



МЕТОДОЛОГИЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

Институт международных отношений МИФИ

Артамонов Алексей Анатольевич

к.т.н.,

зав. кафедрой «Анализ конкурентных систем» НИЯУ МИФИ

Становление ИМО МИФИ

1999 - 2000



ИМО МИФИ был организован в марте 1999 года по инициативе академика **Евгения Максимовича Примакова** при официальной поддержке Минатома России, Министерства иностранных дел, Министерства образования России и ряда других министерств и ведомств.

Основной задачей Института является поддержка интересов экспортно-ориентированных отраслей экономики Российской Федерации. Институт готовит специалистов для выполнения аналитических и управленческих функций с целью повышения конкурентоспособности и продвижения российской продукции на международных рынках.

Направление подготовки «Международные отношения» профиль «Международное научно-технологическое и промышленное сотрудничество».

Созданы две кафедры «Международных отношений» и «Специальной лингвистической подготовки»

2001 - 2005



В 2005 году создается вторая выпускающая кафедра - кафедра «Анализ конкурентных систем» под руководством **Бориса Николаевича Оныкия**.

Б.Н. Оныкий - доктор технических наук, профессор, ректор МИФИ (1997-2007), действительный член Международной академии информатизации.

Одним из направлений деятельности кафедры становится разработка прикладных методов и средств обработки больших массивов информации для аналитических целей и поддержки принятия решений в области поисковых исследований и технологического развития наукоемких производств.

2006 - 2019

Происходит становление и развитие **научной школы по мультязыковым агентным информационно-аналитическим системам и технологиям**.

Ведется подготовка по профилю «Международные научно-технологические и торгово-промышленные отношения». **Реализуются все уровни подготовки специалистов:** бакалавриат, магистратура, аспирантура.

Осуществляется «**пакетная**» подготовка кадров для аналитических и управленческих структур.

Разработка прикладных систем:

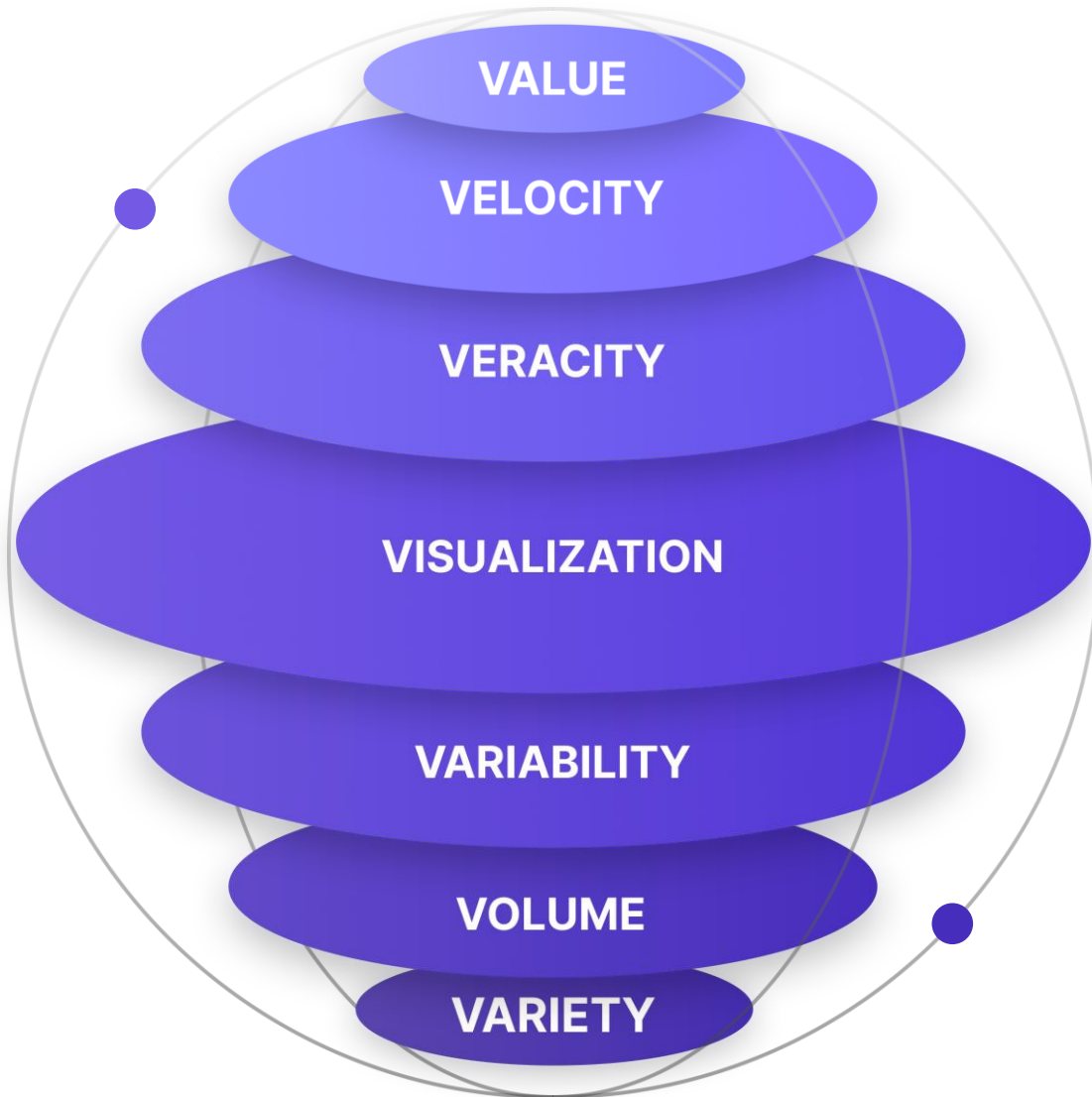
- Мультиагентные информационно-аналитические системы по технологическим направлениям
- Графовые модели анализа данных
- Автоматизированные системы по социально-экономическому состоянию и развитию
- Цифровая платформа совместных исследований

Кафедрами **ведется научная и платная образовательная деятельности:**

- Регионоведческое и аналитическое сопровождение развития бизнеса корпораций и компаний в ключевых секторах экономики
- Факультативные занятия по изучению иностранных языков

РАБОТА С BIG DATA

V7



Дополнительные факторы

Избыточность – присутствуют дублирующие друг друга сведения

Связность – с одной стороны, усложняет аналитические вычисления, с другой – появляется возможности для глубинного анализа сетевых эффектов

Контекстуальность – адекватный учёт контекста существенно повышает точность и ценность получаемых результатов

БАЗОВОЕ ОПИСАНИЕ ЦИФРОВОГО ОБЪЕКТА

$$\text{Obj} = \langle \text{ID}, S, D \rangle,$$

где ID – уникальный идентификатор объекта,

S – совокупность статических характеристик,

D – совокупность динамических (изменяющихся во времени, либо под воздействием) характеристик

Статические характеристики — это характеристики, значения которых либо не меняются в течение всего периода существования объекта, либо меняются крайне редко.

