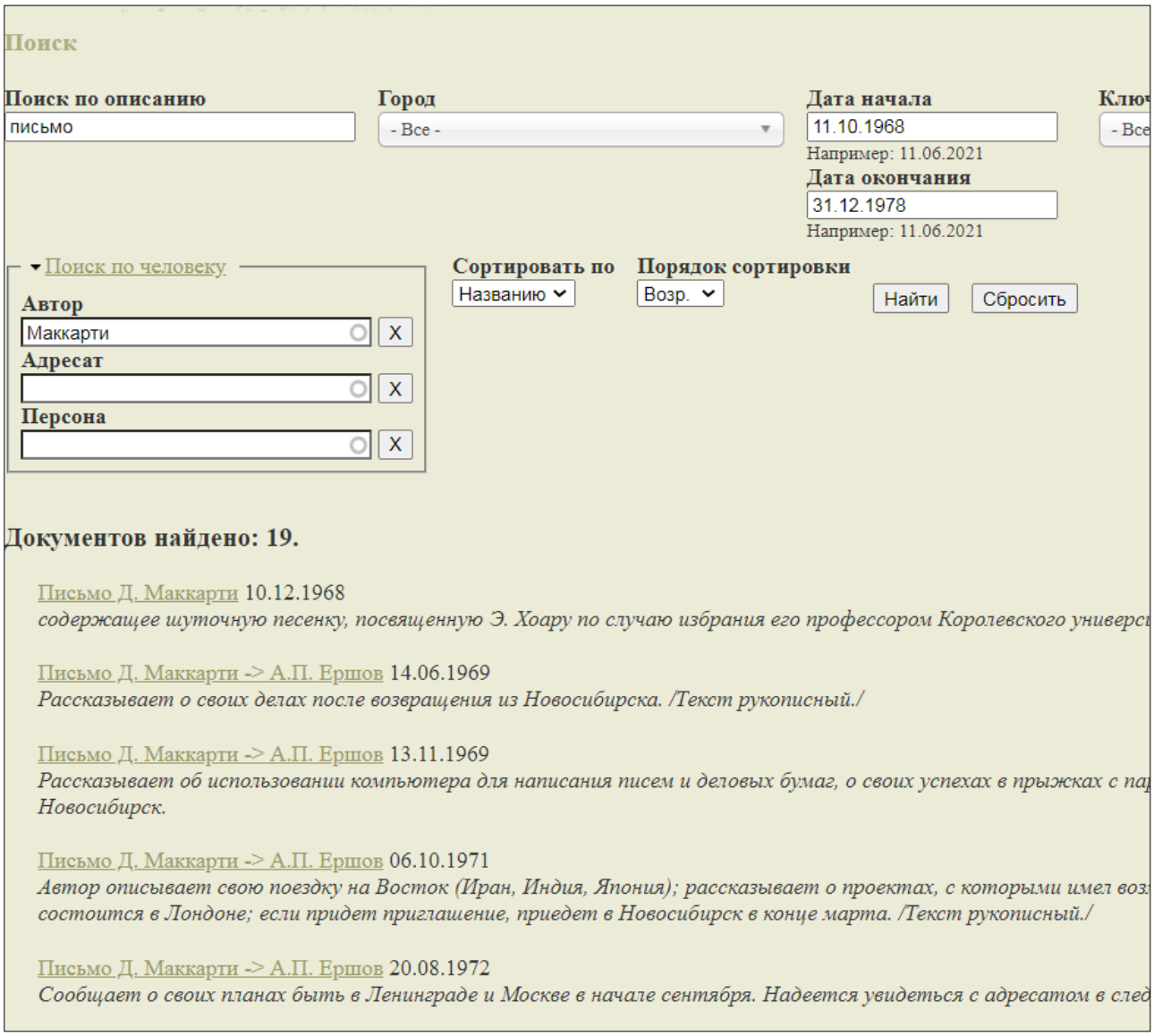


Рис. 2. Пример поискового запроса с указанием параметров автор, тип документа, дата в Электронном архиве академика А.П. Ершова

2. ВКЛАД ИСИ СО РАН

Методологической основой использования информационных технологий в исследовательской работе гуманитариев является междисциплинарность. С течением времени сформирован инструментальный аспект коммуникативных процессов, которые осуществляются при посредстве ИТ, сетевой организации вычислительной техники и поддерживают доступность контента наследия, размещенного в Сети. Рабочий инструментарий привлечения ИТ в архивную работу – электронная историческая фактография, технология и метод – совокупность приемов, на основе которых создаются информационные системы (ИС) для размещения массива разнородных документов в Сети, систематизации их путем установления связей между сущностями, отраженными в документах. Документы могут цитироваться как электронной ссылкой, так и указанием на дело и лист в архиве (в частности, в архиве А.П. Ершова, в других архивах – если документы поступили из государственных хранилищ). Совершенствуется и подход к технике сканирования и визуализации артефактов: в последнее время мы стремимся к визуализации этапов графической обработки документов, в частности, фотографий. На страницу документа в электронном архиве помещаются не только отреставрированный вариант фотографии, но и исходное сканированное изображение, а также изображение обратной стороны фотографии без графической обработки (Рис. 3).

Очевидно, уже нет необходимости доказывать нужность и полезность открытых архивных источников, особенно в условиях значительного ограничения социальной мобильности. И чем полнее будут архивные ИС, тем более продуктивной станет исследовательская работа. В том случае, когда ИС заполняется достаточно репрезентативными и разнообразными источниками, она должна быть снабжена прецизионным поисковым инструментарием, а сами источники, помимо этого, должны быть систематизированы. Открытые архивы СО РАН в данный момент имеют несколько уровней систематизации: фонд, коллекции фонда, подколлекции коллекций (или группы и подгруппы, как в архиве А.П. Ершова). Поисковая же система довольно проста: поиск осуществляется по одному из параметров: типу документа, имени персоны, названию организации, ключевому слову. Поиск по типу документа не дает прецизионного результата, поскольку к

одному типу документов могут относиться артефакты разных фондообразователей. Инструментарий архива академика А.П. Ершова позволяет ограничивать поиск хронологическими рамками, именем персоны, типом документа, то, что и требуется в идеале (Рис. 2). Задача интеграции нескольких архивных ресурсов на одной платформе должна учитывать это обстоятельство, иначе массив станет слабо изучаемым. Чем больше поисковых возможностей будет у исследователя, тем эффективней он сможет использовать возможности открытых архивов.

Рис. 3. Визуализация фотодокументов в Открытом архиве СО РАН

3. ЖИЗНЕННЫЙ ЦИКЛ И МОДЕРНИЗАЦИЯ ИНФОРМАЦИОННЫХ СИСТЕМ

Самой «древней» информационной системой, созданной и развивающейся в нашем коллективе, является информационная система «Библиотека А.П. Ершова» [13]. Она была создана для ЭВМ БЭСМ-6 еще в эпоху перфокарт. И ее создание, и дальнейшие переработки выполнялись силами энтузиастов. Эксплуатация, включая пополнение фондов и исправление ошибок и неточностей, выполнялась сотрудниками Отдела научно-технической информации (ОНТИ) ВЦ АН СССР/ ИСИ СО РАН. В 1990-х годах «Библиотека А.П. Ершова» была перенесе-

на на MS DOS с использованием СУБД FoxPro 2. С помощью этой системы велся каталог книг и журналов, учет читателей, формировались списки новых поступлений. В таком виде она прослужила ИСИ почти 30 лет.

Несмотря на актуальность и функциональность приложения, окружение, в котором оно было создано, устарело, и дальнейшая поддержка и развитие оказались невыполнимы. Кроме того, приложение не поддерживало структуризацию данных и словари. Все данные были строкового типа, заполнялись вручную библиотекарем, что привело к массе ошибочных и дублирующихся данных.

В 2018 был выполнен реинжиниринг приложения «Библиотека А.П. Ершова» http://lib.iis.nsk.su/ на основе свободно распространяемой веб-платформы Drupal. В ходе работ была выполнена полуавтоматическая коррекция ошибок в именах авторов и названиях фондов, осуществлены миграция данных с сохранением модели данных и имплементация удобных современных интерфейсов [12].

Первым интернет-ориентированным электронным архивом, созданным и поддерживаемым в ИСИ СО РАН, был Архив академика А.П. Ершова http://ershov.iis.nsk.su [13]. Он был создан по классической схеме веб-приложения, построенного на реляционной базе данных, имеющего публичный интерфейс (frontend) и интерфейс редактирования (backend). Был проделан значительный объем работ по сканированию и описанию документов, содержащихся в более чем 500 папках «бумажного» архива, хранимого в ИСИ. Технология структуризации внесения, редактирования информации, представления документов и дополнительной информации пользователям оказалась удачной и в дальнейшем. В архив были погружены дополнительные сегменты, уже не связанные напрямую с папками, сформированными А.П. Ершовым. Это архивы ВНТК «Старт», ИСИ СО РАН, включая конференции PSI, а также архив члена-корреспондента АН СССР Святослава Сергеевича Лаврова.

Проект был разработан в 2000 году при поддержке Microsoft Research с использованием технологий Майкрософт и столкнулся с одной из частых проблем для приложений с использованием проприетарного ПО. Спустя 15 лет после разработки проекта актуальность его сохранилась, но дальнейшая поддержка и разработка были затруднены в связи с истечением лицензий на проприе-

тарное ПО. Поэтому в 2016 году было принято решение о миграции приложения на свободно распространяемое ПО. Важным условием миграции было сохранение исходной модели данных архива, представляющей историческую ценность как одной из первых моделей данных для электронных архивов. Миграция на свободно распространяемую веб-платформу Drupal прошла успешно, помимо переноса данных был усовершенствован интерфейс пользователя и библиотекаря, была реализована поддержка массовой загрузки изображений.

Для проекта Фотоархива СО РАН (Фотолетопись) http://soran1957.ru была подготовлена технология, основанная на Semantic Web [14]. Также была сформирована онтология – система структуризации данных, в дальнейшем получившая оформление в виде онтологии неспецифических сущностей [14]. Новыми проблемами, преодолеваемыми в проекте, были не только проблемы нового подхода к структуризации, но и обработка и представление фото- и видеоматериалов, решение задач модульности информации и защиты данных от случайных и злонамеренных искажений. Как и многие другие архивы, фотоархив постоянно пополняется новыми материалами. Это делается с помощью группы backend-технологий и интерфейсов.

Поскольку количество заявок на хранение архивов в цифровом формате стало расти, был сформирован подход и создана технология мультиархивной системы Открытого архива Сибирского отделения СО АН СССР/РАН (http://odasib.ru). Основа технологических решений была прежней, изменились интерфейсы, изменилась в сторону детализации структура описания документов. Например, документы теперь рассматриваются как состоящие из частей (разделы, страницы, сканы страниц), авторство расширено до разделения автор-получатель. Были созданы удобные средства для информационных операторов, которые ведут обработку больших объемов сканированных страниц. Отдельные архивы были квалифицированы как «фонды», это соответствует архивной терминологии. В настоящее время в Открытый архив погружено 25 фондов, ведется работа над еще несколькими. На Рис. 4 изображена схема работы с документами разработанных электронных архивов.

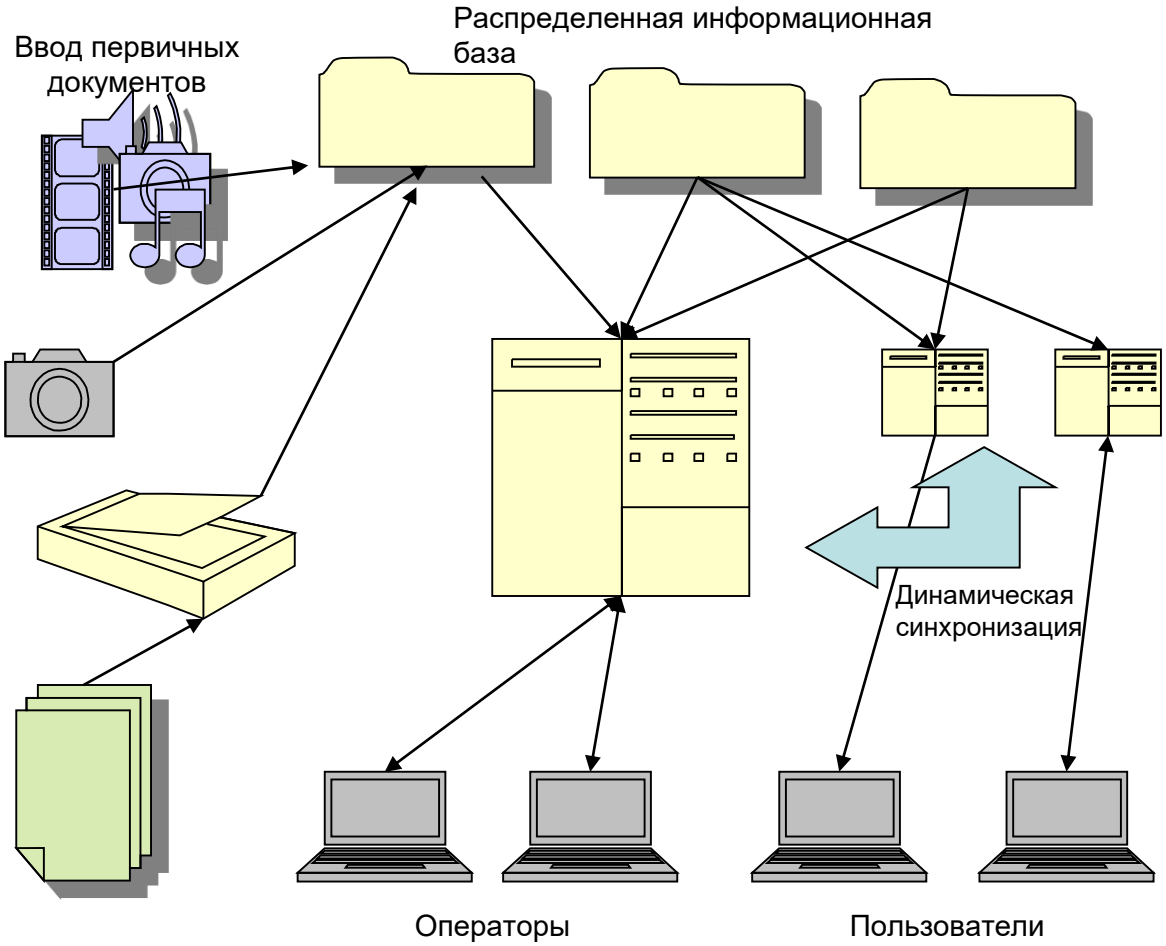


Рис. 4. Схема комплекса работы с электронными документными архивами

Еще одна проблема – физическая сохранность данных. Магнитные носители деградируют, могут быть повреждены в результате физического износа и внешних воздействий. Пока мы не сталкивались с серьезными вирусными заражениями, но это также должно быть учтено. Нельзя сказать, что сохранность данных сегодня обеспечивается нами на современном надежном уровне – принятые решения громоздки и требуют затрат. Основой обеспечения надежности хранения и функционирования электронных архивов является серверный пул машин и устройств, предоставляющий виртуальные машины для разработчиков и эксплуататоров, созданный, развивающийся и функционирующий в ИСИ СО РАН. Собственно, здесь уместно было бы использование специализированных решений или облачной инфраструктуры. Однако пока мы хотим полностью контролировать наши данные, соблюдая необходимую гибкость в решениях. В слу-

чае с модернизацией электронного архива А.П. Ершова описанное решение также не выглядит решающим все проблемы обеспечения долгого срока эксплуатации. Те эксперименты, которые будут описаны ниже, являются, в том числе, попыткой определения направления технологической эволюции архива.

Настоящим полигоном новых решений явился электронный архив Летних школ юных программистов (ЛШЮП), ежегодно проводимых ИСИ для талантливых школьников. При создании данного архива был пройден путь технологических решений от базы данных в XML и интерфейсов, сформированных средствами XSLT (в то время Semantic Web еще был неизвестен) до вполне современных веб-приложений с RDF и OWL, ASP.NET и т. д. В проект внесли свою лепту школьники, обучавшиеся в ЛШЮП. Архив постоянно пополняется и используется (http://mag.iis.nsk.su/syp).

За прошедшие годы проводились различные эксперименты с архивами. Например, к 50-летию Механико-математического факультета НГУ были произведены сбор мультимедийных материалов и данных, их погружение в Фотоархив СО РАН, создан отдельный юбилейный сайт.

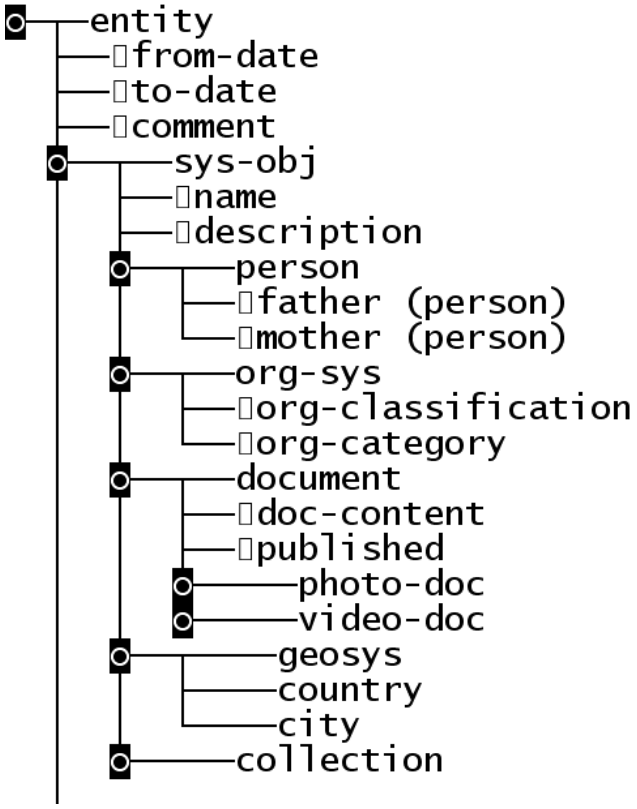
4. ИНТЕГРАЦИЯ: ЦЕЛИ И ЗАДАЧИ

Электронные архивы преследуют цели сбора и хранения данных и информации, а также предоставления удобного доступа к информации для специалистов и общественности. При этом сохранение данных в каком-то смысле является существенно более значимым, чем остальное. Основу архивов составляет тот или иной документный фонд. По информации, содержащейся в документах, архивы уникальны, но заметно пересекаются по таким параметрам, как авторы, адресаты, организации и события. Возникает вопрос о возможности полной, значительной или частичной интеграции разных архивов. Задача интеграции носит не только частный характер относительно упомянутых информационных собраний, но и представляет более общий методический интерес.

Чтобы понять возможный характер интеграции архивов, необходимо изучить интересы их пользователей. Поскольку доступ к архивной информации в электронных архивах осуществляется гораздо быстрее и удобнее, чем в классических «бумажных», круг пользователей существенно расширяется. Если ранее основными пользователями были ученые-историки, получающие в документных

архивах информацию для исследовательских целей, то теперь архивами легко могут пользоваться и те, кто задают вопросы «а кто это такой?», «а что это такое?», «а как это выглядит?». В этом числе могут быть пользователи, интересующиеся историей своей семьи, города, страны и т. д.

В наших архивах для структуризации документов и данных в основном используется вариант онтологии неспецифических сущностей [14]. На Рис. 5 изображена в виде дерева система классов и отношений (properties), составляющих основу этой онтологии.



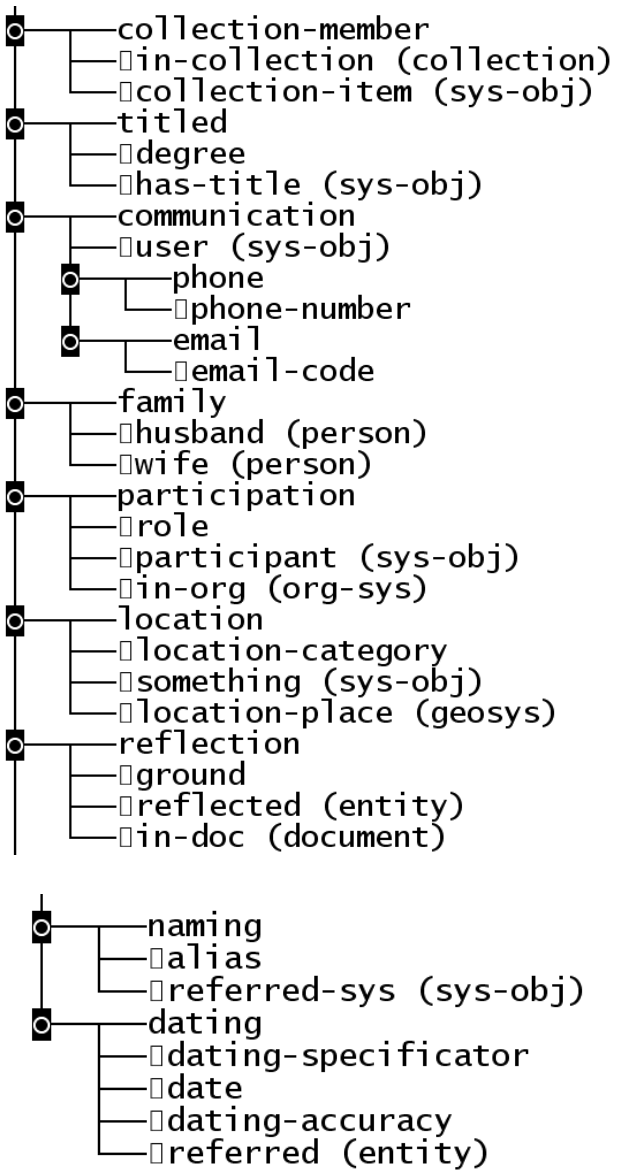


Рис. 5. Основные классы и отношения онтологии неспецифических сущностей

Отталкиваясь от документа той или иной коллекции, создатели архива фиксируют авторов документа, фигурирующие в документе персонажи, организации, географические пункты. Соответственно, через людей, организации и геоинформацию происходит дополнительная структуризация, которая может оказаться полезной для пользователя. Таким образом, интеграция может носить общий (максимальный объем запрашиваемой информации), частный (узкая проблема) и проблемно-ориентированный характер (специальный тематический запрос). Одновременно, как уже было показано, существует потребность в

разбиении единого архива на части, так сказать, «дезинтеграции» архива. Как ни странно, «ломать» в чем-то даже сложнее, чем «строить», т. е. объединять. Дело в том, что объединенный архив интегрирует мультимедиа, документы и базу данных. Несмотря на то, что предприняты значительные усилия для обеспечения модульности хранимых документов через технологию кассет и сегментирования базы данных через fog-сегменты, данные могут пересекаться и реально пересекаются. Технология разбиения так же, как и технология слияния, формируется и оформляется. Задача разбиения архива на части, как правило, включает еще и требование обеспечить дальнейшее развитие архива в части пополнения и изменения документного набора и базы данных. Копия части архива (подколлекции или коллекции, созданной по тематическому запросу) на конкретный момент представляется малополезной в случае «живого» архива, т. е. архива, который продолжает пополняться и редактироваться.

5. ИНТЕГРАЦИЯ КАК ЭКСПЕРИМЕНТ

Целью описанных экспериментов являлось выяснение возможности сопряжения и взаимного обогащения документных и информационных фондов, которые уже по раздельности находятся в эксплуатации. Также интерес представляет методика выполнения той или иной интеграции.

В первом развернутом эксперименте была предпринята попытка расширения информации, предоставляемой пользователю Электронного архива академика А.П. Ершова (ЭАЕ). В ходе эксперимента мы хотели дополнить информацию из ЭАЕ вставками из другой информационной системы. Для этого нужно было в информационную систему-донор добавить веб-сервис, позволяющий динамически отвечать на запросы от веб-приложения или веб-страницы и формировать ответ в виде HTML-разметки с запрашиваемой информацией справки. Аналогичный подход используется в некоторых других системах. В ответное сообщение помещаются фотография, информация об именах, датах жизни, основных степенях и наградах, профессиональном пути. В эксперименте информация берется из фотоархива СО РАН.

Что это дает? С точки зрения документного архива ЭАЕ является (почти) законченной композицией. Там есть информация о персонах, являющихся автора-

ми или адресатами документов. Есть организации и другие самостоятельные (не принадлежащие конкретному архиву) объекты. Информация в этих сущностях, зафиксированная на момент создания архива, может со временем изменяться. Кроме того, в ЭАЕ почти полностью отсутствует фотографический материал о персонажах. Мы предположили, что имеющаяся в ЭАЕ информация должна дополняться в результате интеграции с другими источниками. Пример такого эксперимента можно увидеть на Рис. 6.

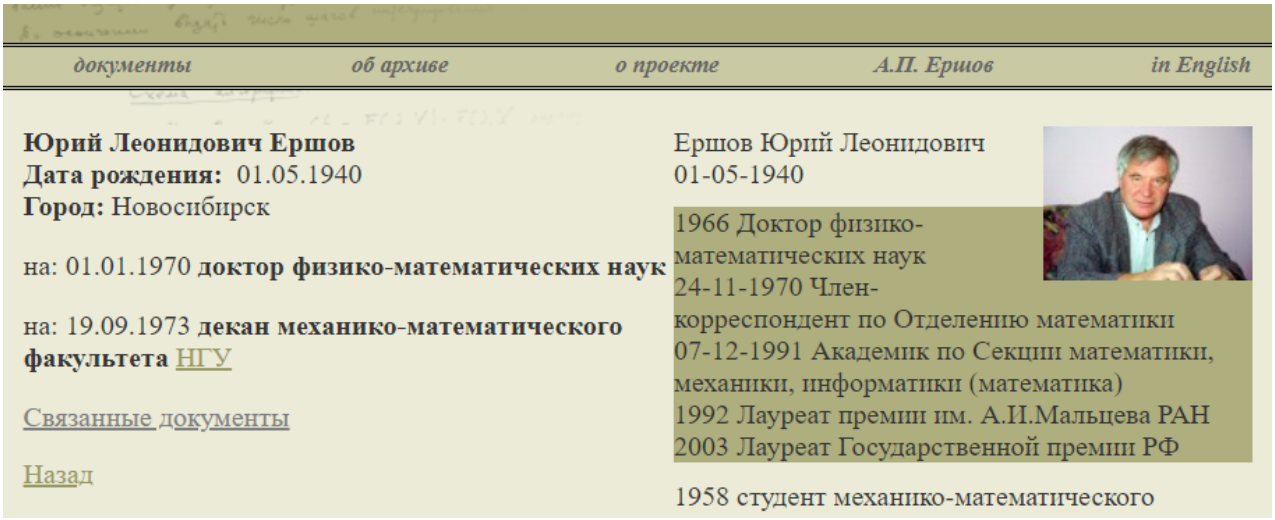


Рис. 6. Информация о персоне из фотоархива СО РАН в Электронном архиве академика А.П. Ершова.

Участником другого эксперимента также являлся ЭАЕ. В этом случае рассматривалась не сама информационная система, а ее база данных. Задачей эксперимента было погружение ЭАЕ в Открытый архив СО РАН (ОА) в качестве отдельного фонда. Если переработать электронный архив на использование принципов и технологий Semantic Web, использовать онтологию неспецифических сущностей, с использованием кассет, лог-файлов базы данных, то подобное погружение осуществляется довольно просто: необходимо было сделать отождествление сущностей и осуществить привязку корня композиции к коллекции «Фонды». На данном этапе эксперимента экстракция базы данных из ЭАЕ в Открытый архив осуществлялась вручную, но этот процесс возможно автоматизировать в дальнейшем.

