МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ДИНАМИКИ СИСТЕМ И ТЕОРИИ УПРАВЛЕНИЯ
ИМЕНИ В.М. МАТРОСОВА
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК
(ИДСТУ СО РАН)

УДК 004.4;004.6;004.75;004.8
Рег. № НИОКТР АААА-А19-119111990037-0
Рег. № ИКРБС
Инв. №

УТВЕРЖДАЮ
Директор ИДСТУ СО РАН
академик РАН
_____ И.В. Бычков
«____» _____декабря 2019 г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:
ИНФОРМАЦИОННО-ТЕЛЕКОММУНИКАЦИОННАЯ ПЛАТФОРМА ЦИФРОВОГО МОНИТОРИНГА ОЗЕРА БАЙКАЛ, НА ОСНОВЕ СКВОЗНЫХ ТЕХНОЛОГИЙ

(промежуточный, этап 1)

Приоритетное направление IV.38. Проблемы создания глобальных и интегрированных информационно-телекоммуникационных систем и сетей. Развитие технологий и стандартов GRID

Программа IV.38.1. Методы и технологии создания и интеграции гетерогенных распределенных информационно-вычислительных ресурсов для поддержки междисциплинарных научных исследований на основе сервис-ориентированной парадигмы

Руководитель НИР,
академик РАН,
д-р техн. наук _____ И.В. Бычков
27.12.2019

Иркутск 2019

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель НИР, академик РАН, д-р технич. наук	27.12.2019 подпись, дата	И.В. Бычков (введение, блок 1, заключение)
Исполнители:		
Зав. отделения д-р технич. наук	подпись, дата 24.12.2019	Г.М. Ружников (блок 3)
Зав. лабораторией канд. технич. наук	24.12.2019 подпись, дата	А.Е. Хмельнов (блок 2)
Вед. науч. сотр. канд. технич. наук	24.12.2019 подпись, дата	Р.К. Фёдоров (блок 3)
Ст. науч. сотр. канд. технич. наук	24.12.2019 подпись, дата	А.О. Шигаров (блок 4)
Ст. науч. сотр. канд. технич. наук	24.12.2019 подпись, дата	А.С. Гаченко (блок 2)
Ст. науч. сотр. канд. технич. наук	24.12.2019 подпись, дата	В.В. Парамонов (блок 4)
Науч. сотр. канд. технич. наук	 подпись, дата	Т.И. Маджара (блок 1)
Науч. сотр. канд. технич. наук	24.12.2019 подпись, дата	Е.С. Фереферов (блок 5)
Науч. сотр. канд. технич. наук	24.12.2019 подпись, дата	А.К. Попова (блок 1)
Науч. сотр. канд. технич. наук	24.12.2019 подпись, дата	А.А. Михайлов (блок 5)
Нормоконтролер канд. технич. наук	25.12.2019 подпись, дата	Е.С. Фереферов

РЕФЕРАТ

Отчет 41 стр., 22 рис., 2 табл., 9 источников, 1 приложение.

ЦИФРОВОЙ МОНИТОРИНГ, МОДЕЛЬ, АЛГОРИТМ, ИНТЕЛЛЕКТУАЛИЗАЦИЯ, МАШИННОЕ ОБУЧЕНИЕ, БОЛЬШИЕ ДАННЫЕ, ЗНАНИЯ, РАСПРЕДЕЛЕННЫЕ ДАННЫЕ, РАСПРЕДЕЛЕННАЯ СЕРВИСНО-ОРИЕНТИРОВАННАЯ СРЕДА

Цель проекта – создание единой сервис-ориентированной платформы цифрового мониторинга озера Байкал, основанной на сквозных технологиях и обеспечивающей сбор, хранение и мониторинг больших массивов разноформатных распределенных междисциплинарных научных данных, их анализ и прогнозирование развития природных и техногенных процессов на основе комплекса математических моделей и методов машинного обучения.

Объект исследования – большие объёмы разноформатных пространственно-временных научных данных (в том числе неструктированных) и знаний, получаемые в результате цифрового мониторинга озера Байкал, а также методы и технологии формирования единой сервис-ориентированной платформы для их обработки.

Методология проведения исследований базируется на:

- разработке подходов к формированию распределенной сервисно-ориентированной информационно-аналитической среды обработки научных данных, в том числе данных цифрового мониторинга озера Байкал;
- создание форматов и инструментальных средств интеграции данных непрерывного мониторинга, поступающих от научного оборудования (сенсоры, измерительные приборы, ДДЗЗ), передачи, хранения, обработки больших объёмов междисциплинарных научных данных и знаний
- создание методологического и инструментального обеспечения поддержки процессов трансформации неструктурированных данных из произвольных таблиц;
- разработке подходов к построению центров высокопроизводительной обработки междисциплинарных данных и их интеграции с распределенной сетью научного оборудования;
- создании методов и технологий сбора, обработки и анализа больших объёмов разноформатных пространственно-временных данных цифрового мониторинга, основанных на интеллектуализации, машинном обучении и использовании конструктивных средств спецификации;

- исследование методов и технологий создания проблемно-ориентированных и интеллектуальных систем поддержки принятия решений, основанных на использовании модельно-управляемого подхода, порождающего программирования и декларативных описаний,
- разработка и интеграция в единую платформу прогнозных моделей для решения задач управления риском техногенных и природных катастроф

Полученные в проекте методы и технологии позволят создать оригинальную распределенную сервисно-ориентированную платформу хранения, обработки больших объёмов разноформатных научных данных и знаний для поддержки процессов непрерывного мониторинга крупных озерных систем, их междисциплинарных исследований и прогнозирования развития возможных событий. Ожидаемые результаты исследований будут способствовать переходу к новым цифровым, интеллектуальным производственным технологиям, автоматизированным системам нового поколения, позволят повысить эффективность создаваемых и внедряемых распределенных информационно-вычислительных технологий, в том числе технологий обработки больших объемов пространственно-временных данных, а также позволят повысить качественный уровень проведения междисциплинарных научных исследований. Созданные методы и сквозные технологии могут найти широкое применение в различных областях человеческой деятельности, в том числе для формировании систем поддержки принятия решений органов государственной власти и местного самоуправления для решения проблем эффективного управления социально-эколого-экономическим развитием территорий, снижения рисков возникновения и сокращения неблагоприятных последствий техногенных и природных катастроф.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

База данных – совокупность данных, организованных по определённым правилам, устанавливающим общие принципы описания, хранения, манипулирования данными.

Базовые пространственные данные – это описания базовых пространственных объектов в заданной системе координат, которые следует использовать для определения координат любых близлежащих пространственных объектов и явлений.

Геопортал – Web-портал, отображающий и предоставляющий доступ к географической информации посредством Web-сервисов.

Геоинформационная система – информационная система, обеспечивающая сбор, хранение, обработку, доступ, отображение и распространение пространственных данных.

Инфраструктура пространственных данных – это иерархически упорядоченная система (информационная среда), построенная с использованием информационных и геоинформационных технологий, основанная на общих стандартах пространственных данных и метаданных, а также сети географических информационных узлов – геопорталов и каталогов метаданных.

Модель пространственных данных – это способ цифрового представления пространственных объектов, тип структуры пространственных данных.

Мониторинг — система постоянного наблюдения за явлениями и процессами, проходящими в окружающей среде и обществе, результаты которого служат для обоснования управленческих решений по обеспечению безопасности людей и объектов.

Метаданные – это описательная информация о структуре и смысле данных, а также приложений и процессов, которые манипулируют данными.

Пространственные данные – данные о географических объектах, которые являются формализованными цифровыми моделями материальных или абстрактных объектов реального или виртуального мира

Хранилище данных – предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений.

Цифровая платформа - группа технологий, которые используются в качестве основы, обеспечивающей создание специализированной системы цифрового взаимодействия.

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

АИР – агрегирование индивидуальных ранжировок

АРМ – автоматизированное рабочее место

БД – база данных

БЗ – база знаний

ГИС – геоинформационная система

ИПД – инфраструктура пространственных данных

ПД – пространственные данные

СХД – система хранения данных

СУБД – система управления БД

ХД – хранилище данных

API – Application Programming Interface - интерфейс прикладного программирования

DNS – Domain Name System — система доменных имён, компьютерная распределённая система для получения информации о доменах.

OGC – Open Geospatial Consortium - консорциум открытых ГИС

RDF – Resource Description Framework - среда описания ресурсов

SOA – Service-oriented architecture - сервис-ориентированная архитектура

SysML – System Modeling Language - язык моделирование систем (язык графического описания различных аспектов структуры систем)

UML – Unified Modeling Language - унифицированный язык моделирования — язык графического описания для объектного моделирования в области разработки программного обеспечения

WMS – Web Map Service стандарт Web-служб OGC (Web-сервис), предоставляющий интерфейс http-запросов для получения браузером или настольным приложением в растровом виде геопривязанных изображений карт в форматах PNG, TIFF, JPEG

WPS – Web Processing Service стандарт Web-служб OGC, (WPS-сервис), описывает правила стандартизации входящих и исходящих данных (запросов и ответов на них), сервисов обработки

WWW – World Wide Web - всемирная информационная паутина

ВВЕДЕНИЕ

Основные направления исследований озера Байкал отмечены в Указе Президента РФ №204 от 7 мая 2018 года «Сохранение уникальных водных объектов, в том числе реализации задач по сохранению озера Байкал», а также определены в рамках Национального проекта (программы) «Экология» (Федеральный проект «Сохранение озера Байкал») Постановлением Правительства РФ от 2 февраля 2015 г. N 85 «Об утверждении Положения о государственном экологическом мониторинге уникальной экологической системы озера Байкал» отмечено, что он является частью государственного экологического мониторинга (государственного мониторинга окружающей среды).

Под экологическим мониторингом понимается комплексная система регулярных наблюдений в пространстве и во времени за состоянием окружающей среды и её изменениями под воздействием природных и антропогенных факторов [1]. Государственный экологический мониторинг озера Байкал состоит из наблюдений за:

- водными объектами;
- атмосферным воздухом;
- водными биологическими объектами;
- состоянием недр;
- состоянием земель;
- радиационной обстановкой;
- объектами животного мира;
- лесом.

В настоящее время экологический мониторинг озера Байкала и Байкальской природной территории (БПТ) осуществляют структуры федеральных органов исполнительной власти Минприроды, Росгидромета, Рослесхоза, Росводресурсов, Росрыболовства, а также органы исполнительной власти Республики Бурятия, Забайкальского края и Иркутской области, в соответствии с их компетенцией в порядке, установленном постановлением Правительства РФ от 9 августа 2013 г. N 681 "О государственном экологическом мониторинге и государственном фонде данных государственного экологического мониторинга". Большую часть работ по государственному мониторингу осуществляют региональные подразделения Федеральной службы по гидрометеорологии и мониторингу окружающей среды (Росгидромет) – Иркутский УГМС, Забайкальский УГМС и Бурятский ЦГМС.

Рисунок 1 - Государственный экологический мониторинг озера Байкал

Дополнительно наблюдения ведут и другие региональные службы. Междисциплинарные научные исследования и частичный мониторинг экосистемы озера Байкала и БПТ, также проводят академические институты РАН, университеты Минобрнауки [2-6].

Наборы данных, формируемые в результате мониторинга, характеризуются разнородностью и разно-структурированностью (реляционные базы данных, электронные таблицы, тематические карты, космоснимки, 3D-модели, фото и видео изображения и др.), пространственно-темпоральной привязкой, мультидисциплинарностью, большим объёмом и высокой скоростью роста объёмов информации. Наличие общей системы управления полученными наборами данных обеспечит выполнение поиска закономерностей и интеллектуального анализа данных, построение прогнозных моделей состояния экосистемы оз. Байкал, создание тематических сервисов обработки, карт и баз данных.