- 1) ФИО автора научной публикации;
- 2) Уникальный идентификатор (например, DOI статьи).

$$S = \{(s_1, \tau_1), (s_2, \tau_2) \dots, (s_k, \tau_k)\},$$

где s_i – наименование i -ой статической характеристики, (например, автор, название организации),

τ_i – тип данной характеристики (строка, число и др.), $i = 1, 2, \dots, k$.

Динамические характеристики — это характеристики, значения которых могут изменяться во времени или под воздействием внешних факторов.

- 1) Количество просмотров сообщения;
- 2) Количество цитирований.

$$D = \{(d_1(t_1), t_1, \tau_1), (d_2(t_2), t_2, \tau_2), \dots, (d_l(t_l), t_l, \tau_l)\},$$

где $d_j(t_j)$ – значение j -ой динамической характеристики в момент времени t_j ,

t_j – значение времени,

τ_j – тип данной характеристики (строка, число),

$j = 1, 2, \dots, l$.

АНАЛИТИЧЕСКАЯ МОДЕЛЬ ЦИФРОВОГО ОБЪЕКТА

$$AObj = \langle ID, S, D, F, Rel \rangle$$

F – совокупность вычисляемых характеристик,

Rel – множество связей (отношений) данного объекта с другими объектами.

Вычисляемые характеристики — это такие значения, которые не собираются непосредственно из информационных ресурсов, а рассчитываются исходя из значений статических и динамических характеристик.

$$F = \{(f_1, \varphi_{f_1}), (f_2, \varphi_{f_2}), \dots, (f_p, \varphi_{f_p})\},$$

где f_j – наименование вычисляемого признака,

φ_{f_j} – формула или алгоритм, определяющий, как вычислять f_j на основе статических (S) и/или динамических полей (D),
 $j = 1, 2, \dots, p$.

Множество связей (отношений) — это такие значения, которые не собираются непосредственно из информационных ресурсов, а рассчитываются исходя из значений статических и динамических характеристик.

$$Rel = \{(r_1, Obj_{j_1}), (r_2, Obj_{j_2}), \dots, (r_m, Obj_{j_m})\},$$

где r_i — тип связи (например, «Автор», «Владелец», «Состоит в группе», «Цитирует»),

Obj_{j_i} — конкретный объект, с которым установлена данная связь,
 $i = 1, 2, \dots, m$.

МОДЕЛЬ КОМПЛЕКСНОГО АНАЛИТИЧЕСКОГО ОБЪЕКТА

$$CAObj = \langle ID, S, D, F, Rel_{complex} \rangle,$$

$$\text{где } Rel_{complex} = \{(r_1, \{AObj_{j_1}, \dots, AObj_{j_n}\}), \dots, (r_m, \{AObj_{j_1}, \dots, AObj_{j_k}\})\},$$

r_i — ТИП СВЯЗИ,

$i = 1, \dots, m.$

Агрегирующие и комбинированные функции для расчета сводных характеристик, таких как рейтинги и иные интегральные метрики, востребованные в задачах интеллектуального анализа научно-технических и социальных данных.

Пусть $LinkedSet(r_i, CAObj) = (r_i, \{AObj_{j_1}, \dots, AObj_{j_n}\})$ — все объекты, связанные с $CAObj$ отношением r_i . В общем случае агрегирующая функция $A(x)$ для характеристики x может быть записана как:

$$A(x, CAObj, r_i) = agg(x(AObj_{j_1}), \dots, x(AObj_{j_n})),$$

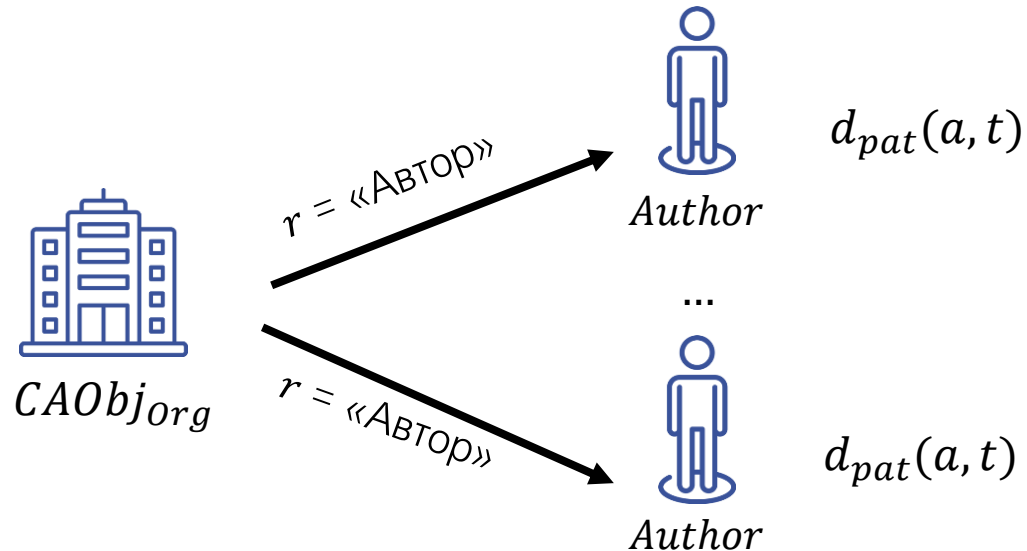
где $x(AObj_{j_i})$ — значение интересующего признака (статического, динамического или вычислимого) у объекта $AObj_{j_i}$,

agg — операция агрегирования (суммирование, среднее, минимум/максимум, конкатенация списков, более сложные операции),

$i = 1, \dots, n.$

ОЦЕНКА ПАТЕНТНОЙ АКТИВНОСТИ ОРГАНИЗАЦИИ

Организация ($CAObj_{org}$) связана со множеством авторов (физических лиц), а каждый автор a имеет динамическое поле $d_{pat}(a, t)$, отражающее число зарегистрированных патентов на момент времени t



$$Authors(CAObj_{org}) = LinkedSet(r, CAObj_{org}).$$

Суммарная патентная активность организации в момент t :

$$PatentActivity(CAObj_{org}, t) = \sum_{a \in Authors(CAObj_{org})} d_{pat}(a, t)$$

