**Веб-ориентированная информационная система управления данными результатов исследования микробиома озера Байкал**

С целью определения функциональных требований и проектирования архитектуры веб-ориентированная информационная система управления данными результатов исследования микробиома озера Байкал (ИС) проведен анализ сложившейся практики в ЛИН СО РАН работы с данными в исследованиях микробиома, охватывая следующие вопросы планирования и исполнения рабочих процессов (вычислительных цепочек) обработки и анализа метагеномных данных: используемое программное обеспечение, объемы и скорость генерации данных, принятие решений, слабо автоматизированные процессы (в которых активно используется «ручная» обработка данных).

Также выполнен обзор известного программного обеспечения, предназначенного для работы с данными в исследованиях микробиома в рамках парадигмы Больших Данных. Выполнен сравнительный обзор наиболее популярных современных платформ конвейерной обработки данных, применяемых в научных исследованиях с интенсивным использованием данных, в том числе, Rapidminer, KNIME, Pipline Pilot, Taverna. Сравнение произведено по таким характеристикам, как модульность, требования к наличию навыков программирования у исследователя, открытость исходного кода, интеграция с языком программирования R, случаи применения в геномике и метагеномике.

В результате подготовлены предложения по включению известного программного обеспечения в планируемый инструментарий, схем и форматов данных, разработке инструментов планирования и исполнения вычислительных технологических цепочек аналитики данных в исследовании микробиома оз. Байкал, хранению и обработки больших объемов метагеномных данных в распределенной среде на основе платформы Hadoop ([https://hadoop.apache.org](https://hadoop.apache.org/)).

Анализ сложившейся практики в ЛИН СО РАН показал, что для реализации многоэтапных рабочих процессов обработки и анализа метагеномных данных (pipline) вовлекаются различные программные комплексы обработки последовательностей, статистические пакеты и скриптовые языки программирования. Для исполнения таких цепочек приходится задействовать высокопроизводительные вычислительные среды (суперкомпьютеры). Для проведения таких исследований специалисту требуются навыки работы со сценариями командной оболочки, запуска пакетов в распределенной вычислительной среде, программирования на языках R или Python. Как правило, для микробиологов реализация подобных цепочек вызывает затруднения.

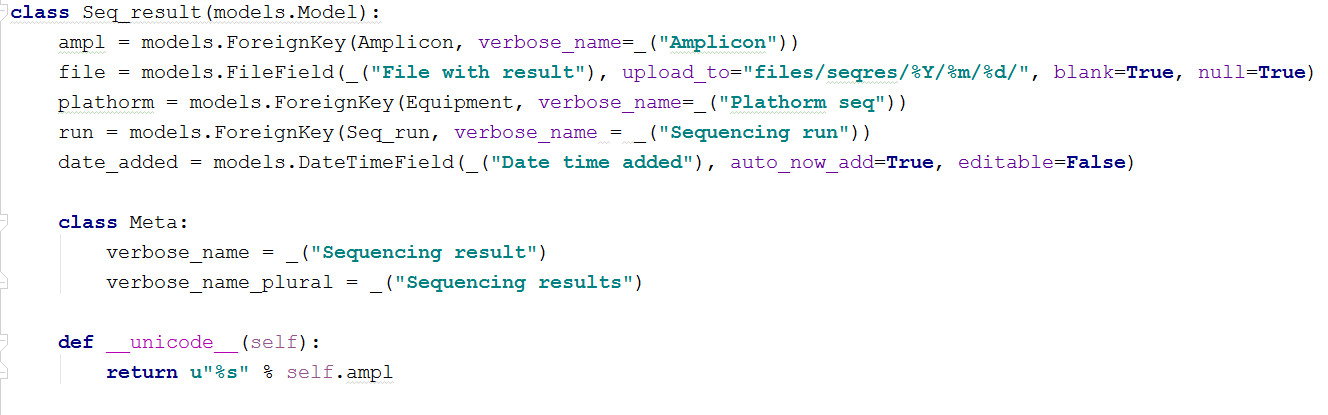
В настоящее время в ЛИН СО РАН для получения таблицы встречаемости используется программный комплекс MOTHUR (<https://www.mothur.org>) с интерфейсом командной строки. Некоторые этапы (например, кластеризация прочтений) требовательны к вычислительным ресурсам. Их выполнение на локальной рабочей машине пользователя занимает продолжительное время. Для решения этой проблемы активно используются ресурсы Иркутского суперкомпьютерного центра СО РАН (https://hpc.icc.ru). Цепочки обработки и анализа метагеномных данных выполняются на вычислительном кластере «Академик В.М. Матросов». Текущая реализация этого процесса сопряжена с рядом сложностей: микробиологам ЛИН СО РАН приходится удаленно подключаться к вычислительному кластеру по протоколу SSH (Secure Shell); каждая команда вводится пользователем вручную в командной строке; хранение получаемых данных не унифицировано, зависит полностью от пользователя, что усложняет доступ к данным и их повторное использование другими исследователями.

