

АНАЛИЗ РАСПРЕДЕЛЕНИЯ КЛЮЧЕВЫХ ТЕРМИНОВ В НАУЧНЫХ СТАТЬЯХ

С. А. Власова¹ [0000-0003-1533-5850], **Н. Е. Каленов**² [0000-0001-5269-0988],
И. Н. Соболевская³ [0000-0002-9461-3750]

^{1–3}Межведомственный суперкомпьютерный центр (МСЦ) РАН – филиал ФГУ ФНЦ
Научно-исследовательский институт системных исследований (НИИСИ) РАН

¹vlas.svetlana2013@yandex.ru, ²nekalenov@mail.ru, ³nik_first@mail.ru

Аннотация

Одними из основных компонентов Единого Цифрового Пространства Научных Знаний (ЕЦПНЗ) являются предметные онтологии отдельных тематических подпространств, включающие в себя основные понятия, относящиеся к данному научному направлению. Задача построения предметных онтологий на первом этапе требует формирования массива ключевых терминов в заданной области науки с последующим установлением связей между ними. Аналогичная задача стоит и при формировании энциклопедий в части определения перечня статей (слотов), определяющего их содержание. Одним из источников формирования массива ключевых терминов могут являться метаданные статей, опубликованных в ведущих научных журналах, а именно, авторские ключевые термины («ключевые слова» – в терминологии редакций журналов), сопровождающие в обязательном порядке эти статьи. Чтобы сделать заключение о возможности использования этого подхода к формированию предметных онтологий, необходимо провести предварительный анализ массива авторских ключевых терминов как с точки зрения реального соответствия основным направлениям исследований в данном разделе науки, так и с точки зрения распределения частоты встречаемости тех или иных терминов. В данной статье приведены результаты частотного анализа встречаемости авторских ключевых терминов на русском и английском языках, проведенного на основе программной обработки нескольких тысяч статей из ведущих российских журналов по математике, информатике и физике, отраженных в базе данных MathNet и на сайтах ряда издательств. Проведена

оценка соответствия распределения ключевых терминов (как словосочетаний) и отдельных слов закону Брэдфорда, выявлены ядра ключевых терминов внутри тематических направлений.

***Ключевые слова:** цифровое пространство научных знаний, предметные онтологии, энциклопедические статьи, ключевые термины, метаданные статей, частотный анализ.*

ВВЕДЕНИЕ

Единое Цифровое Пространство Научных Знаний (ЕЦПНЗ) формируется как интегратор многоаспектной цифровой научной информации, достоверность которой подтверждена научным сообществом¹.

Основными целями создания ЕЦПНЗ являются предоставление различным категориям пользователей нужной им информации и обеспечение сохранности оригиналов артефактов, представляющих историческую ценность, путем создания их цифровых копий или моделей [1, 2].

Одним из основных источников контента ЕЦПНЗ является портал «Знание» [3], создаваемый на базе электронной версии Большой Российской энциклопедии с привлечением других научных энциклопедий, а также ресурсов музеев, архивов библиотек, организаций науки, образования и культуры [4].

Одной из проблем при создании научной составляющей Энциклопедии и портала «Знания» является определение перечня статей (слотов), являющихся «точками входа» в информационную систему. Эта задача, по сути, близка задаче формирования предметной онтологии, поскольку перечень статей научной энциклопедии должен тесно коррелировать с понятийной основой данного научного направления.

Таким образом, идея анализа авторских ключевых терминов с целью формирования фундамента предметной онтологии может оказаться полезной не только при проектировании ЕЦПНЗ, но и при развитии портала «Знание» и его основы – Большой российской энциклопедии.

¹ Таким подтверждением могут служить экспертные оценки, многолетнее использование результатов исследований с положительным эффектом, историческая ценность оригинала цифрового объекта и т. п.