ПРЕИМУЩЕСТВА ИСПОЛЬЗОВАНИЯ МОДЕЛИ КОМПЛЕКСНОГО АНАЛИТИЧЕСКОГО ОБЪЕКТА

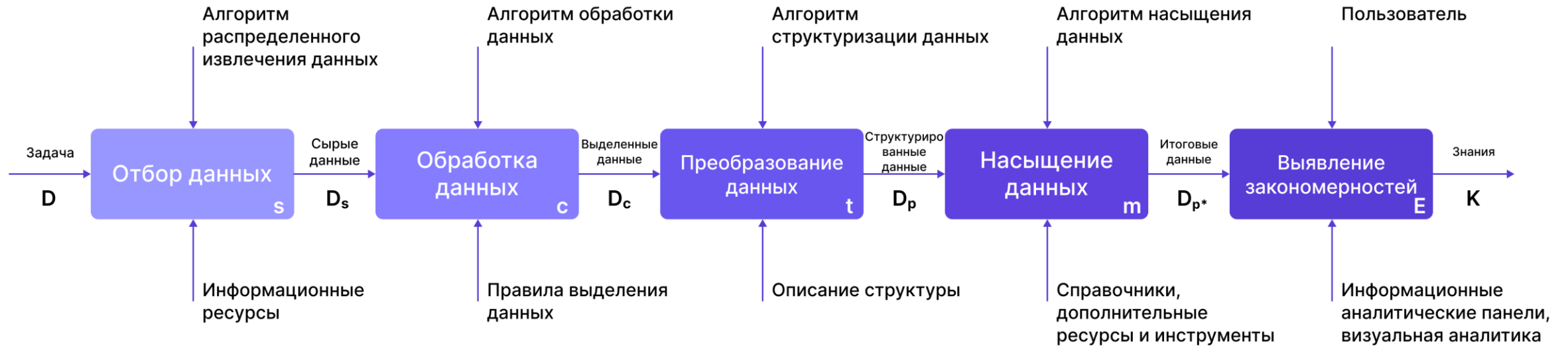
1. Универсальность: формальная структура CAObj не привязана жёстко к одному типу данных, поддерживает как научно-технические (статьи, патенты), так и социальные (профили, сообщества) сущности
2. Модульность в анализе: сложные эксперименты (например, сетевой анализ соавторства) легко запускать, опираясь на готовые связи Rel, а временные аспекты — на D
3. Упрощённая поддержка: при добавлении новых типов объектов (например, технологические отчёты или финансовая отчётность) нужно лишь расширить модель (добавить новые виды характеристик, связей, формул) без полной перестройки анализа
4. Интеграция с разными технологиями: модель может быть реализована в реляционных, NoSQL или графовых системах, причём механизмы динамического расчёта позволяют встраиваться в архитектуры потоковой аналитики, гибридных хранилищ и больших данных (Big Data)

ПРЕИМУЩЕСТВА ИСПОЛЬЗОВАНИЯ МОДЕЛИ КОМПЛЕКСНОГО АНАЛИТИЧЕСКОГО ОБЪЕКТА

5. Гибкость и адаптивность: модель может «подстраиваться» под новые требования без тотальной перестройки существующей системы
6. Сохранение преемственности: учет характеристики времени даёт возможность корректно работать с объектами, созданными в разное время и по разным правилам
7. Удобство для анализа: аналитические алгоритмы могут получать более полную или современную структуру данных, не обрывая доступа к историческим записям
8. Сокращение затрат на миграцию: продуманный механизм эволюции уменьшает необходимость больших миграционных проектов, что критично в крупных БД

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Интеллектуальный анализ данных – это широкий *комплекс математических моделей, алгоритмов и программных средств*, предназначенных для автоматического выявления скрытых закономерностей и генерации новых знаний на основе эмпирических данных.



$s: D \rightarrow D_s, D_s \subseteq D$, где D – исходные данные, s обозначает операцию отбора данных, в которой из исходного набора данных D выбирается подмножество D_s .

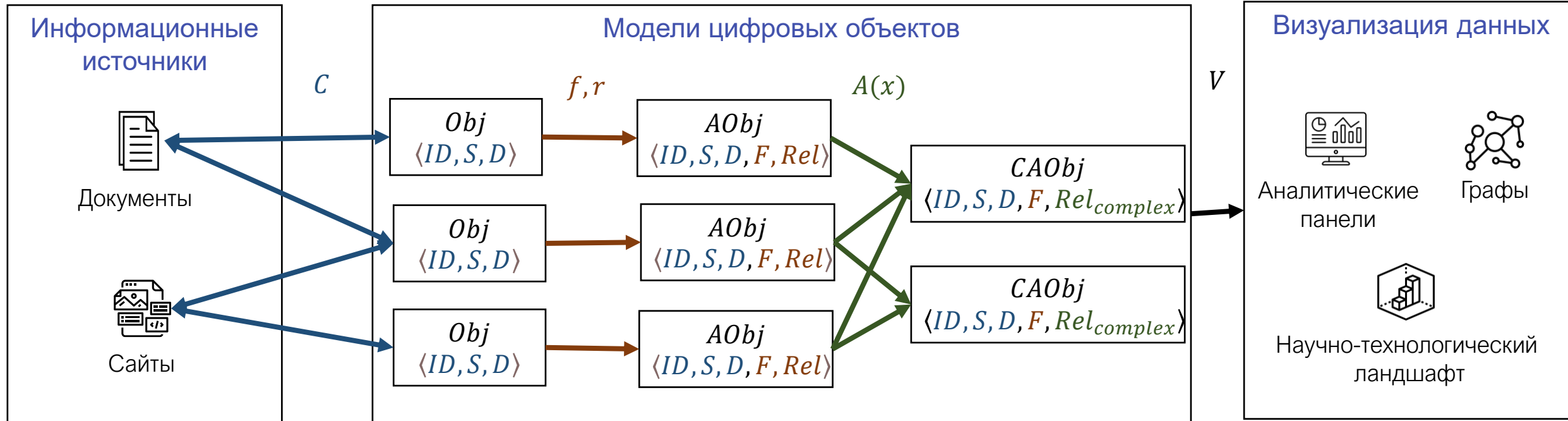
$c: D_s \rightarrow D_c, D_c \subseteq D_s$, c определяет операции обработки выбранных данных D_s к виду D_c для дальнейшего шага процесса.

$t: D_c \rightarrow D_p, D_p = \{p_1, \dots, p_n\}$, где t – операция преобразования и структуризации данных, в результате которых получается набор D_p из извлеченных базовых параметров $p_i, i = 1, \dots, n$, где n – количество базовых параметров.