В рамках проекта разработан прототип веб-ориентированной информационной системы управления данными результатов исследования микробиома озера Байкал (ИС), направленная на преодоление перечисленных недостатков. ИС преследует следующие цели: 1) управление данными (распределенное хранение, обеспечение пользовательского и программного интерфейса доступа) результатов исследования микробиома на протяжение их жизненного цикла; 2) планирование и исполнение технологических цепочек обработки и анализа больших объемов метагеномных данных.



**Рис. 1.** Упрощенное представление схемы данных результатов исследования микробиома, ребра, подписанные (1..n), соответствуют отношениям типа один-ко-многим, подпись (n..m) показывает отношения типа многие-ко-многим.

Спроектирована схема хранения данных, получаемых при исследовании микробиома озера Байкал (Рис. 1). Схема охватывает информацию о сборе проб, анализе физико-химических и биологических параметров этих проб, результатов секвенирования, применяемом оборудовании, таксономических базах, методиках анализа собранных материалов, публикациях результатов исследований, а также исследователях, принимающих участие в получении результатов. Также она позволяет хранить используемые вычислительные цепочки обработки и анализа метагеномных данных, включая информацию о программных инструментах, командах и конфигурационных файлов.



**Рис. 2.** Фрагмент модели ИС для инструментальной платформы Django.

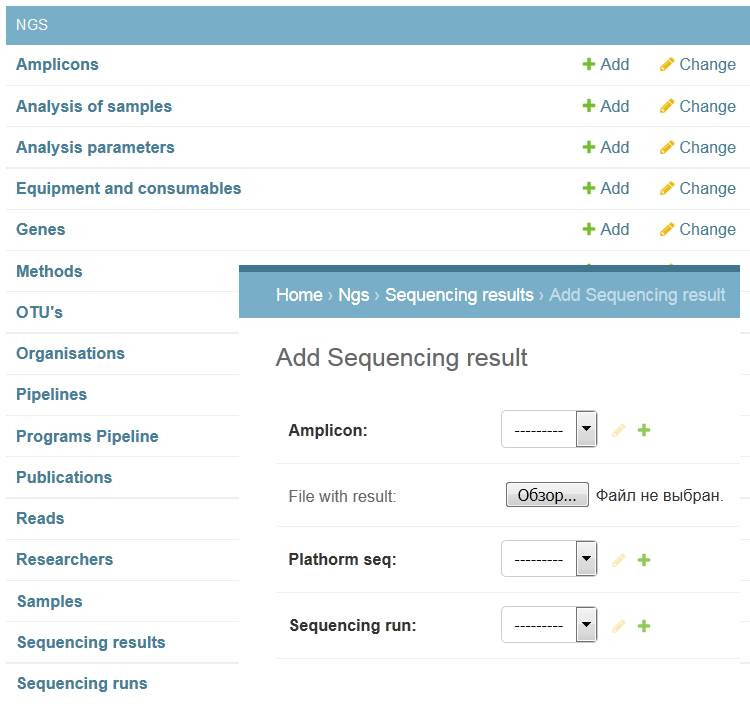
На основе спроектированной схемы разработана информационная модель ИС, обеспечивающая управление данных всего цикла исследования микробиома, начиная от отбора пробы и заканчивая публикацией научных результатов. Модель реализована средствами инструментальной платформы разработки программного обеспечения Django (<https://www.djangoproject.com>). Фрагмент модели представлен на Рис. 2. По модели ИС создана соответствующая структура базы данных в формате системы управления данными PostgreSQL (<https://www.postgresql.org>) и сгенерирован настраиваемый пользовательский веб-интерфейс административной панели, позволяющий на этапе разработки проводить тестирование разработанной модели и тестовое наполнение базы данных результатов исследования микробиома озера Байкал.