Эксперимент показал, что подобное погружение стороннего фонда успеш-

но осуществляется без значительных трудозатрат. С одной стороны, ЭАЕ в качестве фонда дополняется возможностями использования навигации и поиска, заданных для Открытого архива, с другой стороны, в ряде случаев происходит реальное обогащение доступом к документам и данным для других, уже существующих фондов. Например, структурированный в рамках ОА фонд чл.-корр. Алексея Андреевича Ляпунова дополняется документами из архива А.П. Ершова, которых не было в Открытом архиве. На Рис. 7 показан фрагмент списка писем и документов А.А. Ляпунова из Открытого архива, к которому добавились документы из архива А.П. Ершова (выделенные позиции). Копии данных писем в архиве А.А. Ляпунова не сохранились, но они сохранились у адресата, т. е. у А.П. Ершова.

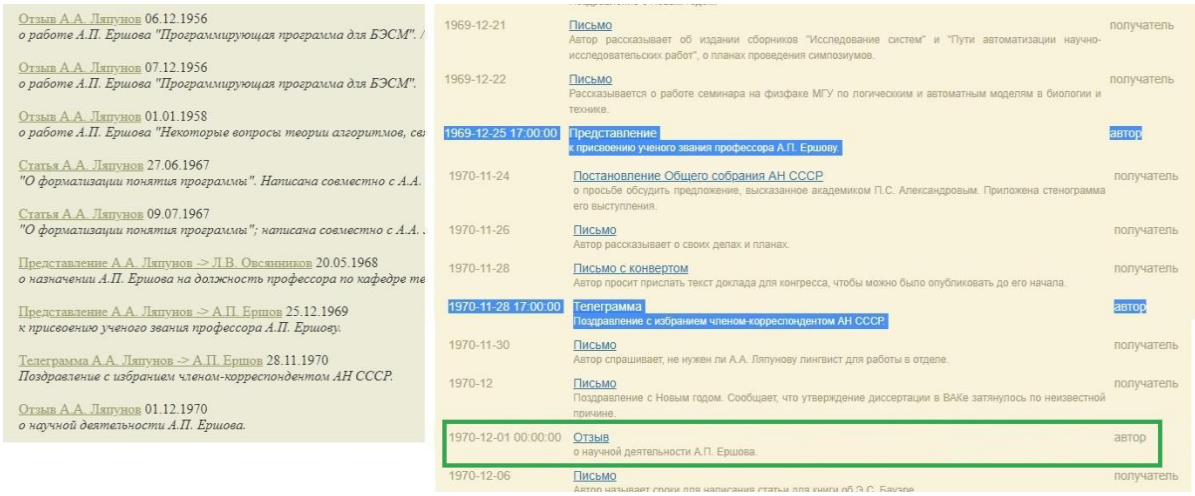


Рис. 7. Слева фрагмент базы данных архива А.П. Ершова с письмами от А.А. Ляпунова. Справа – те же письма, проявившиеся в документах А.А. Ляпунова в Открытом архиве СО РАН

Одной из наиболее существенных проблем практического применения технологии Semantic Web является тот факт, что она не позволяет создать общую систему идентификации объектов реального мира, значит, нет единой системы идентификации для записей базы данных. На практике это означает, что каждая база данных присваивает идентификационные коды самостоятельно, фактически замыкая идентификационное пространство своими рамками. Интеграция таких баз данных требует отождествления идентификаторов записей, соответствующих одним и тем же сущностям. Процесс отождествления можно экспери-

ментально осуществить на сопоставлении (полных) имен сущностей одинаковых или родственных классов. Такой подход обладает определенной эффективностью, особенно когда имена зафиксированы в полном «официальном» варианте. Например, при обработке списка выпускников ММФ НГУ за несколько лет (более 5000 персон) было выявлено несколько вариантов полного совпадения фамилии, имени и отчества у разных людей. Эффективность отождествления быстро теряется при использовании неполных вариантов имен, инициалов, разных вариантов перевода и т. д. В таких случаях надо применять более сложные методы анализа не только имени, но и контекста [15].

ЗАКЛЮЧЕНИЕ

В результате достаточно длительной эволюции сложился подход к обеспечению совершенствования, в первую очередь технологического, долго живущих информационных систем архивной направленности. На повестке дня – выработка принципов сосуществования разных информационных систем, вопросов их интеграции и дезинтеграции. Были проведены содержательные эксперименты по частичному или полному включению ресурсов одного электронного архива в состав другого. Показано, что такое включение может осуществляться без разрушения целостности систем, представляющих авторскую композицию.

СПИСОК ЛИТЕРАТУРЫ

1. Центральный фондовый каталог Архивного фонда РФ URL: http://cfc.rusarchives.ru

2. *Kornienko S., Gagarina D.* Information systems: new methods of Russian history sources study // International Multidisciplinary Scientific Conferences Social Sciences & Arts SGEM 2015. Conference Proceedings. Anthropology, Archaeology, History and Philosophy. Sofia, 2015.

3. Фотоархив СО РАН. URL: http://www.soran1957.ru/

4. Открытый архив СО РАН. URL: http://odasib.ru/

5. Древнерусский берестяные грамоты (Рукописные памятники Древней Руси). URL: http://gramoty.ru/birchbark/

6. Подвиг народа (Электронный банк документов участников Великой отечественной воны 1941–1945). URL: http://podvignaroda.ru/

7. Архив академика В.И. Вернадского.
URL: http://www.ras.ru/vivernadskyarchive/

8. Бессмертный барак (Имена и краткие сведения о более чем 1,9 млн. репрессированных граждан СССР). URL: https://bessmertnybarak.ru/

9. Исторические материалы. URL: http://istmat.info/

10. Прожито (корпус личных дневников и воспоминаний).
URL: https://prozhito.org/

11. Электронный архив Ольги Михайловны Фрейденберг (1890–1955).
URL: http://freidenberg.ru/

12. *Трошков С.Н.* Об опыте миграции приложений на свободно распространяемое программное обеспечение с открытым кодом // Вестник НГУ Серия: Информационные технологии. 2018. Том 16. Вып. 2. С. 86–94.
http://doi.org/10.25205/1818-7900-2018-16-2-86-94. ISSN 1818-7900.

13. *Антюфеев С.В., Марчук А.Г., Немов А.Н., Филиппов В.Э., Черемных Н.А.* Электронный архив академика А.П. Ершова // Электронные библиотеки. 2004. Том 7. № 5. С. 1–13.

14. *Berners-Lee Tim, Hendler James, Lassila Ora.* The Semantic Web// Scientific American. 2001. Vol. 284. No. 5. P. 34–43.

15. *Марчук А.Г., Марчук П.А.* Базовая онтология неспецифических сущностей BONE и её использование для построения информационных систем // Вестник СибГУТИ. 2014. № 4. С. 118–128.

LONG-TERM ELECTRONIC ARCHIVES: MODERNIZATION AND INTEGRATION

A. G. Marchuk[1] [0000-0001-8455-725X], S. N. Troshkov[2] [0000-0003-2952-9509],

I. A. Krayneva[3] [0000-0002-0601-9795]

[1-3]A.P. Ershov Institute of Informatics Systems SB RAS

[1]mag@iis.nsk, [2]kamronis@yandex.ru, [3]cora@iis.nsk

Abstract

Over a period of about twenty years, the IIS SB RAS has created information systems of historical orientation: the Electronic Archive of Academician A.P. Ershov, the Photo Archive of the Siberian Branch of the RAS, the Archive of the newspaper "Science in Siberia", the Open Archive of the SB RAS, etc. Each of the resources has its own specifics but in general their content is based on the general social and territorial basis of scientific and public activities of the SB AS USSR/RAS and the Novosibirsk Akademgorodok. In this report we will look at some of the problems of integrating/disintegrating disparate electronic resources into a common platform using existing and emerging tools.

Keywords*: integration of electronic resources, high-quality information, open archives, interdisciplinarity, history of science, Siberian Branch of the RAS, proprietary software, Semantic Web, Drupal.*

REFERENCES

1. Central Fund Catalog of the Archive Fund of the Russian Federation. URL: http://cfc.rusarchives.ru

2. *Kornienko S., Gagarina D.* Information systems: new methods of Russian history sources study // International Multidisciplinary Scientific Conferences Social Sciences & Arts SGEM 2015. Conference Proceedings. Anthropology, Archaeology, History and Philosophy. Sofia, 2015.

3. Photoarchive of the SB RAS. URL: http://www.soran1957.ru/

4. Open archive of the SB RAS. URL: http://odasib.ru/

5. Old Russian birch bark letters (Manuscripts of Ancient Russia). URL:

http://gramoty.ru/birchbark/

6. The feat of the people (Electronic bank of documents of participants in the Great Patriotic War 1941–1945). URL: http://podvignaroda.ru/

7. Archive of Academician V.I. Vernadsky.
URL: http://www.ras.ru/vivernadskyarchive/

8. Immortal barrack (Names and brief information about more than 1.9 million repressed citizens of the USSR). URL: https://bessmertnybarak.ru/

9. Historical materials. URL: http://istmat.info/

10. Lived (corpus of personal diaries and memoirs).
URL: https://prozhito.org/

11. Electronic archive of Olga Mikhailovna Freidenberg (1890–1955).
URL: http://freidenberg.ru/

12. *Troshkov S.N.* On Expirience in Migrating Applications to the Freely Distributable Open Source Software // Vestnik NSU. Series: Information Technologies. 2018. Vol. 16. No. 2. P. 86–94.

13. *Antufeev S.V., Marchuk A.G., Nemov A.N., Fillipov V.E., Cheremnikh N.A.* Academician A.P. Ershov Electronic Archive // Russian Digital Libraries. 2004. Vol. 7. No. 5. P. 1–13.

14. *Berners-Lee Tim, Hendler James, Lassila Ora*. The Semantic Web // Scientific American. 2001. Vol. 284. No. 5. P. 34–43.

15. *Marchuk A.G., Marchuk P.A.* BONE base ontology for unspecific entities and its use for construction of informatics systems // Vestnik SibSUTIS. 2014. No. 4 (28). P. 118–128.

СВЕДЕНИЯ ОБ АВТОРАХ

МАРЧУК Александр Гурьевич – д. ф.-м. н., зав. лабораторией.

Alexander Gurevich MARCHUK – Doctor of Physical and Mathematical Sciences, Head of the laboratory.

email: mag@iis.nsk.su

ORCID: 0000-0001-8455-725X

ТРОШКОВ Сергей Николаевич – программист 2 кат.

Sergey Nikolaevich TROSHKOV – programmer

email: kamronis@xtech.ru

ORCID: 0000-0003-2952-9509

КРАЙНЕВА Ирина Александровна – д. и. н., с. н. с.

Irina Alexandrovna KRAYNEVA – Doctor of Historical Sciences, Senior Research Officer.

email: cora@iis.nsk.su,

ORCID: 0000-0002-0601-9795

Материал поступил в редакцию 24 октября 2021 года

УДК 02-029, 621.394

ИЗДАНИЯ XIX-XX ВЕКА О ТЕЛЕГРАФЕ (ПО МАТЕРИАЛАМ ЭЛЕКТРОННЫХ БИБЛИОТЕК)

Ю. Е. Поляк[0000-0001-8411-335X]
Федеральное государственное бюджетное учреждение науки Центральный экономико-математический институт Российской академии наук, 117418 Москва, Нахимовский пр., д. 47

polak@cemi.rssi.ru

Аннотация

В позапрошлом столетии произошли революционные изменения в передаче информации. Для функционирования оптического телеграфа, появившегося в конце XVIII века, были необходимы громоздкие башни для прямой видимости сигналов семафора. Сто лет спустя протяжённость телеграфных линий составляла сотни тысяч километров; на рубеже веков начались первые опыты применения беспроводного телеграфа. Информация об этом отражена в многочисленных брошюрах, книгах, периодических изданиях того времени. Ещё через сто лет многие из этих материалов стали общедоступными благодаря развитию интернета и электронных библиотек; они интенсивно сканируются и выкладываются в Сеть. Взрывной рост количества электронных библиотек и их информационного наполнения сделал возможным появление данной работы. Её цель – проследить эволюцию технологий и процессов передачи информации, отражённую в литературе, с помощью самых разнообразных электронных библиотек – от грандиозных проектов Библиотеки Конгресса и Google Books с их миллионами оцифрованных книг до скромных частных собраний, посвящённых локальным темам. Использованы материалы более 20 электронных библиотек.

***Ключевые слова**: электронные библиотеки, история техники, оптический телеграф, электромагнитный телеграф, трансатлантический кабель, радио.*

ВВЕДЕНИЕ

Развивающиеся технологии сканирования многостраничных томов, рост числа и информационного наполнения электронных библиотек открывают доступ к огромному количеству изданий. Понятию «электронная библиотека» в текущем году исполнилось полвека. В 1971 г. Майкл Стерн Харт (Michael Stern Hart, 1947–2011) основал проект «Гутенберг»[1] и сделал электронные книги свободно доступными, в том числе через интернет. Харт мечтал о 10 тысячах произведений мировой литературы в электронном виде; этот рубеж был достигнут при его жизни, в начале века. Сейчас в «Гутенберге» более 60 тысяч документов. Конечно, это выглядит значительно скромнее начатого в 2004 г. проекта Google Books[2] с сервисом полнотекстового поиска по десяткам миллионов книгам, оцифрованным компанией Google, где этим занимается огромный коллектив. Но наряду с ним многие годы поддерживаются усилиями энтузиастов – «первопечатников» интернета другие электронные библиотеки. Среди них старейшая в рунете библиотека М.Е. Мошкова[3], основанная ещё в ноябре 1994 года, когда не было ни Апорта, ни Яндекса, ни Рамблера.

Именно электронные библиотеки сделали возможным появление данной работы. Они содержат колоссальное количество произведений и технической, и художественной литературы. Мы рассмотрим некоторые публикации конца XVIII – начала XX веков, имеющие отношение к передаче информации посредством телеграфа – оптического, а затем электромагнитного. Обращает внимание аналогия между телеграфом и интернетом. В книге [1] автор напоминает: «Во время королевы Виктории была разработана новая коммуникационная технология, которая позволяла людям почти мгновенно общаться на больших расстояниях, фактически сокращая мир быстрее и дальше, чем когда-либо прежде. Всемирная коммуникационная сеть, кабели которой охватывали континенты и океаны, произвела революцию в деловой практике, породила новые формы преступности и затопила своих пользователей потоком информации. Созданная глобальная сеть, по сути,

[1] URL: https://www.gutenberg.org
[2] URL: https://books.google.com
[3] URL: http://lib.ru

была викторианским интернетом. Правительства и регулирующие органы пытались и не смогли контролировать новую среду. Тем временем на проводах формировалась технологическая субкультура со своими обычаями и лексикой.

Сегодня интернет часто описывается как информационная супермагистраль; его предшественник в девятнадцатом веке, электрический телеграф, был назван «магистралью мысли». Оборудование было другим, но влияние телеграфа на жизнь его пользователей было поразительно схожим»[4].

Телеграф произвел величайшую коммуникационную революцию со времён печатного станка. Современные пользователи интернета во многих отношениях являются наследниками телеграфной традиции.

1. НАЧАЛО

К началу XIX-го века в большинстве стран средства оперативной связи не имели существенных качественных отличий от сигнальных костров и набатного звона. Редкое исключение – Франция, где в 1794 году начала действовать первая линия оптического телеграфа между Лиллем и Парижем протяжённостью 225 км (авторы изобретения – братья Клод и Игнатий Шапп, Claude et Ignatius Chappe). За ней последовали линии Париж–Тулон (1100 км), Париж–Страсбург (450 км) и другие. Когда в апреле 1809 г. австрийские войска осадили Мюнхен, Наполеон узнал об этом благодаря телеграфу и быстро очистил Баварию от неприятеля.

Второй страной в мире стала Швеция, которая ввела оптическую телеграфную сеть между Стокгольмом и Вакхольмом (1795), затем Фредриксборгом. Система извещала о движении кораблей, но была полезна и в военное время. Вскоре телеграф появился в Финляндии и Дании. В 1796 г. оптический телеграф был построен в Англии (Лондон–Портсмут); в 1798 г. – в Испании (Кадис–Мадрид). Изобретение сэра Р.Л. Эджворта было во многом вызвано желанием раньше всех получать информацию о результатах скачек. Аналогичные разработки появились и в других странах Британской империи: Канаде (1800), Ирландии (1804), Индии (1810), позднее на Мальте.

[4] URL: https://tomstandage.wordpress.com/books/the-victorian-internet

В России в этот период оперативную информацию доставляли конные курьеры. Правда, об оперативности можно говорить лишь условно. Так, 12 июня 1812 года войска Наполеона вошли в Ковно в 6 часов утра, однако Александр I, находившийся на балу в Вильно, узнал об этом лишь вечером (от Вильнюса до Каунаса 104 километра). Гораздо больше времени потребовалось в 1801 г. курьеру, чтобы после смерти Павла I вернуть казаков атамана Платова из индийского похода (они успели дойти до Саратовской губернии). О смерти Александра I, последовавшей в Таганроге 19 ноября (1 декабря) 1825 года, в Петербурге узнали 27 ноября во время молебна за здравие императора. А на сообщение с поселениями Дальнего востока и «Русской Америки» уходили месяцы. До середины XIX-го века единственным средством связи передачи сообщений между континентами была почта, доставляемая парусными судами, позднее пароходами.

Телеграф изменил эту ситуацию радикально и окончательно. После внедрения в начале 1850-х электрического телеграфа информация стала доходить до адресатов за секунды, а не за недели и месяцы.

В данной работе мы рассмотрим этапы эволюции телеграфной связи в ходе XIX-го столетия на основе материалов, ставших общедоступными благодаря разнообразным электронным библиотекам.

2. ОПТИЧЕСКИЙ ТЕЛЕГРАФ В РУССКИХ ПРОЕКТАХ

Телеграф братьев Шапп нашёл отражение в художественной литературе XIX-го века: «Мистер Карандаш» (1831) Родольфа Тёпфера, «Люсьен Лёвен» (1834) Стендаля, «Ромен Калбрис» (1869) Гектора Мало. Наибольшую популярность получил роман Александра Дюма «Граф Монте-Кристо» (1844). В главе 60 («Le télégraphe») автор подробно описывает функционирование телеграфной линии. Книга Дюма доступна, в частности, в двух замечательных библиотеках: в Проекте Гутенберг – в оригинале[5] и английском переводе[6]; русскоязычный вариант – в Библиотеке Максима Мошкова[7].

Значительно раньше художественных появились технические описания телеграфа. В Вене в 1795 г. была издана 23-страничная брошюра «Ächte und genaue

[5] URL: http://www.gutenberg.org/ebooks/17989
[6] URL: http://www.gutenberg.org/ebooks/1184
[7] URL: http://lib.ru/INOOLD/DUMA/montekristo2.txt

Darstellung der neuerfundenen franzôsischen Fernschreibmaschine, genannt: der Telegraph. Wodurch klar erwiesen wird, dass die in Leipzig herausgekommene, und zu Wien und andern Orten Nachgedruckte Beschreibung des Telegraphen durchaus falsch und ganz unrichtig sei»[8], и в том же году в Москве вышел русский перевод[9] [2]. Мы имеем возможность рассматривать эти издания во всех деталях благодаря проекту Google. Очевидно, в своё время читал их и механик Санкт-Петербургской Академии наук Иван Петрович Кулибин, однако несколько ранее он уже закончил труд по созданию своей оригинальной «дальноизвещающей машины», конструировать которую начал в 1793 году. Кодирование сигналов у Кулибина было предложено удачнее, чем у Шаппа: слова он разбивал на «одинакие и двойные склады», т. е. слоги. Скорость передачи по такому способу была значительно выше. Для оптического телеграфа Кулибин разработал и ряд других оригинальных решений. В 1801 году его модель демонстрировалась Павлу I. Однако, несмотря на отличное качество, правительство не поддержало проект Кулибина; он остался неосуществлённым и был передан в Кунсткамеру. Работы Кулибина подробно описаны в книге [3], с которой можно ознакомиться на сайте Президентской библиотеки имени Б.Н. Ельцина[10] или Открытой электронной архитектурно-строительной библиотеки Totalarch[11]. Популярный биографический очерк о Кулибине [4] имеется в электронной библиотеке ЛитМир[12].

Другой проект «ночного скорого дальнописца или телеграфа о семи фонарях» предложил титулярный советник Межевого департамента геодезист Николай Васильевич Понюхаев (1815). С помощью нескольких фонарей, расположенных в определенной комбинации, можно передавать буквы и цифры, приписав каждой из них сочетание горящих и затемнённых фонарей. К фонарям были приделаны подвижные щитки. Из пункта управления можно было щитком закрывать тот или другой фонарь и тем самым создавать разнообразные сочетания светящихся фонарей. Каждое сочетание соответствовало определённым букве или цифре. На приёмной станции сигналы наблюдали в подзорную трубу, записывали

[8] URL: https://books.google.ru/books?id=yjiTxkM4NhgC
[9] URL: https://play.google.com/books/reader?id=YnppAAAAcAAJ
[10] URL: https://www.prlib.ru/item/1288575
[11] URL: http://science.totalarch.com/book/3601.rar
[12] URL: https://www.litmir.me/br/?b=274482

и расшифровывали. Это оригинальное изобретение выгодно отличалось от предыдущих разработок простотой и возможностью размещения мобильных телеграфных станций. Понюхаев предлагал устраивать не только стационарные установки, но и походные. Изобретатель позаботился и о том, чтобы «ночной дальнописец» работал и при дневном свете. Он считал возможным делать его телеграф и «дневным, складным и возимым парой лошадей». По телеграфу Понюхаева можно было передавать телеграммы на расстояние до 45 километров. По своим характеристикам дальнописец Понюхаева значительно превосходил семафорный телеграф Шаппа. Изобретение рассмотрел Военно-учёный комитет, который, одобрив его идею, счёл механизм слишком сложным, а сам дальнописец мало полезным в дневное время. В итоге проект попал в архив канцелярии Военного министерства, и вспомнили о нём лишь при подготовке юбилейного сборника. Как известно, в 1802 г. Александр I заменил коллегии министерствами. К 100-летию этой реформы были изданы объёмные фолианты с перечислением достижений, и дальнописцу Понюхаева посвящены страницы 227–229 книги [5]. Она теперь доступна на портале «Россия в подлиннике», посвящённого российским истории и культуре – Runivers.ru[13]. Полное описание проекта находится в Российском государственном историческом архиве[14], а также в Российском государственном архиве военно-морского флота (РГАВМФ)[15].

Изобретением оптических телеграфов занимались и другие русские деятели. На том же сайте Runivers.ru[16] и в том же архиве РГАВМФ[17] можно найти информацию о телеграфе капитан-лейтенанта (впоследствии контр-адмирала) Павла Егоровича Чистякова, который признали полезным ввести в войсках. В этом устройстве применялись три шеста, на каждом из которых вверху было по два подвижных крыла. В ночное время на концах крыльев на специальных подвесах укреплялись фонари. Для передачи сообщений использовались два вида кодирования – цифровое и буквенное. Телеграф Чистякова применялся во время русско-

[13] URL: https://runivers.ru/bookreader/book458026/#page/393/mode/1up
[14] URL: https://rgia.su/object/6141917
[15] URL: https://rgavmf.ru/fond/166/fond-166-opis-1/fond-166-opis-1-edhr2517
[16] URL: https://runivers.ru/bookreader/book58653/#page/137/mode/1up
[17] URL: https://rgavmf.ru/fond/166/fond-166-opis-1/fond-166-opis-1-edhr2608

турецкой (1827–1828) и Крымской (1853–1856) войн. 16 сентября 1827 года изобретатель «имел счастье представлять государю императору изобретённый им подвижной телеграф для армии, за что получил бриллиантовый перстень»[18]. Изображение телеграфа Чистякова попало в юбилейный сборник к 100-летию министерства внутренних дел [6], а подробная биография автора помещена в 25-томный «Русский биографический словарь» (1896–1918), представленный, в частности, в библиотеке сайта «Православное духовенство»[19].

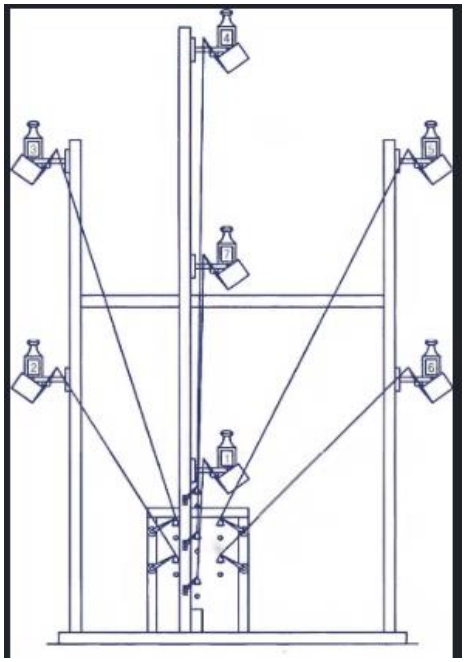


Рис. 1. Книга [5] – обложка и стр. 228 (дальнописец Понюхаева).

Ранее бриллиантового перстня и премии в 1000 рублей был удостоен ещё один капитан-лейтенант (дослужившийся до звания генерал-майора по адмиралтейству) Александр Николаевич Бутаков. В начале карьеры он стажировался в английском флоте (в частности, участвовал в Трафальгарском сражении), где и познакомился с возможностями семафорного телеграфа. В 1808 году, находясь в

[18] Русский биографический словарь: В 25 т. / Изд. под наблюдением пред. Имп. Рус. ист. о-ва А.А. Половцова. СПб.; Типографія И.Н. Скороходова. 1896–1918. Том 22. Стр. 413.
[19] URL: https://pravoslavnoe-duhovenstvo.ru/library/material/8545;
URL: https://runivers.ru/lib/book7666/436346

Лиссабоне, он разработал свод семафорных сигналов на русском языке и собственную систему семафорного (оптического) телеграфа, затем составил «полный словарь» семафорных сигналов, перевёл с английского и дополнил «Морской телеграф». Модель телеграфа Бутакова поместили в адмиралтейском «музеуме». Заметим, что в отличие от описанных выше телеграфных систем Бутаков использовал традиционный для флота семафор сигнальными флажками. В дальнейшем автор вводил дополнительные усовершенствования. По его словам, «предлагаемый телеграф может полезен быть также в армии, если сделать шест складным, наподобие палочек дамских зонтиков. Тогда можно удобно возить его везде и, где представится случай, тотчас поставить и действовать». Бутаков старался внедрить телеграф и в хозяйственную деятельность. В своей брошюре [7], изданной в 1833 г., он писал, что «телеграф мог бы быть полезен помещикам, если бы они решились ввести оный в употребление в своих поместьях, лежащих в виду одно от другого». Доступ к этой работе Бутакова предлагают Национальная электронная библиотека[20], Общество распространения полезных книг[21] и даже Тверская епархия[22].

Оптический телеграф имеет массу недостатков. Н.Я Эйдельман в книге «Твой XIX век» цитирует высказывание Ф.П. Фонтона, относящееся к 1829 году: «теперешние телеграфы при туманной неясной погоде или когда сон нападает на телеграфщиков, что так же часто, как туманы, делаются немыми»[23] (библиотека Vivos voco! А.М. Шкроба). Были необходимы новые идеи.