$m: D_p \rightarrow D_{p^*}, D_{p^*} = \{p_1, \dots, p_i, \dots, p_n, p^*_1, \dots, p^*_j, \dots, p^*_k\}$, где m – правила для насыщения набора параметров новыми параметрами $p^*_j, j = 1, \dots, k$, где k – количество новых параметров. Заметим, что $D_p \subseteq D_{p^*}$.

$$D \xrightarrow{s} D_s \xrightarrow{c} D_c \xrightarrow{t} D_p \xrightarrow{m} D_{p^*} \xrightarrow{E} K.$$

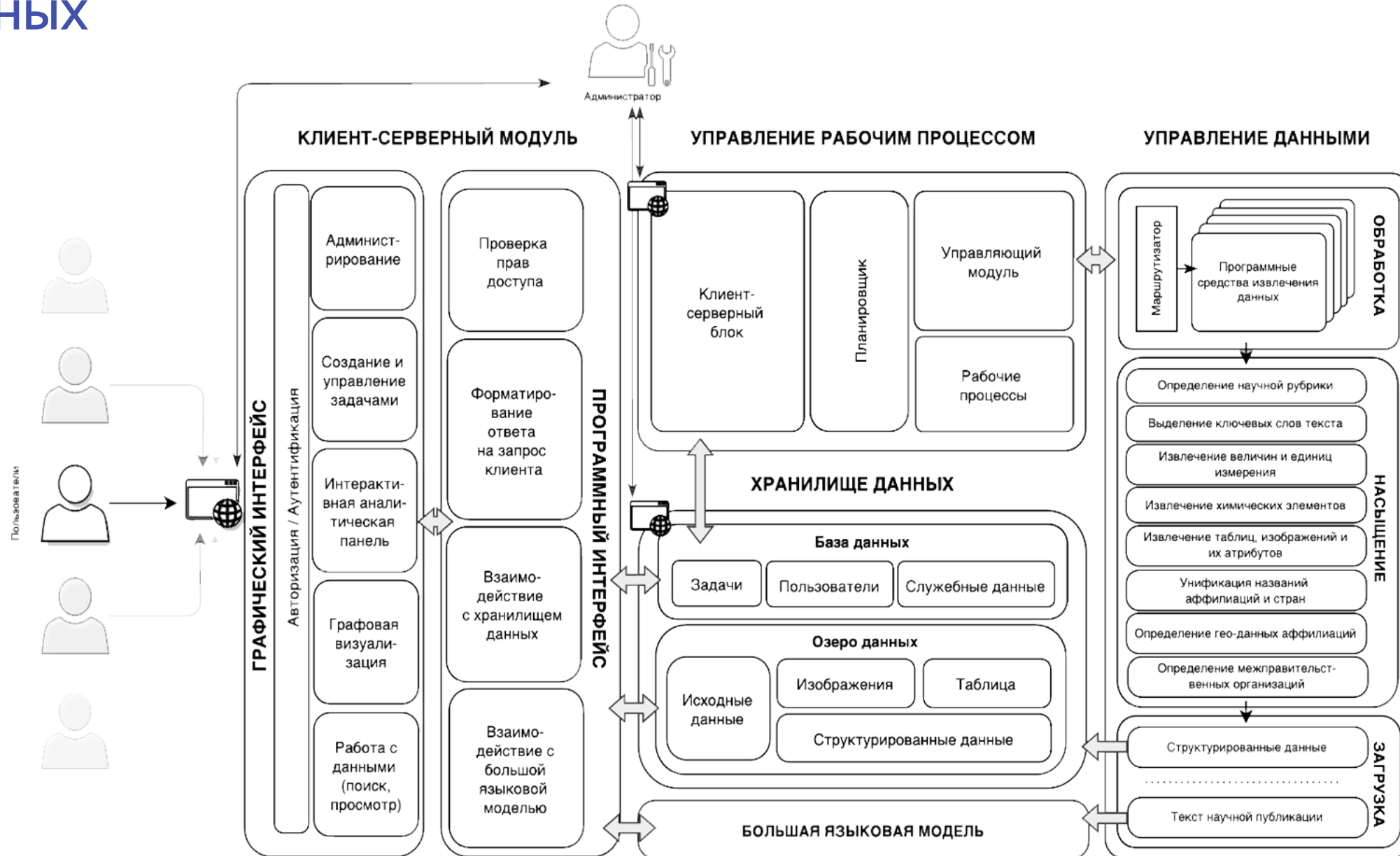
МЕТОДОЛОГИЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ЦИФРОВЫХ ОБЪЕКТОВ В НАУЧНО-ТЕХНИЧЕСКИХ ЗАДАЧАХ



ID – уникальный идентификатор объекта,
 S – совокупность статических характеристик,
 D – совокупность динамических,
 F – совокупность вычисляемых характеристик,
 Rel – множество связей (отношений) данного объекта с другими объектами.

C – методы преобразования исходных данных
 f – функции/алгоритмы расчета характеристик
 r – типы связей между объектами
 $A(x)$ – агрегирующие и комбинированные функции
 V – методы свертки (визуализации) данных