Для получения «устойчивых» результатов, отражающих реальное распределение ключевых терминов, необходимо иметь репрезентативную выборку статей по рассматриваемому научному направлению и, соответственно, достаточно большой массив журналов, содержащих в цифровом виде информацию о ключевых терминах.

Для проведения соответствующих расчетов, касающихся русскоязычных терминов наиболее рационально было бы использовать базы данных РИНЦ или RSCI. Однако РИНЦ не дает возможности выгрузки статей в структурированном виде и закрывает возможности анализа HTML-файлов, содержащих метаданные статей, выдаваемых по запросам. За предоставление возможностей анализа массива данных РИНЦ самим пользователям администрация eLibrary требует почасовую, достаточно высокую, плату. База данных RSCI [5], которая представлена на платформе WEB of Science и содержит, как утверждает руководство РАН, наиболее важные российские журналы, в национальной подписке для российских пользователей недоступна, для работы с ней необходимо коммерческое соглашение с компанией Clarivate.

Поэтому для проведения модельных расчетов нами была выбрана отечественная система MathNet [6], которая позволяет анализировать поддерживаемую ею информацию программным образом. В дополнение к этому были проанализированы сайты журналов, не отражаемых в MathNet, на предмет возможности программного выделения ключевых терминов из метаданных опубликованных в них статей.

Для проведения анализа были разработаны структура соответствующей базы данных, специальные программные средства, обеспечивающие выделение и загрузку в базу данных необходимой информации, а также прикладные программы для анализа данных.

1. СТРУКТУРА БАЗЫ ДАННЫХ

Сформированная база данных поддерживается Microsoft SQL Server и содержит 7 видов объектов – «тематика», «журнал», «статья», ключевые термины (КТ) на русском и английском языках, ключевые слова (КС) (отдельные слова, входящие в состав терминов) на русском и английском языках. Объекты имеют следующие атрибуты.

Тематика

- Идентификатор записи
- Наименование тематики журнала
- Рубрика ГРНТИ журнала

Журнал

- Идентификатор записи
- Название журнала на русском языке
- Название журнала на английском языке

Статья

- Идентификатор записи
- Название статьи на русском языке
- Название статьи на английском языке
- Идентификатор журнала
- Год издания
- Выпуск (том, номер)
- Адрес сайта статьи

Ключевой термин на русском языке

- Идентификатор записи
- Ключевой термин на русском языке
- Идентификатор статьи
- Идентификатор журнала

Ключевой термин на английском языке

- Идентификатор записи
- Ключевой термин на английском языке
- Идентификатор статьи
- Идентификатор журнала

Ключевое слово на русском языке

- Идентификатор записи
- Ключевое слово на русском языке
- Идентификатор ключевого термина

Ключевое слово на английском языке

- Идентификатор записи
- Ключевое слово на английском языке
- Идентификатор ключевого термина

Программная оболочка системы, обеспечивающая работу с базой данных, создана на основе технологии Microsoft ASP.NET на платформе Microsoft.NET Framework в среде разработки Microsoft Visual Studio 2019.

Система представлена в свободном доступе по адресу <http://dirsmc.ru/keyterms/> и предоставляет пользователю следующие возможности.

- ✓ Анализ общего частотного распределения КТ и КС.
- ✓ Хронологический анализ распределения КТ и КС по журналам – по выбранным из ядра КТ или КС можно получить их частотное распределение по годам, а также список журналов, в которых они встречаются (с указанием количества по годам).
- ✓ Анализ КТ и КС, относящихся к конкретным журналам, – по выбранным журналам можно получить списки ядра КТ и КС (с указанием частоты их встречаемости).

2. ОТБОР МАТЕРИАЛА ДЛЯ ПРОВЕДЕНИЯ АНАЛИЗА

Для эксперимента были отобраны следующие журналы по математике, физике и информатике.