3. ЭЛЕКТРОМАГНИТНЫЙ ТЕЛЕГРАФ

В 1820 году датский физик Ханс Кристиан Эрстед (Hans Christian Ørsted) во время лекции в университете демонстрировал нагрев провода электричеством от гальванической батареи. Во время эксперимента на столе лежал морской компас, поверх крышки которого проходил провод. Когда учёный замкнул электрическую цепь, стрелка компаса отклонилась. Эрстед заметил, что отклонение изменяется

[20] URL: https://rusneb.ru/catalog/000199_000009_003558694
[21] URL: https://orpk.org/books/1569
[22] URL: http://old.st-tver.ru/biblioteka-2/b/3467-butakov-a-n/30027-butakov-a-n-telegrafnye-signaly-dlya-gospod-pomeshchikov-1833
[23] URL: http://vivovoco.astronet.ru/VV/PAPERS/NYE/XIX/PART04.HTM

в зависимости от мощности аппарата и расстояния от провода до стрелки. Он проверял эффект на проволоке из разных материалов, пытался экранировать стрелку деревом, стеклом, смолой, помещал в воду. В своём мемуаре 1820 года «Опыты, касающиеся действия электрического конфликта на магнитную стрелку» [8], разосланном коллегам и в журналы, Эрстед пишет: «Главный вывод из этих экспериментов состоит в том, что магнитная стрелка отклоняется от равновесия под действием гальванического аппарата и что этот эффект появляется, когда цепь замкнута, и не проявляется, когда цепь открыта». Под «конфликтом» здесь понимается электрический ток, но этот термин тогда ещё не существовал. Этот оригинальный латинский текст можно видеть на сайте библиотеки Смитсоновского института[24], крупнейшего музея и образовательно-исследовательского комплекса, а датские учёные Поль ла Кур (Poul la Cour) и Якоб Аппель (Jacob Appel) включили его в свой двухтомник «Historisk Fysik»[25] (Копенгаген, 1896–1897). Немецкое издание этой книги «Die Physik auf Grund ihrer geschichtlichen Entwickelung für weitere Kreise in Wort und Bild dargestellt» (Braunschweig 1905) переведено на русский язык и опубликовано в 1908 г. под названием «Историческая физика» одесским издательством Mathesis; книга представлена на сайте «Математические этюды»[26], а также в «Коллекции старинных математических книг»[27] (проект СО РАН).

Андре Мари Ампер (André-Marie Ampère) повторил эксперимент Эрстеда, после чего сформулировал правило для определения воздействия магнитного поля на магнитную стрелку, обнаружил взаимодействие между электрическими токами (закон Ампера) и предложил использовать электромагнитные процессы для передачи сигналов. Таким образом, появились теория электродинамики и электромагнитный телеграф в качестве её приложения.

Первый в мире практически значимый электромагнитный телеграф был изобретён русским учёным и дипломатом, членом-корреспондентом Петербургской академии наук, участником Отечественной войны бароном П.Л. Шиллингом

[24] URL: https://library.si.edu/digital-library/book/experimentacirc00orst
[25] URL: https://www.dba.dk/historisk-fysik-poul-la-cour/id-1055721593/billeder/3
[26] URL: https://www.mathesis.ru/book/lakur
[27] URL: http://books.mathtree.ru/book/lakur2

фон Канштаттом. Исключительно многогранный человек, он, в частности, интересовался экспериментами с электричеством. В Париже, взятом русской армией, Шиллинг общался с Ампером. А после открытия Эрстеда в 1828 г. он построил первый в мире электромагнитный аппарат. Но днём рождения телеграфа считается 21 октября 1832 года, когда состоялась публичная демонстрация работы устройства Шиллинга в его квартире по адресу: Марсово поле, 7. Оно имело стрелочную индикацию сигналов, передаваемых по проводам, которые оператор приёмного устройства расшифровывал согласно специальной кодовой таблице. Расстояние между приборами превышало 100 метров. Демонстрации Шиллинга вызвали большой интерес, их посетили А.С. Пушкин, академик Б.С. Якоби, А.Х. Бенкендорф, Николай I. Император написал текст телеграммы «Je suis charme d'avoir fait ma visite a M-r Schilling» (Я очень рад был посетить господина Шиллинга), которая была практически моментально принята без искажений.

Описывая свой телеграф, Шиллинг указывал «некоторые преимущества оного перед ныне употребляемыми: что быстрота его несравненно больше, что он действует в дождливые и туманные погоды ... что он не требует постройки особых высоких башен и содержится весьма малым числом людей и, наконец, что первоначальное заведение оного стоит меньше, чем в обыкновенных телеграфах»[28].

В 1837 году Шиллинг получил предписание построить линию электрического телеграфа между Санкт-Петербургом и Кронштадтом и приступил к проектированию, но эта работа была прервана смертью. В 1886 г. в России широко отмечалось столетие Шиллинга. Этому событию посвящено издание[29], ставшее общедоступным благодаря электронной библиотеке «Научное наследие России» (проект президиума РАН) [9].

После Шиллинга стали появляться похожие устройства – слегка модернизированные, реже оригинальные. Так, электромагнитный телеграф появился в 1833 году в Германии (Карл Фридрих Гаусс и Вильгельм Эдуард Вебер; Carl Friedrich

[28] Из письма Шиллинга морскому министру князю С.А. Меншикову.
URL: https://polit.ru/news/2016/04/16/shilling

[29] Изобретатель электромагнитного телеграфа барон П.Л. Шиллинг фон-Канштатт [К 100-летию со дня рождения]. СПб.: тип. М-ва внутр. дел, 1886. 40 с.: 1 портр.
URL: http://books.e-heritage.ru/book/10070311

Gauß, Wilhelm Eduard Weber); а также в 1837 году в Великобритании (Уильям Фотергилл Кук и Чарльз Уитстон; William Fothergill Cooke, Charles Wheatstone). В США художник Сэмюэл Финли Бриз Морзе (Samuel Finley Breese Morse) в 1840 году запатентовал электромагнитный телеграф. Огромная заслуга Морзе – изобретение телеграфного кода, где буквы алфавита представлены комбинацией коротких и длинных сигналов – точек и тире.

Все эти и многие другие имена предшественников и создателей телеграфной связи можно видеть в 600-страничной книге по истории телеграфии до 1837 года [10]. Её экземпляр из библиотеки Мичиганского университета отсканирован в рамках проекта Google и загружен[30] в гигантский интернет-архив (Internet Archive). Этот архив содержит более 500 миллиардов веб-страниц, а также десятки миллионов книг, аудиозаписей, видео; сотни тысяч компьютерных программ.

4. ЛИНИИ СВЯЗИ. ТРАНСАТЛАНТИЧЕСКИЙ КАБЕЛЬ

Для передачи информации необходимы не только телеграфная аппаратура, но, разумеется, и средства коммуникации, поэтому параллельно с развитием техники шли интенсивная прокладка новых линий связи, увеличение производительности существующих. Мы уже упоминали об оптических телеграфных линиях XVIII-го века. В 1839 году вступила в строй самая протяжённая в мире (1200 км) линия семафорного телеграфа Петербург–Варшава через Псков, Вильно, Гродно. Но время оптического телеграфа прошло. 7 августа 1837 года была запущена первая в мире электромагнитная телеграфная линия Петербург–Кронштадт (проект П.Л. Шиллинга осуществлён уже после смерти автора). В 1851 году, одновременно со строительством железной дороги между Москвой и Петербургом, был проложен телеграфный кабель с резиновой изоляцией; в 1871 году эта линия была продолжена до Владивостока. Первые подводные кабели в России были проложены в 1852 году через Северную Двину, а в 1853 году – между Ораниенбаумом и Кронштадтом. В 1879 году Красноводск и Баку соединились через Кас-

[30] URL: https://web.archive.org/web/20170715230109/http://www.princeton.edu/ssp/joseph-henry-project/telegraph/A_history_of_electric_telegraphy_to_the.pdf

пийское море. Ранее, в 1842 г., С. Морзе передал сообщение по подводному кабелю, экранированному резиной и свинцовой трубкой, в гавани Нью-Йорка. В 1844 году Морзе построил телеграфную линию Балтимор–Вашингтон протяжённостью 40 миль и 24 мая передал из вашингтонского Капитолия Альфреду Вейлу в Балтимор фразу из Книги Чисел (XXIII, 23) «What hath God wrought!»[31]; изображение этого события попало на фронтиспис книги о наиболее важных событиях XIX-го столетия[32], находящейся в Internet Archive. Многие называют эту дату началом телеграфного века. Телеграфные провода были подвешены на столбах, а в качестве изоляторов использовались горлышки бутылок.

Важным событием стала прокладка бронированного телеграфного кабеля длиной 4500 километров через Атлантический океан. С этой целью в 1856 году была основана Atlantic Telegraph Company. 26 июля 1858 года корабли «Агамемнон» и «Ниагара», каждый со своей частью троса на борту, встретились на полпути между Ирландией и Ньюфаундлендом, соединили трос и спустили его в воду. 5 августа, завершив закладку, корабли достигли места назначения. 16 августа королева Виктория и президент США Д. Бьюкенен обменялись приветственными телеграммами. Но в сентябре 1858 г. кабель был разрушен коррозией; долговременная связь между Европой и Америкой была обеспечена в 1866 году лишь с пятой попытки [11].

Трансатлантический кабель стал источником вдохновения для многих деятелей искусства. Из произведений художественной литературы назовём два. Кабелю посвящена новелла «Первое слово из-за океана (Сайрус Филд, 28 июля 1858 года.)» в сборнике С. Цвейга «Звездные часы человечества» (Stefan Zweig. Das erste Wort über den Ozean. Sternstunden der Menschheit, 1927). Русский перевод[33] помещён на сайте упоминавшейся интернет-библиотеки Vivos voco! – замечательного просветительского проекта А.М. Шкроба (1936–2007), химика и биолога,

[31] «Вот что творит бог!» (другой перевод: «Чудны дела твои, господи!»).
URL: https://bibleonline.ru/bible/rst66/num-23

[32] The Story of the Nineteenth Century of the Christian Era by Elbridge S. Brooks. Lothrop Publishing Company. Boston: 1900, 446 pp. The First Telegram.

[33] URL: http://vivovoco.astronet.ru/VV/BOOKS/VOICE/CYRUS-FIELD.HTM

постоянного участника конференций «Научный сервис в сети Интернет». В оригинале книгу можно прочесть на сайте о Германии[34], где в свободном доступе имеется большая коллекция немецкой литературы. Книга Артура Кларка «Голос через океан» (Arthur Charles Clarke. Voice across the sea, 1958) оцифрована в рамках проекта Google[35], на русском языке она представлена в библиотеке обучающей и информационной литературы Tinlib[36].

Много внимания всем этапам прокладки кабеля уделяла первая в мире иллюстрированная еженедельная газета The Illustrated London News. Она издавалась в Лондоне с 1842 по 2003 годы, публиковала массу интересных материалов с прекрасными гравюрами – от светской хроники и мировых новостей до частных объявлений и шахматных этюдов – и была популярнейшей газетой викторианской Англии с тиражом в 300 000 копий. Среди авторов газеты были Роберт Льюис Стивенсон, Томас Харди, Джеймс Барри, Уилки Коллинз, Джозеф Конрад, Артур Конан Дойль, Редьярд Киплинг, Гилберт Кит Честертон, Агата Кристи. Но в XX-м веке она стала терять своё значение: в 1971 г. стала ежемесячной, в 1994 г. перешла на два выпуска в год. Полные тексты многих выпусков предоставляют английские архивы, американские университеты и частные собрания. Гравюры из 47-го тома (1865) дают представление о ходе работ по прокладке трансатлантического телеграфа[37].

[34] URL: https://deutschland1.ru/books/209-sternstunden-der-menschheit.html
[35] URL: https://books.google.ru/books?id=L2UNAQAAIAAJ
[36] URL: http://www.tinlib.ru/tehnicheskie_nauki/goloc_cherez_okean/index.php
[37] Illustrated London News, No 1331 – Vol. XLVII. August 26, 1865; No. 1332 – Vol. XLVIII, 2nd September 1865. URL: https://archive.org/details/illustratedlondov47lond/

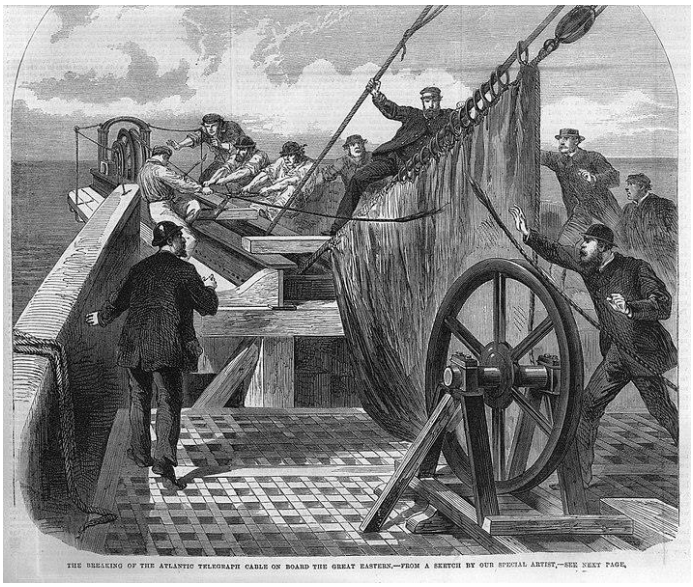


Рис. 2. Разрыв трансатлантического кабеля во время его укладки в 1865 г.
Источник: The Illustrated London News. No. 1331 – August 26, 1865, p. 181.

Большое количество набросков, акварелей и картин выполнил в ходе «кабельных» экспедиций известный художник Роберт Чарльз Дадли в 1865–1870 гг. Его работы можно видеть, в частности, в Метрополитен-музее[38], проекте Google «Искусство и культура»[39], а также на «Викискладе»[40] (Wikimedia Commons) – это репозиторий, содержащий около 70 миллионов бесплатных медиафайлов.

В огромнейшей Библиотеке Конгресса США среди тысяч материалов, посвящённых истории телеграфии, есть любопытный документ, датированный 1 января 1861 г. Это мемуар Генри О'Райли «Russo-American Telegraph»[41], адресованный Российской императорской академии наук. Автор, ссылаясь на свои статьи в American Telegraph Magazine и опыт проектирования телеграфных линий, предлагает построить «русско-американский сухопутный телеграф», продлив телеграфные линии от Миссисипи до Тихого океана для соединения с русскими владениями на этом побережье и далее через Берингов пролив. Именно Российская им-

[38] URL: https://www.metmuseum.org/art/collection/search#!/search?artist=Dudley,%20Robert%20Charles
[39] URL: https://artsandculture.google.com/entity/роберт-чарльз-дадли/g11f6ybg820
[40] URL: https://commons.wikimedia.org/wiki/Category:Robert_Charles_Dudley
[41] URL: https://www.loc.gov/resource/rbpe.2330370a

перия, расположенная в Европе, Азии и Америке, по мнению О'Райли, имела возможность для связного телеграфного сообщения между Старым и Новым светом. В 1865 г. для прокладки линии электрического телеграфа из Калифорнии и Аляски по дну Берингова моря через Сибирь в Москву была создана компания «Русско-американский телеграф», принадлежащая Western Union. Однако в июле 1867 года проект был официально закрыт, а спустя три месяца была оформлена продажа Аляски. Хотя проект не состоялся, он дал толчок развитию многих территорий.

В 1861 году компания Western Union за 112 дней соединила западное и восточное побережья США трансконтинентальной линией. В 1902 году телеграфная линия была проложена через Тихий океан из Канады в Австралию. К началу XX-го века протяжённость телеграфных линий в США составляла 390 990 километров; у идущей на втором месте России – 180 640 км. Эти сведения, наряду со многими другими, содержатся в статистическом атласе 1908 года[42] из электронной факсимильной библиотеки портала «Руниверс», посвящённого российской истории и культуре. Сборник к 100-летию МВД [6] приводит такие данные: телеграфная сеть империи к 1901 году имела линий 154354 версты; проводов – 465011 вёрст, учреждений с приёмом телеграмм 5789, исходящих телеграмм 16.5 миллиона (с. 246–247).

5. РАСЦВЕТ И ЗАКАТ ТЕЛЕГРАФА

Конец XIX-го века ознаменовался давно ожидавшимися открытиями в области беспроводной связи. 25 апреля (7 мая) 1895 г. А.С. Попов прочитал обстоятельный доклад на заседании физического отделения Русского физико-химического общества. Во время доклада профессор Попов продемонстрировал передачу знаков Морзе без помощи проводов. В качестве передатчика была применена катушка Румкорфа с присоединенным к ней вибратором Герца, а в качестве приёмника – разработанная А.С. Поповым оригинальная схема. Протокол заседания РФХО был опубликован в журнале РФХО в августе 1895 г.[43], схема и подробное

[42] Всеобщій географическій и статистическій карманный атласъ – Издание А.Ф. Маркса. СПб. 1908. 130 с. URL: https://runivers.ru/lib/book7609/399065

[43] Попов А.С. Об отношении металлических порошков к электрическим колебаниям // Журнал Русского физико-химического общества. Т. 27, часть физическая. 1895. С. 259–260.

описание прибора появились в журнале РФХО в январе 1896 г.[44] Отметим: «конкурент» Попова Гульельмо Маркони подал заявку на получение патента Великобритании 2 июня 1896 года с формулировкой «Усовершенствования в передаче электрических импульсов и сигналов и в аппаратуре для этого». Дата 7 мая отмечается в России как День радио.

В других странах примерно в то же время многие учёные также работали над созданием подобных устройств. В США Никола Тесла в 1893 г. первым запатентовал радиопередатчик, а спустя два года – радиоприёмник. В Германии изобретателем радио считают Генриха Герца, в Англии – Оливера Джозефа Лоджа, во Франции – Эдуарда Бранли, в Бразилии – Ланделя де Муру, а в Индии – Джагадиша Чандру Боше[45].

В XX-м веке телеграф интенсивно развивался. Но в конце столетия распространение факсимильной связи, появление электронной почты и интернета привели к тому, что во многих странах от телеграфа отказались как от неактуальной, морально устаревшей и нерентабельной технологии. Первую телеграмму в Индии отправили ещё в 1850 году, страна была крупнейшим пользователем телеграфа в мире; однако 15 июля 2013 года Индийский национальный телеграф прекратил существование. Телеграфная связь в Бельгии проработала 171 год, с 1846 г. по 29 декабря 2017 года. Ряд стран отказался от телеграфа ещё раньше: Великобритания в 1982 году, Новая Зеландия – в 1999-м, Швеция – в 2002-м, Нидерланды – в 2004-м, Литва и Словакия (2007), Австралия (2011), Бельгия (2017), Казахстан и Украина (2018) [12].

6. ТЕЛЕГРАФ В РУССКОЙ ЛИТЕРАТУРЕ

Выше было показано, что попытки разработки и применения телеграфной связи в России относятся к первой половине XIX-го века. Параллельно происходило осмысление новых понятий в отечественной культуре. Ещё до появления в стране телеграфа Н.А. Полевой начал издавать первый русский общественно-

URL: http://books.e-heritage.ru/book/10074694

[44] Попов А.С. Прибор для обнаружения и регистрирования электрических колебаний // Журнал Русского физико-химического общества. 1896. Т. 28, часть физическая, отдел I, вып. 1, стр. 1–14. URL: http://library.ruslan.cc/authors/попов-александр-степанович

[45] URL: https://news.rambler.ru/other/39789966

научно-литературный журнал энциклопедического типа «Московскій Телеграфъ», рассчитанный как на образованного, так и на широкого читателя (не следует путать с одноимённой газетой, выходившей в 1881–1883 гг.). В то время само слово «телеграф» было новым, и в русском языке закрепилось во многом благодаря журналу. Выпуски журнала, выходившего с 1825 по 1834 годы, представлены в Internet Archive[46]; на обложке первого номера можно видеть башню семафорного оптического телеграфа, возвышавшуюся на скале вблизи озера. Название «Московский телеграф» подчеркивало нацеленность на злободневность, современность, скорость передачи различных сведений, новых знаний и их практическое применение. На первых порах центральное место в журнале занимал отдел «Науки и искусства»; позднее увеличилась роль литературно-критического направления. Косвенным свидетельством популярности журнала является фраза из монолога гоголевского Хлестакова: «Всё это, что было под именем барона Брамбеуса, Фрегат Надежды и Московский Телеграф ... всё это я написал»[47] (Русская виртуальная библиотека Е. Горного). Однако из-за рецензии, не понравившейся царю, журнал был закрыт.

Телеграф нашёл отражение в творчестве одного из самых оригинальных мыслителей позапрошлого века, князя В.Ф. Одоевского. Возможно, в этом сыграло роль его знакомство с П.Л. Шиллингом. В 1833 г. Одоевский публикует «Опыт о Музыкальном языке, или Телеграфе, могущем посредством музыкальных звуков выражать все то, что выражается словами, и служить пособием для различных сигналов, употребляемых на море и на сухом пути» (с приложением гравированной таблицы «Алфавиты музыкального телеграфа»). Это заметка о способе передачи информации на расстоянии с помощью громких музыкальных инструментов, например, трубы. Информация зашифровывалась с помощью нот. «Опыт …» представлен факсимильным изображением в Президентской библиотеке[48] и

[46] URL: https://archive.org/details/mosk_telegraf/1825_Ch_1__1-4/mode/2up
[47] URL: https://rvb.ru/gogol/01text/vol_04/01_revisor/0085-03.htm
[48] URL: https://www.prlib.ru/item/708756

НЭБ[49], а в современной орфографии – на частном сайте «Противодействие энтропии»[50]. В 1844 году выходит фундаментальное исследование Одоевского «Гальванизм в техническом применении, или искусство гальваническим путем производить типы, покрывать медью жизненные припасы и разные вещи для сохранения их; также делать медные доски для гравирования; изготовлять гравюры ... с объяснением необходимых предварительных понятий о химии и физике» (см. факсимиле в НЭБ[51]). Главы 74–76 этой книги (стр. 201–217) посвящены электрическому телеграфу Кука и Уитстона. К сожалению, осталась неоконченной фантастическая утопия «4338 год. Петербургские письма», где Одоевский, описывая удалённое непосредственное общение, фактически предвосхищает появление интернета и блогов. В электронной библиотеке iknigi.net читаем письмо 4-е: «мы получили домашнюю газету от первого здешнего министра, где, между прочим, и мы приглашены были к нему на вечер. Надобно тебе знать, что во многих домах, особенно между теми, которые имеют большие знакомства, издаются подобные газеты; ими заменяется обыкновенная переписка. Обязанность издавать такой журнал раз в неделю или ежедневно возлагается в каждом доме на столового дворецкого. Это делается очень просто: каждый раз, получив приказание от хозяев, он записывает всё ему сказанное, потом в камер-обскуру снимает нужное число экземпляров и рассылает их по знакомым. Сверх того, для сношений в непредвиденном случае между знакомыми домами устроены магнетические телеграфы, посредством которых живущие на далёком расстоянии разговаривают друг с другом»[52]. Книгу можно найти и в Библиотеке Мошкова[53].

По мере того как телеграф входил в повседневную жизнь, росло количество его упоминаний в художественных произведениях. Как правило, речь шла об отправке (получении) телеграмм либо о столбах и проводах как элементе пейзажа.

[49] URL: https://rusneb.ru/catalog/000199_000009_003558778
[50] URL: http://www.etheroneph.com/audiosophia/38-vfodoevskij-opyt-o-muzykalnom-yazyke.html
[51] URL: https://rusneb.ru/catalog/000200_000018_RU_NLR_DIGIT_113062
[52] URL: https://iknigi.net/avtor-vladimir-odoevskiy/45893-4338-y-god-peterburgskie-pisma-vladimir-odoevskiy/read/page-2.html
[53] URL: http://az.lib.ru/o/odoewskij_w_f/text_0490.shtml

Как указывает ресурс «Карта слов и выражений русского языка»[54], слово «телеграф» в своих произведениях использовали такие авторы, как Аксаков, Л. Андреев, Гиляровский, Гончаров, Горький, Добролюбов, Достоевский, Короленко, Лажечников, Лесков, Мамин-Сибиряк, Некрасов, Островский, Писемский, Салтыков-Щедрин, Л. Толстой, Тургенев, Фет, Чехов, Арсеньев, Булгаков, Грин, Замятин. В базе текстов «Русская классическая литература»[55] (частная электронная библиотека) поиск по слову «телеграф» даёт 850 результатов. В рамках статьи нет возможности все их процитировать, приведём несколько примеров.

Телеграфу в творчестве Ф.И. Тютчева посвящена обстоятельная статья Р.Г. Лейбова [13]. Автор указывает, что 13 августа 1855 года Тютчев впервые обратился к теме телеграфа.

Вот от моря и до моря / Нить железная скользит,
Много славы, много горя / Эта нить порой гласит.
И, за ней следя глазами, / Путник видит, как порой
Птицы вещие садятся / Вдоль по нити вестовой.
Вот с поляны ворон черный / Прилетел и сел на ней,
Сел и каркнул, и крылами / Замахал он веселей.
И кричит он, и ликует, / И кружится все над ней:
Уж не кровь ли ворон чует / Севастопольских вестей?[56]

Стихотворение связано с тревожным ожиданием вести о сдаче Севастополя. Телеграф выступает в стихотворении в виде магического и зловещего средства связи, смысл сообщений которого скрыт от путника, но явлен вещей птице. Интересно, что ворон садится на провода и в стихотворении Некрасова «Как празднуют трусу» (1870):

Я обругал его грубо невежею. / На телеграфную нить
Он пересел. «Не донос ли депешею / Хочет в столицу пустить?»
Глупая мысль, но я, долго не думая, / Метко прицелился. Выстрел гремит:
Падает замертво птица угрюмая, / Нить телеграфа дрожит ...[57]

[54] URL: https://kartaslov.ru/цитаты-из-русской-классики/со-словом/телеграф
[55] URL: http://www.lit-info.ru
[56] URL: https://www.bibliofond.ru/view.aspx?id=82522
[57] URL: http://nekrasov-lit.ru/nekrasov/stihi/375.htm

В той же статье Лейбова приводится отрывок из поэмы Я.П. Полонского «В конце сороковых годов»:

> Какие-то огни ... они играли, / Качались, подымались и опять
> Кувыркались. То телеграфы были, / И ум его впотьмах они дразнили:
> Условные огни во все концы / Переносили вести, все дворцы
> Их ожидали с жадным нетерпеньем[58]

Огни оптического телеграфа у Полонского превращаются в беседу «демонов глухонемых».

Из прозаических произведений отметим рассказ А.И. Куприна «Телеграфист» (1911), электронная библиотека ЛитМир[59]. Вспомним также персонажа другой повести Куприна «Гранатовый браслет» Желткова, «влюблённого телеграфиста» (сервер Классика.ру)[60].

В эмоциональный контекст помещают слово «телеграф» поэты русского авангарда. В стихотворении А. Крученых «Лунатизм вокзала» (1920) «стершиеся надписи / в остывающем пару / перепрыгивают на фаянсовые гнезда / телеграфных столбов»[61]. В «Поэме событий» К. Большакова 2-я глава называется «Город в телеграммах»; в ней «город сутулится, закиданный выкриками телеграммных вестей». В сборнике Б. Земенкова «Стеарин с проседью: Военные стихи экспрессиониста» «телеграфный с неба точно столб опущен, как бы лот». В «Заклятье вечера» Д. Петровского «висит на телеграфе лапоть»[62]. Эти авторы цитируются по книгам из серии «Библиотека поэта» петербургского издательства «Академический проект», которые доступны в «Электронной библиотеке русской и советской классики». Библиотека содержит более 50 тысяч произведений 200 авторов, представленных как в виде текста, так и сканированных изображений оригинальных изданий. Электронная библиотека русской литературы RuLit предлагает более

[58] URL: https://www.bibliofond.ru/view.aspx?id=82522
[59] URL: https://www.litmir.me/br/?b=48587
[60] https://klassika.ru/read.html?proza/kuprin/garnet.txt
[61] URL: https://ruslit.traumlibrary.net/book/kruchenih-poems/kruchenih-poems.html
[62] URL: https://ruslit.traumlibrary.net/book/futuristy-poetry/futuristy-poetry.html

600 тысяч книг, в том числе составленный в 2017 г. сборник произведений И. Соколова столетней давности «Бунт экспрессиониста. Стихи и манифесты»[63]. Вот несколько цитат: «дни бегут, как телеграфные столбы»; «радиотелеграф своими длинными пальцами прощупывает весь мир»; «телеграфные провода земного шара – нервная система апокалиптического зверя». Однако гораздо более известны фразы экс-футуриста В. Маяковского (приводятся по Библиотеке Мошкова»): «телеграммой лети, строфа», «это время гудит телеграфной струной»[64], «телеграф охрип от траурного гуда»[65].