Это обосновывает актуальность комплексного цифрового мониторинга экосистемы озера Байкал для информационно-коммуникационной поддержки процессов сбора, хранения и управления получаемых данных.

Целью проекта является создание единой сервис-ориентированной платформы цифрового мониторинга озера Байкал, основанной на сквозных технологиях и обеспечивающей сбор, хранение и мониторинг больших массивов разноформатных распределенных междисциплинарных научных данных, их анализ и прогнозирование развития природных и техногенных процессов на основе комплекса математических моделей и методов машинного обучения.

Цели и задачи этапа НИР в 2019 году:

- Разработка концепции доступа к высокопроизводительным вычислительным ресурсам и ресурсам хранения данных центров коллективного пользования.
- Исследование и разработка эффективных методов и технологий сбора, обработки и анализа больших объёмов разноформатных пространственно-временных данных, основанных на интеллектуализации, машинном обучении и использовании конструктивных средств спецификации.
- Разработка сервисов предобработки и обработки больших объемов данных ДЗЗ.
- Разработка сервисов обработки неструктурированных данных.
- Разработка сервисов управления жизненным циклом научных данных.

Разработка единой сервис-ориентированной платформы цифрового мониторинга озера Байкал основывается на использовании современных подходов к организации инфо-коммуникационных инфраструктур, интеллектуализации, обработке больших объемов данных, создания информационно-аналитических систем цифрового мониторинга.

Проект соответствует:

- приоритетным направлениям развития науки, технологий и техники в РФ: Информационно-телекоммуникационные системы (3);

- критическим технологиям РФ: Технологии и программное обеспечение распределенных и высокопроизводительных вычислительных систем (18), Технологии информационных, управляющих, навигационных систем (13), Технологии предупреждения и ликвидации чрезвычайных ситуаций природного и техногенного характера (21).

- технологической платформе - Национальная программная платформа.

1 Разработка концепции доступа к высокопроизводительным вычислительным ресурсам и ресурсам хранения данных центров коллективного пользования

В настоящее время нет комплексных полнофункциональных систем мониторинга экологических параметров природных территорий, применимых для исследования уникальных районов Сибири, таких как участок мирового природного наследия ЮНЕСКО озеро Байкал, крупный и экологически нагруженный горнодобывающий комплекс Кузбасса, сибирские бореальные леса и другие, от состояния и процессов в которых существенно зависят погодные условия и экологическая обстановка в Сибири в целом, а также на Урале и Дальнем Востоке. В основном акцент делается на создание и поддержку информационно-образовательных сайтов. Мало внимания уделяется формированию научных цифровых информационных ресурсов и аналитических систем. В связи с активным развитием по всему миру цифровизации, её внедрение в решение проблем экологии и рационального природопользования становится приоритетом. Это также связано с большими объемами тематических и пространственно-временных данных экологического мониторинга и с количеством существующих программно-аппаратных комплексов, а также развитием систем передачи данных и сети Интернет. Под цифровизацией понимают процесс внедрения киберфизических систем, интеграцию датчиков во все компоненты оборудования, замену физических или аналоговых ресурсов информационными или цифровыми данными. Таким образом, цифровизация предполагает системный подход к использованию цифровых ресурсов и внедрению цифровых технологий для достижения конкретных целей.

В рамках исследований 2019 года, был проведен анализ существующего состояния компонентов государственного экологического мониторинга озера Байкал и БПТ. Для всех участников мониторинга оценивался ряд показателей - организация-исполнитель, точки наблюдения, состав показателей наблюдений, периодичность сбора, программно-аппаратные комплексы, системы передачи данных и т.д.

Анализ выявил следующие особенности мониторинга озера Байкал. Мониторинг водных объектов включает сбор основных параметров – гидрохимические и гидрологические характеристики, состояние берегов и дна, водопотребление, использование водоохранных зон. В его проведении участвуют:

- МУП «Водоканал» - оценивает чистоту забираемых поверхностных вод;
- ФБУЗ «Центр гигиены и эпидемиологии в Иркутской области» несколько раз в год проверяет химическое и биологическое загрязнение в водозаборах, скважинах и рекреационных зонах;

- ФГУ "Востсибрегионводхоз" следит за гидрохимическими и гидрофизическими показателями в период навигации с помощью судового комплекса;
- Енисейское ВБУ измеряет объем вод при водопотреблении и сбросе стоков.

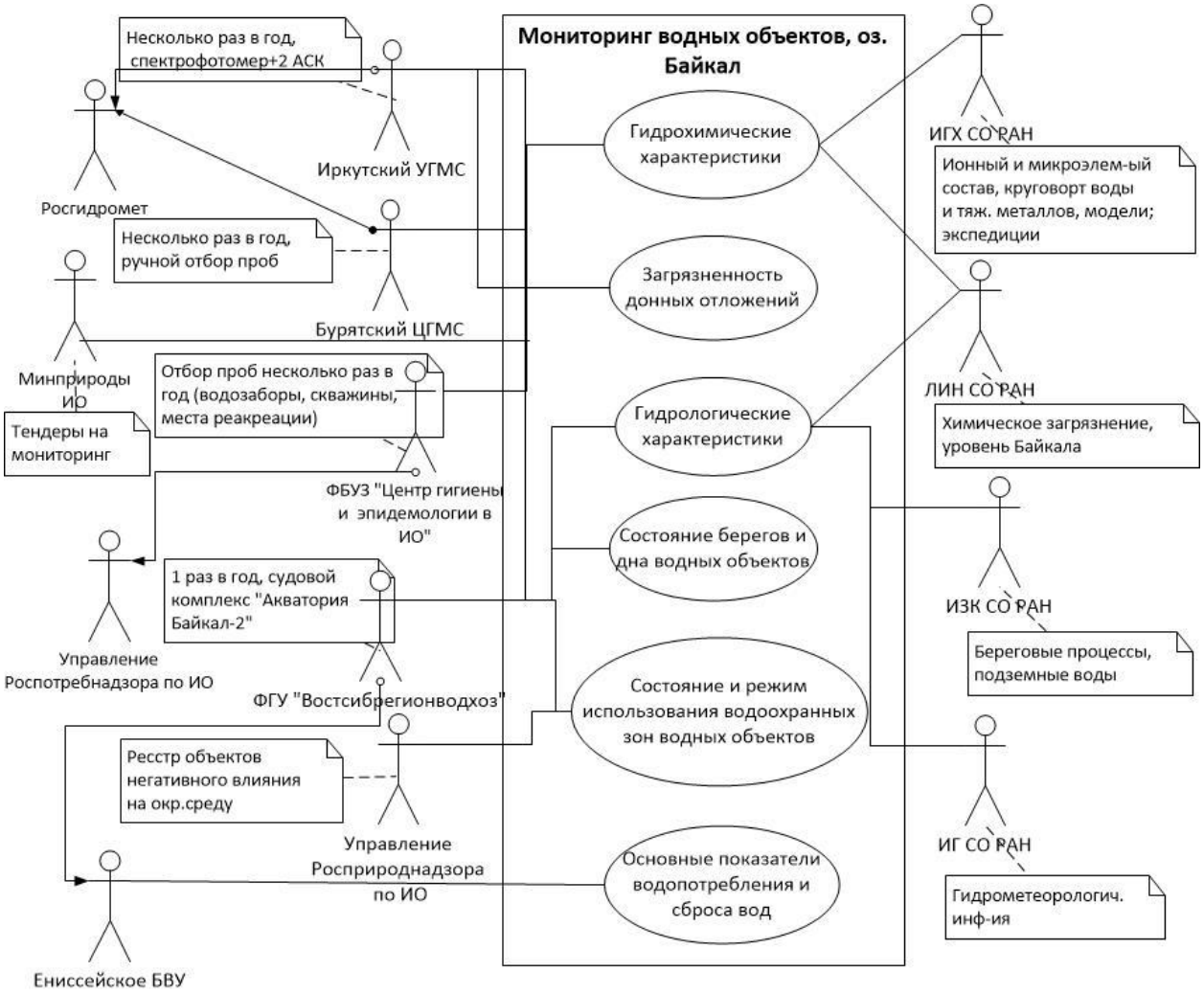

Рисунок 2 - Мониторинг водных объектов озера Байкал

На весь Байкал есть только два автоматических поста мониторинга в южной части озера (в г. Байкальске и п. Листвянка), измеряющих 7 гидрохимических показателей воды (температура, мутность, растворенный кислород, pH, аммонийный азот, окислительно-восстановительный потенциал, удельная электропроводность). Остальные наблюдения на озере и прилегающих реках ведутся с помощью ручного забора проб раз в месяц или реже. Судовой комплекс работает только в период летней навигации.

Чуть больше автоматизирован мониторинг атмосферного воздуха – у Росгидромета есть 23 автоматических станции контроля атмосферы, однако большинство из них расположены Иркутская зоогалерея продает каймановых черепах для покрытия долгов

Это вынужденная мера из-за средств для оплаты электроэнергии. В августе зоопарк переехал в два новых помещения, расходов стало на две трети больше.не непосредственно на Байкале, а в крупных городах (Иркутск, Ангарск, Чита, Улан-Удэ). Управления Роспотребнадзора и Росприроднадзора измеряют загрязнения реже – от 1 раза в неделю до 1 раза в месяц в зависимости от показателя.

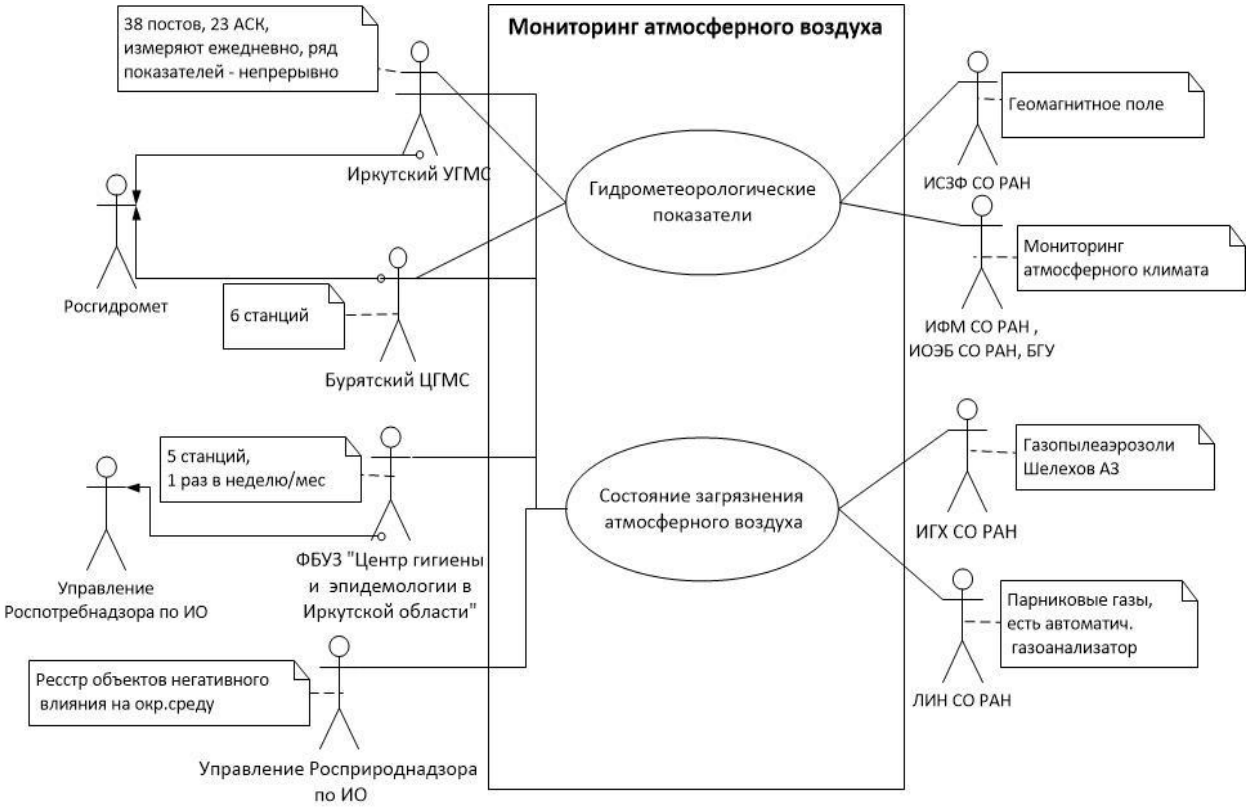

Рисунок 3 - Мониторинг атмосферного воздуха озера Байкал

Отсутствует комплексный подход и при ведении других видов мониторинга. За государственный мониторинг земель отвечает Росреестр вместе с Министерством сельского хозяйства, за состояние недр – Роснедра, водные биологические объекты отслеживает Росрыболовство, лесопатологический мониторинг курирует Рослесхоз. Также регулярными наблюдениями не охвачена прибрежная мелководная зона Байкала, которая испытывает наибольшую антропогенную нагрузку. Особенно это касается о. Ольхон и района Малого моря, где летом в разы возрастает поток туристов.