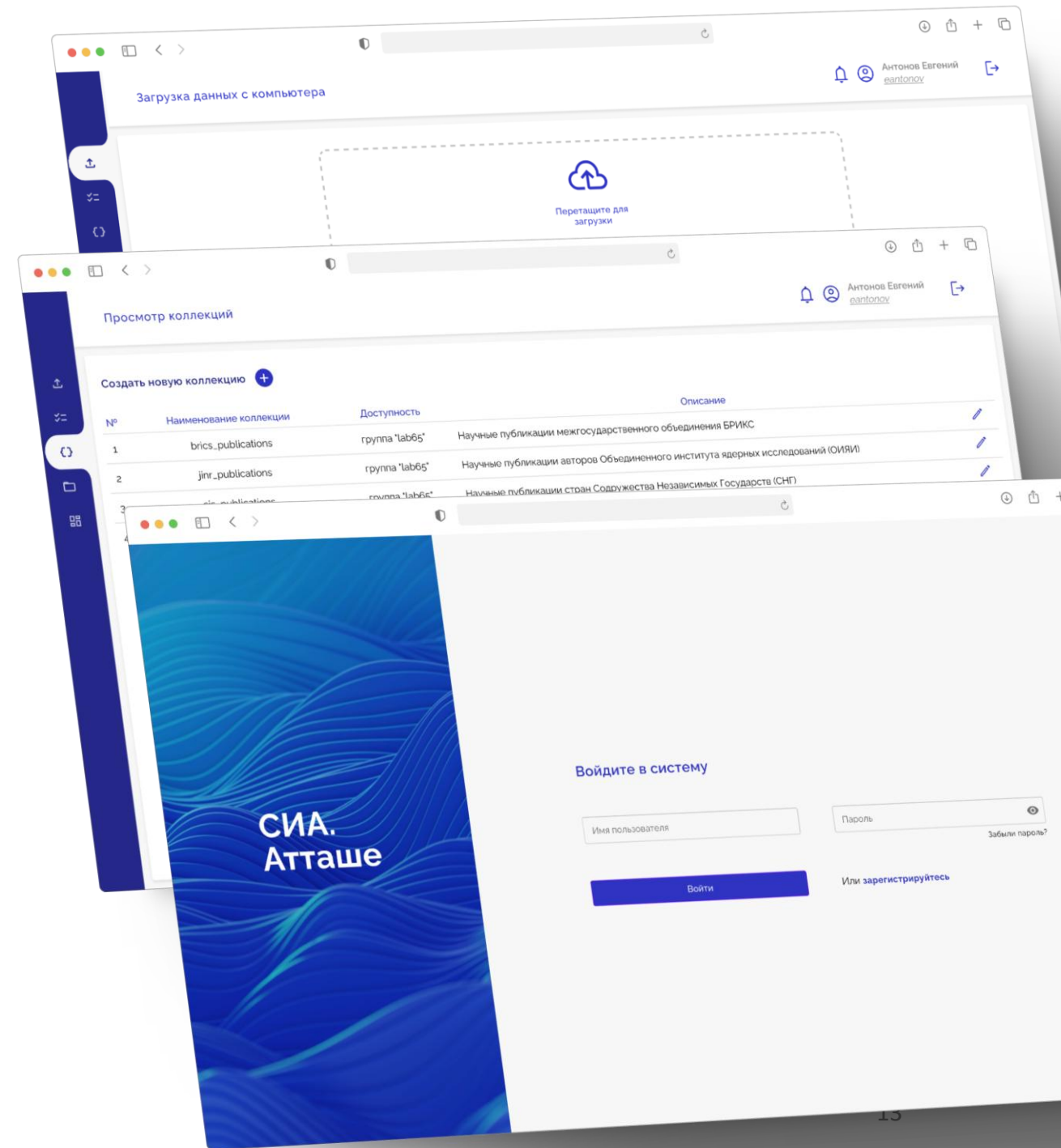
КОМПОНЕНТНАЯ АРХИТЕКТУРА СИСТЕМЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ



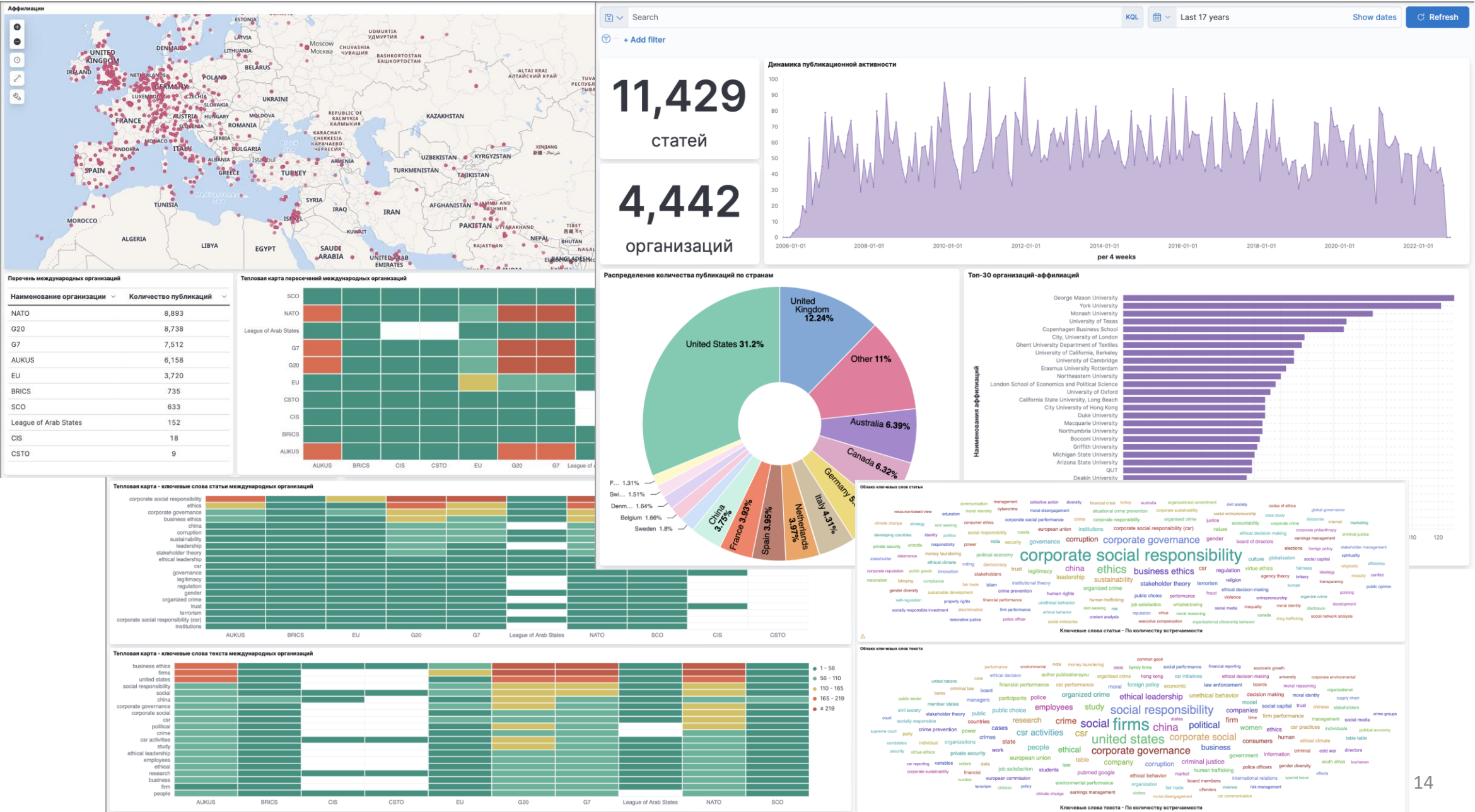
МУЛЬТИАГЕНТНАЯ СИСТЕМА ОБРАБОТКИ И ПРЕДСТАВЛЕНИЯ НЕСТРУКТУРИРОВАННЫХ МАССИВОВ ДАННЫХ

Система позволяет:

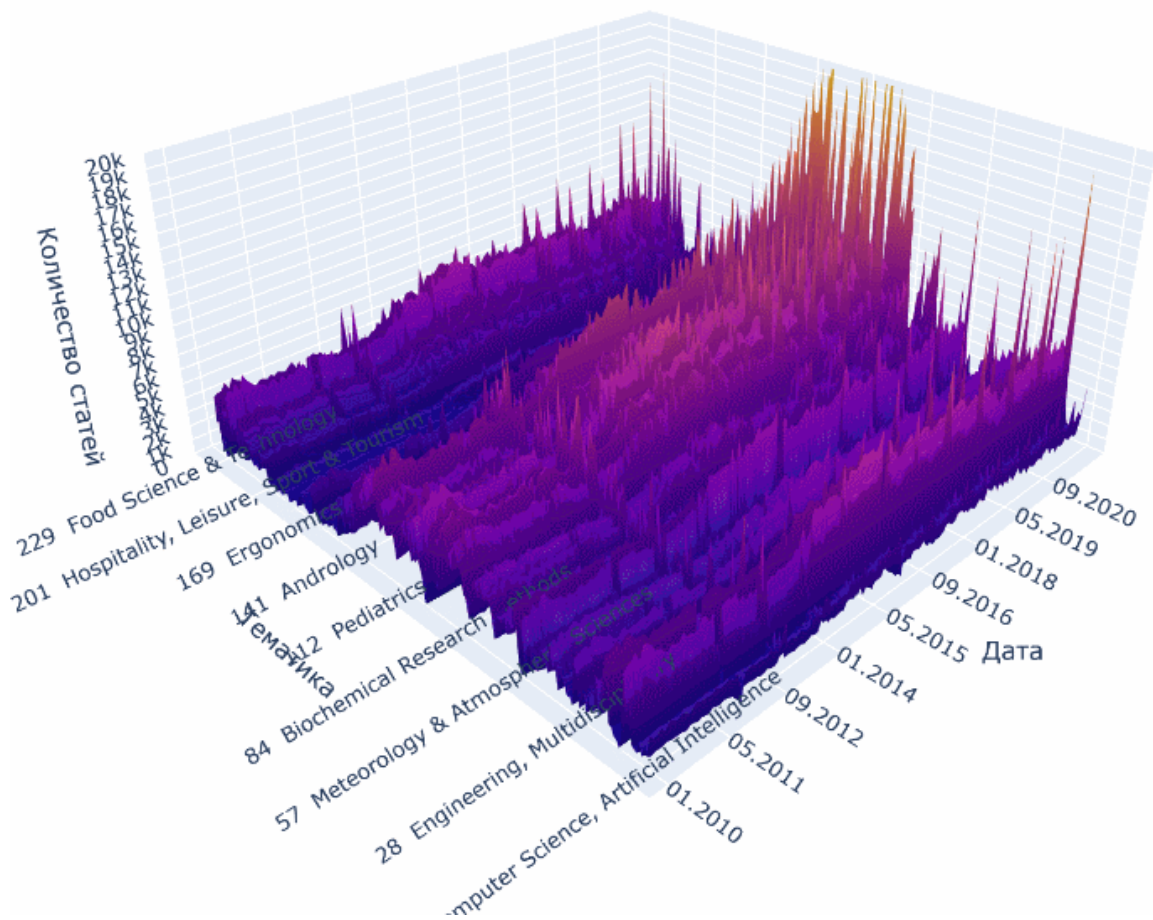
- выделять в автоматизированном режиме из файлов сущности (авторы, ключевые слова, технологии),
- проводить систематизацию набора документов по выделенным сущностям,
- анализировать и систематизировать документ или группу документов,
- формировать статистический срез по хранилищу для определения встречаемости сущностей,
- создавать сложные запросы и проводить поиск,
- выводить статистику и результаты поиска на интерактивные дашборды.



ИНТЕРАКТИВНАЯ АНАЛИТИЧЕСКАЯ ПАНЕЛЬ ПО ПУБЛИКАЦИЯМ В ОБЛАСТИ ФИНАНСОВОЙ БЕЗОПАСНОСТИ



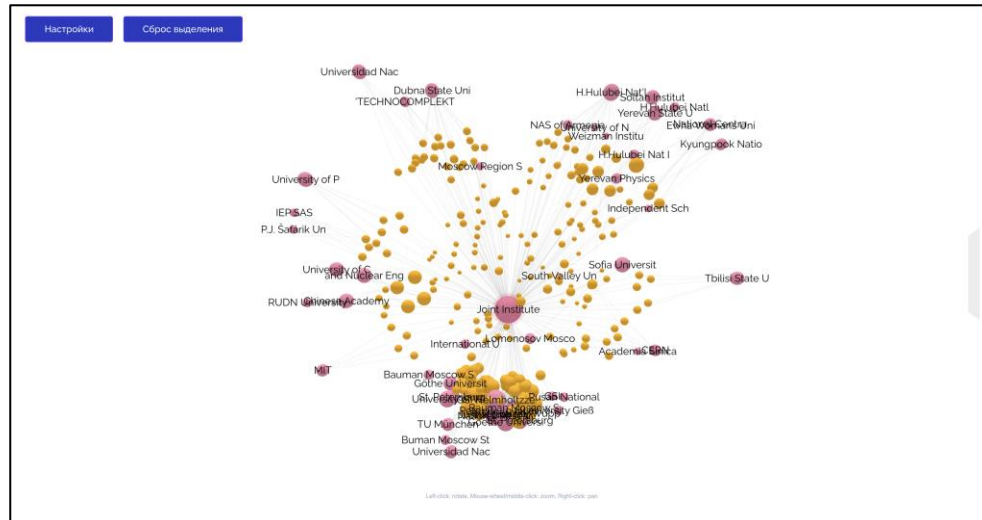
НАУЧНО-ТЕХНИЧЕСКИЙ ЛАНДШАФТ СТРАНЫ (ЯПОНИЯ)



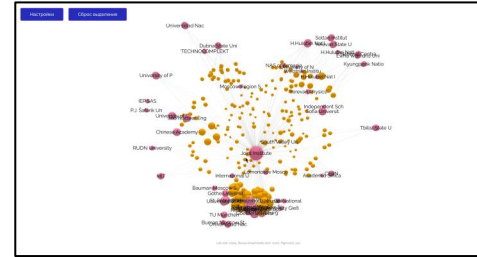
	 Япония
Временной период	2010 – 2021 гг.
Количество научных публикаций	1 050 000
Объем данных	24.3 Гб

ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ

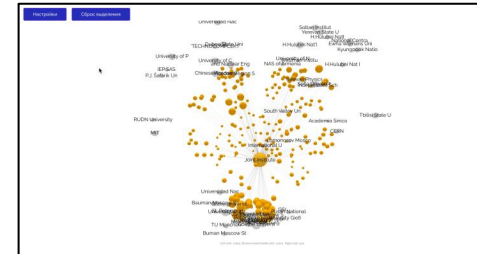
Реализованный граф



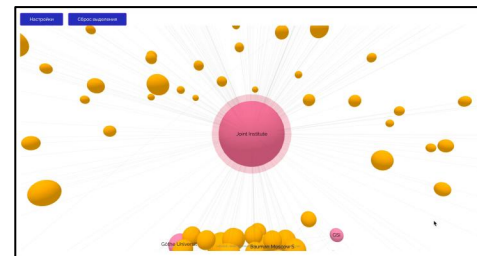
Выделение узла



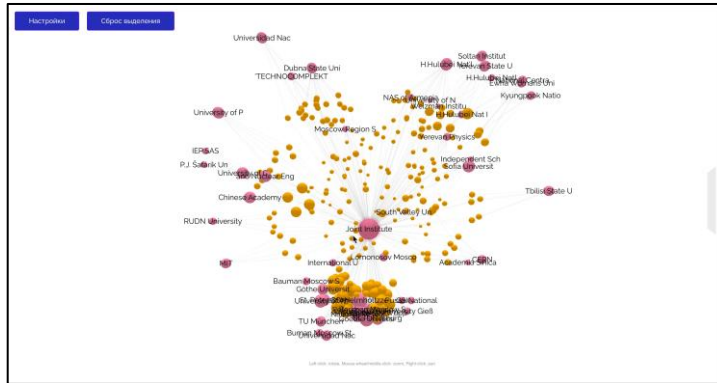
Сброс выделения узла



Подсвечивание узла

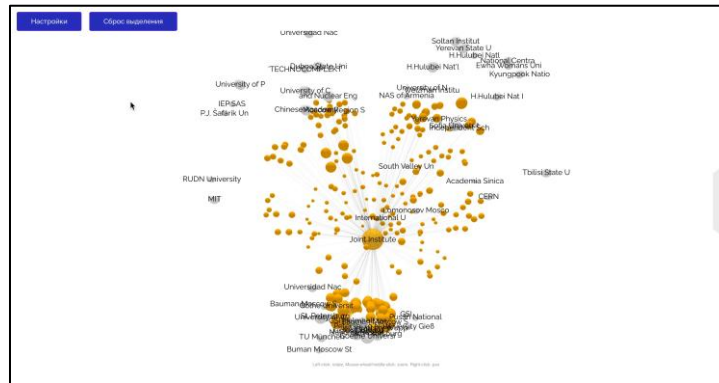


ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ



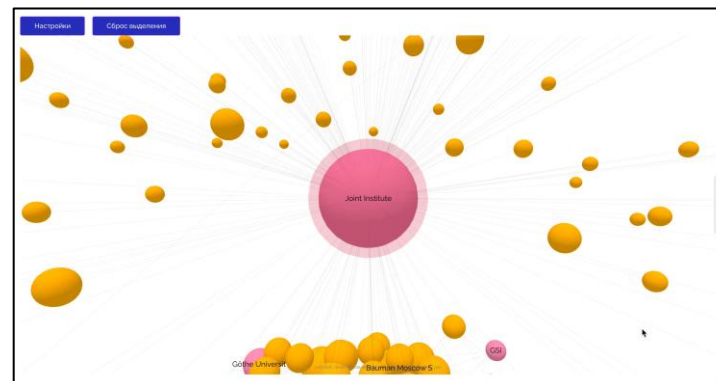
Выделение узла

- Перекрашивает выбранный узел
- Связанные узлы не изменяются
- Обесцвечивает несвязанные узлы и ребра
- Выводится информация по узлу



Сброс выделения узла

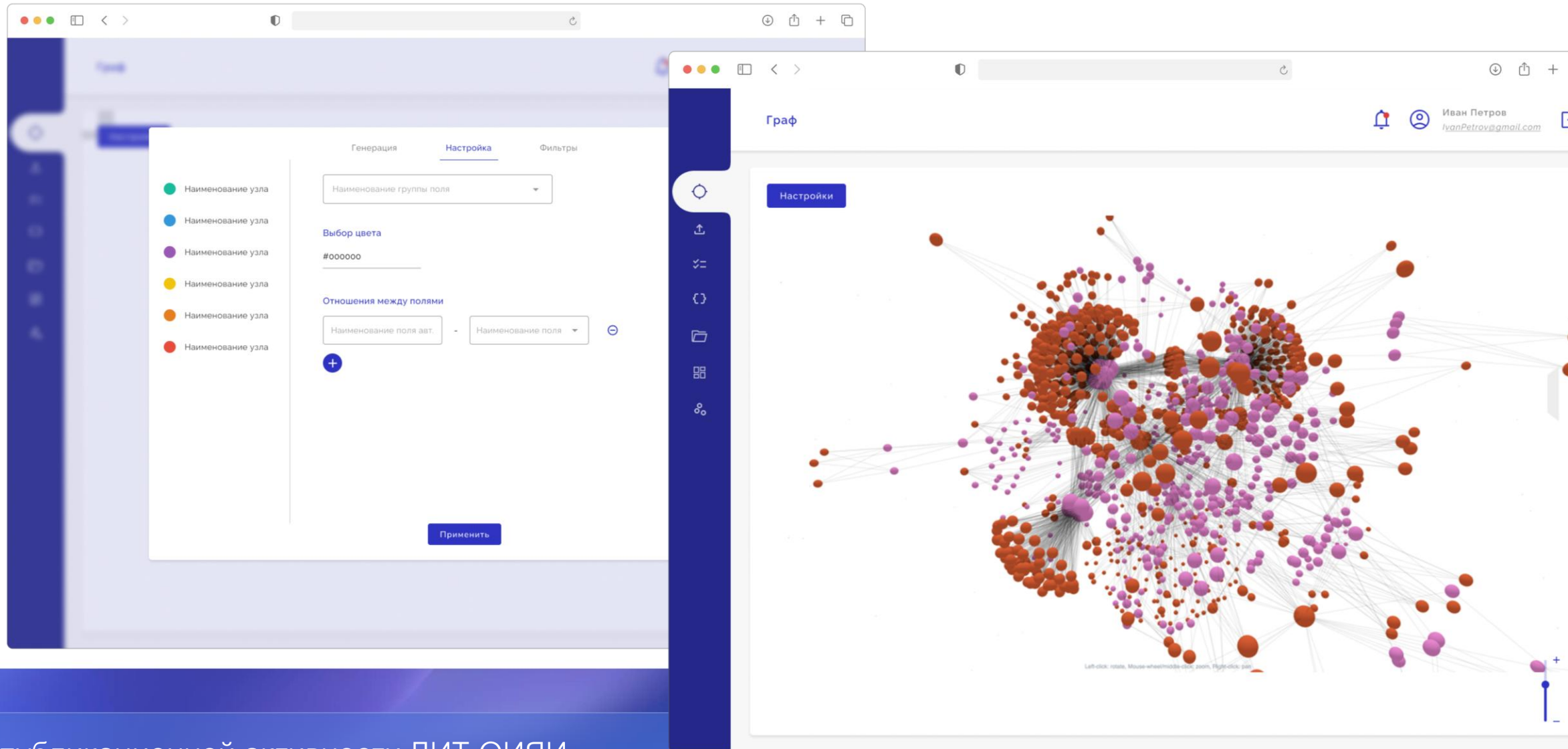
- Возвращает граф в первоначальное состояние