Приведём ещё две цитаты известных в своё время публицистов. «Арена исторических действий становится необозримо великой, а земной шар — обидно малым. Чугунные полосы рельс и проволока телеграфа одели весь земной шар в искусственную сеть, точно школьный глобус» (Л.Д. Троцкий, «Наше Отечество во времени. Культура старого мира». В кн.: Л. Троцкий. Сочинения. Том 20. Москва–Ленинград, 1926, сайт The Marxists Internet Archive)[66]. А В.И. Ленин в «Советах постороннего» (1917) требует, «чтобы непременно были заняты и ценой каких угодно потерь были удержаны: а) телефон, б) телеграф, в) железнодорожные станции, г) мосты в первую голову»[67].

Итак, к 1920-м годам телеграф стал привычным атрибутом повседневного быта. «На волоколамском базаре побили нескольких милиционеров, отнимавших кур у баб, да выбили стёкла в местном почтово-телеграфном отделении. По счастью, расторопные волоколамские власти приняли меры, в результате которых, во-первых, пророк прекратил свою деятельность, во-вторых, стёкла на телеграфе вставили»[68] (М. Булгаков, «Роковые яйца», 1924). Остап Бендер возмущается: «Проклятый телеграф всюду понапихал свои столбы с проволоками»[69] (И. Ильф, Е. Петров, «Золотой теленок», 1931). Даже в детской литературе телеграмма – вполне обыденный предмет.

[63] URL: https://www.rulit.me/data/programs/resources/pdf/Ippolit_Vasilevich_Cokolov_Bunt_ekspressionista_ RuLit_Me_487947.pdf
[64] URL: http://az.lib.ru/m/majakowskij_w_w/text_0600.shtml
[65] URL: http://az.lib.ru/m/majakowskij_w_w/text_0480.shtml
[66] URL: https://www.marxists.org/russkij/trotsky/1926/trotl482.htm
[67] URL: https://www.marxists.org/russkij/lenin/works/lenin006.htm
[68] URL: http://lib.ru/BULGAKOW/eggs.txt
[69] URL: http://az.lib.ru/i/ilfpetrov/text_0130.shtml

Мы так давно с тобой в пути. / Скучают наши мамы.
На телеграф бы нам пойти, / Послать им телеграммы!..
За телеграмму с нас берут / Положенную плату.
А через несколько минут, Стуча ключом, передают / Слова по аппарату[70].
(С. Маршак, «Весёлое путешествие от А до Я»).

Вдруг откуда-то шакал / На кобыле прискакал:
«Вот вам телеграмма / От Гиппопотама!»
«Приезжайте, доктор, / В Африку скорей
И спасите, доктор, / Наших малышей!»[71]
(К. Чуковский, «Айболит», 1929, сайт «Русская поэзия»).

В рассказе А. Гайдара «Чук и Гек» нераспечатанная телеграмма серьёзно повлияла на развитие сюжета[72] (портал Litrus.Net).

Технический прогресс продолжался. В 1929 г. начинает работу первая в Советском Союзе факсимильная связь между Москвой и Ленинградом. А к началу 1940 г. между Москвой и крупными городами СССР было задействовано уже более 20 фототелеграфных связей. В то «доинтернетовское» время это была предтеча современного факса. При этом технология была настолько интересной, что даже попала в роман М.А. Булгакова «Мастер и Маргарита». Читаем в первой публикации (журнал «Москва» за 1966, №11, с. 68): «В дверях появилась всё та же женщина, и оба — и Римский, и Варенуха — поднялись ей навстречу, а она вынула из сумки уже не белый, а какой-то тёмный листок ... На темном фоне фотографической бумаги отчётливо выделялись черные писаные строки ... Затем он достал из письменного стола кипу бумаг и начал тщательно сличать жирные, с наклоном влево буквы в фотограмме с буквами в Стёпиных резолюциях и в его же подписях, снабжённых винтовой закорючкой»[73]. Некоммерческая электронная библиотека ImWerden (что означает «в развитии», «в становлении») содержит, в частности, собрание авторских чтений своих произведений в аудио- и видеоформатах.

Закончим этот раздел строчками из альбома Б.Б. Гребенщикова «Письма капитана Воронина»: «Наверно, только птицы в небе и рыбы в море знают, кто прав; но мы знаем, что о главном не пишут в газетах, и о главном молчит телеграф»[74].

[70] URL: https://strana-skazki.ru/stihi-marshaka.html
[71] URL: https://rupoem.ru/chukovskij/dobryj-doktor-ajbolit.aspx
[72] URL: https://litrus.net/book/read/12374
[73] URL: https://imwerden.de/pdf/bulgakov_master_i_margarita_moskva_1966_11__ocr.pdf
[74] URL: http://old.aquarium.ru/discography/pisma_kapi230.html#@647

ЗАКЛЮЧЕНИЕ

В России в 1982 г. было отправлено 540 млн. телеграмм[75]. После этого пика начался спад. Телеграфная связь продолжает существовать, переместившись в электронную среду; сообщения передаются и принимаются с помощью телеграфных модемов, подключенных к персональным компьютерам. Физические лица используют телеграммы для поздравлений, денежных переводов, соболезнований; среди потребителей этой услуги они составляют 8%. Остальные 92% приходятся на юридические лица; в месяц они отправляют свыше 100 тысяч сообщений[76]. Телеграммы принимают суды, нотариусы, госорганы. Заверенная телеграмма является юридическим документом. В нашей стране с её особенностями территории телеграфная связь важна для государственного управления. В ситуациях, когда факт отправки и получения сообщений должен быть юридически зафиксирован, телеграф может составить конкуренцию более удобным средствам связи. Рано или поздно телеграф заменят новые информационные технологии, но в нише передачи документированной информации он может существовать ещё долгое время.

СПИСОК ЛИТЕРАТУРЫ

1. *Standage T.* Victorian Internet: The Remarkable Story of the Telegraph and the Nineteenth Century's On-line Pioneers. N.Y., Walker & Company, 1998. 227 p.

2. Точное и подробное описаніе телеграфа или новоизобрѣтенной дальноизвѣщающей машины, помощію которой въ самое кратчайшее время можно доставлять и получать извѣстия изъ самыхъ отдаленнѣйшихъ мѣстъ. Москва, типографія И. Зеленникова. 1795.

3. *Данилевский В.В.* Русская техника. Ленинград: Лениздат, 1947. С. 516.

4. *Артоболевский И.И.* Русский изобретатель и конструктор Кулибин. Научно-популярная библиотека солдата и матроса. М.: Воениздат, 1948. 36 с.

5. Столѣтіе военнаго министерства. 1802–1902. Главное инженерное управленіе. Историческій очеркъ. СПб: Типографія «Слово», 1902. 675 с. См. также:

[75] URL: Островский А.В. История мировой и отечественной связи: учебное пособие. СПб.: СПбГУТ, 2011. 312 с.

[76] URL: https://rg.ru/2016/04/26/pochemu-rossiiane-do-sih-por-polzuiutsia-telegrafom.html

Российский государственный архив военно-морского флота (РГАВМФ), фонд 166, опись 1, ед. хр. 2517. О изобретенном титулярным советником Понюхаевым телеграфе. 1815. 18 страниц; Российский государственный исторический архив, ф. 398 оп. 81 д. 410. Описи дел, поступивших в архив Третьего департамента Министерства государственных имуществ из Хозяйственного департамента Министерства внутренних дел за 1763–1836 гг. По проектам и сочинениям (1797–1836 гг.). Департамент земледелия Министерства земледелия. О изобретенном титулярным советником Понюхаевым ночном телеграфе. 12 февраля 1815 – 30 апреля 1825. 18 листов.

6. Министерство внутреннихъ дѣлъ. 1802–1902. Историческій очеркъ. Приложеніе второе. Почта и телеграфъ въ XIX столѣтіи. СПб: Тип. Министерства внутреннихъ дѣлъ. 1901. 335 с. См. также: Российский государственный архив военно-морского флота (РГАВМФ), фонд 166, опись 1, ед. хр. 2608. Об изобретенном капитан-лейтенантом Чистяковым телеграфе. 1824. 19 листов.

7. *Бутаковъ А.* Телеграфные сигналы для господъ помѣщиковъ. СПб: тип. Врем. деп. воен. поселений, 1833. 12 с.

8. *Oersted H.Ch.* Experimenta circa efficaciam conflictus electrici in acum magneticam. Hafniae, 1820.

9. *Поляк Ю.Е.* О мониторинге сетевых научных ресурсов (ЭБ «Научное наследие России») // XIX конференция представителей региональных научно-образовательных сетей «RELARN-2012». Нижний Новгород, 2012. С. 19–21.

10. *Fahie John Joseph.* A history of electric telegraphy to the year 1837. New York: E.&F.N. Spon, 1884.

11. *Поляк Ю.Е.* К истории интернета: первые полвека // История науки и техники. 2018. № 12. С. 3–16. https://doi.org/10.25791/intstg.12.2018.285

12. *Polak Y.* The Bicentennial History of the Electromagnetic Telegraph (from Ørsted's Experiments to Social Networks) // International Conference Engineering Technologies and Computer Science EnT. 2020. P. 91–95.

13. *Лейбов Р.Г.* Телеграф в поэтическом мире Тютчева: тема и жанр // Лотмановский сборник. М., 2004. Вып. 3. С. 346–356.

PUBLICATIONS OF THE XIX-XX CENTURIES ABOUT THE TELEGRAPH (BASED ON MATERIALS FROM ELECTRONIC LIBRARIES)

Y. E. Polyak[0000-0001-8411-335X]

Central Economics and Mathematics Institute of the Russian Academy of Sciences, 47 Nakhimovski Pr. Moscow, 117418 Russia

polak@cemi.rssi.ru

Abstract

The century before last saw revolutionary changes in the transmission of information. For the functioning of the optical telegraph, which appeared in the late 18th century, cumbersome towers were necessary for the line of sight of semaphore signals. A hundred years later, the length of telegraph lines was hundreds of thousands of kilometers. At the turn of the century, the first experiments with the use of wireless telegraph began. This is reflected in numerous brochures, books, periodicals of that time. After another hundred years, many of these materials have become publicly available thanks to the development of the Internet and electronic libraries; they are intensively viewed and posted online. The rapid growth in the number of digital libraries and their content made this paper possible. Its goal is to trace the evolution of technologies and processes of information transfer reflected in literature, using a wide variety of electronic libraries - from the ambitious projects of the Library of Congress or Google Books with their millions of digitized books to modest private collections. dedicated to local topics. Materials from more than 20 electronic libraries have been used.

Keywords*: digital libraries, history of technology, optical telegraph, electromagnetic telegraph, trans-Atlantic cable, radio.*

REFERENCES

1. *Standage T.* Victorian Internet: The Remarkable Story of the Telegraph and the Nineteenth Century's On-line Pioneers. N.Y., Walker & Company, 1998. 227 p.

2. Tochnoe i podrobnoe opisanie telegrafa ili novoizobretennoj dal'noizveshchayushchej mashiny, pomoshchiyu kotoroj v samoe kratchajshee vremya mozhno

dostavlyat' i poluchat' izvestiya iz samyh otdalennejshih mest. Moskva, tipografiya I. Zelennikova. 1795.

3. *Danilevskij V.V.* Russkaya tekhnika. Leningrad: Lenizdat, 1947. S. 516.

4. *Artobolevskij I.I.* Russkij izobretatel' i konstruktor Kulibin. Nauchno-popularnaya biblioteka soldata i matrosa. M.: Voenizdat, 1948. 36 s.

5. Stoletie voennogo ministerstva. 1802–1902. Glavnoe inzhenernoe upravlenie. Istoricheskij ocherk.- SPb: Tipografiya «Slovo», 1902. 675 s.

6. Ministerstvo vnutrennih del. 1802–1902. Istoricheskij ocherk. Prilozhenie vtoroe. Pochta i telegraf v XIX stoletii. SPb: Tip. Ministerstva vnutrennih del. 1901. 335 s.

7. *Butakov A.* Telegrafnye signaly dlya gospod pomeshchikov. SPb: tip. Vrem. dep. voen. poselenij, 1833. 12 s.

8. *Oersted H.Ch.* Experimenta circa efficaciam conflictus electrici in acum magneticam. Hafniae, 1820.

9. *Polyak Y.E.* O monitoringe setevyh nauchnyh resursov // XIX konferenciya predstavitelej regional'nyh nauchno-obrazovatel'nyh setej «RELARN-2012». Nizhnij Novgorod, 2012. S. 19–21.

10. *Fahie John Joseph.* A history of electric telegraphy to the year 1837. New York: E.&F.N. Spon, 1884.

11. *Polyak Y.E*. K istorii interneta: pervye polveka // Istoriya nauki i tekhniki. 2018. No. 12. S. 3–16. https://doi.org/10.25791/intstg.12.2018.285

12. *Polak Y.* The Bicentennial History of the Electromagnetic Telegraph (from Ørsted's Experiments to Social Networks) // International Conference Engineering Technologies and Computer Science EnT. 2020. P.91–95.

13. *Lejbov R.G.* Telegraf v poeticheskom mire Tyutcheva: tema i zhanr // Lotmanovskij sbornik. M., 2004. Vyp. 3. S. 346–356.

СВЕДЕНИЯ ОБ АВТОРЕ

ПОЛЯК Юрий Евгеньевич – ведущий научный сотрудник Центрального экономико-математического института РАН (Москва). Подробнее: http://computer-museum.ru/articles/sovet-muzeya/561/

Yuri Evgenievich POLYAK – Candidate of Economic Sciences, Leading Researcher, Central Economics and Mathematics Institute. Moscow, Russia. More detailed: http://computer-museum.ru/articles/sovet-muzeya/561/

email: polak@cemi.rssi.ru
ORCID: 0000-0001-8411-335X

Материал поступил в редакцию 9 ноября 2021 года

УДК 004.415.25

РЕДАКТОР ИНТЕРАКТИВНОЙ СТРУКТУРЫ ДЛЯ ИНСТРУМЕНТА ГЕНЕРАЦИИ СЦЕНАРНЫХ ПРОТОТИПОВ

Г. Ф. Сахибгареева[1] [0000-0003-4673-3253], В. В. Кугуракова[2] [0000-0002-1552-4910]

[1,2] Институт информационных технологий и интеллектуальных систем Казанского федерального университета, ул. Кремлевская, 35, г. Казань, 420008
[1]gulnara.sahibgareeva42@gmail.com, [2]vlada.kugurakova@gmail.com

Аннотация

Задача автоматизации рутинной работы сценаристов компьютерных игр, нарративных дизайнеров, поставленная в ранних работах, получила свое продолжение в настоящей работе. Рассмотрены вопросы визуализации разветвленных структур повествования компьютерных игр, проведен анализ различных подходов визуализации сюжета и других важных составляющих видеоигры, выбран технологический стек и приведены конкретные решения для хранения в виде структурированного сценария, позволяющего генерацию продолжения сюжетных веток и тестирование этапа повествовательного прототипирования при помощи автоматически генерируемой текстовой новеллы.

Ключевые слова*: интерактивное повествование, компьютерные игры, сценарий игры, визуализация, тональность текста, разветвленные структуры, повествовательное прототипирование, прототип сценария, структурированный сценарий, GPT-2, ruGPT3, python, unity.*

ВВЕДЕНИЕ

Несомненно, наблюдается бурный рост рынка видеоигр [1], это один из трендов нашего времени. Компьютерные игры выпускаются в огромном количестве, и естественно растет количество интеллектуальных систем для их разработки. Процесс разработки компьютерных игр – длительный и дорогостоящий процесс. При создании компьютерных игр идет разработка в большом количестве направлений: 2D/3D визуализация, UI/UX дизайн, программирование и дизайн игрового процесса, программирование искусственного интеллекта, дизайн персонажей, уровня и окружения, создание сценария, нарративный дизайн, зву-

ковое и музыкальное сопровождение. Поэтому для индустрии актуально создание новых эффективных инструментов автоматизации рутинных процессов.

Анализ таких инструментов [2] показывает, что они способствуют увеличению уровня вариативности сюжета игр. «*Использование искусственного интеллекта в реализации интерактивных повествовательных систем увеличивает выразительные возможности системы, частично принимая на себя творческую ответственность за повествовательный опыт пользователя. Это, в свою очередь, может обеспечить большую отзывчивость и разнообразие повествований без уменьшения самостоятельности игрока».*

Ниже мы сфокусируем наше внимание на визуализации разветвленной структуры сюжета, проверке построенных графов сюжета на непротиворечивость, возможности хранения упрощенного (по сравнению с натуральным текстом) структурированного сценария в формате JSON и автоматическом формировании продолжений сюжетных веток. Все эти новые функции должны быть интегрированы в общее решение для работы над интерактивным повествованием компьютерных игр [3, 4].

1. ГРАФИЧЕСКОЕ ОТОБРАЖЕНИЕ ПОВЕСТВОВАНИЯ

Для автоматизации процесса утверждения сценариев компьютерной игры необходима качественная визуализация её разветвленной структуры с возможностью не только её автоматического построения, но и автоматической проверки на логичность.

Ряд приложений, таких как Twine [5], Articy:Draft [6], Fungus [7], Storybricks Engine [8], реализует в той или иной мере функционал управления игровым контентом, включая отображение структур. Так, например, Storybricks Engine – механизм с элементами искусственного интеллекта – формализует возможность создания нарратива истории, ложащейся в основу будущей компьютерной игры, с крайне сложными, ветвящимися сюжетными арками.

В качестве примера сложности интерактивных структур можно привести (см. рис. 1) фрагмента сценария из рабочего проекта компании Quantic Dream над игрой Detroit: Become Human [9]. Совершенно нечитаемое представление тем не менее несет массу смысла, но явно необходимо изменить отображение в

сторону большей иллюстративности.

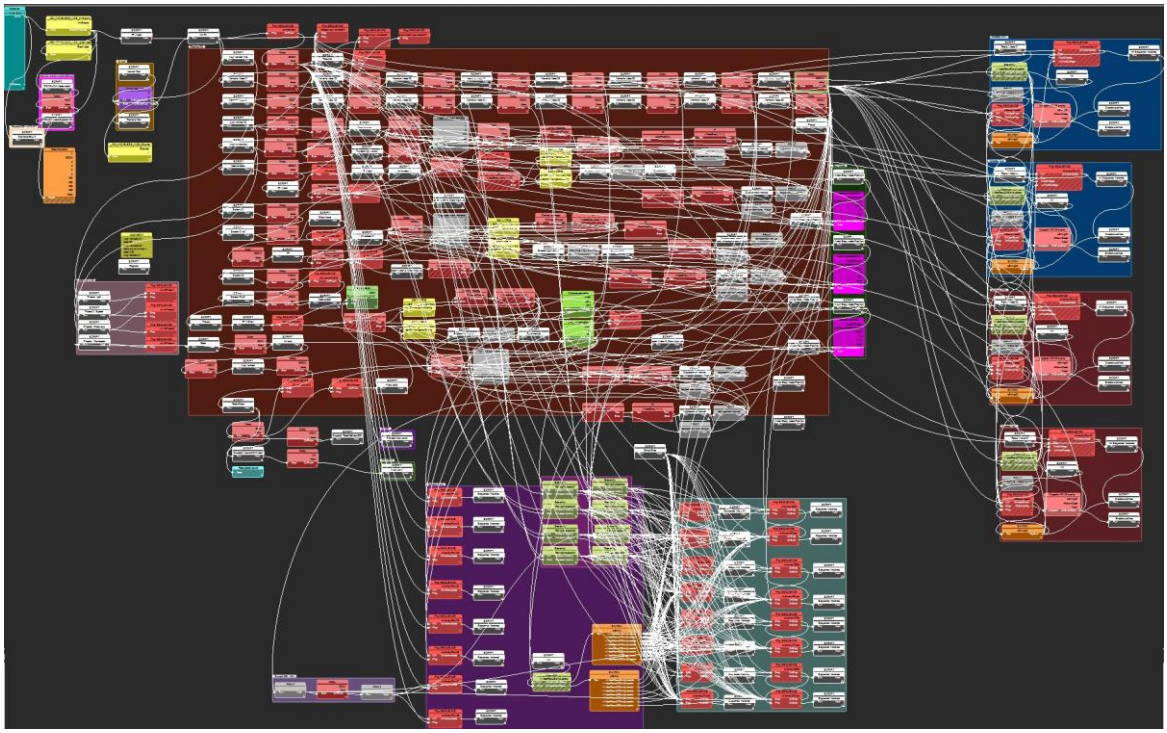


Рис. 1. Фрагмент графа сценария игры Detroit: Become Human [10].

К слову, визуализация разветвленных структур хорошо представлена в научной литературе, зачастую совсем не относящейся к теме разработки игр. Например, StoryFlow [11] используется для уточнения конкретизации хронологических событий (см. рис. 2) в книгах или киносериалах – и эти структуры представлены в виде нитей (англ., yarn). В таких нитях можно хорошо отразить вариативность происходящих событий или взаимодействие персонажей в конкретных временных промежутках.

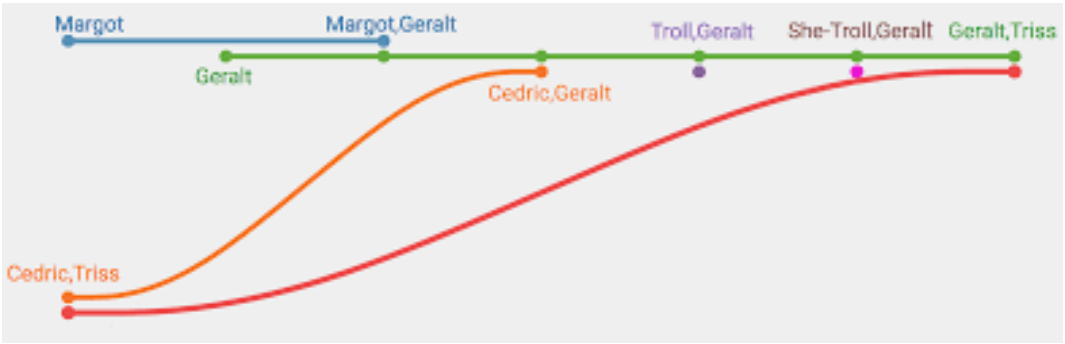


Рис. 2. Структура в виде нитей.

Другой интересный способ визуализации структур – это диаграммы пото-

ков (рис. 3), в которых ширина стрелок пропорциональна скорости потока, так называемая *диаграмма Sankey* [12]. Такое представление может помочь отразить динамику специфических данных. Как пример использования в разработке видеоигр, можно предложить отслеживание изменения характеристик субъектов по ходу сюжета.

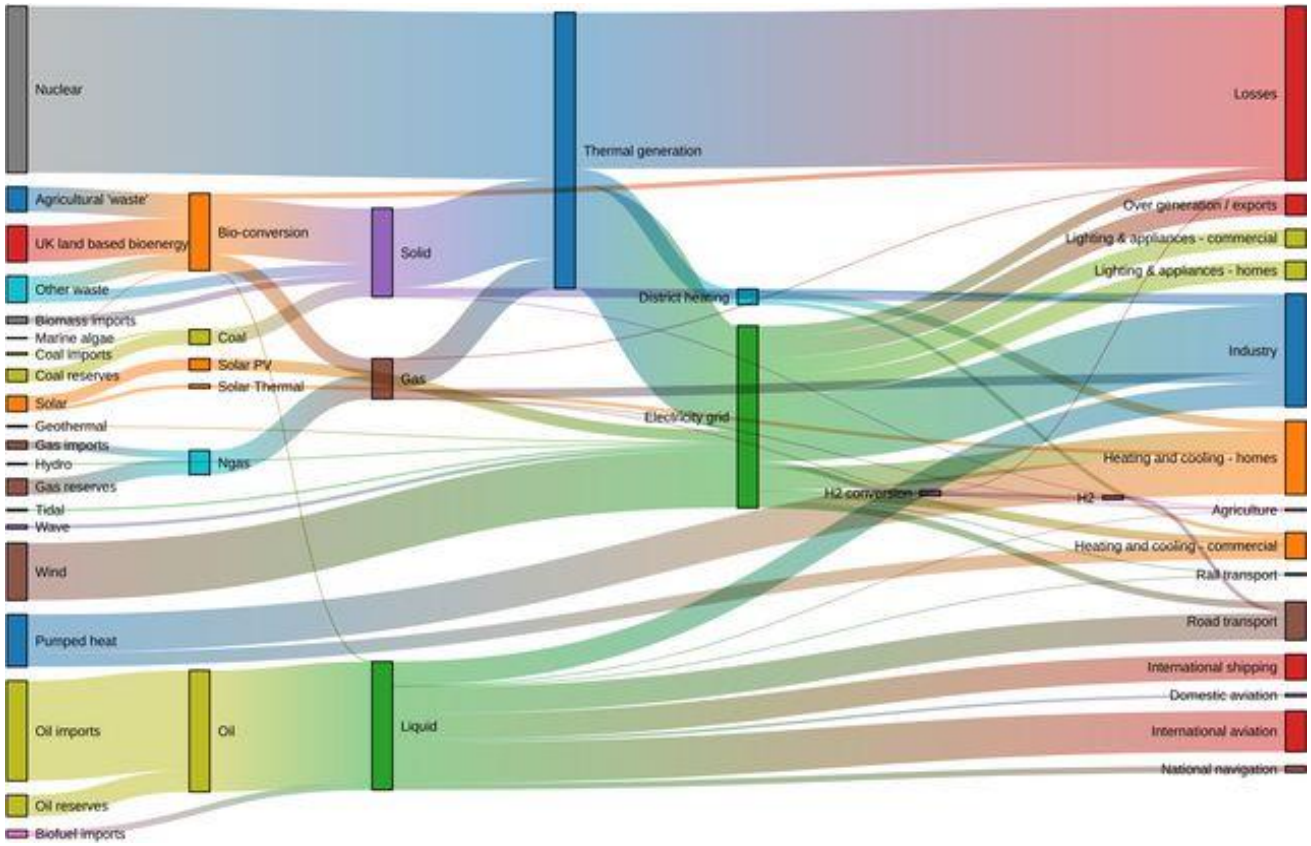


Рис. 3. Пример диаграммы Sankey.

Другим интересным решением для художественных произведений, которое следует интегрировать в инструментарий нарративного проектирования повествования, является визуализация тональностей и сущностей, взятых из текста. Мы реализовали некоторые такие подходы, используя python-библиотеку *Spacy*, проведя эксперименты на текстах из шеститомника Дж.Р.Р. Толкина «Властелин Колец», в процессе чего были извлечены триплеты “субъект → отношение → объект”, именованные сущности текста (персонажи, локации, артефакты и т. д.) и проведена фильтрация триплетов на основе найденных сущностей. Библиотека networkx была использована для построения графа, а библиотека *matplotlib* – для его отрисовки. Из набора триплетов, полученного на последнем этапе обра-

ботки, формировался список вершин-сущностей и рёбер-отношений (см. рис. 4).