В рамках ведения Единого государственного фонда данных о состоянии окружающей среды (ЕГФД) Росгидромет собирает сведения о загрязнении Байкала от сторонних организаций – отчеты по экологическим экспертизам, о выполненных гидрометеорологических работах, инженерных изысканиях. Данные передаются в разных форматах, частично – только на бумажных носителях.

Поэтому можно сделать выводы:

- Каждая из организаций, исследующих экосистему озера Байкал и БПТ, придерживаясь своей схемы мониторинга, формирует и использует большие объёмы пространственных и тематических данных, которые, как правило, локализованы и не скоординированы между собой, в параметрическом, хронологическом и других аспектах. Это осложняет проведение комплексных оценок, прогнозирования, принятие управленческих решений на базе имеющихся ведомственных данных.
- В открытый доступ выложены только разрозненные данные о состоянии и загрязнении озера Байкала и БПТ.
- Нет единой системы с оперативной мониторинговой информацией, позволяющей вносить, хранить и обмениваться данными о состоянии экологической системы озера в реальном времени.
- Отсутствует единая система идентификации пространственных объектов, как универсальных элементов связи ведомственных пространственных и тематических баз данных.
- Не проведена оценка системной достаточности выбора «информативных показателей» мониторинга экосистемы озера Байкал.
- Не ведется полноценный государственный мониторинг по всей акватории озера в режиме реального времени за гидрохимическими, гидрофизическими и биологическими параметрами, а присутствуют только сезонные наблюдения.

Отсюда следует, что существующие ведомственные системы мониторинга экосистемы озера Байкал и БПТ не позволяют оперативно реагировать на изменения природного и антропогенного характера, выявлять в них компоненты локального или глобального генезиса. Это обосновывает необходимость перехода на комплексный цифровой мониторинг [3, 5], особенностью которого является интеграционный характер, непрерывность (режим 24/7/365) и распределённость наблюдений, большие объёмы разноформатных пространственно-временных данных о состоянии озера с сенсоров и измерительных приборов.

Фундаментальные исследования уникальных экосистем, проводимые в мире и России, базируются на мониторинге, хранении и обработке больших объёмов научных данных и знаний о системе, имеющих пространственно-временной характер, а также на использовании распределённых информационно-вычислительных технологий и их приложений, современных сетей передачи данных [7-9].

Современный этап цифрового мониторинга экосистемы озера Байкал должен базироваться на создании информационно-телекоммуникационной платформы цифрового мониторинга экосистемы озера Байкала и БПТ, обеспечивающей координацию межведомственного (Росгидромет, Минприроды, Рослесхоз, Росводресурсов, Росрыболовства РАН, Минобрнауки) и межрегионального (Республика Бурятия, Забайкальский край и Иркутская область) взаимодействия. На базе этого центра (ЦОД) должно осуществляться хранение и обработки пространственно-временных тематических данных цифрового экологического мониторинга, с возможностью прямого информационного доступа. Сборка распределённых междисциплинарных данных (Big Data) в ЦОД позволит значительно повысить качество прогнозных моделей по развитию экосистемы Байкала.

Ключевыми составляющими разработанной концепции доступа к высокопроизводительным вычислительным ресурсам и ресурсам хранения данных центров коллективного пользования являются: инфраструктуры центров коллективного пользования ИДСТУ СО РАН «Интегрированная информационно-вычислительная сеть Иркутского научно-образовательного комплекса» и «Иркутский суперкомпьютерный центр СО РАН», а также «Информационно-аналитическая среда» (ИАС), которые обеспечат сбор, передачу, поиск, хранение, и параллельную обработку больших объемов данных, возможность онлайн-доступа к данным, каталогам, сервисам и информационно-аналитическим системам, возможность проведения на основе полученных данных оценки, моделирования и прогноза экологических и климатических изменений Байкала и прилегающих территорий с применением средств суперкомпьютерного моделирования и облачных вычислений.

Для проведения мониторинга озера Байкал разрабатывается распределенная сервис-ориентированная информационно-аналитическая платформа (ИАП) геопортального типа, включающая подсистемы сбора, передачи, хранения, поиска и обработки больших объемов разноформатных пространственно-временных данных и знаний [2].

Следует учитывать, что в цифровизации самой важной и определяющей технологией является платформа. Цифровая платформа как программный продукт аккумулирует в себе все остальные необходимые технологии, предоставляя пользователям доступ к информации (данным), сервисам их обработки, аналитике и, самое главное, к клиентам, производителям, пользователям. Это система алгоритмизированных взаимоотношений участников, объединенных единой информационной средой, обеспечиваемая архитектурным стеком программно-аппаратного обеспечения,

необходимого для хранения и анализа цифровых данных, приводящая к снижению транзакционных издержек, за счет применения пакета цифровых технологий и изменения организации разделения труда. Каждая развитая цифровая платформа строится вокруг какого-либо массового процесса (Facebook, Uber, AliExpress и др.), обеспечивая взаимодействие потребителей и поставщиков информационных услуг. Одно из важнейших свойств процессов на платформе, отличающее их от привычных форм взаимодействий, – алгоритмизированность. Платформа ограничивает вариативность действий пользователей своим текущим функционалом. В качестве такого процесса можно рассматривать - ведение экологического мониторинга. Являясь участниками единой информационной среды, поддерживаемой цифровой платформой, исследователи получают новые формы организации научных исследований, которые были невозможны ранее.

Методология исследований по формированию цифровой платформы базируется на комплексном применении сервис-ориентированной парадигмы и современных технологий распределенной обработки данных, использовании декларативных спецификаций и интеллектуализации с использованием методов и технологий глубокого обучения. При этом декларативные спецификации обеспечивают компактность, выразительность и предметную ориентированность, включая возможность интерпретации трансформационными и другими процедурами. В свою очередь, использование сервис-ориентированного подхода позволяет проводить полноценный учет распределенных информационных ресурсов в сочетании с легкостью тестирования, масштабируемостью и возможностью повторного использования создаваемых сервисов [2].

Для организации комплексного цифрового мониторинга экологических систем озера Байкал сервисы тематического экологического мониторинга интегрируются с помощью логических конструкций для решения задач обработки пространственно-временных данных, управления потоком выполнения и т.д. Разнообразие коллекций разрабатываемых сервисов тематического экологического мониторинга позволит передавать данные между ними, согласовывать форматы данных, запускать асинхронные вычислительные процессы.

ИАП должна с заданной степенью надежности сохранять любые виды исходных данных экологического мониторинга: временные ряды с результатами измерений различных датчиков, материалы экспедиций, космические снимки, векторные карты и т.д. (рисунок 4). Все эти данные, характеризующиеся пространственной и временной привязкой, могут иметь ряд дополнительных атрибутов, специфичных для конкретного вида информации.

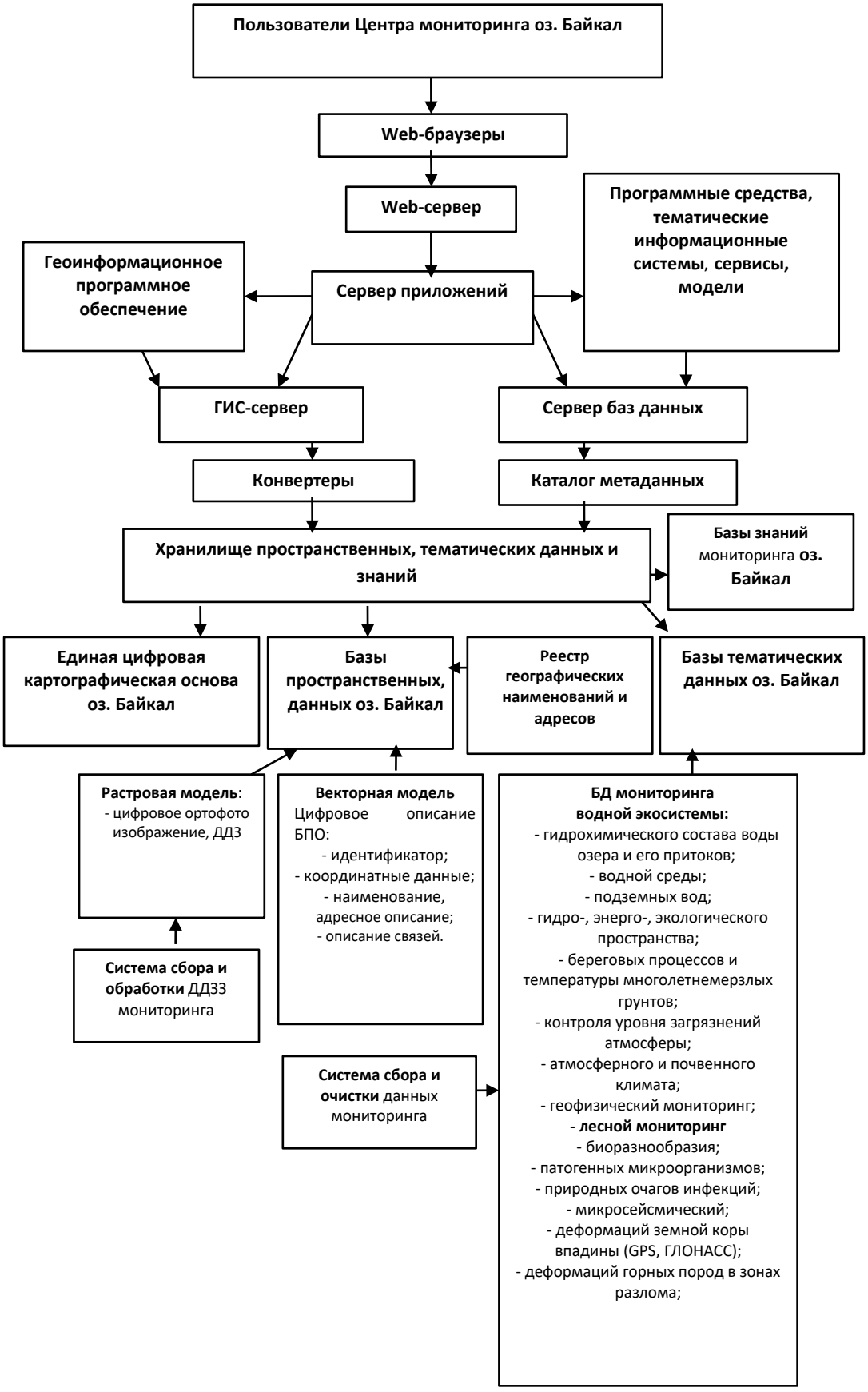

Рисунок 4 - Структура ИАП

Геопортальная сервис-ориентированная информационно-аналитическая платформа комплексного цифрового мониторинга озера Байкал должна обеспечивать:

1. онлайн-доступ к распределенным сенсорам;
2. доступ к архивным данным сенсоров;
3. высокую скорость обработки данных;
4. доступ к высокопроизводительным вычислительным ресурсам и ресурсам хранения данных центров коллективного пользования;
5. масштабирование вычислительных ресурсов и ресурсов хранения и обработки данных центров коллективного пользования с учетом роста числа задач и объемов данных мониторинга;
6. возможность использования различных методов и технологий распределённой обработки данных.