Математика [6]

Известия Российской академии наук. Математическая серия. Количество статей – 573 за период 2009–2021 гг.;

Математический сборник. Количество статей – 873 за период 2009–2020 гг.;

Дискретная математика. Количество статей – 249 за период 2014–2021 гг.;

Успехи математических наук. Количество статей – 251 за период 2010–

2021 гг.;

Функциональный анализ и его приложения. Количество статей – 861 за период 2003–2021 гг.;

Алгебра и анализ. Количество статей – 578 за период 2010–2021 гг.;

Алгебра и логика. Количество статей – 724 за период 2001–2020 гг.

Физика [6]

Вестник Самарского государственного технического университета. Серия «Физико-математические науки». Количество статей – 806 за период 2008–2021 гг.;

Теоретическая и математическая физика. Количество статей – 2626 за период 2002–2021 гг.

Информатика

Вычислительные методы и программирование. Количество статей – 1152 за период 2000–2021 гг. [7];

Программные продукты и системы. Количество статей – 1815 за период 2008–2021 гг. [8];

Информатика и ее приложения. Количество статей – 573 за период 2007–2021 гг. [6].

Таким образом, по математике было проанализировано более 3700 статей, в среднем, за 12-летний период; по физике – около 3500 статей, в среднем, за 16-летний период; по информатике – более 2500 статей за 13-летний период.

3. РЕЗУЛЬТАТЫ ПРОВЕДЕННОГО АНАЛИЗА

По загруженным данным для различных тематических направлений были отдельно проанализированы русскоязычные и англоязычные ключевые термины, а также входящие в них ключевые слова. Результаты анализа ключевых терминов приведены в таблицах 1, 3, 5, результаты анализа ключевых слов – в таблицах 2, 4, 6.

Таблица 1. Математика. Ключевые термины

№	Наименование	Русские	Английские
1	Всего Ключевых терминов	15949	14924
2	Различных КТ	10135	9690
3	20% из них (наиболее повторяющихся)	2027	1938
4	Всего КТ для выбранных 20% (с повторениями)	7301	6747
5	Процент повторяющихся КС из 20% от всех	45,78%	45,2%

Таблица 2. Математика. Ключевые слова

№	Наименование	Русские	Английские
1	Всего Ключевых слов	36286	34849
2	Различных КС	7358	4509
3	20% из них (наиболее повторяющихся)	1471	901
4	Всего КС для выбранных 20%	26786	27668
5	Процент повторяющихся КС из 20% от всех	73,8%	79,4%

Таблица 3. Физика. Ключевые термины

№	Наименование	Русские	Английские
1	Всего Ключевых терминов	14103	14038
2	Различных КТ	8172	8096
3	20% из них (наиболее повторяющихся)	1634	1619
4	Всего КТ для выбранных 20%	6799	6847
5	Процент повторяющихся КТ из 20% от всех	48,2%	48,8%

Таблица 4. Физика. Ключевые слова

№	Наименование	Русские	Английские
1	Всего Ключевых слов	31211	31694
2	Различных КС	6642	4216
3	20% из них (наиболее повторяющихся)	1328	843
4	Всего КС для выбранных 20%	22657	25019
5	Процент повторяющихся КС из 20% от всех	72,6%	79%

Таблица 5. Информатика. Ключевые термины

№	Наименование	Русские	Английские
1	Всего Ключевых терминов	19050	16689
2	Различных КТ	11341	9688
3	20% из них (наиболее повторяющихся)	2268	1937
4	Всего КТ для выбранных 20%	9672	8432
5	Процент повторяющихся КТ из 20% от всех	50,77%	50,52%

Таблица 6. Информатика. Ключевые слова

№	Наименование	Русские	Английские
1	Всего Ключевых слов	40913	36339
2	Различных КС	9578	5425
3	20% из них (наиболее повторяющихся)	1914	1085
4	Всего КС для выбранных 20%	29774	28406
5	Процент повторяющихся КС из 20% от всех	72,77%	78,17%

Графики распределения частоты встречаемости русских ключевых терминов и русских ключевых слов в математических журналах представлены на рис. 1 и 2 соответственно.