В текущей версии модели ИС реализованы следующие сущности данных:

* *Организации* – справочник научных организаций, проводящих исследования.
* *Исследователи* – справочник исследователей с привязкой к аффилированным с ними организациям.
* *Водные объекты* – справочник исследуемых водных объектов.
* *Зоны* – справочник зон водных объектов (например, котловин озера Байкала) с привязкой к водным объектам.
* *Станции* – пространственный список станций (мест отбора проб) с привязкой к водным объектам и их зонам.
* *Типы исследования* – справочник типов исследования (физические, химические и др.).
* *Параметры* – справочник физических, химических и других измеряемых параметров с привязкой к типам исследования.
* *Гены* – список генов.
* *Публикации* – список публикаций с результатами исследований, на которые могут ссылаться другие сущности с привязкой к исследователям, являющимися авторами данных публикаций.
* *Методы исследования* – список методов, используемых в исследованиях с привязкой к используемому оборудованию и расходным материалам, а также публикациям с описанием данных методов.
* *Категории методов* – поисковый справочник категориями методов исследования.
* *Программное обеспечение* – список программных пакетов, используемых в рабочих процессах обработки и анализа метагеномных данных.
* *Таксономические базы данных* – список баз данных, используемых для определения таксономических элементов.
* *Идентификаторы таксонов* – список идентификаторов таксонов с привязкой к таксономическим базам данных.
* *Типы пробы* – справочник типов пробы, (например, донные осадки, подлёдная проба, и др.).
* *Пробы* – список отобранных проб с привязкой к станциям с привязкой к типам проб и станциям отбора проб.
* *Анализы проб* – список результатов анализов проб по определённом параметрам (для одной пробы может быть проведено много разных анализов) с привязкой к собранным пробам, использованному оборудованию, расходным материалам и методам исследования, исследуемым параметрам, а также задействованным исследователям.
* *Типы оборудования* – древовидный классификатор типов оборудования, используемого в исследовании.
* *Оборудование и расходные материалы* – список оборудования и расходных материалов, используемых в исследовании (например, NGS платформы секвенирования), с привязкой к типам оборудования.
* *Рабочие процессы* – список цепочек (pipeline) обработки и анализа метагеномных данных, обеспечивающий их воспроизведение, с привязкой к программному обеспечению.
* *Ампликоны* – список ампликонов с привязкой к пробам, рабочим процессам, методам исследования и генам.
* *Запуски секвенирования* – список выполненных запусков секвенирования с привязкой к исследователям и организациям.
* *Результаты секвенирования* – список результатов секвенирования (файлов с данными секвенирования) с привязкой к запускам секвенирования, оборудованию и расходным материалам.
* *OTU* – список полученных матриц OTU по результатам секвенирования с привязкой к рабочим процессам и идентификаторам таксонов.
* *Прочтения* – список прочтений с привязкой к ампликонам и матрицам OTU.

Пользовательский веб-интерфейс доступа к данным разработанного прототипа ИС настроен с использованием встроенных возможностей инструментальной платформы Django. В пользовательском интерфейсе были реализованы возможности добавления, редактирования и удаления для всех представленных сущностей. Так же настроены элементы управления для поиска, фильтрации, сортировки добавленных данных. Кроме того, при добавлении данных пользователями перед отправкой данных на сервер, осуществляется проверка вводимых данных, согласно заданным правилам, и в случае ошибки ввода данных пользователь видит, где именно произошла ошибка.

В настоящее время прототип ИС развернут в облачной вычислительной инфраструктуре ИДСТУ СО РАН. Опытная версия ИС доступна по адресу <http://ngs.icc.ru:88>. Прототип развёрнут в серверном дистрибутиве операционной системы GNU/Linux CentOS версии 7.3. В качестве СУБД используется PostgreSQL версии 9.4, а веб сервер сконфигурирован в связке инструментов nginx+gunicorn. Такая реализация ИС позволила организовать удаленный доступ к данным исследований микробиома, и в дальнейшем позволит микробиологам самостоятельно управлять данными рабочих процессов обработки и анализа метагеномных данных при исследовании микробиома озера Байкал. В настоящее время начат тестовый ввод имеющихся данных ЛИН СО РАН по исследованию микробиома озера Байкал.



**Рис. 3.** Снимок экрана пользовательского веб-интерфейса опытной версии ИС: форма доступа к данным результатов исследования микробиома озера Байкал.

**Публикации**

1. Морозов А.А., Шигаров А.О., Малков Ф.С., Паскал К.К., Черкашин Е.А., Лихошвай Е.В. Информационно-вычислительная инфраструктура для поддержки метагеномного анализа // Вестник ИГУ. – 2017. *(принята в печать)*
2. Малков Ф.С., Галачьянц Ю.П., Шигаров А.О., Морозов А.А., Михайлов И.С., Ломакина А.В., Захаренко А.С. Управление данными в исследовании микробиома оз. Байкал // Тезисы докладов 18-й Всеросс. конф. молодых ученых по математическому моделированию и информационным технологиям. Иркутск, 2017. С. 82.
3. Малков Ф.С., Черкашин Е.А., Шигаров А.О. Перспективы применения системы управления рабочим процессом Taverna в задачах обработки метагеномных данных при исследовании микробиома озера Байкал // Материалы конференции «Ляпуновские чтения». 2017. С. 32.