В качестве программно-аппаратной платформы, обеспечивающей непрерывную работу ИАП предлагается использовать зарекомендовавшую схему построения центров обработки данных (рисунок 5), в основе которой лежат основополагающие принципы:

1. Полное резервирование инженерной инфраструктуры и аппаратного комплекса.
2. Эффективное использование оборудования за счет организации пулов вычислительных ресурсов.
3. Виртуализация ресурсов и приложений.
4. Системы резервного копирования и восстановления.

Создание центра обработки данных на начальном этапе включает:

- реконструкцию инженерной инфраструктуры – систем бесперебойного питания и охлаждения;
- модернизацию сетевой инфраструктуры;
- развертывание серверов и систем хранения и обработки данных на базе оборудования ведущих в отрасли вендоров.

В качестве стартовой площадки для развертывания ЦОД будут использованы существующие и успешно функционирующие центры коллективного пользования ИИВС ИрНОК и ИСКЦ, что обеспечит повышение доступности высокопроизводительных вычислительных ресурсов для обработки данных цифрового мониторинга, в том числе, с применением средств суперкомпьютерного моделирования.

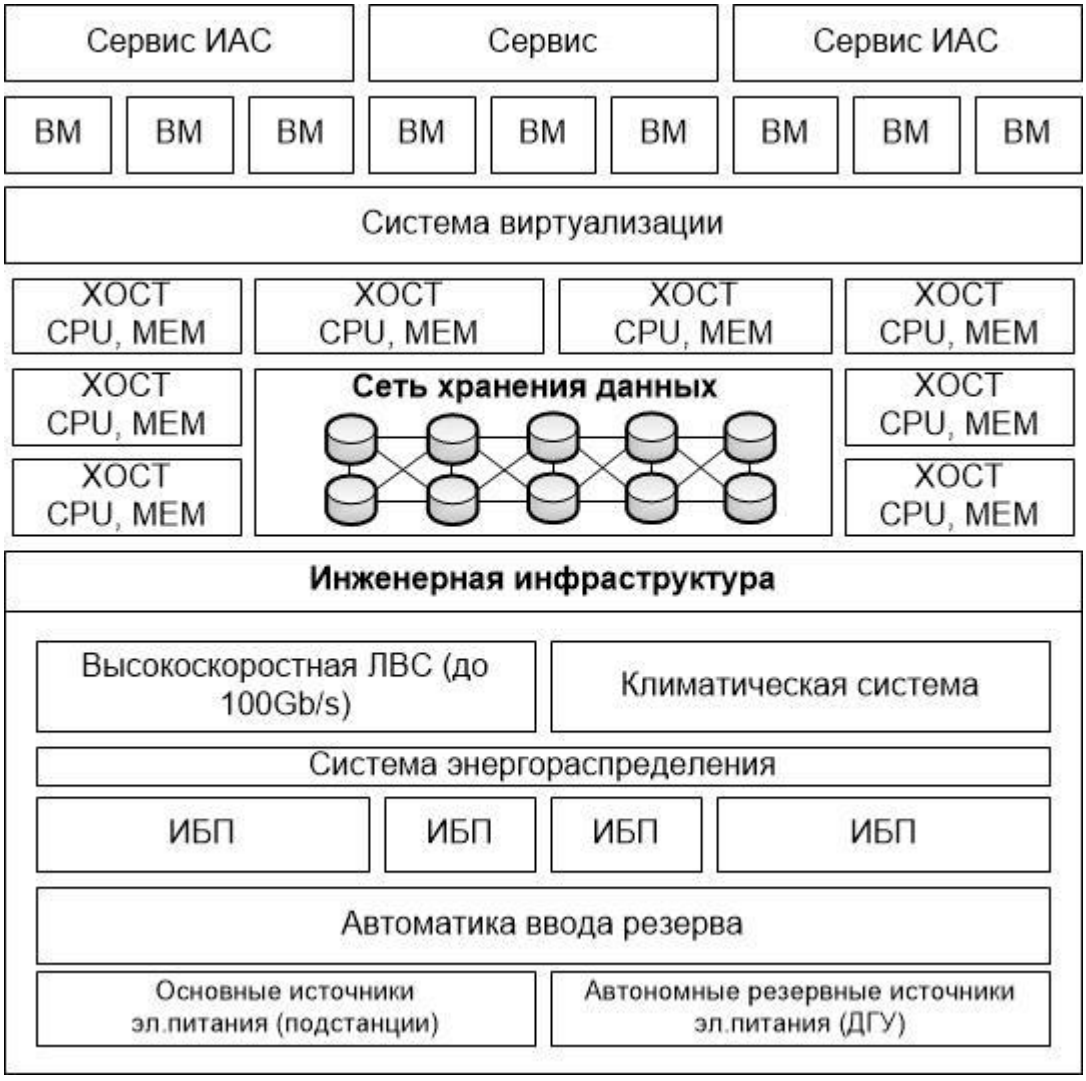

Рисунок 5 - Инфраструктура ЦОД

2 Исследование и разработка эффективных методов и технологий сбора, обработки и анализа больших объёмов разноформатных пространственно-временных данных, основанных на интеллектуализации, машинном обучении и использовании конструктивных средств спецификации

Технологии дистанционного зондирования Земли (ДЗЗ) из космоса — эффективный способ изучения и постоянного мониторинга нашей планеты, помогающий эффективно использовать и управлять ее ресурсами. Современное использование данных ДЗЗ находят применение практически во всех сферах нашей жизни. Сегодня разработанные предприятиями Роскосмоса методы позволяют выработать уникальные решения для обеспечения безопасности, повышения эффективности разведки и добычи природных ресурсов, внедрения новейших практик в сельское хозяйство, предупреждения чрезвычайных ситуаций и устранении их последствий, охраны окружающей среды и контроля над изменением климата. В настоящее время российская орбитальная группировка дистанционного зондирования Земли состоит из космических аппаратов серий «Ресурс-П», «Канопус-В», «Метеор-М» и «Электро-Л».

В рамках разрабатываемой платформы цифрового мониторинга создан каталог цифровых спутниковых снимков (рисунок 6), обеспечивающий сбор и эффективное хранение разноформатных данных большого объема. Каталог наполнен данными дистанционного зондирования Земли из Космоса за период 2018 – 2019 год по Иркутской области и югу озера Байкала.

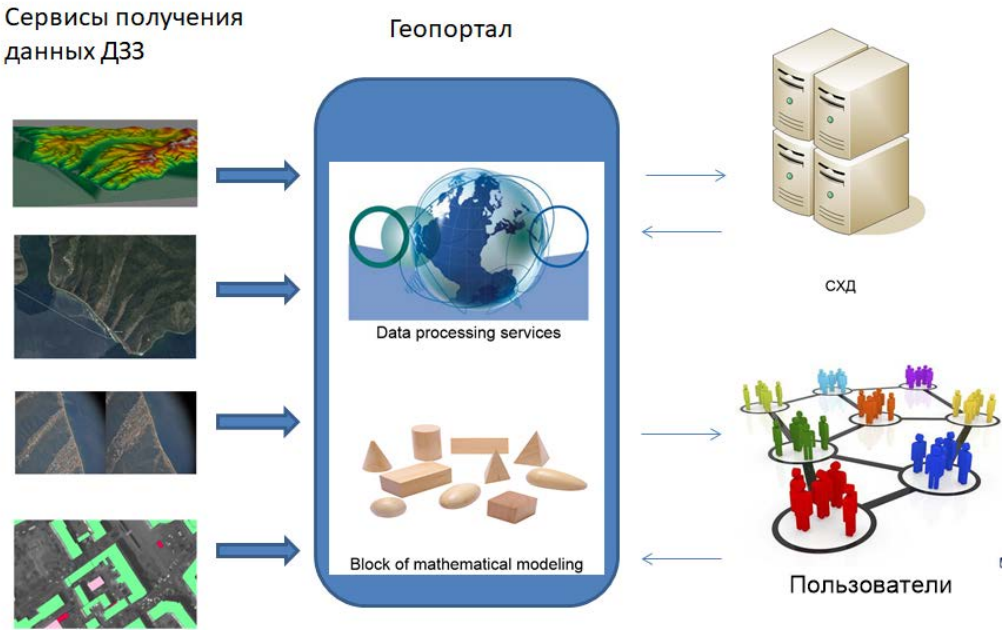

Рисунок 6 - Архитектура каталога данных ДЗЗ

Каталог включает в себя данные группировки Российских космических аппаратов (КА). В каталоге собрано порядка 200 космических снимков на территорию Иркутской области в разном разрешении и временном интервале. Исходные данные получены с геопортала Роскосмоса на основании соглашения между Министерства науки и высшего образования РФ и государственной корпорации Роскосмос. Данные съемки получены с двух космических аппаратов «Канопус-В» и «Ресурс-П».

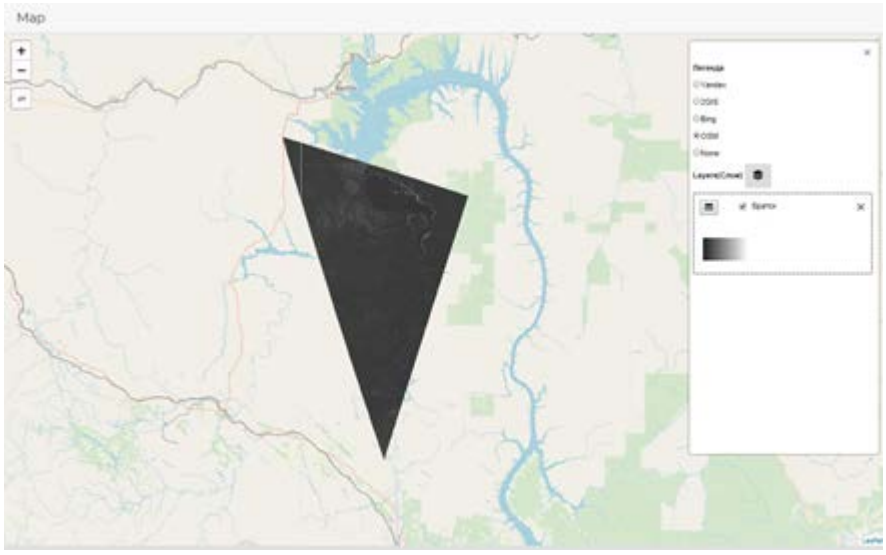

Рисунок 7 - Подсистема хранения и визуализации данных ДЗЗ.

Рисунок 8 - Спутниковые снимки космических аппаратов «Канопус-В» и «Ресурс-П».