Рис. 1. Распределение русских КТ в математических журналах

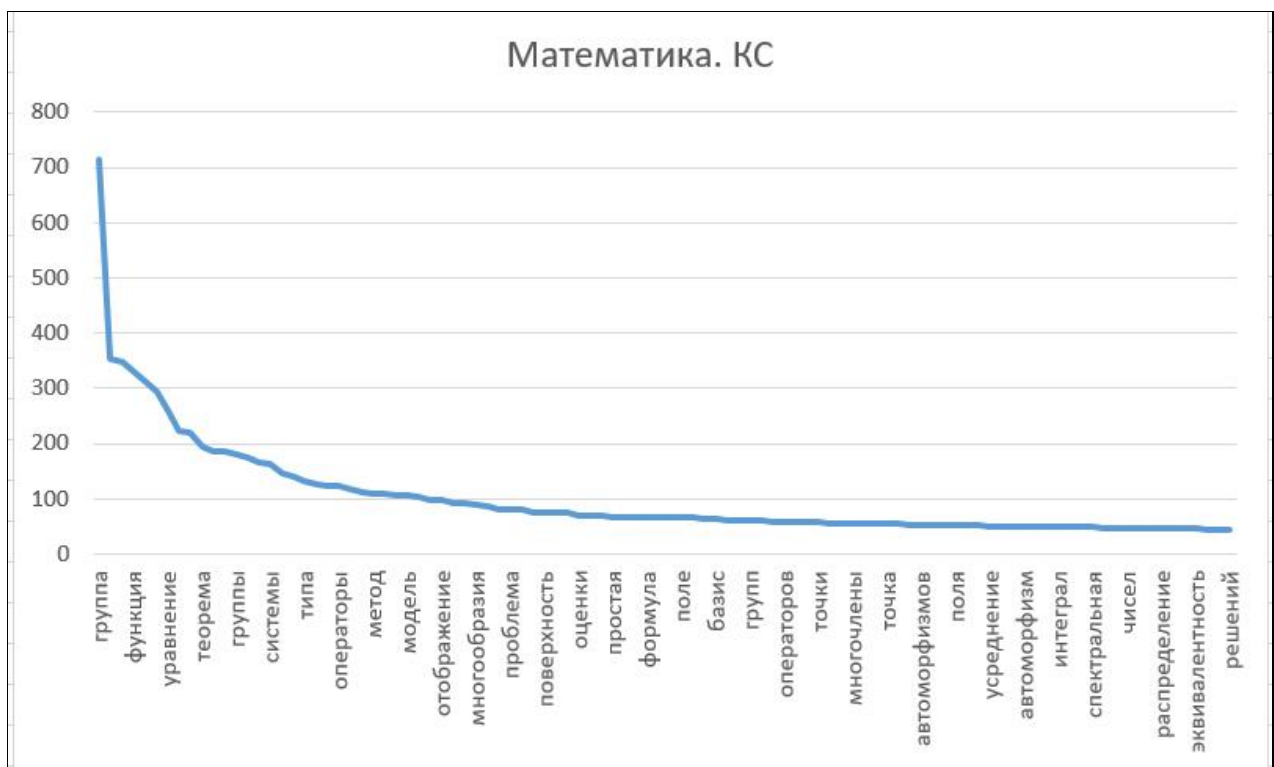


Рис. 2. Распределение русских КС в математических журналах

Если мы проанализируем список ключевых терминов, выделенных в русскоязычных журналах по информатике (его фрагмент приведен в таблице 7), то увидим, что в нем встречаются термин «параллельные алгоритмы» 29 раз, «параллельный алгоритм» 22 раза. Аналогично, в списке англоязычных терминов по информатике КТ «parallel algorithms» встречается 23 раза, а КТ «parallel algorithm» – 21 раз.