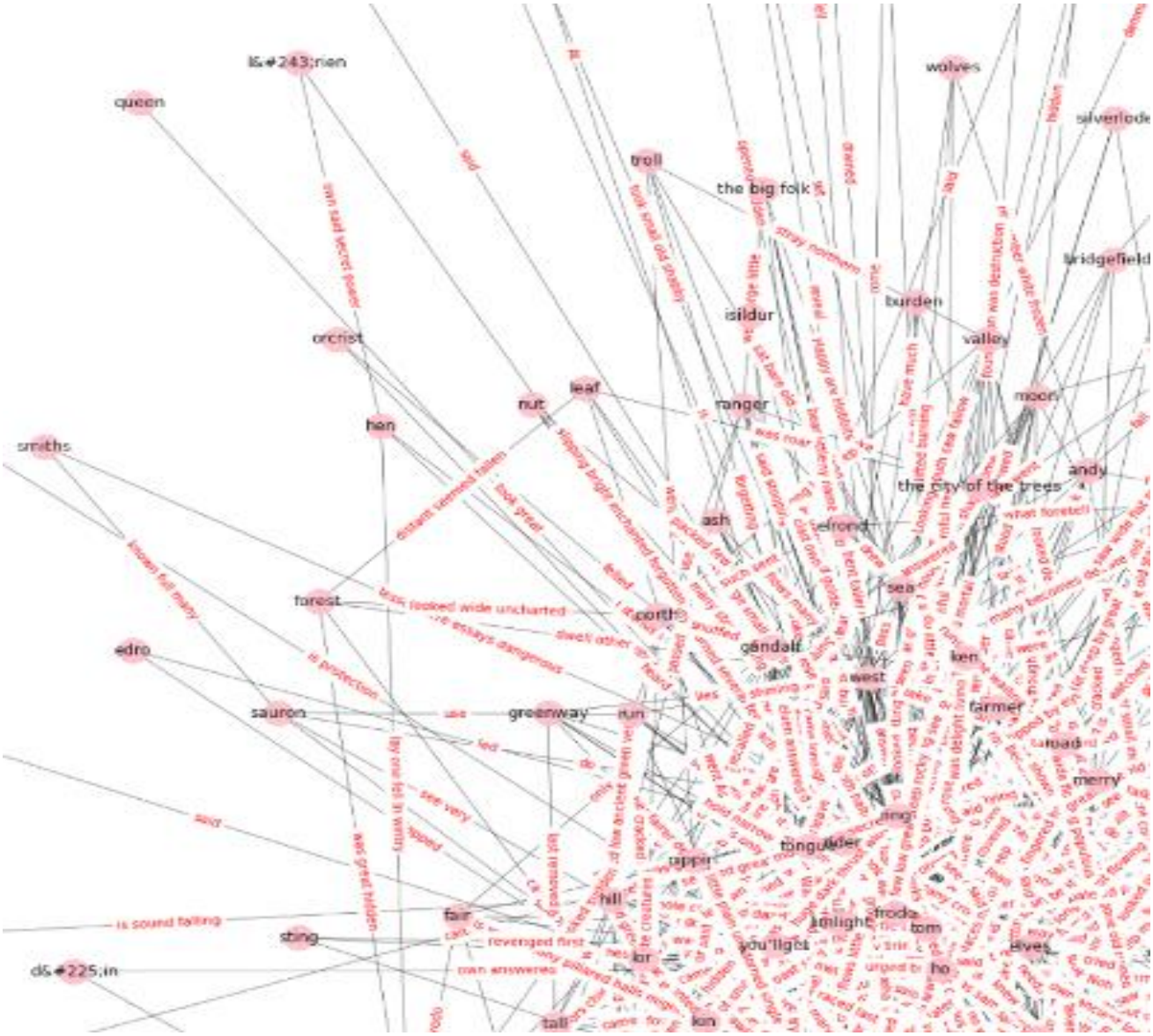


Рис. 4. Фрагмент графа сущностей произведения «Властелин Колец».

При помощи библиотеки *TextBlob* была выявлена тональность текста для каждого предложения. Полученные значений тональности и субъективности использовались как параметры визуализации. Тональность (значения от –1 до 1) интерпретировалась как угол отклонения линии графика, значение субъективности – как длина линии графика. На основе этих параметров происходил расчет координат точек графика по следующим формулам:

$$d = d + T\ (pi / 2);\ x = x0 + cos(d)\ p\ S;\ y = y0 + sin(d)\ p\ S,$$

где *d* – направление (в радианах, по умолчанию 0), *x0* и *y0* – координаты предыдущей точки, *p* – длина линии (константа), *S* – субъективность (значения от 0 до 1), *T* – тональность (значения от –1 до 1). По полученным точкам, с помощью библиотеки *matplotlib* строился график тональностей текста (см. рис. 5).

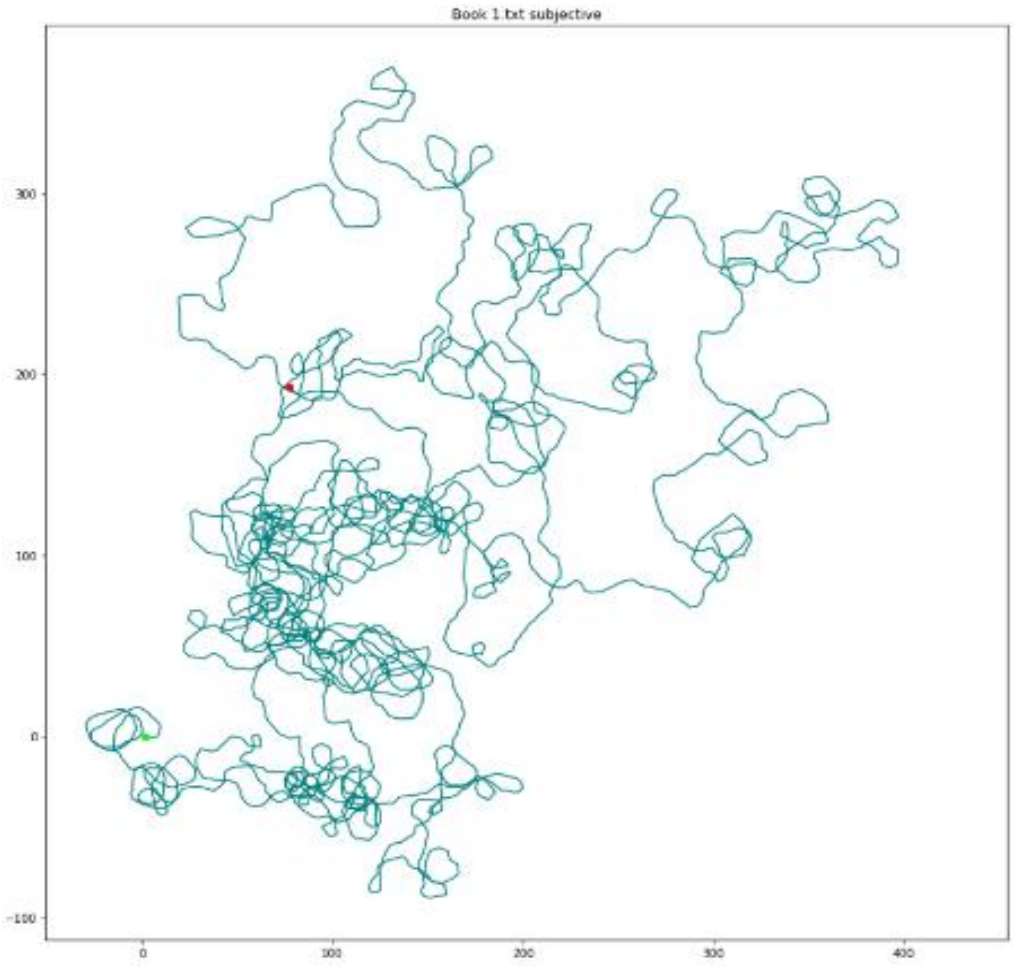


Рис. 5. Пример визуализации тональности текста первой книги «Властелин колец».

Кроме того, необходимо отметить, что существует довольно обширная классификация разнообразных структур [13] сценариев видеоигр, впрочем, не ограничивая общности, можно сказать, что эти структуры общие для любого интерактивного опыта (см., напр., рис. 6). Логично, чтобы шаблоны таких структур были доступны нарративным дизайнерам при проектировании повествования видеоигры.

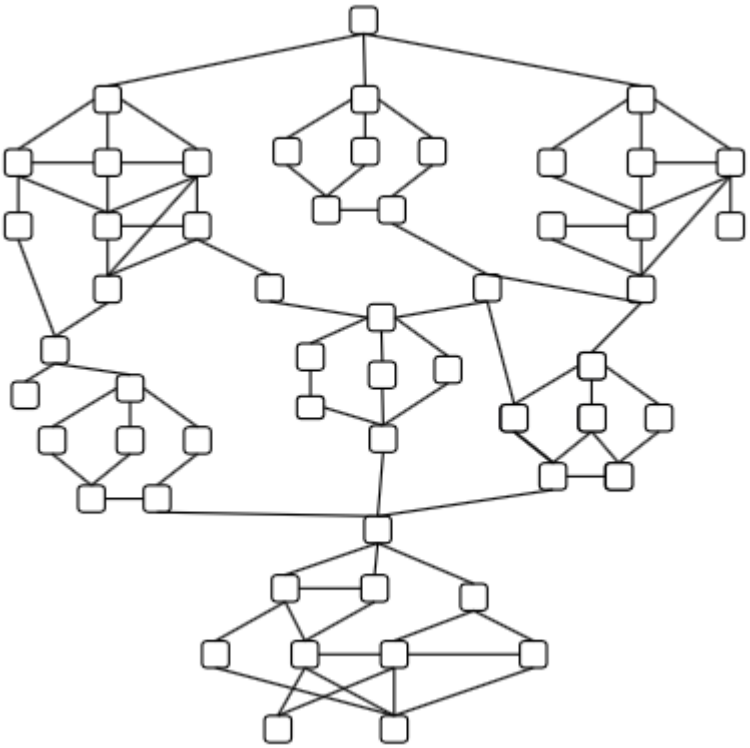


Рис. 6. Квест – одна из разветвленных структур.

Не отказываясь от реализации в дальнейшем этих и других форм визуализации, которые могут оказаться уместными для отображения данных для частных задач, мы выбрали в качестве представления направленный граф, реализация сюжета при помощи которого, однако, пока не решает проблемы, возникающие для более сложных структур, как, например, на рис. 1.

2. ОБОБЩЕНИЕ ОПЫТА

Предшествующие работы описывали общую концепцию инструмента генерации сценарного прототипа [3, 4, 14, 15]. Также отдельной работой был представлен функционал генерации положения камеры и объектов относительно друг друга по текстовому запросу [16].

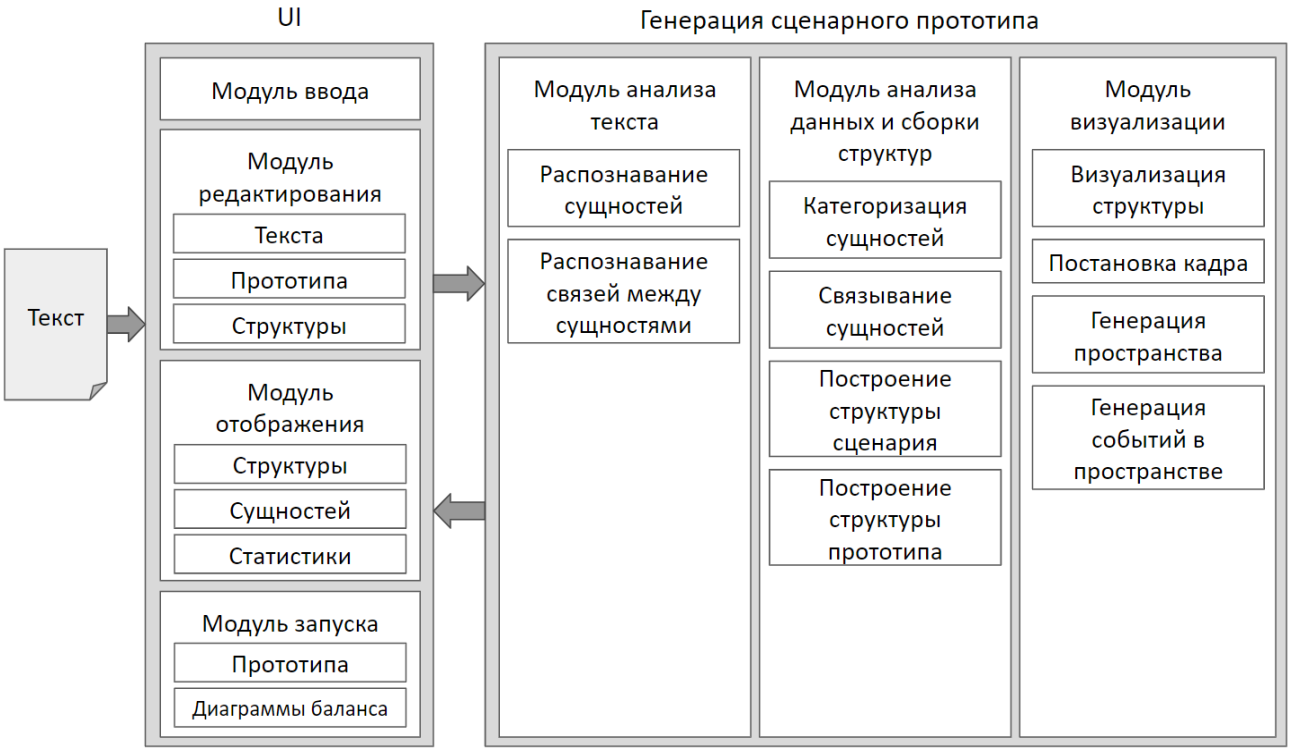


Рис. 7. Архитектура инструмента генерации сценарного прототипа.

Итак, обобщенный конвейер работы генератора (см. рис. 7) выглядит следующим образом:

1. Текст сценария на естественном языке анализируют алгоритмы, которые извлекают из них информацию о внутриигровых сущностях: имена и характеристики персонажей, их реплики, описание локации, основные события.
2. Информация о разветвленной структуре визуализируется в удобной форме, приводится статистика.
3. На основе полученной информации генерируется трехмерная сцена, автоматически подбираются трехмерные модели и анимации.
4. Генерируется программная возможность интерактива в форме предоставления возможности для выбора перехода к тому или иному событию.

Сборка проекта завершается формированием установочного файла, который является *сценарным прототипом*, иными словами, интерактивным проектом, которые игроки и все заинтересованные лица могут пройти или протестировать.

Продолжая разработку, обозначим необходимый функционал для отработки новых возможностей: (1) формализация подхода хранения особой разветвленной структуры, отражающей сюжет видеоигры; (2) возможность автоматического продолжения сюжета.

3. ВЕТВЛЕНИЕ СЮЖЕТА

Опишем, как мы реализовали функцию визуализации разветвленной структуры повествования при помощи создания и редактирования направленного графа с проверкой получаемой структуры на целостность и непротиворечивость, что позволит также запускать игровой проект для проигрывания полученных событий.

Примечательно, что граф проходит проверку различными алгоритмами, которые освобождают пользователя от необходимости тщательно вычитывать проект самостоятельно.

Отдельные части реализации детально представлены в ряде работ [17–19], выполненных при активном участии обоих авторов, ряд других практических работ позволил выбрать эффективные тактики для реализации желаемого функционала.

Граф сценария любого проекта начинается со стартовой вершины. При взаимодействии пользователя с ней открывается редактор персонажей и их свойств. В любой момент пользователь может создавать вершины и связывать их между собой. Условия связывания повторяют определения направленных графов (стартовая вершина должна быть связана минимум с одной вершиной; любая вершина должна иметь родительскую и дочернюю вершины, если это не последние вершины в структуре). Обычные свойства редактора: вершины можно создавать, дублировать, удалять, редактировать, а также изменять положение относительно друг друга для лучшей репрезентативности. Введенные отличия: при взаимодействии с *вершиной действия* становится доступен редактор внутриигровых событий. Такой вершиной может стать любая, кроме стартовой.

Такой редактор графа сюжета (см. рис. 8) необходим только для данной реализации инструмента, отработки поставленных гипотез. Наша цель, чтобы граф создавался автоматически из текста на естественном языке, причем все характеристики персонажей, локаций, событий и связи, балансные коэффициенты

между ними были сформированы также автоматически и непротиворечиво.

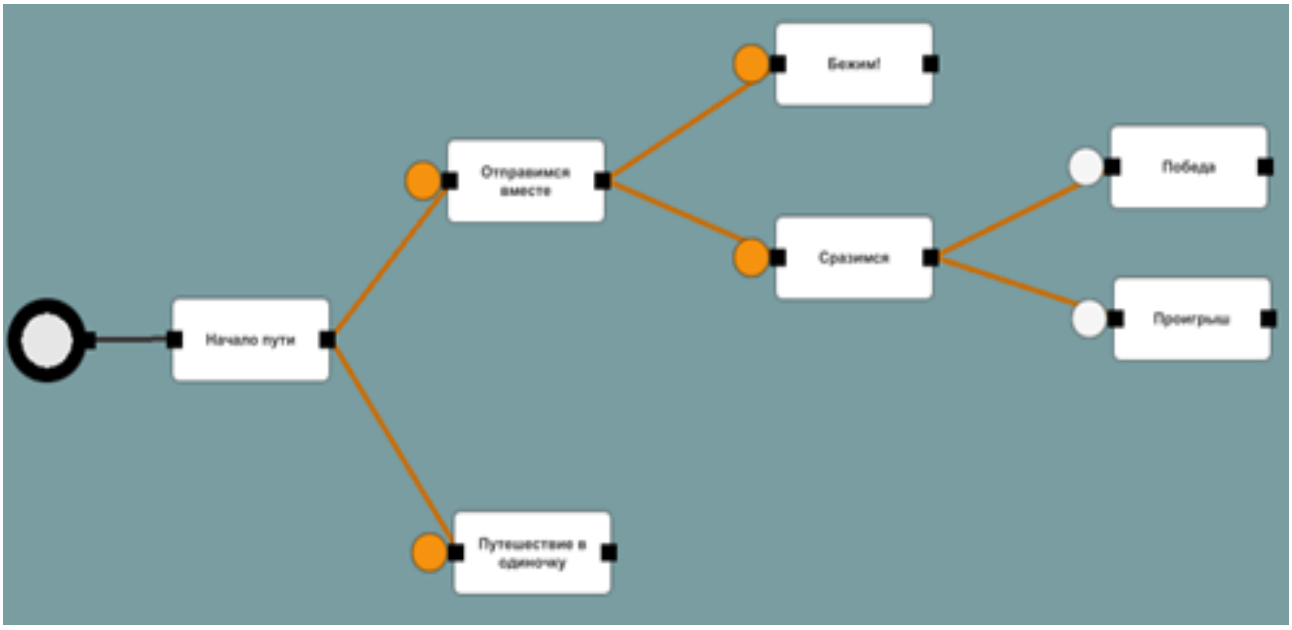


Рис. 8. Визуализация графа.

Полученный файл сюжета представляется в формате *JSON*, что не ограничивает сущности, и содержит в себе введенную (пока) вручную информацию, такую как список персонажей, перечень и значения их свойств; саму структуру направленного графа; условия и производимые эффекты для каждой вершины графа; текстовый отрывок для вершины и выбора действия или реплики; условия перехода из вершины в вершину. По сути это и есть *структурированный сценарий*.

Импортируя данные из структурированного сценария, можно отобразить его в виде графа, изменяя по нуждам нарративного дизайнера ход интерактивного повествования, а для тестирования отдельных этапов нарративного проектирования можно воспроизводить игровую сессию в виде текстовой новеллы.

Отметим существующее ограничение текущей реализации – низкую степень вложенности используемых разветвленных структур сюжета [13], одностороннюю направленность рёбер. Только простейшие из этих структур могут быть реализованы в представленном решении. Получить возможность распознавать в тексте устойчивые крупномасштабные разветвленные структуры и использовать это как шаблон – одна из будущих задач развития инструмента.

4. ПРОДОЛЖЕНИЕ СЮЖЕТА

Еще один модуль позволяет генерировать на основе существующих текстовых отрывков и доступных игроку действий дополнительные вершины, т. е. продолжение развития событий.

Идея была позаимствована из инструмента Storybricks Engine [8], для которого одно время активно развивали собственный технологический подход в формировании динамического нарративного повествования. Был предложен механизм повествования, использующий искусственный интеллект, который дает разработчикам возможность создавать и контролировать повествование с очень сложными разветвленными сюжетными дугами. Впервые анонсированный на конференции Game AI Conference в 2014 году, проходящей в Вене, движок имел заделы на довольно практичные решения некоторых давних проблем проектирования интерактивных историй: в частности, реиграбельности – ведь большинство повествовательных игр (например, BioShock Infinite [20]) следуют по единственному сюжетному пути, который практически не дает возможности повторного прохождения.

Традиционное повествование в видеоиграх работает следующим образом: разработчики пытаются создать историю, которая разворачивается на основе действий, совершаемых игроками. В этой модели игроки в игре изменяют мир, определяя условия запуска, которые показывают следующий шаг в сюжете.

Используя процедурную генерацию контента для автоматического создания ответвлений там, где они необходимы, из кирпичиков-мотиваций персонажей и влияние этой мотивации на мир игры, ИИ-инструмент StoryBricks генерировал новую сюжетную линию на лету, основываясь на том, как выбор игрока влияет на мотивацию других. Мы были впечатлены этой идеей, кроме того, развитие нейронных сетей позволило легко реализовать похожий подход.

Для генерации продолжения развития сюжета используется результат работы *GPT-2* (Generative Pre-trained Transformer) [21, 22] – это мощная языковая модель, которая может быть адаптирована к широкому спектру задач НЛП с использованием очень небольшого набора данных, специфичных для конкретной задачи. До недавнего времени таких моделей для русского языка не существовало. В рамках конкурса AI Journey (https://ai-journey.ru, доступ 2 сентября 2021

года) была выпущена модель ruGPT3, способная генерировать связные и осмысленные тексты и на русском языке.

Главной особенностью *GPT-2* и *ruGPT3*, в частности, является то, что нейросеть не нужно дообучать под конкретную задачу, чтобы та показывала результаты, нужные пользователю. Нейросеть приспосабливается к стилю и содержанию текста, что позволяет ей генерировать реалистичные отрывки, продолжающие исходные фразы. Сразу после обучения нейросеть уже готова генерировать текст со всеми логическими вставками: повторное упоминание имен героев, цитаты, отсылки, выдержка одного стиля на протяжении всего текста, связанное повествование.

Стоит отметить, что алгоритм всякий раз выдает уникальный текст даже при одинаковом запросе (см. табл. 1). В примере показаны варианты продолжения без семантической осмысленности.

Таблица 1. Разные варианты сгенерированного продолжения текста

Художественный отрывок	**Сгенерированное продолжение**
Камера медленно поворачивается вправо, показывая пожилого джентльмена, стоящего в пещере. Раздаются быстрые шаги, когда фигура в белой мантии с капюшоном подбегает к мужчине сзади и заставляет его встать на колени, после чего погружает спрятанное лезвие в его затылок, убивая его.	Камера поворачивается обратно, но уже под другим углом — теперь появляется тень мужчины, стоящего на коленях рядом с трупом.
	Рука убийцы взлетает вверх, и камера поворачивается вверх, показывая высокого седобородого священника, поднимающего вверх один из черепов на длинной рукояти.
	Камера идет влево, показывая другой подземный тоннель, освещенный факелами.
	падает на пол.

В будущем должна быть проработана более эффективная и функциональная генерация, основанная на всех взаимоотношениях персонажей, в том числе временные параметры, которые можно извлечь из полноценного сценария, а не его отрывка.

5. ПОИСК НЕСООТВЕТСТВИЙ

Отдельное внимание стоит уделить функционалу поиска несоответствий в построенном графе сценария. Это могут быть несоответствия свойств персонажей и переходов между вершинами.

Каждый персонаж – это его имя, а также значения свойств, составляющих описание данного персонажа. Свойства могут представлять собой целочисленные, логические или текстовые значения.

Доступен автоматический переход в вершину при выполнении действия или самостоятельный переход игрока в зависимости от выбора.

В вершине возможна проверка значений свойств на выполнение условий. Если условие не выполнено – вершина для игрока недоступна. *Пример*: персонаж имеет текстовое значение класса лучник. Если у персонажа другой класс, то ему доступен другой пул вариантов развития событий.

Кроме того, в каждой вершине возможно применение функций изменения значений свойств в зависимости от происходящих событий. *Пример*: лучник имеет целочисленное свойство «здоровье». Если персонаж попал в неприятность, данное значение снижается, что производится за счет соответствующих вычислений в вершине.

ЗАКЛЮЧЕНИЕ

Представлены простая, но эффективная визуализация разветвленной структуры сюжетов видеоигр и варианты автоматической генерации продолжений для сюжетных веток, позволяющие повысить реиграбельность конечного продукта. Данное решение станет очередной частью одного большого инструмента, призванного упростить работу игровых сценаристов и повысить качество игрового повествование. После объединения функционала в общий конвейер обработки и визуализации информации можно надеяться на создание полноценного инструмента для повествовательного прототипирования. В целом комплексный инструмент должен представлять собой набор редакторов различных аспектов игрового проекта.

Ближайшие планы на будущее: разработка функционала создания и отслеживания квестов; реализация поддержки более сложных и массивных раз-

ветвленных структур для уточнения сценария игры; работа с балансом, для чего уместна интеграция с функционалом инструмента Machination [23].

БЛАГОДАРНОСТИ

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета.

СПИСОК ЛИТЕРАТУРЫ

1. *Седых И.А.* Индустрия компьютерных игр // Национальный исследовательский университет Высшая школа экономики, 2020. 74 с.

2. *Riedl M.O., Bulitko V.* Interactive Narrative: An Intelligent Systems Approach // AI Magazine. 2013. V. 34. 67 p.

3. *Сахибгареева Г.Ф., Кугуракова В.В.* Концепт инструмента автоматического создания сценарного прототипа компьютерной игры // Электронные библиотеки. 2018. Т. 21. № 3-4. С. 235–249.

4. *Сахибгареева Г.Ф., Бедрин О.А., Кугуракова В.В.* Разработка компонента генерации визуализации сценарного прототипа видеоигр // Научный сервис в сети Интернет: труды XXII Всероссийской научной конференции. 2020. С. 581–603.

5. Twine.URL: https://twinery.org/, last accessed 2021/10/21.

6. Articy:draft. URL: https://www.articy.com/en/, last accessed 2021/10/21.

7. Fungus. URL: https://fungusgames.com/, last accessed 2021/10/21.

8. Storybricks Engine.
URL: https://www.youtube.com/watch?v=id-3sUo_DFU&ab_channel=Storybricks, last accessed 2021/10/21.

9. *Cage D.*, Twitter blog.
URL: https://twitter.com/David__Cage/status/1034374760392794112, last accessed 2021/10/21.

10. Detroit: Become Human.
URL: http://www.quanticdream.com/en#!/en/category/detroit, last accessed 2021/10/21.

11. *Padia K., Bandara K., Healey C.* A system for generating storyline visuali-

zations using hierarchical task network planning // Computers & Graphics. 2019. P. 64–75.

12. Sankey Diagram.
URL: https://observablehq.com/@d3/sankey-diagram, last accessed 2021/10/21.

13. *Сахибгареева Г.Ф.* Применимость разветвленных структур для генерации сценарных прототипов видеоигр // 65-я Международная научная конференция Астраханского государственного технического университета. 2021.

14. *Сахибгареева Г.Ф., Бедрин О.А., Кугуракова В.В.* Раскадровка как одно из представлений сценарного прототипа компьютерных игр // Электронные библиотеки. 2021. Т. 24. №2. С. 408–444.

15. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Visualization Component for the Scenario Prototype Generator as a Video Game Development Tool // CEUR. Proceedings of the 22nd Conference on Scientific Services & Internet (SSI-2020). 2020. P. 267–282.

16. *Кугуракова В.В., Сахибгареева Г.Ф., Нгуен А.З., Астафьев А.М.* Пространственная ориентация объектов на основе обработки текстов на естественном языке для генерации раскадровок // Электронные библиотеки. 2020. Т. 23. №6. С. 1213–1238.

17. *Вакатов С.А.* Разработка инструмента вариативности сюжета с запуском прототипа в виде текстовой игры // Казанский (Приволжский) федеральный университет. 2021. 36 с.

18. *Вакатова Э.С.* Разработка функционала генерации продолжения сюжета для инструмента прототипирования сюжета в компьютерных играх // Казанский (Приволжский) федеральный университет. 2021. 33 с.