Существующий объем хранимых данных ДЗЗ 509 ГБ. Для оптимизации хранилища данных разработан и используется специальный формат хранения данных – MRG. Данный формат предназначен для хранения больших объемов целочисленных растровых данных. Он поддерживает быструю выборку фрагментов растра с различными уровнями детализации за счет представления вместе с наиболее детальными данными, растров более низкого разрешения. Предложенная структура данных позволяет не только не увеличить расход дисковой памяти на хранение дополнительных растров более низкого разрешения, но, и наоборот, в целом сократить ее использование. Для упаковки данных разработан оригинальный алгоритм сжатия разностных целочисленных последовательностей, позволяющий, как увеличить коэффициент сжатия, так и сократить время, необходимое для упаковки данных (по сравнению с популярной библиотекой ZLib)/ Для чтения информации из файлов MRG реализована динамическая библиотека, позволяющая использовать представленные в них данные сторонних программ. Программное обеспечение для чтения форматов MRG способно эффективно работать при использовании ограниченного объема памяти (от 16 Mb).

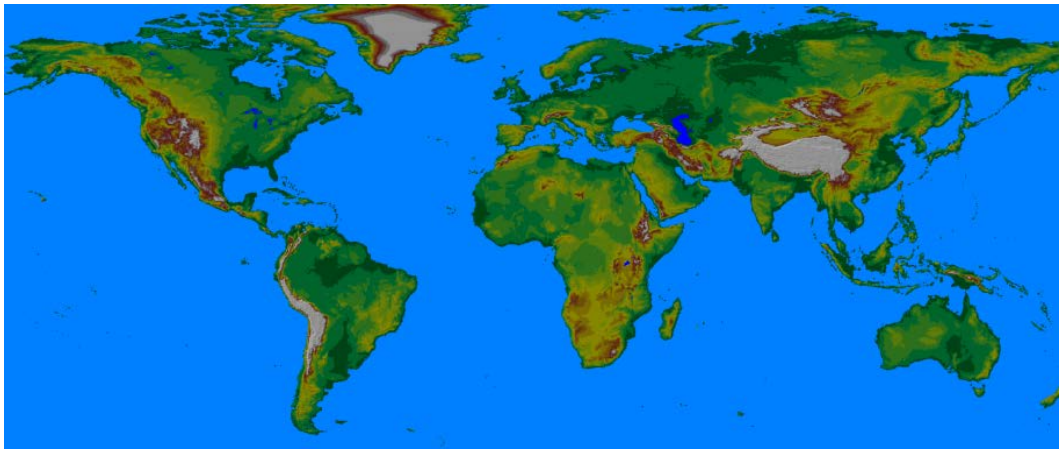

Рисунок 9 - Все данные SRTM3 занимают 10,8 Гб.

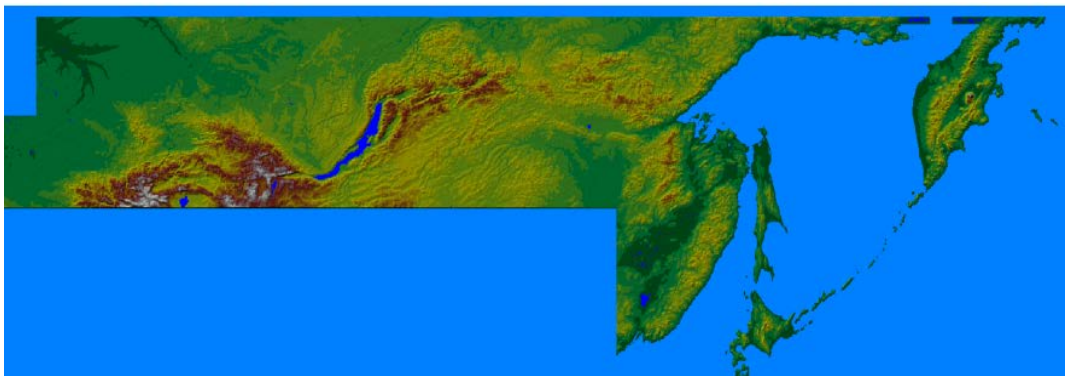

Рисунок 10 - Данные в формате MRG, которые помещаются на один компакт-диск (700 Mb).

3 Разработка сервисов предобработки и обработки больших объемов данных ДЗЗ

Разработана технология обработки больших данных ДЗЗ (рисунок 11) в виде совокупности WPS сервисов. Разработанная технология состоит из следующих этапов:

1) получение снимков, на этом этапе пользователь производит выбор подходящего снимка, на котором имеется ледник;

2) формирование обучающей выборки, на этом этапе пользователь указывает границы ледников. Данный этап является опциональным;

3) создание классификатора, на текущий момент применяются пока только нейронные сети. Данный этап является опциональным;

4) классификация снимков, пользователь может использовать готовые классификаторы или создать свои с помощью предыдущих этапов;

5) включение результатов классификации в каталог, который позволяет быстро найти нужные объекты и сохранить в одном из форматов.

Рисунок 11 - Основные этапы технологии обработки данных ДЗЗ

Рассмотрим эти этапы подробнее.

Для формирования обучающей выборки в рамках геопортала создана таблица (рисунок 12), в которой пользователь указывает границы объектов на различных снимках.

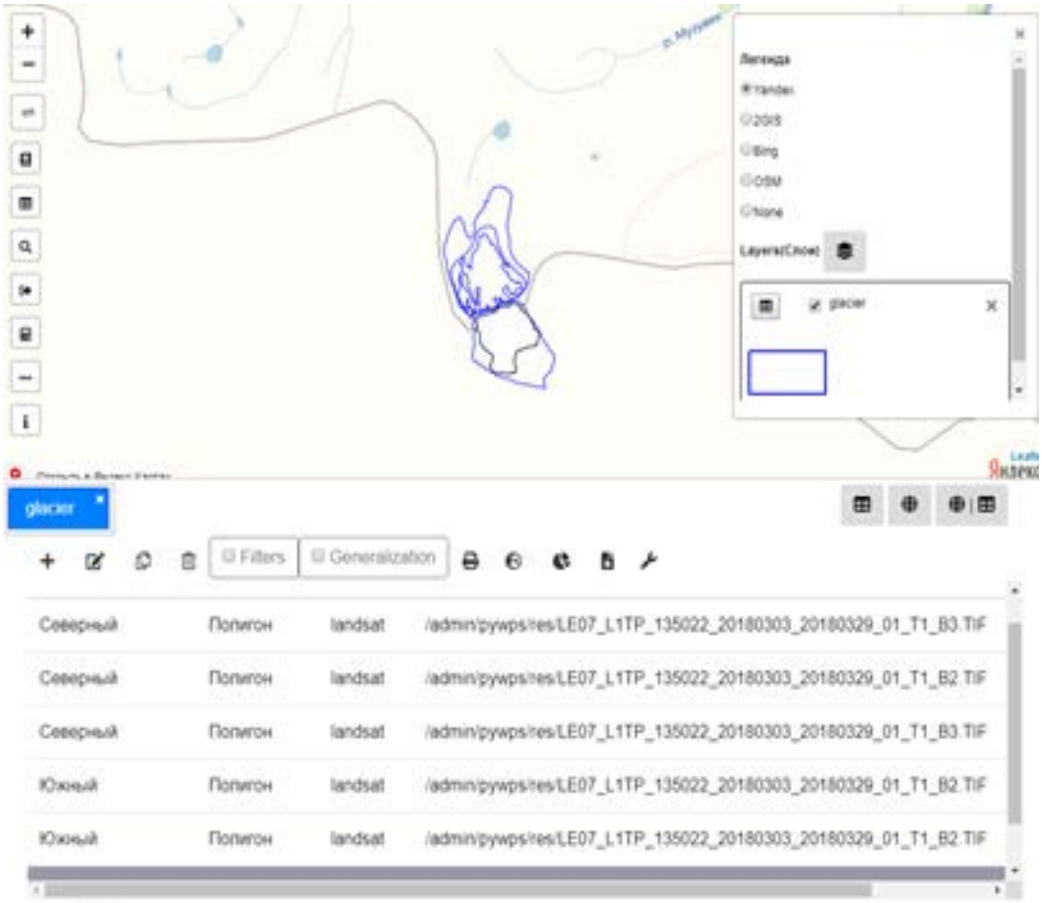

Рисунок 12 - Таблица обучающей выборки

Таблица состоит из следующих атрибутов:

1) name – название ледника;
2) boundaries – границы объекта в виде полигонального объекта;
3) date – дата снимка;
4) imagename – путь к файлу снимка в директории системы хранения данных геопортала;
5) imprecisedate – неточная дата снимка.

В каждой записи таблицы пользователь определяет границы одного объекта на конкретном снимке. Вся область снимка делится на два класса, область, покрытая полигонами из таблицы для этого снимка, считается объектом. Вся оставшаяся область изображения противоположным классом. Соответственно в процессе составления обучающей выборки необходимо полностью покрыть объекты полигонами.

Космоснимки, полученные со спутника Landsat, представлены в виде набора файлов в формате GEOTIFF, в каждом из которых находится отдельный канал (монохромное изображение). Для отображения на карте геопортала необходимо совокупность файлов преобразовать в RGB представление. Для этого используется сервис Bands_to_RGB, который принимает на вход три монохромных изображения,

соответствующих каналам модели RGB и строит из них цветное. В основе сервиса используется модуль gdal_merge_simple (https://github.com/gina-alaska/dans-gdal-scripts#gdal_merge_simple).

Для создания классификатора на основе обучающей выборки разработан набор сервисов, на вход которых подается таблица и на выходе получаем файл классификатора. В таблице можно отфильтровать записи по всем атрибутам, в том числе указать область обработки. Например, задать диапазон дат или определенное название объекта. Для каждого изображения в обучающей выборке формируется набор записей. Затем используется сервис Rasterize, который принимает на вход векторный файл в формате SHAPE, определяющий положения прецедентов, исходное изображение в формате GEOTIFF и область прямоугольную область интереса. На рисунке 13 представлен пример запуска сервиса Rasterize.

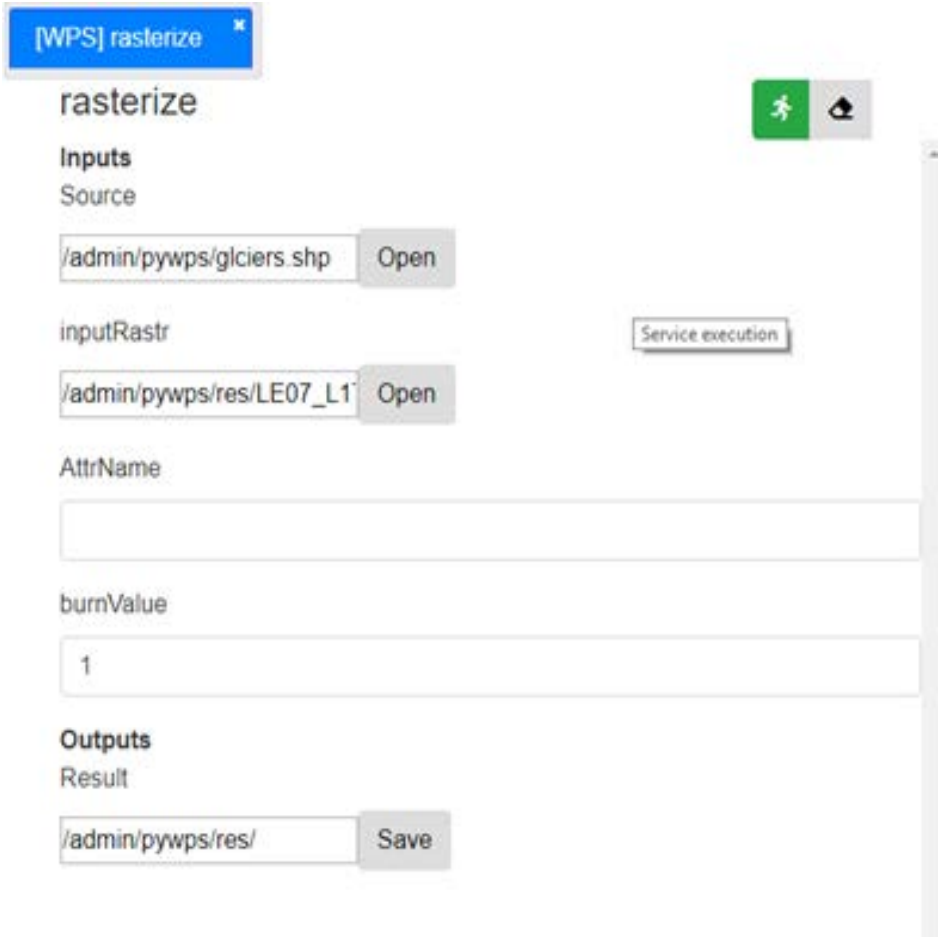

Рисунок 13 - Запуск сервиса Rasterize

Результатом работы сервиса является растровое изображение – бинарная маска (рисунок 14), содержащая положения прецедентов по заданной области интереса. В основе сервиса используется инструмент gdal_rasterize (https://gdal.org/programs/gdal_rasterize.html). Сервис позволяет задать размер выходного изображения в пикселях.