Если исключить из рассмотрения такие общие понятия, как «алгоритм», «вычисления», «компьютеры» и т. п., то, анализируя ядро перечня КТ по информатике, можно сделать вывод, что к наиболее актуальным проблемам относятся направления, связанные с:

– моделированием (в списке из 200 наиболее часто используемых КТ термины, связанные с этой проблемой, встречаются 436 раз (математическое моделирование – 104 раза, моделирование – 96 раз, численное моделирование – 46 раз, модель – 45 раз, имитационное моделирование – 44 раза, математическая модель – 32 раза, компьютерное моделирование – 22 раза, модель данных – 10 раз, суперкомпьютерное моделирование – 10 раз, информационная модель – 9 раз, имитационная модель – 9 раз, аналитическое моделирование – 9 раз);

– параллельными вычислениями (в списке из 200 наиболее часто используемых КТ термины, связанные с этой проблемой, встречаются 221 раз: параллельные вычисления – 132 раза, параллельные алгоритмы – 51 раз, параллельное программирование – 38 раз);

– оптимизацией в списке из 200 наиболее часто используемых КТ термины, связанные с этой проблемой, встречаются 68 раз (оптимизация – 50 раз, глобальная оптимизация – 10 раз, многокритериальная оптимизация – 8 раз).

Таблица 7. Фрагмент рейтингового списка КТ по информатике

КТ информатика	Частота встречаемости
Параллельные вычисления	132
Математическое моделирование	104
Численные методы	97
Моделирование	96
Оптимизация	50
Алгоритм	47

Высокопроизводительные вычисления	46
Программный комплекс	46
Численное моделирование	46
Модель	45
Нейронные сети	44
Имитационное моделирование	44
Прогнозирование	43
Информационная система	41
Параллельное программирование	38
Информационная безопасность	37
Управление	36
Обратные задачи	34
Программное обеспечение	33
Принятие решений	32
Математическая модель	32
Краевые задачи	31
База знаний	31
Система массового обслуживания	31
Автоматизация	30
Машинное обучение	30
Метод конечных элементов	30
Искусственный интеллект	30
Параллельные алгоритмы	29
Мониторинг	27
Визуализация	27
Генетический алгоритм	27
Устойчивость	27
Численный анализ	26
Информационные технологии	26
Кластеризация	26
Надежность	26
Суперкомпьютер	26

САПР	25
Нечеткая логика	25
Верификация	25
Распределенные вычисления	24
Экспертная система	24
Обыкновенные дифференциальные уравнения	23
МРІ ²	23
Компьютерное моделирование	22
Итерационные методы	22
Параллельный алгоритм	22
Сходимость	22
Эффективность	22

ЗАКЛЮЧЕНИЕ

Результаты анализа показывают, что распределение ключевых терминов в том виде, как они представлены авторами, достаточно далеко от распределения Брэдфорда, в то время как распределение ключевых слов вполне ему соответствует. Более подробный анализ рейтингового списка ключевых терминов объясняет причину этого, которая в значительной степени обусловлена разной последовательностью одних и тех же слов, входящих в состав ключевого термина.

Очевидно, что для более точной картины при обработке КТ необходимо применять методы лингвистического анализа, что на данном этапе в нашу задачу не входило. Однако сформированная база данных и простой «ручной» анализ полученного «ядра» КТ позволяют сформировать список наиболее значимых терминов для последующего их включения в Единое цифровое пространство научных знаний [9].

Разработанные методика и программная оболочка позволяют проводить анализ динамики развития той или иной области науки, а также могут служить инструментом для развития и корректировки политематических и специальных

² Message Passing Interface

научных энциклопедий. Сравнение приведенного в табл. 7 списка из 50-ти наиболее употребительных авторских ключевых терминов с электронной версией Большой российской энциклопедии показало, что в ней отсутствуют статьи, посвященные таким терминам, как «высокопроизводительные вычисления», «имитационное моделирование», «обратные задачи», «генетический алгоритм», «MPI» и др. В БРЭ отсутствуют лидирующие в рейтинге авторских ключевых терминов «параллельный алгоритм» и «параллельные вычисления», но присутствует термин «параллельное программирование», который не используют авторы статей. Вместо распространенного термина «машинное обучение» (26-е место в рейтинге) в БРЭ приведен термин «программированное обучение».