19. *Каюмов Б.И.* Проблемы визуализации разветвленных сюжетов компьютерных игр // Казанский (Приволжский) федеральный университет. 2021. 79 с.

20. BioShock Infinite. URL: https://2k.com/en-US/game/bioshock-infinite/, last accessed 2021/10/21.

21. *Radford A., Wu J., Child R., Luan D., Amodei D., & Sutskever I.* Language models are unsupervised multitask learners // OpenAI Blog. 2019. V. 1. 9 p.

22. GPT-2. URL: https://openai.com/blog/better-language-models/, last ac-

cessed 2021/10/21.

23. *Adams E., Joris D.* The Designer's Notebook: Machinations, A New Way to Design Game Mechanics.
URL: https://www.gamasutra.com/view/feature/176033/the_designers_notebook, last accessed 2021/10/21.

INTERACTIVE STRUCTURE EDITOR FOR SCENARIO PROTOTYPING TOOL

G. F. Sahibgareeva[1] [0000-0003-4673-3253], V. V. Kugurakova[2] [0000-0002-1552-4910]

[1, 2] Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008
[1]gulnara.sahibgareeva42@gmail.com, [2]vlada.kugurakova@gmail.com

Abstract

The task of automating the routine work of computer game writers and narrative designers, set forth in earlier works, has been continued in the presented work. The issues of visualization of branching narrative structures of computer games are considered, the analysis of various approaches to visualization of the plot and other important components of a video game is performed, a technological stack is selected and specific solutions for storing in the form of a structured script, allowing the generation of continuing narrative branches and testing of the narrative prototyping stage using the automatically generated text novelette are given.

Keywords*: interactive storytelling, computer games, game script, visualization, branched structures, graphs, narrative prototyping, script prototype, GPT-2, ruGPT3, python, unity.*

REFERENCES

1. *Sedyh I.A.* Industriya komp'yuternyh igr // Nacional'nyĭ issledovatel'skiĭ universitet Vysshaya shkola ekonomiki. 2020. 74 s.

2. *Riedl M.O., Bulitko V.* Interactive Narrative: An Intelligent Systems Approach // AI Magazine. 2013. V. 34. 67 p.

3. *Sahibgareeva G.F., Kugurakova V.V.* Koncept instrumenta

avtomaticheskogo sozdaniya scenarnogo prototipa komp'yuternoj igry // Elektronnye biblioteki. 2018. T. 21. № 3-4. S. 235–249.

4. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Razrabotka komponenta generacii vizualizacii scenarnogo prototipa videoigr // Nauchnyj servis v seti Internet: trudy XXII Vserossijskoj nauchnoj konferencii. 2020. S. 581–603.

5. Twine. URL: https://twinery.org/, last accessed 2021/10/21.

6. Articy:draft. URL: https://www.articy.com/en/, last accessed 2021/10/21.

7. Fungus. URL: https://fungusgames.com/, last accessed 2021/10/21.

8. Storybricks Engine.
URL: https://www.youtube.com/watch?v=id-3sUo_DFU&ab_channel=Storybricks, last accessed 2021/10/21.

9. *Cage D.* Twitter blog.
URL: https://twitter.com/David__Cage/status/1034374760392794112, last accessed 2021/10/21.

10. Detroit: Become Human.
URL: http://www.quanticdream.com/en#!/en/category/detroit, last accessed 2021/10/21.

11. *Padia K., Bandara K., Healey C.* A system for generating storyline visualizations using hierarchical task network planning // Computers & Graphics. 2019. P. 64–75.

12. Sankey Diagram. URL: https://observablehq.com/@d3/sankey-diagram, last accessed 2021/10/21.

13. *Sahibgareeva G.F.* Primenimost' razvetvlennyh struktur dlya generacii scenarnyh prototipov videoigr // 65-ya Mezhdunarodnaya nauchnaya konferenciya Astrahanskogo gosudarstvennogo tekhnicheskogo universiteta. 2021.

14. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Raskadrovka kak odno iz predstavlenij scenarnogo prototipa komp'yuternyh igr // Elektronnye biblioteki. 2021. T. 24. №2. S. 408–444.

15. *Sahibgareeva G.F., Bedrin O.A., Kugurakova V.V.* Visualization Component for the Scenario Prototype Generator as a Video Game Development Tool // CEUR. Proceedings of the 22nd Conference on Scientific Services & Internet (SSI-

2020). 2020. P. 267–282.

16. *Kugurakova V.V., Sahibgareeva G.F., Nguen A.Z., Astaf'ev A.M.* Prostranstvennaya orientaciya ob"ektov na osnove obrabotki tekstov na estestvennom yazyke dlya generacii raskadrovok // Elektronnye biblioteki. 2020. T. 23. №6. C. 1213–1238.

17. *Vakatov S.A.* Razrabotka instrumenta variativnosti syuzheta s zapuskom prototipa v vide tekstovoj igry // Kazanskij (Privolzhskij) federal'nyj universitet. 2021. 36 s.

18. *Vakatova E.S.* Razrabotka funkcionala generacii prodolzheniya syuzheta dlya instrumenta prototipirovaniya syuzheta v komp'yuternyh igrah // Kazanskij (Privolzhskij) federal'nyj universitet. 2021. 33 s.

19. *Kayumov B.I.* Problemy vizualizacii razvetvlennyh syuzhetov komp'yuternyh igr // Kazanskij (Privolzhskij) federal'nyj universitet. 2021. 79 s.

20. BioShock Infinite. URL: https://2k.com/en-US/game/bioshock-infinite/, last accessed 2021/10/21.

21. *Radford A., Wu J., Child R., Luan D., Amodei D., & Sutskever I.* Language models are unsupervised multitask learners // OpenAI Blog. 2019. V. 1. 9 p.

22. GPT-2. URL: https://openai.com/blog/better-language-models/, last accessed 2021/10/21.

23. *Adams E., Joris D.* The Designer's Notebook: Machinations, A New Way to Design Game Mechanics.
URL: https://www.gamasutra.com/view/feature/176033/the_designers_notebook, last accessed 2021/10/21.

СВЕДЕНИЯ ОБ АВТОРАХ

САХИБГАРЕЕВА Гульнара Фаритовна – ассистент кафедры программной инженерии Института ИТИС КФУ. Сфера научных интересов – игровая сценаристика, нарративный дизайн, изучение вопроса эффективности создания сценарного прототипа и возможности автоматизации данного процесса.

Gulnara Faritovna SAHIBGAREEVA – assistant of the Department of Software Engineering of the Institute ITIS KFU. Research interests - game scripting, narrative design, studying the issue of the effectiveness of creating a scenario prototype and the possibility of automating this process.

email: gulnara.sahibgareeva42@gmail.com
ORCID: 0000-0003-4673-3253

КУГУРАКОВА Влада Владимировна – к. т. н., доцент кафедры программной инженерии Института ИТИС КФУ, руководитель НИЛ разработки AR/VR приложений и компьютерных игр. Сфера научных интересов – иммерсивность виртуальных сред, проблемы генерации реалистичной визуализации, различные аспекты проектирования игр, AR/VR, подходы к интерпретации UX.

Vlada Vladimirovna KUGURAKOVA, PhD., Docent of the Institute ITIS KFU, Head of Laboratory «AR/VR applications and Gamedev». Research interests include immersiveness of virtual environments, problems of generating realistic visualization, various aspects of game design, AR/VR, approaches to UX interpretation.

email: vlada.kugurakova@gmail.com
ORCID: 0000-0002-1552-4910

Материал поступил в редакцию 25 октября 2021 года

УДК 004.65, 544.33

ЭЛЕКТРОННАЯ БАЗА ДАННЫХ ПО ЭКСПЕРИМЕНТАЛЬНЫМ ЭНЕРГИЯМ ДИССОЦИАЦИИ СВЯЗЕЙ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

В. Е. Туманов[1] [0000-0003-0443-5346], А. И. Прохоров[2] [0000-0002-2009-4265]

[1] Ногинский филиал ГБПОУ МО «Московский областной медицинский колледж № 3», г. Ногинск

[2] Институт проблем химической физики Российской академии наук, г. Черноголовка

[1]tve@icp.ac.ru, [2]aipro@icp.ac.ru

Аннотация

Представленная веб-база данных по экспериментальным гомолитическим энергиям диссоциации связей в органических соединениях предназначена для использования широким кругом химиков теоретиков и практиков в свободном доступе. В работе приведены краткий обзор источников значений энергии диссоциации связей органических молекул, которые вычисляются теоретически, измеряются экспериментально и оцениваются по кинетическим и термохимическим экспериментальным данным, и их представление в базе данных в интернете. Представлена веб база данных по гомолитическим энергиям диссоциации связей органических соединений. Приводимые значения энергий диссоциации связей вычислены по экспериментальным кинетическим и термохимическим данным. Приведены описания источников экспериментальных данных, классов органических соединений и методов расчета. Приведена логическая структура базы данных и дано описание основных полей ее таблиц. Представлена главная поисковая форма интерфейса базы данных и приведен пример результата поиска для конкретного органического соединения. Энергии диссоциации связи снижены до температуры 298,15 К, которая обычно отсутствует в большинстве источников. Аналоги настоящей базы уступают последней в учете температурных корреляций. В настоящее время ведутся работы по анализу и анализу опубликованных данных с учетом энтропийных эффектов.

***Ключевые слова**: электронный справочник, органические соединения, энергия диссоциации связи, база данных, интернет.*

ВВЕДЕНИЕ

Энергия диссоциации связи является одной из фундаментальных характеристик органической молекулы. Она играет важную роль в оценке реакционной способности молекулы в химических превращениях и может быть использована при разработке новых соединений. Значения энергии диссоциации связей органических молекул могут быть вычислены теоретически, измерены экспериментально и оценены по кинетическим и термохимическим экспериментальным данным.

Экспериментальные данные по энергиям диссоциации связи собраны в справочниках [1, 2] и обзорах [3–5].

Вероятно, одно из первых упоминаний о реализации базы данных по экспериментальным значениям энергии диссоциации связей появилось в конце прошлого века Y.-R. Luo (Molecular Structure & Bond Dissociation Energies), позже ссылка http://www.molenergetics.com/construc.htm была удалена. Использование технологии хранилищ данных для таких электронных ресурсов было предложено в [6].

Использование интернет-технологий привело к созданию большого количество электронных ресурсов по термодинамике и термохимии, которые включают значения энергии диссоциации связей органических молекул [7]. Гомолитические энергии диссоциации связей небольших молекул собраны в наборе данных BDE-db [8]. Далее электронные ресурсы в интернет, представленные в виде таблиц не рассматриваются.

Остановимся на веб базах данных, содержащих значения энергии диссоциации связей органических соединений.

Веб база данных iBonD (Китай) содержит экспериментальные и теоретические значения энергий диссоциации гомолитических и гетеролитических связей [9]. Содержит сведения о более чем 7500 соединений, что соответствует [1].

База данных ИВТАНТЕРМО (Россия) в веб редакции «Термические константы веществ» представлена небольшим числом экспериментальных значений [10].

BDE Estimator (США, Великобритания) – веб-ресурс для вычисления энергий диссоциации связей химических соединений с использованием квантово-химических методов и машинного обучения [11, 12]. Позволяет теоретически оценить энергию диссоциацию связей, возможно, в более чем 200000 соединений.

Несмотря на широкое представление термохимических данных в электронных ресурсах интернета, веб-баз данных, ориентированных непосредственно на энергии диссоциации связей органических соединений, явно недостаточно, что делает разработку и публикацию таких баз данных актуальной научно-практической задачей.

Целью настоящей работы является представление электронной справочной базы данных по экспериментальным гомолитическим энергиям диссоциации связей органических соединений. Значения энергии диссоциации связей вычислены по экспериментальным термохимическим и кинетическим данным радикальных реакций в жидкой и газовой фазах.

1. ИСТОЧНИКИ ДАННЫХ, КЛАССЫ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ И МЕТОДЫ РАСЧЕТА

Источниками экспериментальных данных являются научные публикации по кинетике реакций радикального отрыва в жидкой и газовой фазах, термического радикального распада и констант равновесия бимолекулярных радикальных реакций. Отобран небольшой набор экспериментальных термохимических данных в качестве реперных, который представлен в Табл. 1 (фрагмент).

Таблица 1. Значения энергии диссоциации связей реперных соединений.

Соединение	**Связь**	**Энергия диссоциации связи, кДж/моль**
CH_4	C-H	440.0±1.1 [13]
RCH_3	C-H	422.0±2.1 [13]
Me_3CH	C-H	400±2.9 [13]
PhH	C-H	474.0±4.0 [13]
$PhC^{\bullet}H_3$	C-H	375±4.0 [13]
Вторичные алкилпероксиды	O-H	365.5 [14]
Третичные алкилпероксиды	O-H	358.6 [15]
….		

В базе данных представлены значения энергий диссоциации C-H, O-H, N-H, O-O, C-O, S-H, C-S, S-S, C-I, C-Cl, C-Br, C-F связи органических соединений. Основным методами расчета является метод пересекающихся парабол [16], а также законы Гесса и Кирхгофа [17].

Экспериментальные значения энергий диссоциации связей органических соединений отбирались по следующим методам их определения [18].

Химическое равновесие со стабильным радикалом (CHE): использование связи между константой равновесия и энергией Гиббса с учетом зависимости энтальпии реакции от температуры.

Кинетический метод пересекающихся парабол (MIP). Этот метод использует регрессионное соотношение между классическим потенциальным барьером радикальной реакции и классической энтальпией реакции. Данное соотношение при использовании опорной реакции (для которой все параметры известны) в реакционной серии позволяет во многих случаях вычислить энергию диссоциации связей по кинетическим данных исследуемой реакции. Для различных классов реакций были вычислены теоретические ошибки данного метода.

Энергии диссоциации связей вычислялись при той температуре, которая была в эксперименте, после чего значение приводились к температуре 298,15 К. Важно отметить, что если энтропийный эффект для некоторых классов органических соединений является незначительным [5], то, например, для тиофенолов он имеет существенное значение даже в небольшом интервале температур.

Метод: кислотность-окислительный потенциал (AOP): использование термохимического цикла.

Метод фотоакустической калориметрии (PAC): термодинамический метод определения энергий диссоциации связей в растворе.

2. СТРУКТУРА БАЗЫ ДАННЫХ

Таблицы базы данных содержат информацию об идентификации молекулы, значения энергии диссоциации связей в молекуле, экспериментальные или полученные после обработки экспериментальных кинетических данных, литературный источник для каждой связи, метод измерения или оценки, классификатор

органических соединений, а также техническую информацию. Логическая структура базы данных (без таблиц технической информации) приведена на Рис. 1. Описание полей таблиц дано в Таблице 2.

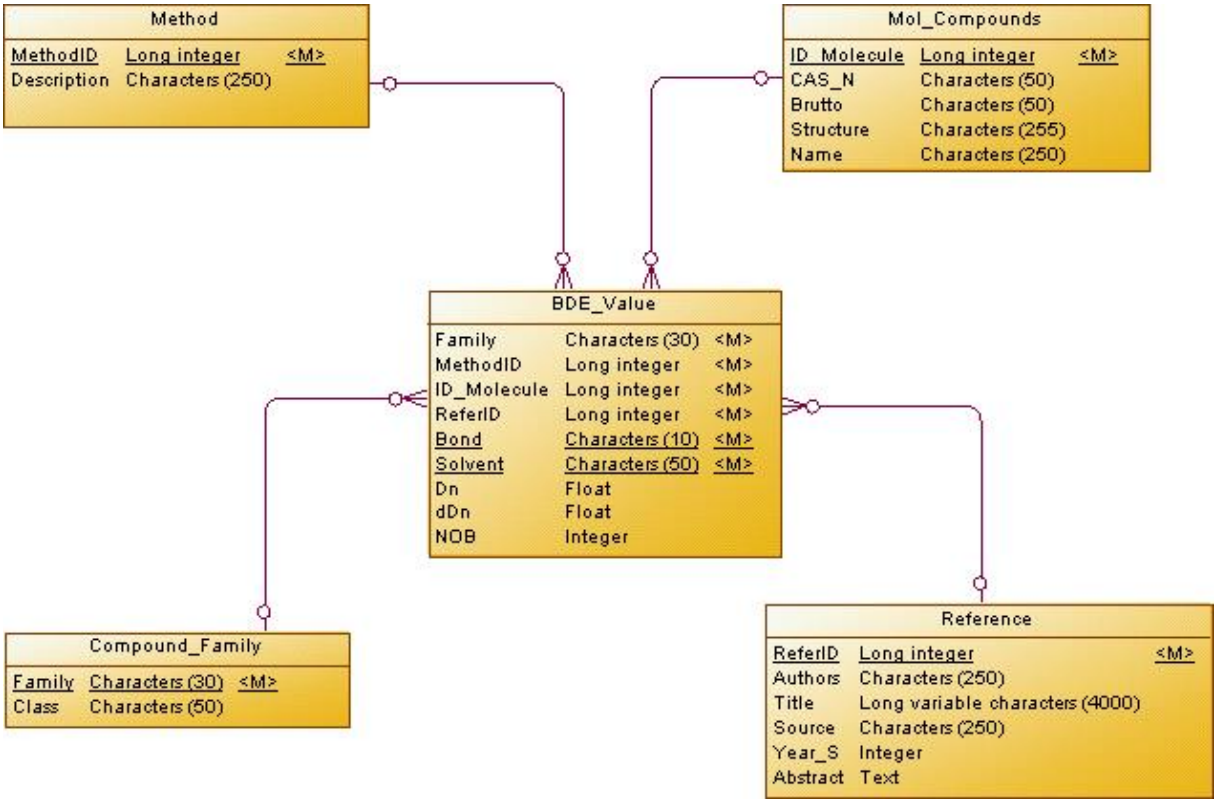


Рис. 1. Логическая структура веб базы данных по энергиям диссоциации органических соединений.

Таблицы технической информации содержат наборы данных, необходимые для пересчета значений энергий диссоциации связей в случаях обоснованного изменения (уточнения) этих значений для опорных соединений. Например, для тиофенола долгое время рекомендуемым значением энергии диссоциации S-H связи считалось значение, равное 330 кДж/моль, в настоящее время это значение равно 349,0 кДж/моль [1].

Таблица 2. Поля таблиц базы данных по энергиям диссоциации связей органических молекул.

Поле	**Описание поля**
Измерение *Mol_Compounds* – содержит перечень реагентов молекул	
ID_Molecule	Идентификатор молекулы
CAS_N	Регистрационный номер молекулы
Brutto	Атомарная формула молекулы
Structure	Полулинейная структурная формула
Name	Наименование молекулы
Измерение *Compound_Family* – содержит классификацию органических молекул	
Family	Класс молекул
SubFamily	Подкласс молекул
Измерение *Method* – содержит описание методов определения энергии диссоциации связи	
MethodID	Идентификатор метода измерения или оценки
Description	Описание метода
Таблица фактов *BDE_value* содержит данные о энергии диссоциации связи	
ID_Molecule	Идентификатор молекулы
Bond	Тип связи
Fragment	Фрагмент молекулы
ReferID	Идентификатор первоисточника
MethodID	Идентификатор метода измерения или оценки
Dn	Значение энергии диссоциации связи
dDn	Величина погрешности
NOB	Число связей данного типа в молекуле

Каждая молекула отнесена к определенному классу органических соединений, а для некоторых классов и определенному подклассу. Так, для углеводородов определены подклассы n-Alkanes, t-Alkanes, q-Alkanes и Cycloalkanes. Классификация органических соединений, принятая в базе данных, близка к классификации, используемой в [1, 19].

3. ИНТЕРФЕЙС БАЗЫ ДАННЫИ

На Рис. 2 показана главная электронная форма для организации поиска в базе данных. Она состоит из двух окон: первое представляет поисковое дерево для выбора органической молекулы, второе – поисковое окно по полям, идентифицирующим молекулу. На Рис. 3 показан результат поиска с выбором по дереву соединений.

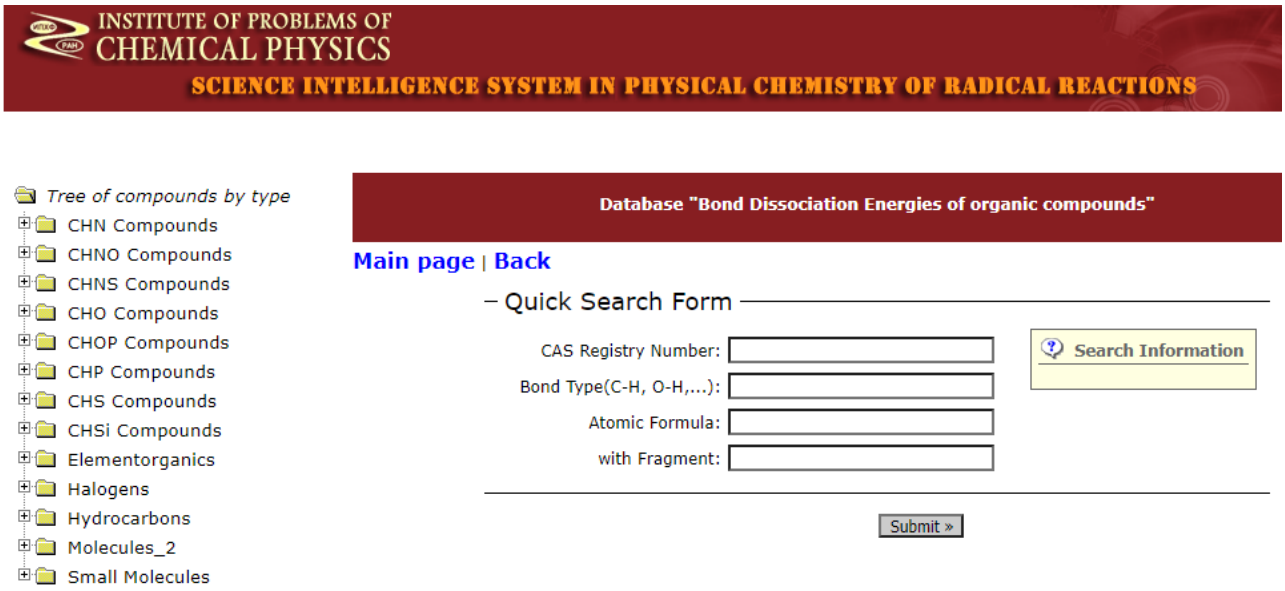


Рис. 2. Главная электронная форма базы данных по энергиям диссоциации связей органических молекул.

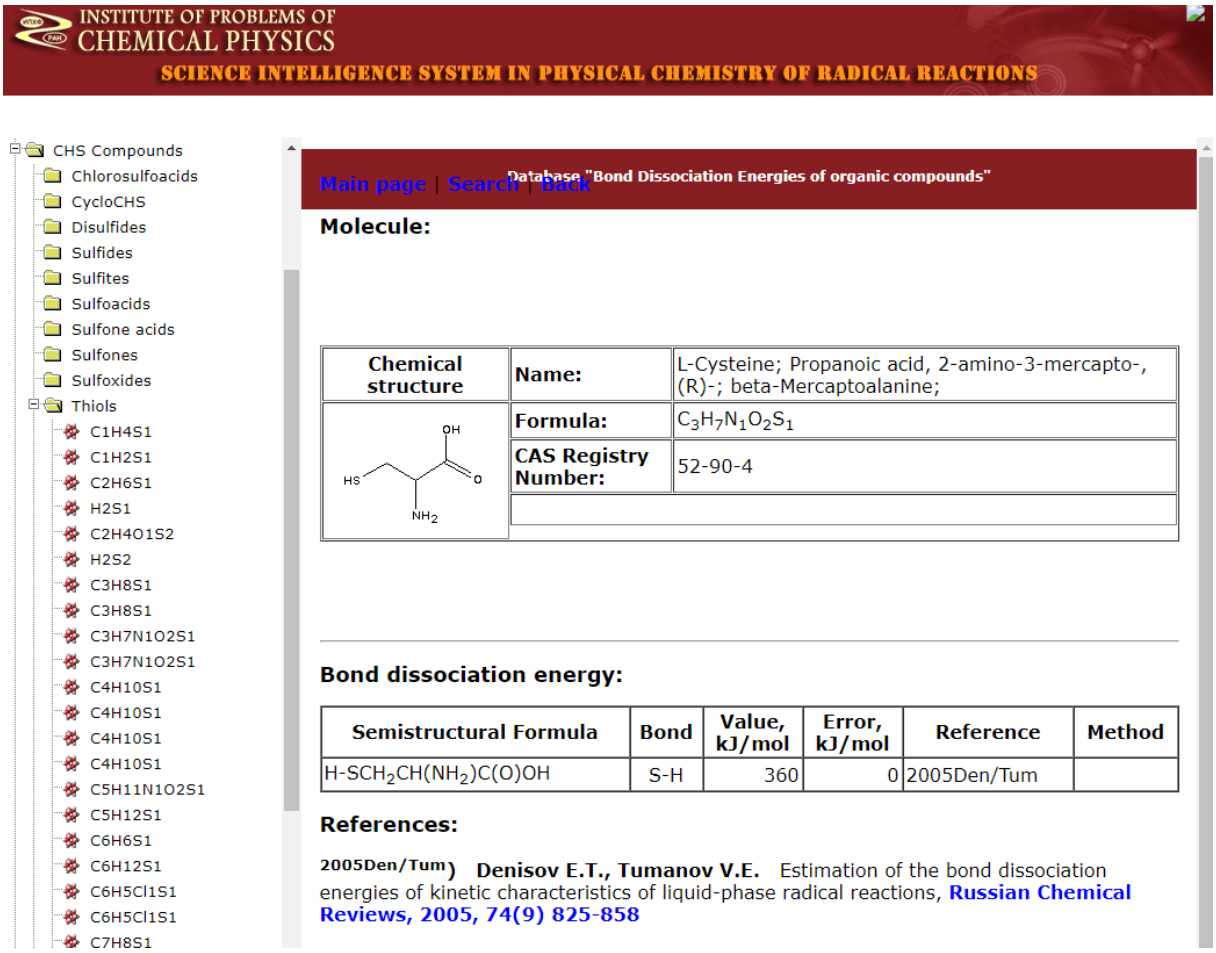


Рис. 3. Результат поиска энергии диссоциации S-H-связи в конкретном соединении.

ЗАКЛЮЧЕНИЕ

Представленная веб база данных по экспериментальным энергиям диссоциации связей в органических соединениях предназначена для использования широким кругом химиков теоретиков и практиков в свободном доступе. Основная часть данных получена из экспериментальных кинетических данных. Аналоги настоящей базы данных уступают последней в учете температурных корреляций.

Значения энергий диссоциации связей приведены к температуре 298,15 К, что, как правило отсутствует в большинстве источников.

В настоящее время проводится работа по анализу и экспертизе публикуемых данных с учетом энтропийных эффектов.

Планируется дополнить базу данных информацией о теплоемкости соединений, а также функционалом для расчета значений энергии диссоциации связей на заданную температуру в определенном интервале температур.

Основная часть работы выполнена при поддержке Российского фонда фундаментальных исследований, проекты 15-07-08645-а и 09-07-00297-а.

СПИСОК ЛИТЕРАТУРЫ

1. *Yu-Ran L.* Comprehensive Handbook of Chemical Bond Energies. Boca Raton: CRC Press, 2007. 1655 p. https://doi.org/10.1201/9781420007282

2. *Гурвич Л.В., Караченцев Г.В., Кондратьев В.Н., Лебедев Ю.А., Медведев В.А., Потапов В.К., Ходеев Ю.С.* Энергии разрыва химических связей. Потенциалы ионизации и сродство к электрону. М.: Наука, 1974. 351 с.