Рисунок 14 - Маска ледника, полученная на основе таблицы

После сервиса Rasterize используется сервис Learning, который непосредственно производит обучение нейронной сети. На вход сервису подается растровое изображение модели RGB и бинарная маска, с указанием положений ледников. На выходе формируется файл модели. В основе сервиса используется проект Segmentation models (https://github.com/qubvel/segmentation_models). Данный проект позволяет использовать одну из предобученных нейросетей, распространенных архитектур Unet, FPN, Linknet или PSPNet. Принцип работы сервиса следующий. Из изображения и маски методом скользящего окна формируется обучающая выборка. После чего вызывается метод обучения модели. Затем полученная модель возвращается пользователю в качестве результата. Так как процесс обучения достаточно долгий, то данный сервис вызывается в асинхронном режиме. Пользователь может следить за статусом его выполнения.

Классификация осуществляется с помощью сервиса Segmentation. На вход сервису подается растровое изображение модели RGB. На выходе получается бинарная маска, с отмеченными положениями объектов. В основе сервиса используется проект Segmentation models (https://github.com/qubvel/segmentation_models). Сервис получает на вход изображение, после чего пробегается по нему скользящим окном. В результате чего изображение делится на строки и столбцы. Из каждой строки формируется набор данных и передается на вход нейронной сети. Затем результат работы записывается на

соответствующие позиции копии исходного изображения и так далее пока не будет обработано все изображение целиком.

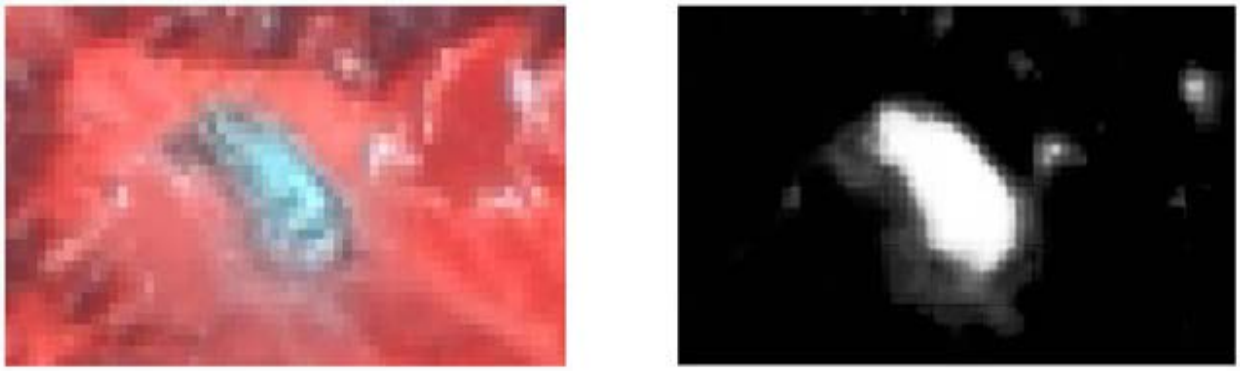

Рисунок 15 - Снимок Landsat и результат классификации

4 Разработка сервисов обработки неструктурированных данных

Предложена концепция микросервисной архитектуры веб-ориентированной системы очистки не стандартизированной структуры и не гармонизированного содержания табличных данных, базирующаяся на современном архитектурном стиле взаимодействия компонентов распределённого приложения в сети — REST (Representational State Transfer). Предлагаемая архитектура предназначена для реализации следующих операций:

- Передача исходных табличных данных в формате CSV (Comma Separated Values), декларативных программ очистки структуры и содержания таблиц на языке правил общего назначения DRL (Drools Rule Language), спецификацию типов данных в формате YAML (YAML Ain't Markup Language).
- Конфигурирование и выполнения процессов очистки табличных данных на основе исполнения DRL-программ в системе «Drools Expert».
- Формирование выходных табличных данных и сопровождающих их метаданных в формате JSON (JavaScript Object Notation).
- Кэширование сессионных данных и предоставление программного доступа к ним.

```
{"$schema": "http://json-schema.org/draft-07/schema#",
"$id": "http://td.icc.ru/cell.schema.json",
"title": "Cell",
"description": "A cell from CTable",
"type": "CCell",
"properties": {
"id": {"type": "Integer"},
"cl": {"description": "a left column position", "type": "Integer"},
"cr": {"description": "a right column position", "type": "Integer"},
"rt": {"description": "a top row position", "type": "Integer"},
"rb": {"description": "a bottom row position", "type": "Integer"},
"height": {"description": "the height of cell", "type": "Integer"},
"width": {"description": "the width of cell", "type": "Integer"},
"text": {"description": "a text ( a processed text )", "type": "String"},
"indent": {"description": "", "type": "Integer"},
"typeTag": {"description": "a data type tag provided by a apreadsheet cell
(e.g. NUMERIC, STRING, DATE, FORMULA)",
"type": "TypeTag", "$ref": "http://td.icc.ru/type-tag.schema.json"},
"entry": {"type": "CEntry", "$ref": "http://td.icc.ru/entry.schema.json" }
```

Рисунок 16 - Фрагмент JSON-схемы ответа микросервиса очистки табличных данных.

Предлагаемая архитектура определяет следующий рабочий процесс очистки «грязных» табличных данных. Исходные данные загружаются и отображаются в Java объекты. Структуры данных, представляющие ячейки, вхождения, метки и категории, реализованы как Java-классы в соответствии с соглашениями спецификации JavaBeans. Они добавляются как факты в рабочую память системы исполнения правил. Опционально одна или несколько категорий, описанных с помощью языка YAML, могут загружаться в систему, также формируя соответствующие факты в рабочей памяти системы исполнения правил. При этом синтаксический разбор YAML описаний выполняется в свободной системе SnakeYAML. При этом каждый факт является экземпляром одного из этих классов. Правила анализа и интерпретации таблиц выражаются на языке DRL. В качестве системы исполнения правил используется свободное программное обеспечение «Drools Expert». Факты, загруженные в рабочую память, сопоставляются скомпилированным правилам. В результате создаются новые факты, описывающие семантику таблицы: вхождения, метки, категории и отношения между ними. Из них генерируется каноническая таблица. В результате очистки табличных данных формируются JSON объекты, представляющие таблицу, ее ячейки, вхождения и метки в канонической форме (рисунок 16).

Очистка данных позволяет повысить качество информации, представленной в табличных документах. Java-класс, получающий на вход объекты таблицы, представленные в виде JSON-формата. Очистка включает в себя несколько стадий, схематически представленных на рисуноке 17.

На первой стадии проводится идентификация типа данных. Наличие семантики и отношений между элементами таблицы, представленными в виде объектов, дает возможность проводить анализ по диапазону данных и меток, образующих столбцы и строки. При определении типа данных рассматривается их представление (символы, формат). Также в расчет берутся данных, представленные в соседних столбцах и (или) строках. По результатам их совокупного анализа дается заключение о принадлежности информации к определенному типу данных. Идентификация основана на применении механизма регулярных выражений. Также, для анализа в расчет берется информация о типе данных ячейки в электронной таблице, если таковая присутствует и совпадает с представлением данных. В случае, если из меток или категорий есть возможность извлечь информацию об единицах измерения - она используется для нормализации единиц измерения в соответствии с целевыми единицами. Также для данных, идентифицированных как числовые разработаны подходы поиска вычисляемых значений. Предлагается использовать комплексный подход включающий как анализа содержания,

так и структуры табличного документа. Для этого проводится анализ на наличие в метках и категориях ключевых слов (“итого”, “всего”, “среднее”, “нарастающий итог” и т.п.), обозначающих их возможное вхождение вычисляемого значения. Проводится анализ окружающих ячейку вхождений с целью выявления зависимостей. Для анализ строится на возможных гипотезах относительно агрегации данных.

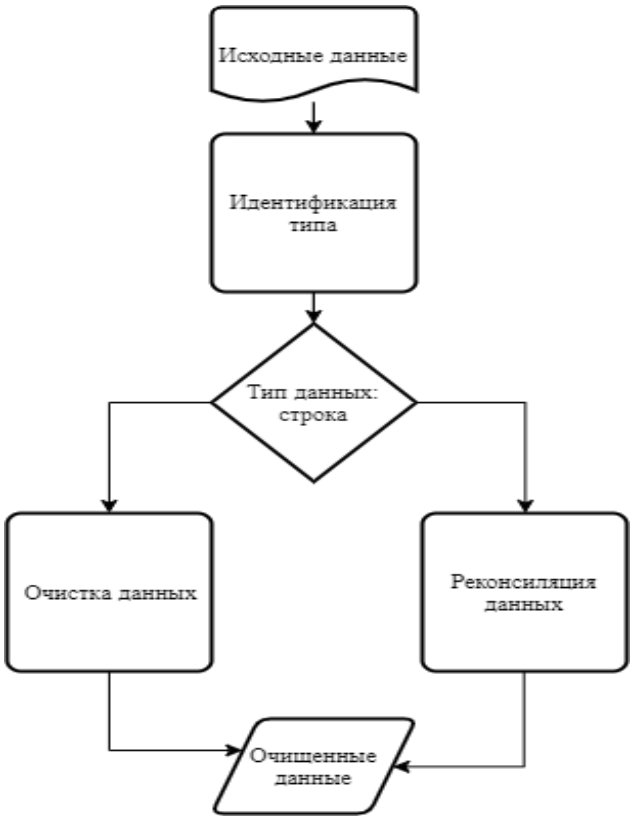

Рисунок 17 - Схема процессов очистки табличных данных

Для данных, которые идентифицируются как «строки» и которые могут быть соотнесены с каким-либо значением словаря (город, регион, марка и т.п.) будет проведена их реконсиляция. Принадлежность данных к словарю обуславливается пользовательской спецификацией данных перед загрузкой. Для таких данных реконсиляция обеспечивает устранение опечаток и установление взаимно-однозначных связей с классификаторами. Для устранения опечаток применяются методы нечеткого сравнения строк и фонетического кодирования.

Очищенные данные могут быть сохранены в виде электронной таблицы или переданы на последующую обработку как JSON-объекты.

Для тестирования предлагаемой архитектуры спроектировано веб-ориентированное приложение, предоставляющее пользовательский интерфейс доступа к микросервисам очистки табличных данных, обеспечивающий визуализацию и интерактивную работу с получаемыми выходными JSON данными.

5 Разработка сервисов управления жизненным циклом научных данных

Предложена концепция модели курирования больших объемов исследовательских данных, накапливаемых в результате цифрового мониторинга озера Байкал. Модель предполагает активное и непрерывное управление ими на протяжении всего их жизненного цикла, для того чтобы обеспечить и увеличить их ценность, как в настоящем, так и в будущем.