В качестве следующего шага исследований в данном направлении планируется использовать сформированную базу данных для анализа динамики изменения состава «ядра» ключевых терминов, что представляет интерес для задач наукометрии, характеризуя, в определенной степени, динамику развития отдельных областей рассматриваемых наук.

Работа выполнена в МСЦ РАН в рамках государственного задания по теме FNEF-2023-0014.

СПИСОК ЛИТЕРАТУРЫ

1. Антопольский А.Б., Каленов Н.Е., Серебряков В.А., Сотников А.Н. О едином цифровом пространстве научных знаний // Вестник Российской академии наук. 2019. Т. 89 (7). С. 728–735.

URL: <https://doi.org/10.31857/S0869-5873897728-735>.

2. Савин Г.И. Единое цифровое пространство научных знаний: цели и задачи // Информационные ресурсы России. 2020. № 5. С. 3–5.

URL: <https://doi.org/10.51218/0204-3653-2020-5-3-5>.

3. Большая российская энциклопедия. URL: <https://bigenc.ru/> (дата обращения: 22.12.2022).

4. Kalenov N., Savin G., Sotnikov A. Fundamentals of Common Digital Space of Scientific Knowledge Building // CEUR Workshop Proceedings (CEUR-WS.org). 2021. Vol. 2990. P. 93–99. URL: <https://doi.org/10.51218/1613-0073-2990-93-99>

5. Михайлов О.В. Новая платформа журналов RSCI в WEB of Science Вестник

Российской академии наук. 2017. Т. 87. № 2. С. 177–180.

6. Общероссийский портал Math-Net.ru. URL: <http://www.mathnet.ru/> (дата обращения: 22.12.2022).

7. Вычислительные методы и программирование.
URL: <https://num-meth.ru/index.php/journal/issue/archive> (дата обращения: 22.12.2022).

8. Программные продукты и системы.
URL: <http://www.swsys.ru/index.php?page=10&lang=> (дата обращения: 22.12.2022)

9. *Власова С.А., Каленов Н.Е., Сотников А.Н.* Web-ориентированная система формирования контента единого цифрового пространства научных знаний // Программные продукты и системы. 2020. № 3. С. 365–374.

URL: <https://doi.org/10.15827/0236-235X.131.365-374>.

ANALYSIS OF THE DISTRIBUTION OF KEY TERMS IN SCIENTIFIC ARTICLES

S. A. Vlasova¹ [0000-0003-1533-5850], **N. E. Kalenov**² [0000-0001-5269-0988],

I. N. Sobolevskaya³ [0000-0002-9461-3750]

¹⁻³Joint Supercomputer Center of the Russian Academy of Sciences – JSC

¹vlas.svetlana2013@yandex.ru, ²nekalenov@mail.ru, ³nik_first@mail.ru

Abstract

One of the Common Digital Space of Scientific Knowledge (CDSSK) main components are the subject ontologies of individual thematic subspaces, which include the basic concepts related to this scientific area. The constructing subject ontologies task at the initial phase requires the array of key terms formation in a given scientific area with the subsequent establishment of links between them. A similar task is in the encyclopedias formation in terms of the articles (slots) list generating that determines their content. One of the sources for the formation of the key terms array can be the metadata of articles published in the leading scientific journals. Namely, the author's key terms ("keywords" in the terminology of the journals editors) quoted by the article. To make a conclusion about the possibility of using this approach to the subject ontologies formation, it is necessary to conduct the author's key terms array preanalysis,

both in terms of real correspondence to the main areas of research in this science branch and in terms of the distribution of the certain terms occurrence frequency. This article presents the results of the occurrence frequency analysis of the author's key terms in Russian and English, carried out on the software processing basis of several thousand articles from leading Russian journals in mathematics, computer science and physics, reflected in the MathNet database. An assessment was made of the distribution of key terms correspondence (as phrases) and individual words to the Bradford's law, and the key terms cores within the thematic direction were identified.