3. *Денисов Е.Т., Туманов В.Е.* Оценка энергий диссоциации связей по кинетическим характеристикам радикальных жидкофазных реакций // Успехи химии. 2005. Т. 74. № 9. С. 905–938. http://dx.doi.org/10.1070/RC2005v074n09ABEH001177

4. *McMillen J.F., Golden D.M.* Hydrocarbon Bond Dissociation Energies // Ann. Rev. Chem. 1982. V. 33. P. 493–532. https://doi.org/10.1146/annurev.pc.33.100182.002425

5. *Blanksby S., Ellison G.* Bond dissociation energies of organic molecules // Accounts of chemical research. 2003. V. *36.* No. 4. P. 255–263. https://doi.org/10.1021/ar020230d

6. *Туманов В.Е., Денисов Е.Т.* База данных по энергиям диссоциации связей углеводородов и их производных // Нефтехимия. 2003. Т. 43. № 1. С. 65–67.

7. Thermochemical databases for pure substances and solutions including alloys, oxides, sulfides, ceramics, aqueous, nuclear and inorganic systems. URL: https://www.crct.polymtl.ca/thermo_databases.html other

8. *John P.S., Guan Y., Kim Y., Kim S., Paton R.S.* BDE-db: A collection of 290,664 Homolytic Bond Dissociation Enthalpies for Small Organic Molecules. figshare. 2019. Dataset. https://doi.org/10.6084/m9.figshare.10248932.v1

9. iBonD 2.0. URL: http://ibond.nankai.edu.cn/

10. *Иориш В.С., Юнгман В.С.* База данных «Термические константы веществ» [Электронный ресурс].
URL: http://www.chem.msu.ru/cgi-bin/tkv.pl?show=welcome.html/welcome.html

11. *John P.C., Guan Y., Kim Y., Kim S., Paton R.S.* Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost // Nat. Commun. 2020. V. 11. P. 2328.
https://doi.org/10.1038/s41467-020-16201-z

12. *John P.C., Guan Y., Kim Y., Etz B.D., Kim S., Paton R.S*. Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules // Sci 2020. Data 7, 244. https://doi.org/10.1038/s41597-020-00588-x

13. *Tsang W.* Heats of Formation of Organic Free Radicals by Kinetic Methods. // In: Martinho Simões J.A., Greenberg A., Liebman J.F. (eds) Energetics of Organic Free Radicals. Structure Energetics and Reactivity in Chemistry Series (SEARCH series). 1996. V. 4. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-0099-8_2

14. *Денисов Е.Т.* Оценка энергии диссоциации O–H-связи фенолов на основании кинетических измерений // Журнал физической химии. 1995. Т. 69. № 4. С. 623–631.

15. *Денисов Е.Т.*, *Денисова Т.Г.* Кинетические параметры реакций $RO_2^{\bullet}$ + RH в рамках параболической модели переходного состояния // Кинетика и катализ. 1993. Т. 34. № 2. С. 199–206.

16. *Денисов Е.Т.* Новые эмпирические модели реакций радикального отрыва // Успехи химии. 1997. Т. 66. № 10. С. 953–971.
http://dx.doi.org/10.1070/RC1997v066n10ABEH000364

17. *Benson S.W*. Thermochemical Kinetics, 2nd ed.; Wiley-Interscience: New York, 1976.

18. *Денисов Е.Т.*, *Денисова Т.Г.* Энергии диссоциации N–H-связей в ароматических аминах (обзор) // Нефтехимия. 2015. Т. 22. № 2. С. 91–109.
https://doi.org/10.7868/S0028242115020070

19. *Domalski E.S.*, *Hearing E.D.* Estimation of Thermodynamic Properties of Organic Compounds in the Gas, Liquid, and Solid Phases at 298.15 K // In: Jochum C., Hicks M.G., Sunkel J. (eds) Physical Property Prediction in Organic Chemistry. Springer, Berlin, Heidelberg. 1988. https://doi.org/10.1007/978-3-642-74140-1_10

ELECTRONIC DATABASE ON EXPERIMENTAL BOND DISSOCIATION ENERGIES OF ORGANIC COMPOUNDS

V. E. Tumanov[1] [0000-0003-0443-5346], A. I. Prokhorov[2] [0000-0002-2009-4265]

[1] Noginsky branch of GBPOU MO "Moscow Regional Medical College № 3", Noginsk

[2] Institute of Problems of Chemical Physics of RAS, Chernogolovka

[1]tve@icp.ac.ru, [2]aipro@icp.ac.ru

Abstract

The presented web database on experimental homolytic bond dissociation energies in organic compounds is intended for use by a wide range of theoreticians and practitioners in free access. The paper provides a brief overview of the sources of the dissociation energies of bonds of organic molecules, which are calculated theoretically, measured experimentally and estimated from kinetic and thermochemical experimental data, their presentation in the Internet database. A web database on homolytic bond dissociation energies of organic compounds is presented. The reported bond dissociation energies are calculated from experimental kinetic and thermochemical data. Descriptions of experimental data sources, classes of organic compounds and calculation methods are given. The logical structure of the database and the description of the main fields of its tables are given. The main search form of the database interface is presented and an example of a search result for a specific organic compound is given. Bond dissociation energies are calculated at a temperature of 298.15 K, which is usually absent in most sources. The analogs of the present base are inferior to the latter in taking into account temperature correlations. Currently, work is underway to analyze and analyze the published data taking into account the entropy effects.

Keywords*: electronic directory, organic compounds, bond dissociation energy, database, internet.*

REFERENCES

1. *Yu-Ran L.* Comprehensive Handbook of Chemical Bond Energies. Boca Raton: CRC Press, 2007. 1655 p. https://doi.org/10.1201/9781420007282

2. *Gurvich L.V.* Energii razryva himicheskih svyazej. Potencialy ionizacii i srodstvo k elektronu / L.V. Gurvich, G.V. Karachencev, V.N. Kondrat'ev, Yu.A. Lebedev, V.A. Medvedev, V.K. Potapov, Yu.S. Hodeev. M.: Nauka, 1974. 351 s.

3. *Denisov E.T., Tumanov V.E.* Estimation of the bond dissociation energies from the kinetic characteristics of liquid-phase radical reactions // Russian Chemical Reviews. 2005. V. 74. No. 9. P. 825–858.
http://dx.doi.org/10.1070/RC2005v074n09ABEH001177

4. *McMillen J.F., Golden D.M.* Hydrocarbon Bond Dissociation Energies // Ann. Rev. Chem. 1982. V. 33. P. 493–532.
https://doi.org/10.1146/annurev.pc.33.100182.002425

5. *Blanksby S., Ellison G.* Bond dissociation energies of organic molecules // Accounts of chemical research. 2003. *V.* 36. No. 4. P. 255–263.
https://doi.org/10.1021/ar020230d

6. *Tumanov V.E., Denisov E.T.* A bond energy database for hydrocarbons and related compounds // Petroleum chemistry. 2003. V. 43. No. 1. P. 62–64.

7. Thermochemical databases for pure substances and solutions including alloys, oxides, sulfides, ceramics, aqueous, nuclear and inorganic systems.
URL: https://www.crct.polymtl.ca/thermo_databases.html other

8. *John P.S., Guan Y., Kim Y., Kim S., Paton R.S.* BDE-db: A collection of 290,664 Homolytic Bond Dissociation Enthalpies for Small Organic Molecules. figshare. 2019. Dataset. https://doi.org/10.6084/m9.figshare.10248932.v1

9. iBonD 2.0. URL: http://ibond.nankai.edu.cn/

10. *Iorish V.S., Ungman V.S.* Baza dannyh “Termicheskie konstanty veshchestv”.
URL: http://www.chem.msu.ru/cgi-bin/tkv.pl?show=welcome.html/welcome.html

11. *John P.C., Guan Y., Kim Y., Kim S., Paton R.S.* Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost // Nat. Commun. 2020. V. 11. P. 2328.
https://doi.org/10.1038/s41467-020-16201-z

12. *John P.C., Guan Y., Kim Y., Etz B.D., Kim S., Paton R.S.* Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules // Sci 2020. Data 7, 244. https://doi.org/10.1038/s41597-020-00588-x

13. *Tsang W.* Heats of Formation of Organic Free Radicals by Kinetic Methods. // In: Martinho Simões J.A., Greenberg A., Liebman J.F. (eds). Energetics of Organic Free Radicals. Structure Energetics and Reactivity in Chemistry Series (SEARCH series). 1996. V. 4. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-0099-8_2

14. *Denisov E.T.* Evaluation of the dissociation energies of the O-H-bond in phenols on the basis of kinetic measurements // Russian Journal of Physical Chemistry A. 1995. V. 69. P. 563–571.

15. *Denisov E.T., Denisova T.G.* Kinetic parameters of the reactions $RO_2^{\bullet}$ + RH in the framework of the parabolic model of transition state // Kinetics and Catalysis. 1993. V. 34. P. 173–179.

16. *Denisov E.T.* New empirical models of radical abstraction reactions // Russ. Chem. Rev. 1997. V. 66. No. 10. P. 859–876. http://dx.doi.org/10.1070/RC1997v066n10ABEH000364

17. *Benson S.W.* Thermochemical Kinetics, 2nd ed.; Wiley-Interscience: New York, 1976.

18. *Denisov E.T., Denisova T.G.* Dissociation Energies of N—H Bonds in Aromatic Amines // Petroleum Chemistry. 2015. V. 55. No. 2. P. 85–103. https://doi.org/10.1134/S0965544115020073

19. *Domalski E.S., Hearing E.D.* Estimation of Thermodynamic Properties of Organic Compounds in the Gas, Liquid, and Solid Phases at 298.15 K // In: Jochum C., Hicks M.G., Sunkel J. (eds). Physical Property Prediction in Organic Chemistry. Springer, Berlin, Heidelberg. 1988. https://doi.org/10.1007/978-3-642-74140-1_10

СВЕДЕНИЯ ОБ АВТОРАХ

ТУМАНОВ Владимир Евгеньевич – преподаватель естественно-научных дисциплин Ногинского филиала ГБПОУ МО «Московский областной медицинский колледж № 3**»**, г. Ногинск, учитель физики ЧОУ «Православная классическая гимназия им. Константина Богородского», к. х. н. Сфера научных интересов – научные базы данных, хемометрика, термохимия, математическое моделирование.

Vladimir Evgen'evicn TUMANOV – lecturer of natural sciences at the Noginsk branch of the Moscow Regional Medical College No. 3, Noginsk; Physics teacher, ChOU «Orthodox Classical Gymnasium. Konstantin Bogorodsky», Candidate Chem. Sci. Research interests include scientific databases, chemometrics, thermochemistry, mathematical modeling.

email: tve@icp.ac.ru,

ORCID: 0000-0003-0443-5346

ПРОХОРОВ Андрей Иванович – зам. начальника административно-организационного отдела Института проблем химической физики РАН. Сфера научных интересов – научные базы данных, хемометрика, термохимия, математическое моделирование.

Andrey Ivanovich PROKHOROV – Deputy Head of the Administrative and Organizational Department of the Institute of Problems of Chemical Physics of the Russian Academy of Sciences. Research interests include scientific databases, chemometrics, thermochemistry, mathematical modeling.

email: aipro@icp.ac.ru,

ORCID: 0000-0002-2009-4265

Материал поступил в редакцию 12 октября 2021 года

УДК 519.6, 519.2

ДАЛЬНЕЙШЕЕ РАЗВИТИЕ ИССЛЕДОВАНИЙ ПОЛЕЙ ДАВЛЕНИЯ В АРКТИЧЕСКОМ РЕГИОНЕ РОССИИ

Н. П. Тучкова[1] [0000-0001-5357-9640], К. П. Беляев[2] [0000-0003-2111-2709], Г. М. Михайлов[3] [0000-0002-4535-7180], А. Н. Сальников[4] [0000-0001-8669-9905]

[1-4]Вычислительный центр им. А.А. Дородницына ФИЦ Информатика и управление РАН, г. Москва

[2]Институт океанологии им. П.П. Ширшова РАН, г. Москва

[4]ФОУ ВПО «Московский государственный университет имени М.В. Ломоносова», факультет ВМК, г. Москва

[1]natalia_tuchkova@mail.ru, [2]kosbel55@gmail.com, [3]gmickail@ccas.ru, [4]salnikov@angel.cs.msu.ru

Аннотация

Представлены результаты исследований атмосферного давления в Арктическом регионе России в период с 1948 по 2008 годы. Проведен анализ климатического сезонного хода полей атмосферного давления. В качестве основного метода исследования использован вероятностный и статистический анализ временных рядов поля давления длиной в 60 лет в фиксированных точках области Арктической зоны России. Всего было исследовано около 90000 ежедневных (с шестичасовым шагом) значений давления. На основе этих данных построен климатический сезонный ход как осреднение значений данного временного ряда в каждой точке пространства и для фиксированной даты. Изучены характеристики сезонного хода, его амплитуда и фаза. Эти характеристики были проанализированы, проведена их геофизическая интерпретация. В частности, определены минимальное и максимальное значения ряда по всей области и построены временные ряды этих характеристик. Показано, что отклонение носит несимметричный характер, это составляет неочевидный результат исследований. Для максимума и минимума построены наилучшие аппроксимации, и эти аппроксимации протестированы известными методами статистического анализа, включая методы максимального правдоподобия, наименьших квадратов и методы (критерии) согласия,

в частности, χ^2-критерий. Проведенное исследование имеет приложение как чисто физическое (позволяет объяснить природу, генезис и распространение крупномасштабных атмосферных образований в климатическом году), так и прогностическое (позволяет понять и отследить тенденции в климате, а также количественно оценить масштабы и изменчивость крупномасштабных атмосферных процессов). Численные расчеты выполнялись на суперкомпьютере Ломоносов-2 Московского государственного университета имени М.В. Ломоносова.

Ключевые слова*: анализ временных рядов, климатический сезонный ход, максимальные и минимальные значения давления внутри климатического года.*

ВВЕДЕНИЕ

В работе использованы методы анализа временных рядов, в частности, разбиение временного ряда на периодическую и непериодическую составляющие. Такие методы успешно используются при анализе финансового рынка [1] и многолетней изменчивости геофизических характеристик, таких как температура воздуха или воды [2], и в более сложных моделях и схемах [3]. В геофизических схемах часто используется понятие климатического сезонного хода, когда строятся средние значения всего временного ряда на каждую дату года и в каждой фиксированной точке пространства. Например, все значения на 1 января в конкретной точке пространства за весь период наблюдений усредняются, и в результате строится среднее значение ряда на 1 января, которое считается климатическим значением. Эту процедуру осуществляют на каждое число внутри года, таким образом, строится климатический сезонный ход конкретной физической характеристики. Подробно этот метод описан, например, в работе [4]. Далее в исследованиях можно более подробно проанализировать полученный ряд наблюдений, например, выделить максимальные и (или) минимальные значения этой физической характеристики по заданной области и изучить изменчивость только этих максимума или минимума. Заметим важность названных характеристик, поскольку, например, в поле атмосферного давления эти экстремумы связаны с такими физическими процессами, как циклоны и антициклоны, их локализацией и изменчивостью.

Изучению Арктического региона посвящено колоссальное количество современных работ, поскольку климат этого региона оказался особенно подвержен

изменениям в связи с глобальным потеплением последних десятилетий, что привело к значительному уменьшению снежного и ледяного покрова [5, 6]. Например, в работе [7] представлены результаты анализа полей давления Арктического региона в «ранний инструментальный» период 1801–1920 гг. Эти данные относятся ко времени начала сбора метеорологических данных из сети регулярных станций. Массив наблюдений [7] за 20 лет недостаточен для оценки климатического сезонного хода, однако их анализ позволил авторам получить представление о состоянии давления в Арктике и выяснить, что, в целом, оно было ниже современного.

В настоящей работе продолжены исследования, опубликованные ранее в работах [8, 9]. Так же, как в этих работах, здесь использовалось поле атмосферного давления в области, ограниченной координатами 62°с.ш. – 80°с.ш. и 15°в.д. – 60°в.д. С одной стороны, эта область достаточно широка, чтобы пренебречь локальными особенностями атмосферных процессов, с другой – достаточно однородна, так как размеры крупных атмосферных образований сопоставимы с размерами всей области. По времени данные по давлению записаны с 1 января 1948 г. по 31 декабря 2008 г. ежесуточно в одноградусной сетке. Эти данные получены в Гидрометцентре России[1] и использовались ранее в некоторых работах, например, [10].

Ниже приведены также результаты вероятностного анализа полей атмосферного давления, выполненного на основе разбиения всего ряда на периодическую и непериодическую составляющие. Отдельно такой анализ осуществлен для максимальных и минимальных значений полей давления по области и внутри климатического года. Показано, что изученные процессы могут быть представлены в виде суммы, где одно слагаемое характеризует регулярный периодический сигнал, а другое – случайный процесс, независимый от первого. Характеристики этого случайного процесса можно определить из полученной выборки. При этом и периодический сигнал, и характеристики случайного процесса для максимума и минимума давления различны, имеют свои принципиальные особенности и требуют дополнительного изучения.

[1] http://meteoinfo.ru

Для поля атмосферного давления в регионе, который мы рассматриваем, максимальные и минимальные значения заметно отличаются. Если минимумы по пространству изменяются от 980 *гПа* до 995 *гПа*, то максимумы изменяются от 1010 *гПа* до 1025 *гПа*, и эти значения в течение года «мигрируют» внутри области. Это связано с поведением циклонов и антициклонов, их передвижением, углублением и перестройкой. В Арктической области России циклонов в течение климатического года наблюдается заметно больше, чем антициклонов, поэтому распределения их характеристик различны. Отметим, что размер циклонического атмосферного образования, которое, в основном, и формирует поле давления, сопоставим с размерами всей рассматриваемой области. Другими словами, мы имеем дело с одним, максимум, двумя циклонами и антициклонами одновременно.

В настоящей работе выполнено следующее исследование:

- построен климатический сезонный процесс для поля атмосферного давления в районе Арктической области России, описаны его особенности для максимальных и минимальных значений по области;

- построены временные графики этих характеристик, проведен их анализ;

- проведено разбиение этих процессов на периодические и апериодические составляющие, оценены амплитуды и фазы периодических составляющих;

- для апериодических составляющих подобраны оптимальные в смысле минимума дисперсии аппроксимации наблюдаемых величин, показано их согласие с аппроксимирующими распределениями.

1. ДАННЫЕ НАБЛЮДЕНИЯ ПОЛЕЙ ДАВЛЕНИЯ

Рассмотрим поле атмосферного давления в области, ограниченной координатами 62°с.ш. – 80°с.ш. и 15°в.д. – 60°в.д., то есть Европейскую часть Арктической зоны России, включая побережье Балтики, акваторию Белого, Баренцева морей до Карского моря и полуострова Ямал. Это достаточно широкая область для того, чтобы пренебречь локальными особенностями атмосферных процессов, а также достаточно однородная, поскольку размеры крупных атмосферных образований сопоставимы с размерами этой области. Данные наблюдений за давлением записаны в период с 01.01.1948 по 31.12.2008 гг. ежесуточно с интервалом в 6 часов в одноградусной сетке.

Для наглядности представим сами поля давления, характерные для этого региона. На рис. 1 показано среднее поле давления для каждой точки Европейской части Арктической зоны РФ за 60 лет с 1948 до 2008 гг. Видно, что поле представляет собой достаточно гладкую по пространству поверхность, хотя, если представить сами значения в виде кривой (рис.2), то становятся заметны скачки и аномалии внутри годового сезонного хода.

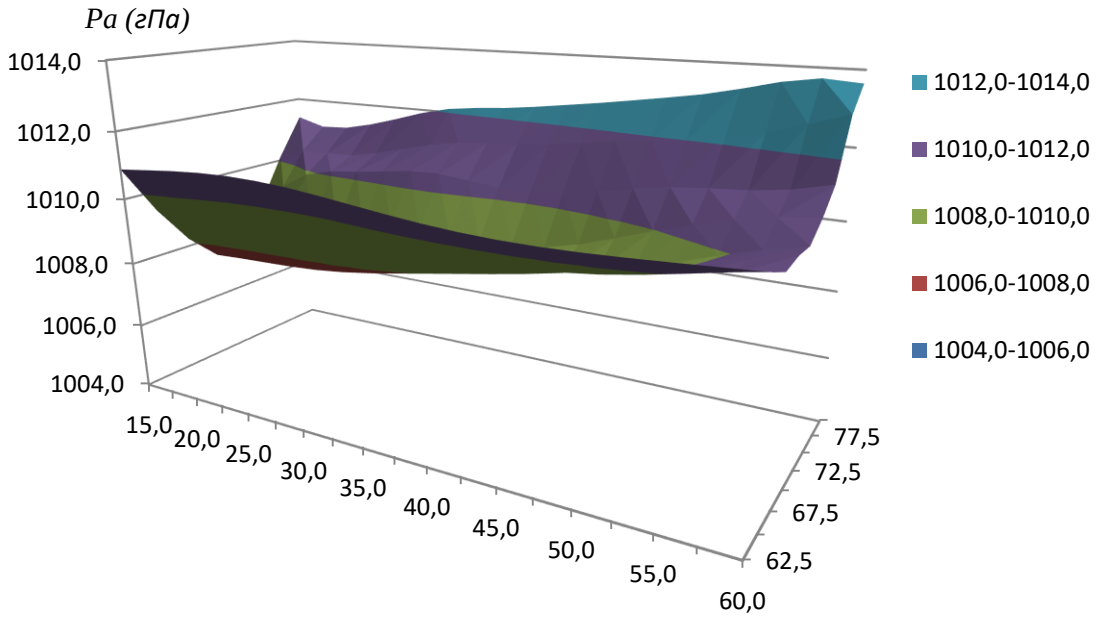


Рис. 1. Поле средних значений давления за 60 лет для каждой точки области, ограниченной координатами 62°с.ш. – 80°с.ш. и 15°в.д. – 60°в.д.

На рис. 2 показана кривая изменения средних значений за весь временной период наблюдений на каждый день. Можно увидеть, что нет видимых закономерностей, но есть периоды наибольших скачков значений, например, в 1958 г. и 1968–1988 гг. Начиная с 1975 г., по данным Гидрометцентра РФ, намечается тренд на увеличение скачков температуры, что показано в работе [11] на картине аномалий среднегодовой температуры воздуха в Арктическом регионе.

Разница аномалий средних значений в период 1948–2008 гг. для Европейской части Арктической зоны России составляет 8,04 *гПа*, а за период 1882–1990 гг. для всего Арктического региона в соответствии с исследованиями [7, Table

VII] – 1,4 *гПа*. Это сравнение дает представление об изменении тренда наблюдаемых значений.

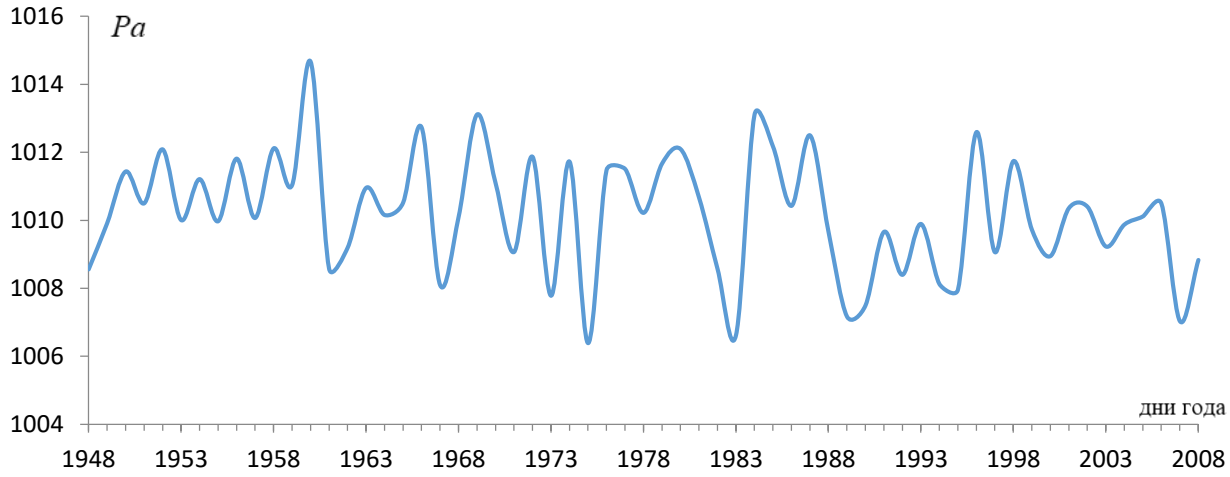


Рис. 2. Кривая средних значений давления для каждого дня за 60 лет с 01.01.1948 по 31.12.2008 в Арктической зоне РФ

2. РЕЗУЛЬТАТЫ АНАЛИЗА ПОЛЕЙ ДАВЛЕНИЯ

В рассматриваемых полях выделялись области минимального и максимального давления за каждые сутки и строились графики их поведения для значений, средних за климатический год.

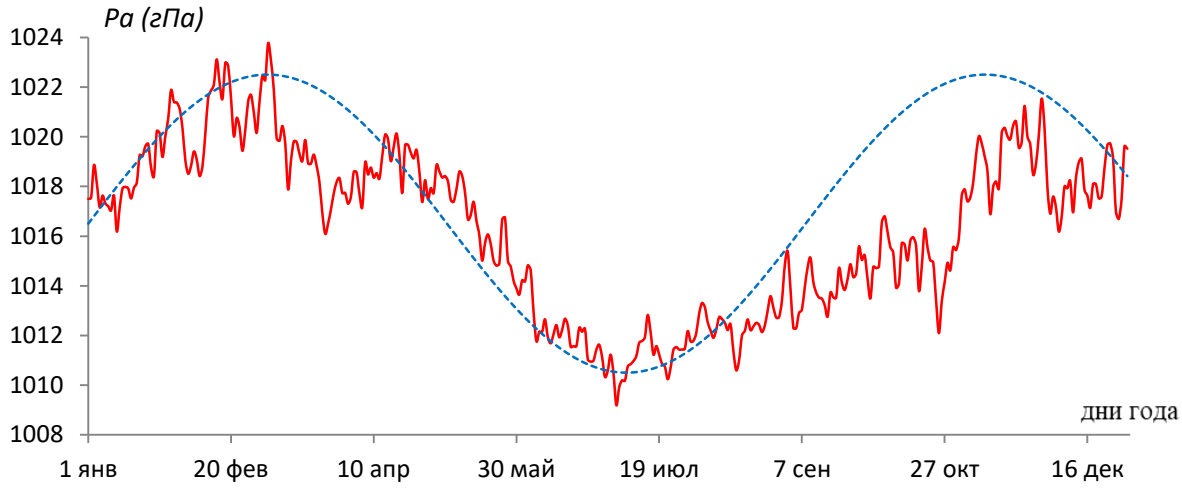


Рис. 3. Поведение среднегодового максимума давления – красная линия; аппроксимация – синий пунктир.

На рис. 3 приведены графики среднего максимума климатического годового хода за рассматриваемый период и аппроксимирующей функции. Показано

поведение среднегодового максимума давления *Pa* (y_i) по всей области за весь период с 1948 по 2008 годы и на всем пространстве 62,5°с.ш. – 80°с.ш. и 15°в.д. – 60°в.д. в сравнении с тригонометрической функцией $f_i = 6\sin(0{,}01 y_i) + 1016{,}5$ (год – *усредненный по массиву данных наблюдений год*). Из рис. 3 очень хорошо видно, что поведение максимума можно представить в виде суммы тригонометрической функции и случайного остатка. Методом наименьших квадратов можно найти параметры этой тригонометрической функции, которая оказывается равной $f_i = A + B\sin \omega y_i$, где *A*=1016,5 (*гПа*), *B*=6 (*гПа*), *ω*=0,01(*год*$^{-1}$). После вычитания из «красной линии» ее аппроксимации получается кривая (*ΔPa*), показанная на рис. 4 (начало координат перенесено для лучшей презентации).