Предлагаемая модель жизненного цикла данных цифрового мониторинга охватывает процессы планирования, сбора и обработки, моделирования и анализа, хранения, каталогизации и архивирования, публикации, открытого доступа и повторного использования.

Рисунок 18 - Модель жизненного цикла научных данных цифрового мониторинга озера Байкал.

Реализация этой модели предполагает разработку детального плана (включая стратегию и политики) управления данными в соответствие с современными стандартами и рекомендациями центров цифрового курирования. Сбор и обработка включает решение в рамках проекта вопросов автоматизации проектирования и исполнения рабочих процессов конвейерной обработки данных, упаковки получаемых исследовательских данных в современные обменные форматы (NetCDF, HDF5, CERN ROOT и др.). Для краткосрочного хранения и долгосрочного архивирования планируется создание инфраструктуры репозиториев данных, каталогов метаданных и резервного копирования. Для поддержки моделирования и анализа данных предполагается внедрить современную платформу аналитики научных данных (Data Science Platform). Эта платформа обеспечит исследование озера Байкал необходимыми инструментами моделирования (в т. ч., методами Монте-Карло) и интеллектуального анализа (в т. ч., методами обнаружения знаний и поиска скрытых закономерностей) с интенсивным использованием

междисциплинарных данных. Для публикации наборов исследовательских данных планируется создание и наполнение портала открытых данных озера Байкал. Такой портал позволит распространять и повторно использовать результаты цифрового мониторинга, в т.ч., в других научных проектах и образовательных курсах.

Современные исследовательские программные системы ориентированы на решение задач сбора, хранения, обработки и анализа данных. При этом, как правило, осуществляется работа с большими объемами информации сложной структуры, что обуславливает необходимость использования баз данных (БД) и реализации специализированных функций для решения задач конкретной предметной области. Кроме того, такие системы часто требуют модернизации, т.к. в процессе исследований появляются новые структуры данных или видоизменяются существующие.

Для автоматизации создания и последующего развития информационных систем мы разработали технологию создания ИС на основе декларативных спецификаций, которые являются средством представления и хранения моделей этих систем. Предлагаемые спецификации содержат минимально необходимую информацию о структуре данных ИС, которой, однако, оказывается достаточно для автоматической реализации приложения, и, в частности, создания пользовательских интерфейсов, обеспечения выполнения CRUD-функций, построения пользовательских запросов, поддержки взаимодействия с пространственными данными, а также реализации взаимодействия с внешними подсистемами для решения специфических задач. Декларативные спецификации удобны своей компактностью, при этом обладают предметной ориентированностью и выразительностью, а также широкой возможностью интерпретации различными трансформационными и другими процедурами. Кроме того, представление моделей в виде спецификаций позволяет поддержать модульную разработку ИС – интегрировать готовые спецификации отдельных приложений в одну более масштабную систему.

Для формирования спецификаций приложений разработан декларативный язык, конструкции которого, обеспечивают достаточно детальное и в тоже время компактное описание всех элементов приложения БД. Каждый элемент приложения описывается отдельным предложением, кроме того существуют конфигурационные предложения для описания общих системных настроек ИС.

Конструкции языка позволяют описать: способ подключения к БД, общие параметры ИС, таблицы, представления, подключаемые модули (plugin), подключение других спецификаций, способ взаимодействия с пространственными данными. В способе подключения к БД содержится информация о технологии доступа к БД (поддерживается

BDE, ADO, FDAC) и о способе аутентификации. В общих параметрах задаются настройки внешнего вида приложения, а также общие для всех таблиц правила взаимодействия с БД (схема БД, способ формирования имён полей и таблиц для запросов). При создании спецификации ПБД в качестве входных данных используется метаданные о структуре уже созданной БД (схема БД) хранящиеся в СУБД. Схема данных является уже структурированными знаниями о сущностях и связях между ними, которые необходимо расширить знаниями о способах применения этих сущностей и связей в ПБД.

Описания таблиц в спецификации (Рисунок 19) содержат информацию обеспечивающую автоматическое создание пользовательского интерфейса и механизмов взаимодействия с соответствующими таблицами БД. Так, для полей таблиц указывается их вид, который кроме типа данных, известного СУБД, определяет способ использования данного поля в приложении, который, например, влияет на выбор элементов управления и доступных в построителе запросов операций сравнения. Например, строковому полю могут быть сопоставлены следующие виды: «строковый», «содержащий имя файла», «именованный», «списочный». «Строковый» вид указывает, что в таком поле может содержаться произвольный набор символов, а при формировании запроса пользователь будет иметь возможность задать условия на значения строки или её подстрок. Вид поля, «содержащий имя файла», применяется для указания путей к файлам большого размера, хранение которых непосредственно в БД сказывается на скорости доступа к данным. «Именованное» поле – это строковое поле, уникально идентифицирующее запись данной таблицы. При формировании условий запроса пользователь будет иметь возможность выбрать интересующие его значения такого поля из списка. Кроме того, если такое поле используется как поле-подстановка (lookup) в представлении, то при выборе записи будет задействован механизм контекстной фильтрации значений поля (функция autocomplete). «Списочный» вид указывает, что строковое поле может принимать ограниченное число значений (например, «да», «нет»). Необходимость использования таких полей возникает при описании БД, в которых разработчик поленился организовать справочник для хранения списка возможных значений (при разработке ПБД на основе унаследованных БД). При формировании условий запроса пользователь, тем не менее, будет иметь возможность выбрать интересующие его значения такого поля из списка.

Для описания blob-полей также существует два вида: «графический» и «документный». Если вид поля таблицы определён как «графический», то при просмотре записи в виде формы пользователю будет доступно изображение из файла хранящегося в данном поле. Для «документных» полей дополнительно требуется указание расширения

файла для реализации загрузки и просмотра содержимого поля соответствующими приложениями.

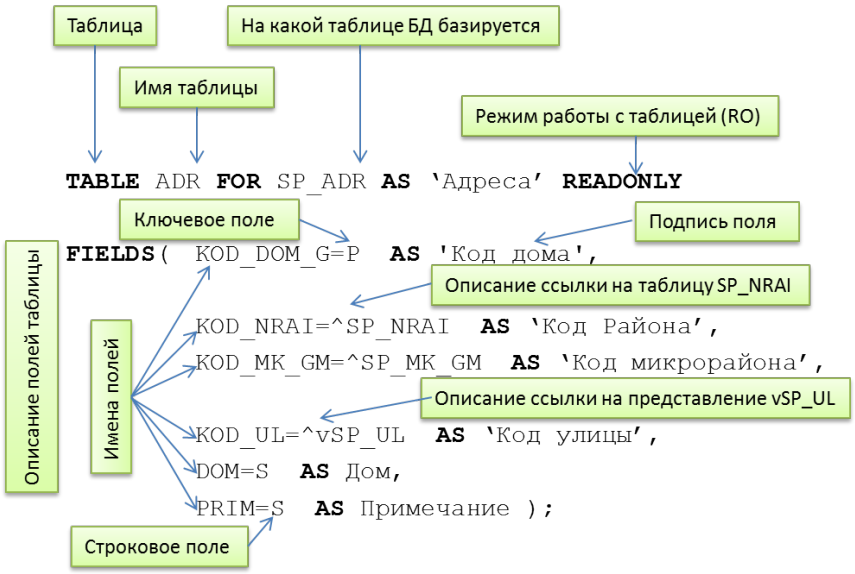

Рисунок 19 - Пример описания таблицы

Кроме вида в описании полей может быть включено указание на их роль во взаимодействии с пространственными объектами (объектами цифровых карт). Например, группа полей таблицы (представления) может быть описана как адресная (поля содержащие значения «Улица» и «Дом»). Тогда пользователю при работе с записями такой таблицы будет доступна возможность поиска и отображения пространственных объектов с соответствующими значениями семантик. Для полей таблиц, содержащих координаты пространственных объектов, в спецификации можно определить способ создания соответствующих объектов на цифровой топооснове, что позволяет реализовать автоматическое создание цифровых карт на основе записей таблиц БД.

Все современные СУБД предоставляют информацию о наличии связей между таблицами, но при создании приложений её недостаточно для понимания того, как должна быть реализована работа пользователя со связанными таблицами. В спецификации при описании таблиц, кроме наличия внешних ключей (ссылок на справочники), могут быть указаны связи типа «мастер-детали», что позволяет автоматически создать формы обеспечивающие при просмотре записей справочников доступ к записям-деталям (записям из таблиц содержащих ссылку на данную запись справочника). При этом поддерживается описание связей типа «мастер-детали» через несколько уровней ссылок.

В рамках управления жизненным циклом научных данных построение предлагаемых спецификаций позволяет автоматизировать создание систем сбора данных, а также организовать курирование научных данных, за счет консолидации метаданных, как структурных, так и содержательных.

Бинарные форматы данных являются существенно более эффективными, чем текстовые, как по расходу памяти, так и по скорости обращения к данным. Поэтому научные данные, особенно те, которые собираются в больших объёмах, часто могут быть представлены в бинарных форматах. Это могут быть, как широко известные форматы (например, TIFF, HDF5), так и оригинальные форматы, разрабатываемые специально для уникальной научной установки. Для последнего случая характерно отсутствие документации о формате или её отставание от последних версий формата. Часто информацией о формате владеет чрезвычайно узкая группа исследователей, прекращение работы которой может привести к утрате возможности работать с соответствующими данными. Для решения этой проблемы разработан подход, основанный на использовании спецификаций.

Для спецификации форматов бинарных данных в ИДСТУ СО РАН разработан декларативный язык FlexT (сокращение от Flexible Types). Основными конструкциями языка FlexT являются определения типов данных, которые напоминают определения типов в традиционных языках программирования, но являются более гибкими. Так, типы данных в языке FlexT могут содержать составляющие переменного размера и иметь параметры,

Основным назначением интерпретатора языка FlexT является отображение содержимого бинарных данных в соответствии со спецификацией формата в понятном для человека виде. Чаще всего следующим шагом при изучении некоторого формата данных является написание кода для работы с ним. Поэтому был разработан генератор кода для чтения данных, позволяющий полностью автоматизировать этот процесс для значительной доли описанных форматов. Автоматически генерируется, как модуль чтения данных, так и тестовая программа, демонстрирующая его правильное использование для отображения данных. Примеры работы генератора кода для описания формата STL на рисунке 20 показаны на рисунках 21, 22.

Использование спецификаций бинарных форматов данных облегчает создание кода для работы с ними, а для оригинальных научных форматов является необходимым условием сохранения собираемых в этих форматах данных для будущих поколений исследователей.