Keywords: *digital space of scientific knowledge, subject ontologies, encyclopedia articles, key terms, article metadata, frequency analysis.*

REFERENCES

1. Antopol'skij A.B., Kalenov N.E., Serebryakov V.A., Sotnikov A.N. O edinom cifrovom prostranstve nauchnyh znaniy // Vestnik Rossijskoj akademii nauk, 2019. V. 89 (7). S. 728–735. URL: <https://doi.org/10.31857/S0869-5873897728-735>.
2. Savin G.I. Edinoe cifrovoe prostranstvo nauchnyh znaniy: celi i zadachi // Informacionnye resursy Rossii. 2020. № 5. S. 3–5. URL: <https://doi.org/10.51218/0204-3653-2020-5-3-5>.
3. Bol'shaya rossijskaya enciklopediya. URL: <https://bigenc.ru/> (accessed 22 December 2022).
4. Kalenov N., Savin G., Sotnikov A. Fundamentals of Common Digital Space of Scientific Knowledge Building // CEUR Workshop Proceedings (CEUR-WS.org). 2021. V. 2990. P. 93–99. <https://doi.org/10.51218/1613-0073-2990-93-99>.
5. Mikhailov O.V. Novaya platforma zhurnalov RSCI on WEB of Science // Vestnik Rossijskoj akademii nauk Вестник. 2017. V. 87. № 2. S. 177–180.
6. Obshcherossijskij portal Math-Net.ru. URL: <http://www.mathnet.ru/> (accessed 22 December 2022).
7. Vychislitel'nye metody i programmirovaniye. URL: <https://num-meth.ru/index.php/journal/issue/archive> (accessed 22 December 2022).
8. Programmnye produkty i sistemy. URL: <http://www.swsys.ru/index.php?page=10&lang=> (accessed 22 December 2022).

9. *Vlasova S.A., Kalenov N.E., Sotnikov A.N.* Web-orientirovannaya sistema formirovaniya kontenta edinogo cifrovogo prostranstva nauchnyh znaniy // Programmnye produkty i sistemy. 2020. № 3. S. 365–374.

URL: <https://doi.org/10.15827/0236-235X.131.365-374>.

СВЕДЕНИЯ ОБ АВТОРАХ



ВЛАСОВА Светлана Александровна – ведущий научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», кандидат технических наук.

Svetlana Aleksandrovna VLASOVA – Leading Researcher of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution «Scientific Research Institute for System Analysis of the Russian Academy of Sciences», Candidate of Technical Sciences

email: vlas.svetlana2013@yandex.ru;

ORCID: 0000-0003-1533-5850.



КАЛЕНОВ Николай Евгеньевич – главный научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», доктор технических наук, профессор.

Nikolay Evgenievich KALENOV – Chief Researcher of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution «Scientific Research Institute for System Analysis of the Russian Academy of Sciences», Doctor of Technical Sciences, Professor.

email: nekalenov@mail.ru;

ORCID: 0000-0001-5269-0988.

Соболевская Ирина Николаевна – старший научный



сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», кандидат физико-математических наук.

Sobolevskaya Irina Nikolaevna – higher senior officer of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution «Scientific Research Institute for System Analysis of the Russian Academy of Sciences», Candidate of Physics and Math Sciences.

email: nik_first@mail.ru;

ORCID: 0000-0002-9461-3750

Материал поступил в редакцию 3 января 2023 года