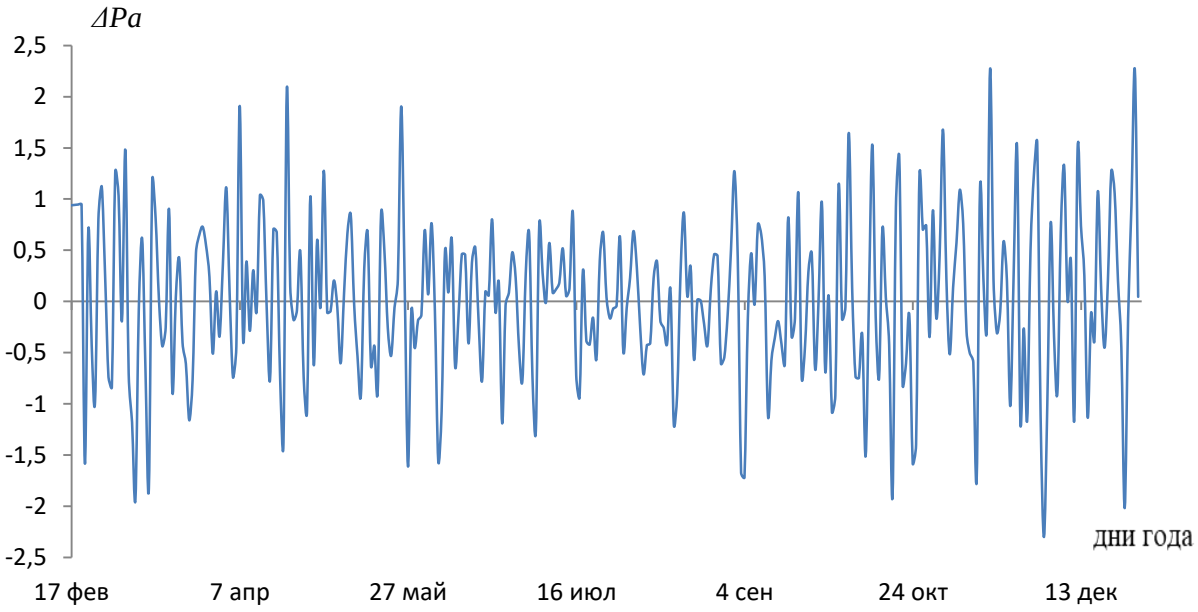


Рис. 4. Разность между наблюдаемой и аппроксимирующей кривой для среднегодового максимума давления.

Из рис. 4 видно, что величина разности *ΔPa* не имеет заметной регулярной изменчивости и может рассматриваться как случайная величина, не зависящая от аппроксимирующей функции (см. рис. 3). Имеет смысл известными статистическими методами, в частности, методом наибольшего правдоподобия, подобрать вероятностное распределение этой величины. Соответствующая аппроксимации

гистограмма частот показана на рис. 5. Предварительно была проведена центровка остатка поля давления относительно величины *A*=1016,5 (*гПа*).

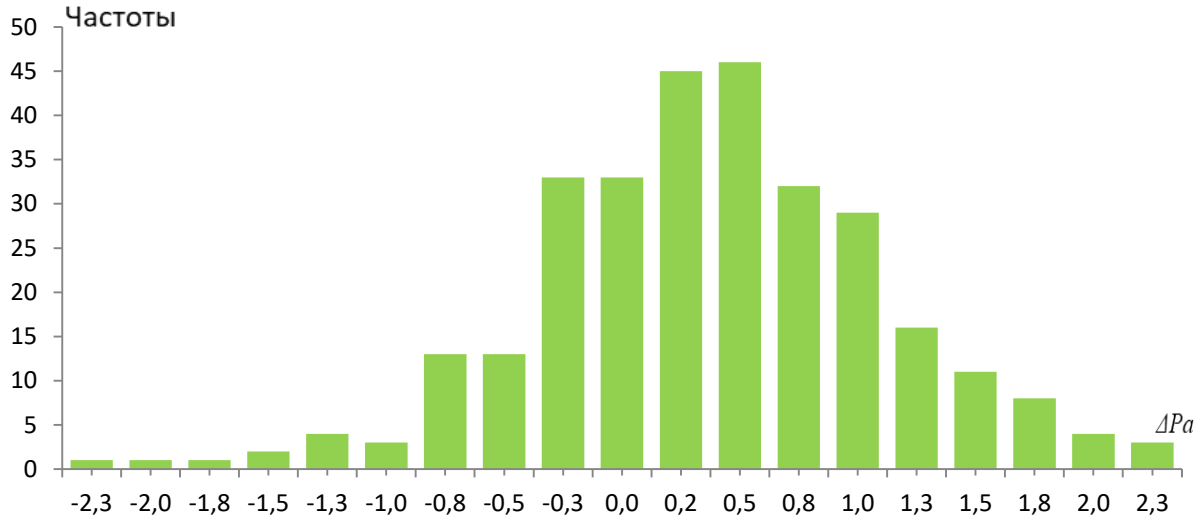


Рис. 5. Гистограмма максимумов для *ΔPa* среднегодового климатического хода.

Эта гистограмма достаточно хорошо, с нужной степенью вероятности относительно критерия χ^2 согласуется с двумя распределениями: Гаусса (Pd_1) и Максвелла (Pd_2), что показано на рис. 6 (а, б).

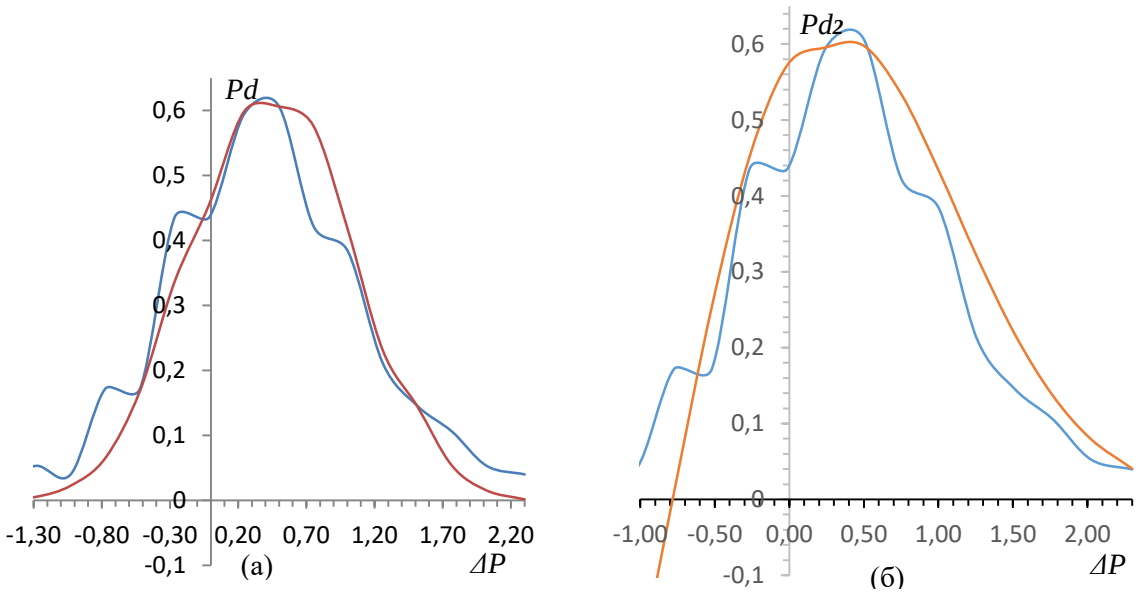


Рис. 6. Аппроксимация случайного остатка максимумов (синяя кривая) распределением: а) Гаусса (красная кривая); б) Максвелла (красная кривая).

Для распределения Гаусса $F_1(x) = \beta \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$ оптимальными параметрами будут α=0,277, σ=0,5, β=0,78. При этом аппроксимация происходит равномерно по всей области, но не очень хорошо в области малых вероятностей («хвостов» распределения). Распределение Максвелла $F_2(x) = \frac{(x-\alpha)}{\sigma^2} e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$ при параметрах α=1, σ=0,8 гораздо лучше аппроксимирует «хвосты», но плохо аппроксимирует область малых отрицательных значений давления.

Аналогичное исследование было выполнено и для минимумов давления. Поведение климатических сезонных минимумов показано на рис. 7 (год – *усредненный по массиву данных наблюдений год*).

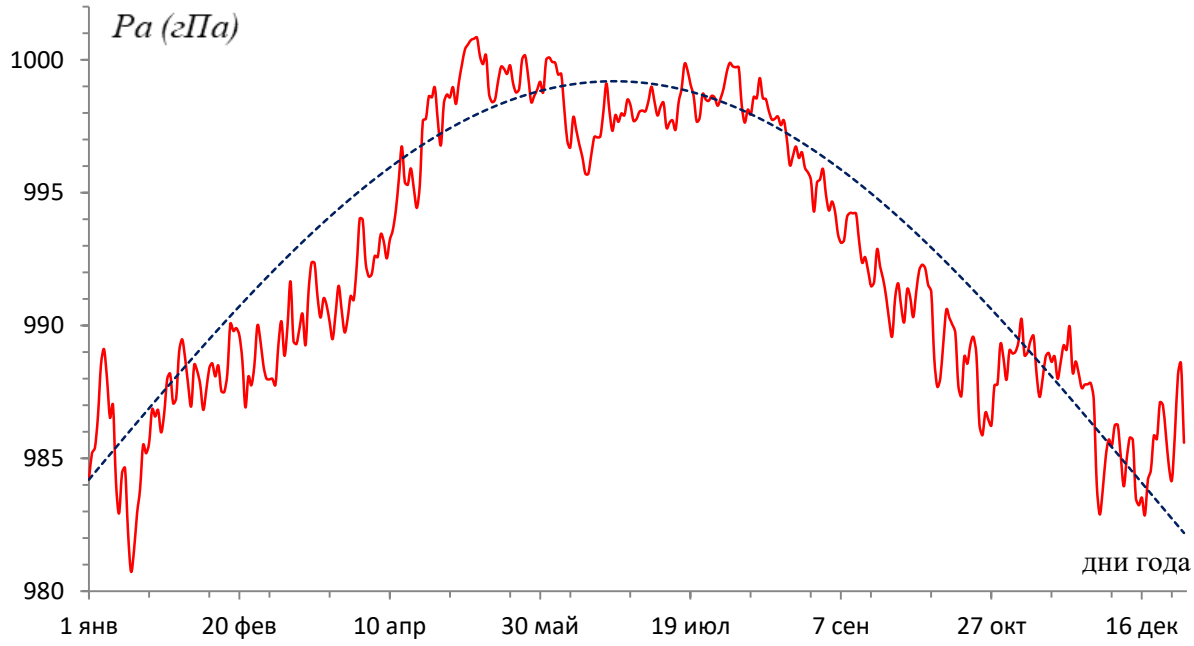


Рис. 7. Кривая средних ежедневных минимумов (красная линия) и аппроксимирующая кривая (синий пунктир).

Кривая средних ежедневных минимумов (1948–2008 гг., во всем регионе) и аппроксимирующая кривая $f_i = A + B\sin\omega x_i$, A=984,2, B=15, ω=0,01 (x_i=0,9*k, k=0,1,...,368) (рис. 7), а гистограмма представлена на рис. 8.

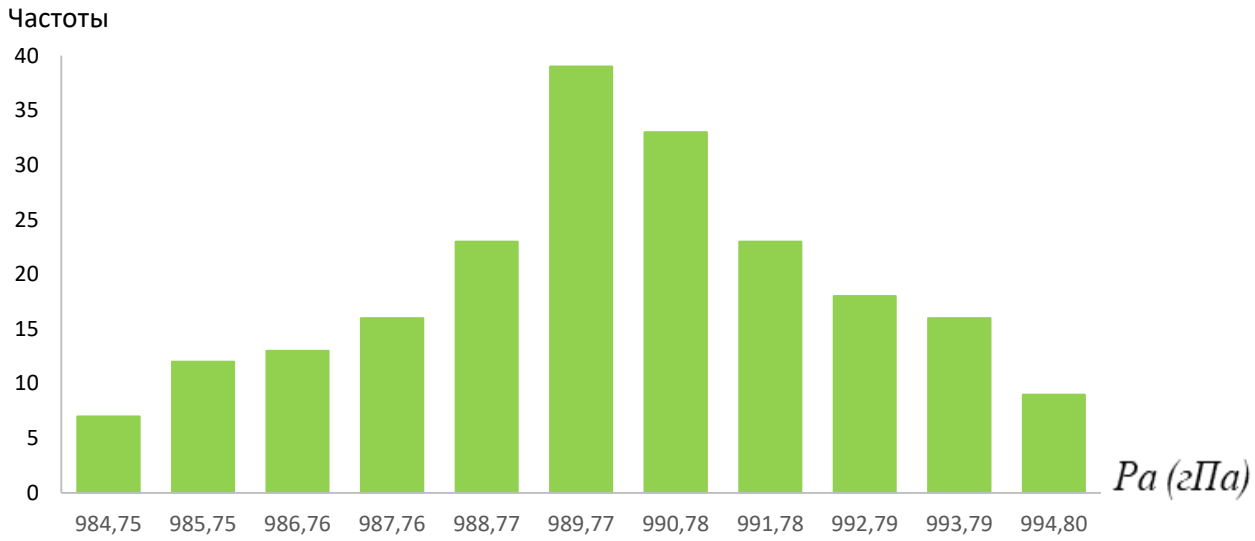


Рис. 8. Гистограмма минимумов среднегодового климатического хода.

По принятому методу соответсвующая функция плотности вероятностей (*Pd*) аппроксимирована распределением Максвелла $F_2(x) = \frac{(x-\alpha)}{\sigma^2} e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$, где α=0, σ=1 (рис. 9).

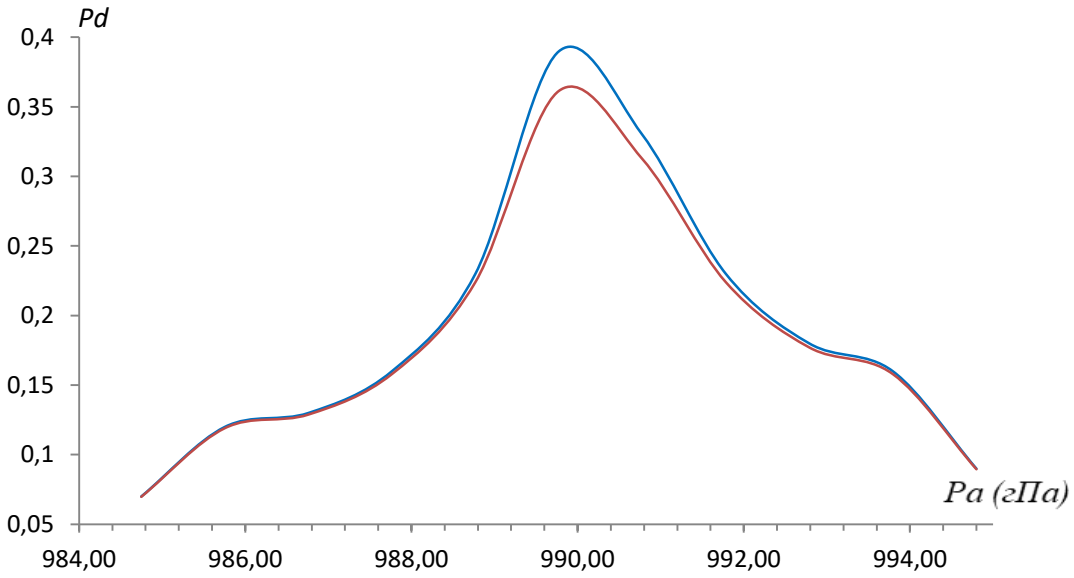


Рис. 9. Аппроксимация случайного остатка минимумов (синяя линия) распределением Максвелла (красная линия).

Из этого рисунка (рис. 9) видно, что распределение Максвелла с приведенными выше параметрами, построенными по методу наибольшего правдоподобия и протестированными по критерию χ^2 с 12 степенями свободы (разбиение интервала давления на 14 подынтервалов и 2 параметра оценивалось по выборке) очень хорошо согласуется с наблюдениями (с доверительным уровнем 95%). Можно отметить, что для минимумов модель суммы фиксированных тригонометрических функций плюс случайный остаток лучше согласуется с наблюдениями, чем для максимумов. Это можно объяснить тем фактом, что для полярной зоны России циклоны, их физические особенности более характерны, чем антициклоны и их изменчивость.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

В работе предложены и реализованы методы вероятностного анализа временного ряда атмосферного давления за 60 лет. Показано, что построенный по такому ряду климатический сезонный ход и построенные по этому сезонному ходу максимальные и минимальные значения, хорошо аппроксимируются суммой регулярного и стохастического слагаемых. При этом дана количественная оценка отклонения реального давления от сезонного хода в виде распределения плотности вероятностей. Показано, что отклонение носит несимметричный характер, что составляет неочевидный результат исследований. Для максимума и минимума построены наилучшие аппроксимации, и эти аппроксимации протестированы известными методами статистического анализа, включая метод максимального правдоподобия, метод наименьших квадратов и методы (критерии) согласия, в частности, χ^2-критерий.

Такое исследование имеет приложение как чисто физическое, то есть позволяет объяснить природу, генезис и распространение крупномасштабных атмосферных образований в климатическом году, так и прогностическое, то есть позволяет понять и отследить тенденции в климате, а также количественно оценить масштабы и изменчивость крупномасштабных атмосферных процессов.

БЛАГОДАРНОСТИ

Работа выполнена в рамках тем Минобрнауки РФ 0128-2021-0002 ИО РАН и «Математические методы анализа данных и прогнозирования» ФИЦ ИУ РАН.

СПИСОК ЛИТЕРАТУРЫ

1. *Kendall M., Stuart A., Ord J.K.* The Advanced Theory of Statistics. Volume 3: Design and Analysis, and Time-Series. Fourth edition Hardcover – March 13, 1983.

2. *Murphy J.* Technical analysis of the futures markets. A Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance, 1986. 556 p.

3. *Привальский В.Е.* Статистическая предсказуемость средней годовой температуры воздуха северного полушария // Докл. АН СССР. 1981. Т. 257. № 6. С. 1342–1345.

4. *Зверяев И.И., Яшаяев И.М.* Сезонная изменчивость полей давления, температуры воды и воздуха в Северной Атлантике по данным COADS // Известия АН СССР. Физика атмосферы и океана. 1996. № 2. С. 222–239.

5. *Фролов И.Е., Гудкович З.М., Карклин В.П., Ковалев Е.Г., Смоляницкий В.М.* Климатические изменения ледовых условий в арктических морях евразийского шельфа // Проблемы Арктики и Антарктики. 2007. № 75. С. 149–160.

6. Environmental Working Group. Edited by F. Fetterer and V.F. Radionov. 2000. *Environmental Working Group Arctic Meteorology and Climate Atlas, Version 1* [Indicate subset used]. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. https://doi.org/10.7265/N5MS3QNJ (доступно 26.10.2021)

7. *Przybylak R., Wyszyński P., Vízi Z., Jankowska J.* Atmospheric pressure changes in the Arctic from 1801 to 1920 // The International Journal of Climatology. 2013. V. 33. P. 1730–1760. https://doi.org/10.1002/joc.3546

8. *Belyaev K., Mikhaylov G., Salnikov A., Tuchkova N.* Seasonal and Decadal Variability of Atmosphere Pressure in Arctic, its Statistical and Temporal Analysis // CEUR Workshop Proceedings, M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen (Aachen, Germany). 2020. V. 2784. P. 51–61.
URL: http://ceur-ws.org/Vol-2784/rpaper05.pdf urn:nbn:de:0074-2784-8.

9. *Беляев К.П., Михайлов Г.М., Сальников А.Н., Тучкова Н.П.* Сезонная и многолетняя изменчивость атмосферного давления в Арктике, статистический и временной анализ // Электронные библиотеки, 2021. Т. 24. № 1. С. 57–73.

10. *Попов С.К.* Влияние морского льда на приливные колебания уровня моря и скорости течений в Баренцевом и Белом морях // Труды Гидрометцентра РФ, Гидрометеорологические исследования и прогнозы. 2018. № 4 (370). С. 137–155.

11. *Бирман Б.А., Бережная Т.В., Голубев А.Д.* Основные погодно-климатические особенности, наблюдавшиеся на Северном полушарии Земли в 2017 г. Аналитический обзор. М: ФГБУ «Гидрометцентр России». URL: http://www.meteorf.ru/upload/iblock/dc0/Бирман%20Климат_2017.pdf (доступно 26.10.2021)

FURTHER DEVELOPMENT OF STUDIES OF PRESSURE FIELDS IN THE ARCTIC REGION OF RUSSIA

N. P. Tuchkova[1] [0000-0001-5357-9640], K. P. Belyaev[2] [0000-0003-2111-2709],
G. M. Mickailov[3] [0000-0002-4535-7180], A. N. Salnikov[4] [0000-0001-8669-9905]

[1-4]Dorodnicyn Computing Center FRC CSC of RAS, Vavilov str., 40, 11933, Moscow
[2]Shirshov Institute of Oceanology of RAS, Nahimovskiy pr., 36, 117218, Moscow
[4]Lomonosov Moscow State University, GSP-1, Leninskie Gory, 11999, Moscow

[1]natalia_tuchkova@mail.ru, [2]kosbel55@gmail.com, [3]gmickail@ccas.ru, [4]salnikov@angel.cs.msu.ru

Abstract

The results of studies of atmospheric pressure in the Arctic region of Russia in the period from 1948 to 2008 are presented. The analysis of the climatic seasonal variation of the atmospheric pressure fields is carried out. As the main research method, the probabilistic and statistical analysis of the time series of the pressure field 60 years long at fixed points in the region of the Arctic zone of Russia was used. In total, about 90,000 daily (in six-hour increments) pressure values were examined. On the basis of these data, a climatic seasonal variation was constructed as an averaging of the values of a given time series at each point in space and for a fixed date. The characteristics of the seasonal course, its amplitude and phase have been studied. These characteristics were analyzed and their geophysical interpretation was carried out. In particular, the minimum and maximum values of the series were determined for the entire region and the time series of these characteristics were constructed. It is shown that the deviation is asymmetric, this is an unobvious research result. For the maximum and minimum, the best approximations were constructed, and these approximations were tested by known methods of statistical analysis, including maximum likelihood, least squares and

goodness of fit methods (tests), in particular, the χ^2-criterion. The conducted research has applications both purely physical (allows to explain the nature, genesis and distribution of large-scale atmospheric formations in a climatic year) and prognostic (allows understanding and tracking trends in climate, as well as quantitatively assessing the scale and variability of large-scale atmospheric processes). Numerical calculations were performed on the Lomonosov-2 supercomputer of the Lomonosov Moscow State University.

Keywords*: time series analysis, climatic seasonal cycle, maximum and minimum pressure values within a climatic year.*

REFERENCES

1. *Kendall M., Stuart A., Ord J.K.* The Advanced Theory of Statistics. Volume 3: Design and Analysis, and Time-Series. Fourth edition Hardcover – March 13, 1983.

2. *Murphy J.* Technical analysis of the futures markets. A Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance, 1986. 556 p.

3. *Prival'skij V.E.* Statisticheskaya predskazuemost' srednej godovoj temperatury vozduha severnogo polushariya // Dokl. AN SSSR. 1981. T. 257. № 6. S. 1342–1345.

4. *Zveryaev I.I., Yashayaev I.M.* Sezonnaya izmenchivost' polej davleniya, temperatury vody i vozduha v Severnoj Atlantike po dannym COADS // Izvestiya AN SSSR. Fizika atmosfery i okeana. 1996. № 2. S. 222-239.

5. *Frolov I.E., Gudkovich Z.M., Karklin V.P., Kovalev E.G., Smolyanickij V.M.* Klimaticheskie izmeneniya ledovyh uslovij v arkticheskih moryah evrazijskogo shel'fa // Problemy Arktiki i Antarktiki. 2007. № 75. S. 149–160.

6. Environmental Working Group. Edited by F. Fetterer and V.F. Radionov. 2000. *Environmental Working Group Arctic Meteorology and Climate Atlas, Version 1* [Indicate subset used]. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. https://doi.org/10.7265/N5MS3QNJ

7. *Przybylak R., Wyszyński P., Vízi Z., Jankowska J.* Atmospheric pressure changes in the Arctic from 1801 to 1920 // The International Journal of Climatology. 2013. V. 33. P. 1730–1760. https://doi.org/10.1002/joc.3546

8. *Belyaev K., Mikhaylov G., Salnikov A., Tuchkova N.* Seasonal and Decadal Variability of Atmosphere Pressure in Arctic, its Statistical and Temporal Analysis // CEUR

Workshop Proceedings, M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen (Aachen, Germany), 2020. V. 2784. P. 51–61.
URL: http://ceur-ws.org/Vol-2784/rpaper05.pdf urn:nbn:de:0074-2784-8.

9. *Belyaev K., Mikhaylov G., Salnikov A., Tuchkova N.* Sezonnaya i mnogoletnyaya izmenchivost' atmosfernogo davleniya v Arktike, statisticheskij i vremennoj analiz // Russian Digital Libraries Journal. 2021. T. 24. № 1. S. 57–73.
https://doi.org/10.26907/1562-5419-2021-24-1-57-73.

10. *Popov S.K.* Influence of sea ice on the harmonic tidal oscillations of sea level and currents in the Barents and White seas // Trudy Gidrometcentra RF, Gidrometeorologicheskie issledovaniya i prognozy. 2018. #4 (370). P. 137–155.

11. *Birman B.A., Berezhnaya T.V., Golubev A.D.* Osnovnye pogodno-klimaticheskie osobennosti, nablyudavshiesya na Severnom polusharii Zemli v 2017 g. Analiticheskij obzor. M: FGBU «Gidrometcentr Rossii».
URL: http://www.meteorf.ru/upload/iblock/dc0/Бирман%20Климат_2017.pdf (access 25.10.2021)

СВЕДЕНИЯ ОБ АВТОРАХ

ТУЧКОВА Наталия Павловна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-мат. наук, окончила ВМиК МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – senior researcher of Dorodnicyn computing center FRC SCS RAS, PhD in physics with a math degree, graduated from CS Faculty of Lomonosov MSU. The expert in the field of algorithmic languages and information technologies.

email: natalia_tuchkova@mail.ru
ORCID: 0000-0001-5357-9640

БЕЛЯЕВ Константин Павлович – ведущий научный сотрудник Института океанологии им. П.П. Ширшова РАН и ФИЦ ИУ, доктор физ.-мат. наук, профессор кафедры теории вероятностей и статистики МГУ им. М.В. Ломоносова. Сфера научных интересов – математическое моделирование и усвоение данных наблюдений, статистический анализ натурных данных.

Konstantin Pavlovich BELYAEV – leading scientist of Shirshov Institute of Oceanology, Russian Academy of Science. Doctor of science, professor of Dept. of Appiied Math and Cybernetics, Lomonosov Moscow State University. Research interests – math. modelling and data assimilation, statistical analysis of natural data.

email: kosbel55@gmail.com

ORCID: 0000-0003-2111-2709

МИХАЙЛОВ Гурий Михайлович – ведущий научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-мат. наук. Сфера научных интересов – архитектура вычислительных систем и сетей, вычислительные и информационные технологии.

Gury Mickailovich MICKAILOV – leading scientist of Dorodnicyn computing center FRC SCS RAS, PhD in physics with a math degree. Research interests include architecture of computing systems and networks, computing and information technology.

email: gmickail@ccas.ru

ORCID 0000-0002-4535-7180

САЛЬНИКОВ Алексей Николаевич – ведущий научный сотрудник кафедры математической физики факультета ВМиК МГУ им. М.В. Ломоносова, кандидат физ.-мат. наук. Сфера научных интересов – параллельное программирование, биоинформатика, суперкомпьютеры.

Alexey Nikolaevich SALNIKOV – leading researcher Dept. of Appiied Math. and Cybernetics, Lomonosov Moscow State University, PhD in physics with a math degree. Research interests include bioinformatics, parallel and supercomputing programming

email: salnikov@angel.cs.msu.ru

ORCID 0000-0001-8669-9905

Материал поступил в редакцию 26 октября 2021 года