Развернута инфраструктура проекта, включая систему совместной разработки программного обеспечения (доступна по адресу https://git.icc.ru), систему управления проектом (доступна адресу https://redmine.icc.ru).

```
                Data
        0 array[5] of char Hdr0

assert not(Hdr0='solid'); //ignore text
                 STL

           include Float.rfi

                type

      TSTLPoint array[3]of TSingle

            TSTLFace struc
     TSTLPoint Normal //Normal vector
       array[3]of TSTLPoint Vertex
     Word Attr //Attribute byte count
                ends

                data
         5 array[75] of char Hdr1
              80 ulong Count

assert 84+Count*TSTLFace:Size=FileSize;

                data
    84 array[Count] of TSTLFace Faces
```

Рисунок 20 - Описание формата STL.

```
Unit STL;
interface
uses
  SysUtils, FmtSys;

type
  PUlong = ^ulong;
  ulong = LongWord;

  PHdr0 = ^THdr0;
  THdr0 = array[0..4]of AnsiChar;

  PSTLPoint = ^TSTLPoint;
  TSTLPoint = array[0..2]of Single;

  PSTLFaceVertex = ^TSTLFaceVertex;
  TSTLFaceVertex = array[0..2]of TSTLPoint;

  PSTLFace = ^TSTLFace;
  TSTLFace = packed record
          Normal: TSTLPoint;
          Vertex: TSTLFaceVertex;
          Attr: word;
  end;

  PHdr1 = ^THdr1;
  THdr1 = array[0..74]of AnsiChar;

  PFaces = ^TFaces;
  TFaces = array[0..0]of TSTLFace;

  TTFacesAccessor = class(TSimpleArrayAccessor)
   protected
          class function GetItemSize: Integer; override;
          function GetTbl: PFaces;
   public
          function Fetch(AIndex: Integer): PSTLFace;
          constructor Create(AOwner: TComplexDataAccessor;
           AOfs: TOffset; AIndex: Integer; Ap_Count: Integer);
          property Tbl: PFaces read GetTbl;
          property p_Count: Integer read GetCount;
  end;

  TSTLReader = class(TBaseDataReader)
   protected
          FHdr0Ptr: PHdr0;
          FHdr1Ptr: PHdr1;
          FCountPtr: PUlong;
          FFacesPtr: PFaces;
          FFaces: TTFacesAccessor;
          procedure SetItem(AIndex: Integer;
           V: TDataAccessor); override;
          function GetItem(AIndex: Integer):
      TDataAccessor; override;
          function GetItemCount: Integer; override;
   public
          function Hdr0: PHdr0;
          function Hdr1: PHdr1;
          function Count: LongWord;
          function Faces: TTFacesAccessor;
          constructor Create(const AFileName: String);
          property CountPtr: PUlong read FCountPtr;
          property FacesPtr: PFaces read FFacesPtr;
  end;

function THdr0ToStr(V: PHdr0): String;
function THdr1ToStr(V: PHdr1): String;
```

Рисунок 21. - Интерфейсная часть модуля чтения данных формата STL.

```
#include <typeinfo>
#include <iostream>
#include <memory>
#include "FmtSys.h"
#include "STL.h"
#pragma hdrstop

using namespace std;
using namespace STL;

int main(int argc, char* argv[]){
      TSTLReader * Reader;
        std::string FN;
          int i;
          int i0;
          int i1;
          int i2;
        PSTLPoint V;
        PSTLFace V0;
        if (argc-1<=0) {
          cout<<"Usage:"<<endl;
    cout<<"  "<<extractFileName(argv[0])<<
            " <STL file>"<<endl;
                exit(1);
          }
        FN = argv[1];
          try {
                {
          std::unique_ptr<TSTLReader>
  must_free_Reader(new TSTLReader(FN));
          Reader = must_free_Reader.get();
cout<<"Hdr0: "<<THdr0ToStr(Reader->Hdr0())<<endl;
cout<<"Hdr1: "<<THdr1ToStr(Reader->Hdr1())<<endl;
      cout<<"Count: "<<Reader->Count()<<endl;
          cout<<"Faces:"<<endl;
    for (i=0; i<Reader->Faces()->Count(); i++) {
        V0 = Reader->Faces()->Fetch(i);
         cout<<"  ["<<i<<"]:"<<endl;
        cout<<"        Normal:"<<endl;
         for (i0=0; i0<3; i0++)
         cout<<"        ["<<i0<<"]: "<<
            V0->Normal[i0]<<endl;
        cout<<"        Vertex:"<<endl;
            for (i1=0; i1<3; i1++) {
             V = &V0->Vertex[i1];
      cout<<"        ["<<i1<<"]:"<<endl;
            for (i2=0; i2<3; i2++)
  cout<<"        ["<<i2<<"]: "<< *V[i2]<<endl;
                }
    cout<<"        Attr: "<<V0->Attr<<endl;
                }
                }
} catch(const std::exception& E) {
          cout<<E.what()<<endl;
          }
  std::cin.ignore(0x100,'\x0A');
          }
```

Рисунок 22 - Тестовая программа отображения содержимого файла в формате STL.

ЗАКЛЮЧЕНИЕ

Проект был включен в госзадание в сентябре 2019 г. Все запланированные исследования в рамках этапа НИР 2019 г. выполнены в соответствии с государственным заданием ИДСТУ СО РАН на 2019-2021 гг. по теме «Информационно-телекоммуникационная платформа цифрового мониторинга озера Байкал, на основе сквозных технологий». Содержание НИР раскрыто в Планах научно-исследовательских работ ИДСТУ СО РАН на 2019 год

В процессе выполнения работ на этапе НИР 2019 г. получены следующие

- Разработана концепция доступа к высокопроизводительным вычислительным ресурсам и ресурсам хранения данных центров коллективного пользования.
- Исследованы и разработаны эффективные методы и технологии сбора, обработки и анализа больших объёмов разноформатных пространственно-временных данных, основанных на интеллектуализации, машинном обучении и использовании конструктивных средств спецификации.
- Разработаны сервисы предобработки и обработки больших объемов данных ДЗЗ.
- Разработаны сервисы обработки неструктурированных данных.
- Разработаны сервисы управления жизненным циклом научных данных.

Запланированные на 2019 год задачи выполнены полностью и создают основу фундаментальных исследований методов и технологий для разработки единой сервис-ориентированной платформы цифрового мониторинга озера Байкал, основанной на сквозных технологиях и обеспечивающей сбор, хранение и мониторинг больших массивов разноформатных распределенных междисциплинарных научных данных, их анализ и прогнозирование развития природных и техногенных процессов на основе комплекса математических моделей и методов машинного обучения.

Полученные в проекте методы и технологии позволят создать оригинальную распределенную сервисно-ориентированную платформу хранения, обработки больших объёмов разноформатных научных данных и знаний для поддержки процессов непрерывного мониторинга крупных озерных систем, их междисциплинарных исследований и прогнозирования развития возможных событий. Ожидаемые результаты исследований будут способствовать переходу к новым цифровым, интеллектуальным производственным технологиям, автоматизированным системам нового поколения, позволят повысить эффективность создаваемых и внедряемых распределенных информационно-вычислительных технологий, в том числе технологий обработки больших

объемов пространственно-временных данных, а также позволят повысить качественный уровень проведения междисциплинарных научных исследований. Разрабатываемые методы и сквозные технологии найдут широкое применение в различных областях человеческой деятельности, в том числе для формировании систем поддержки принятия решений органов государственной власти и местного самоуправления для решения проблем эффективного управления социально-эколого-экономическим развитием территорий, снижения рисков возникновения и сокращения неблагоприятных последствий техногенных и природных катастроф.

По результатам этапа НИР 2019 года опубликовано **9** публикаций. Из них 2 статьи в изданиях, включенных в международную базу цитирования Scopus (Приложение А).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Израэль Ю. А. Экология и контроль состояния природной среды. // – Л.: Гидрометеоиздат, 1979, – 376 с. 5.
2. Бычков И.В., Ружников Г.М., Хмельнов А.Е. [и др.] Инфраструктура информационных ресурсов и технологии создания информационно-аналитических систем территориального управления. – Новосибирск: Издательство СО РАН, – 2016. – 240 с.
3. Сутурин А.Н., Чевыркин Е.П., Мальник В.В., Ханаев И.В., Минаев А.В., Минаев В.В. Роль антропосферных факторов в развитии экологического стресса в литорали озера Байкал (акватория пос. Листвянка) // География и природные ресурсы, – 2016, №6, – С.43-54.
4. Кравцова Л.С. Ижболдина Л.А., Ханаев И.В. Помазкина Г.В. Домышева В.М., Кравченко О.С., Грачёв М.А. Нарушение вертикальной зональности зелёных водорослей в открытом Лиственничном заливе озера Байкал, как следствие антропогенного воздействия // Док. РАН. – 2012. – Т.447, №2. – С. 227-229.
5. Грачёв М.А. О современном состоянии экологической системы озера Байкал // Новосибирск: Изд-во СО РАН, 2002. – 156 с.
6. Самсонов Д.П., Кочетков А.И., Пасынкова Е.М. [и др.] Содержание стойких органических загрязнителей в компонентах уникальной экологической системы озера Байкал // Метеорология и гидрология. – 2017. – №5. – С. 105-115.
7. NOAA`s Nartional Centers for Environmental Information. https://www.ngdc.noaa.gov (дата обращения 04.12.2019).
8. Center for International Earth Science Information Network (CIESIN) // http://www.ciesin.org (дата обращения 04.12.2019).
9. Биоразнообразие и динамика экосистем: информационные технологии и моделирование. Отв. ред. акад. В.К. Шумный, акад. РАН Ю.И. Шокин. – Новосибирск: Изд-во СО РАН. – 2006. – 648 с.

ПРИЛОЖЕНИЕ А

Список публикаций по проекту

1. Igor V. Bychkov, Gennadii M. Ruzhnikov, Alexey E. Hmelnov, Roman F. Fedorov, Taras I. Madzhara, Anastasia K. Popova Digital Monitoring of Lake Baikal and its Coastal Area // 2nd Information Technologies: Algorithms, Models, Systems (ITAMS 2019), CEUR Workshop Proceedings, 2019, Vol.2463, pp. 13-23. **(Scopus)**
2. Roman K. Fedorov Building Service Composition based on Statistics of the Services Use // 2nd Information Technologies: Algorithms, Models, Systems (ITAMS 2019), CEUR Workshop Proceedings, 2019, Vol.2463, pp. 40-46. **(Scopus)**
3. Бычков И.В., Ружников Г.М., Хмельнов А.Е., Фёдоров Р.К., Маджара Т.И. Цифровой мониторинг экосистемы озера Байкал // Труды всероссийской конференции с международным участием Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов. Бердск. 26-30 августа 2019. С. 8-14. **(РИНЦ)**
4. Фёдоров Р.К, Китов А.Д., Авраменко Ю.В. Автоматизация формирования базы данных ледников на основе ДЗЗ // Труды всероссийской конференции с международным участием Обработка пространственных данных в задачах мониторинга природных и антропогенных процессов. Бердск. 26-30 августа 2019. С. 207-211. **(РИНЦ)**
5. Попова А.К. Разработка веб-сервиса прогнозирования динамики лесных ресурсов // Восьмая Международная конференция «Системный анализ и информационные технологии» САИТ – 2019 (8 – 14 июля 2019 г., г. Иркутск - Листвянка, Россия): Труды конференции. М.: ФИЦ ИУ РАН, 2019. С. 573-578.
6. Бычков И.В., Ружников Г.М., Хмельнов А.Е., Фёдоров Р.К., Маджара Т.И., Шигаров А.О., Попова А.К. Информационно-телекоммуникационная платформа цифрового мониторинга озера Байкал // Восьмая Международная конференция «Системный анализ и информационные технологии» САИТ – 2019 (8 – 14 июля 2019 г., г. Иркутск - Листвянка, Россия): Труды конференции. М.: ФИЦ ИУ РАН, 2019. С. 26-33.
7. Popova A.K. Online service for modeling forest dynamics on the geoportal // Proceedings of IX International conference on Applied Internet and Information Technologies, Zrenjanin, Republic of Serbia, October 3-4, 2019. Pp. 283-288.
8. Федоров Р.К. Сервисы публикации картографических и реляционных баз // Материалы Междунар. научно-практ. конф., посвященной памяти чл.-корр. РАН А.Н. Антипова (Иркутск, 23-27 сентября 2019 г.). 2019. С. 72-76.
9. Гаченко А.С., Хмельнов А.Е. Технология цифрового моделирования фрагментов рельефа озера Байкал и прилегающей акватории // Тр. VI Междунар. научной конф. "Региональные проблемы дистанционного зондирования Земли" (Красноярск, 10-13 сентября 2019 г.). 2019. С. 205-207. **(РИНЦ)**