Подсвечивание узла

- Появление мигающей сферы на интересующем узле
- Настраивается с помощью правил управления графом или первоначальной

ВИЗУАЛЬНЫЙ АНАЛИЗ ПУБЛИКАЦИОННОЙ АКТИВНОСТИ НАУЧНОЙ ОРГАНИЗАЦИИ



Граф публикационной активности ЛИТ ОИЯИ

СЕМАНТИЧЕСКИЕ СЕТИ НА ОСНОВЕ ЗАПРОСА В SCOPUS

Выявление на основе собранных по тематическим направлениям публикаций:

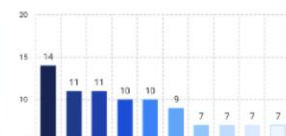
- явных и скрытых связей между объектами;
- поиск потенциально-интересных персон/организаций/конференций
- визуализация взаимосвязей в рамках публикационной активности

Введение

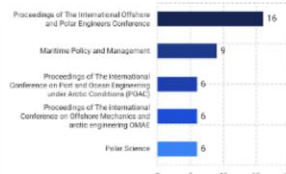
Данная семантическая сеть была построена на основе запроса в базу данных Scopus, в результате которого 2024 по 2024 было выявлено 180 публикаций.

Запрос включал фильтрацию по дружественным Российской Федерации странам, исключая при этом такие страны как США и Великобритания.

На семантической сети рассмотрены некоторые из топ-10 организаций с наибольшим количеством публикаций по результатам запроса.

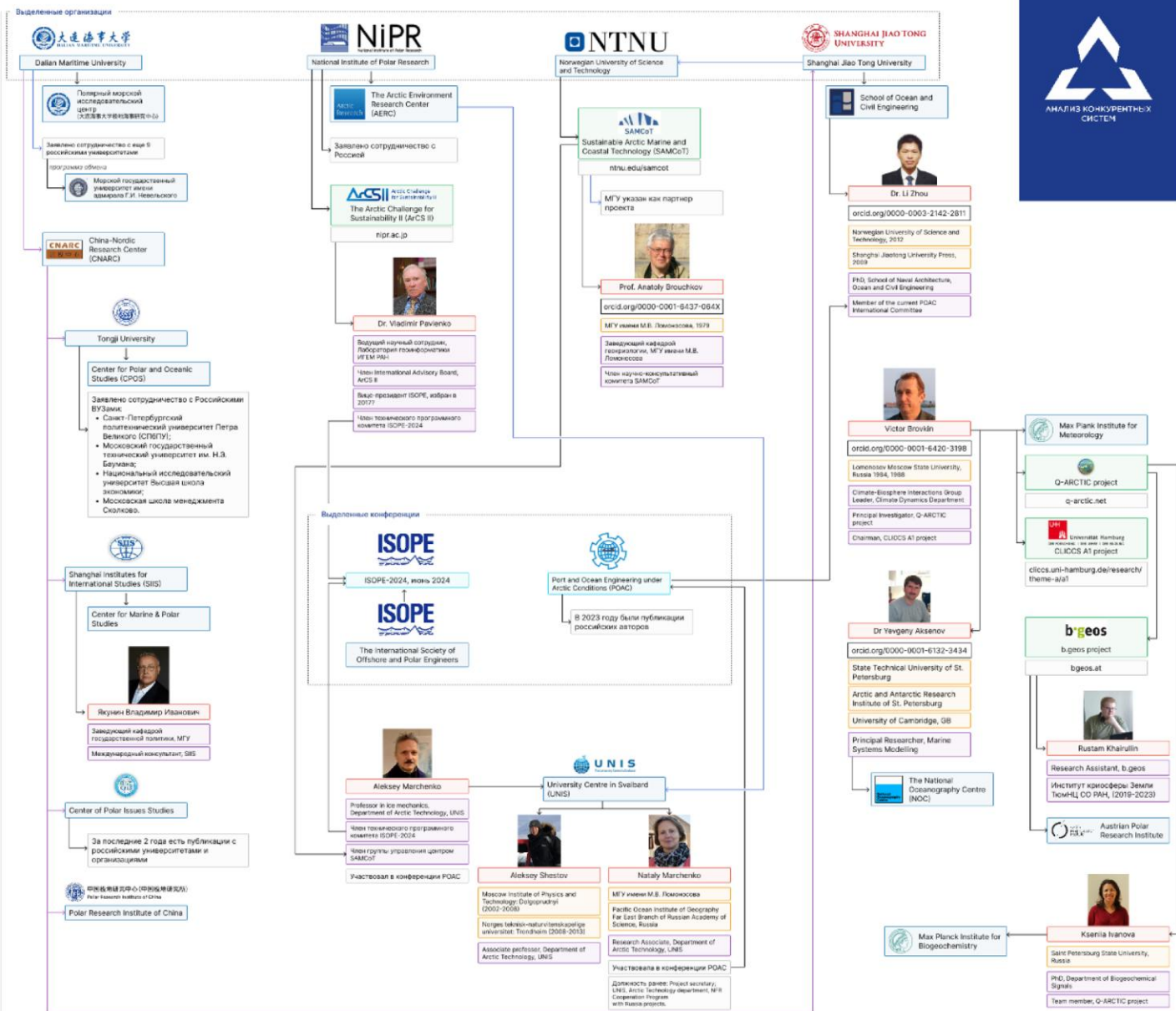


Также были рассмотрены конференции из перечня топ-5 изданий с наибольшим количеством публикаций, продолжающих взаимодействие с российскими авторами.



Легенда

- Персона
- Организация
- Должность
- Проект
- Конференция
- Образование
- Дополнительные сведения
- Принадлежность (дочерний элемент)
- Сотрудничество
- Член проекта / организации



Спасибо за внимание!

Артамонов Алексей Анатольевич, к.т.н.,
Зав. кафедрой «Анализ конкурентных
систем» МИФИ №65