Вавиловский журнал генетики и селекции. 2023;27(7):898-905
DOI 10.18699/VJGB-23-104

Перевод на английский язык https://vavilov.elpub.ru/jour

Лабораторные информационные системы для управления исследовательскими работами в биологии

А.М. Мухин[1, 2, 3] ✉, Ф.В. Казанцев[1, 2, 3], С.А. Лашин[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ mukhin@bionet.nsc.ru

Аннотация. Современная исследовательская работа в биологии нередко требует усилий одной или нескольких групп исследователей. Часто это группы специалистов из смежных областей, которые генерируют и обмениваются данными разных форматов и размеров. Без применения современных подходов автоматизации работы и версионирования данных (когда данные от разных сотрудников сохраняются в разные моменты времени) коллективная работа быстро переходит в неуправляемый хаос. В настоящем обзоре приведен ряд информационных систем, предназначенных для решения озвученных задач. Их применение для организации научной деятельности позволяет управлять потоком действий и данных, добиваясь работы всех участников с актуальной информацией, и решением вопроса воспроизводимости как экспериментальных, так и вычислительных результатов. Описаны методики по организации потоков данных в рамках работы коллектива, принципы по организации метаданных и онтологий. Рассмотрены информационные системы Trello, Git, Redmine, SEEK, OpenBIS и Galaxy. Описана их функциональность и сфера использования. Выбирая те или иные инструменты, важно понимать цель внедрения, определить набор задач, которые они должны решать, и исходя из этого формулировать требования и отслеживать применение рекомендаций на местах. Задачи по созданию структуры онтологий, метаданных, схем хранения данных и программных систем являются ключевыми для коллектива, который решился на проведение работ по автоматизации оборота данных. Не всегда возможно внедрить такие системы целиком, но все же следует стремиться к этому через поэтапное внедрение принципов по организации данных и задач с освоением отдельных программных инструментов. Следует отметить, что системы Trello, Git и Redmine проще в использовании, настройке и поддержке для малых исследовательских групп. В то же время SEEK, OpenBIS и Galaxy более специфичные, их применение целесообразно в случае, если возможностей простых систем уже недостаточно.
Ключевые слова: управление; LIMS; ELN; FAIR; системы контроля версий; Trello; GitHub; Redmine; SEEK; OpenBIS; Galaxy.

Для цитирования: Мухин А.М., Казанцев Ф.В., Лашин С.А. Лабораторные информационные системы для управления исследовательскими работами в биологии. *Вавиловский журнал генетики и селекции*. 2023;27(7):898-905. DOI 10.18699/VJGB-23-104

Laboratory information systems for research management in biology

A.M. Mukhin[1, 2, 3] ✉, F.V. Kazantsev[1, 2, 3], S.A. Lashin[1, 2, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
✉ mukhin@bionet.nsc.ru

Abstract. Modern investigations in biology often require the efforts of one or more groups of researchers. Often these are groups of specialists from various scientific fields who generate and share data of different formats and sizes. Without modern approaches to work automation and data versioning (where data from different collaborators are stored at different points in time), teamwork quickly devolves into unmanageable confusion. In this review, we present a number of information systems designed to solve these problems. Their application to the organization of scientific activity helps to manage the flow of actions and data, allowing all participants to work with relevant information and solving the issue of reproducibility of both experimental and computational results. The article describes methods for organizing data flows within a team, principles for organizing metadata and ontologies. The information systems Trello, Git, Redmine, SEEK, OpenBIS and Galaxy are considered. Their functionality and scope of use are described. Before using any tools, it is important to understand the purpose of implementation, to define the set of tasks they should solve, and, based on this, to formulate requirements and finally to monitor the application of recommendations in the field. The tasks of creating a framework of ontologies, metadata, data

warehousing schemas and software systems are key for a team that has decided to undertake work to automate data circulation. It is not always possible to implement such systems in their entirety, but one should still strive to do so through a step-by-step introduction of principles for organizing data and tasks with the mastery of individual software tools. It is worth noting that Trello, Git, and Redmine are easier to use, customize, and support for small research groups. At the same time, SEEK, OpenBIS, and Galaxy are more specific and their use is advisable if the capabilities of simple systems are no longer sufficient.
Key words: management; LIMS; ELN; FAIR; version control systems; Trello; GitHub; Redmine; SEEK; OpenBIS; Galaxy.

For citation: Mukhin A.M., Kazantsev F.V., Lashin S.A. Laboratory information systems for research management in biology. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):898-905. DOI 10.18699/VJGB-23-104

Введение

Современная исследовательская работа в биологии нередко требует усилий одной или нескольких групп исследователей. Часто это группы специалистов из смежных областей, которые генерируют и обмениваются данными разных форматов и размеров. Для автоматизации и компьютерной поддержки этой работы используют различные инструменты каталогизации, протоколирования хода протекания экспериментов и фиксации результатов: бумажные блокноты и лабораторные журналы, программы ведения электронных таблиц, составление отчетов в разных текстовых редакторах. Без применения современных подходов автоматизации работы и версионирования данных в коллективе быстро наступает «неуправляемый хаос». Критическим местом организации взаимодействия в коллективе является сложность процедуры передачи знаний от одного члена команды другому, так как такие знания не формализованы и часто содержат пометки, понятные только автору. Все это приводит к задержкам в проведении следующих этапов исследования или в оформлении публикаций. Иногда сотрудники забывают записывать новые факты и заметки либо вообще не ведут никакого учета промежуточных этапов работы. Это приводит к безвозвратным потерям знаний и тратам ресурсов на повторные эксперименты и наблюдения.

При сборе первичных данных исследователи могут также допускать ошибки в обработке значений или приписывании их к той или иной категории. Например, транскриптомные данные могут быть ошибочно приписаны к организму, отличному от того, откуда они получены; данные могут быть записаны не в унифицированном виде, с использованием значений разных типов (целое число, число с плавающей точкой, строка, дата и т. п.). При работе с Excel может произойти ошибочное преобразование строк в числа с плавающей точкой, что критично для интерпретации результатов исследования (Zeeberg et al., 2004), поэтому неявных преобразований данных необходимо избегать. В статье (Roche et al., 2015) были проанализированы биоресурсные коллекции (БРК) в области Экологии и Эволюции. Выяснилось, что 56 % этих БРК были неполными, т. е. в табличных данных были пустые значения, а 64 % собраны таким образом, что повторно использовать хранящиеся данные невозможно ввиду ошибок записи значений.

Поэтому перед каждым коллективом стоит задача по грамотной формализации процессов управления данными и обмена знаниями между сотрудниками. Далее мы рассмотрим конкретные методологии организации данных и реализующие их информационные системы и программные инструменты, которые используются научными организациями для распределения задач и автоматизации потока рабочих данных.

Методологии организации данных и процессов

Для решения задачи организации потоков научных данных существует несколько путей, но все они требуют от коллектива исследователей создания систем договоренностей по управлению, обработке и передаче научной информации. Системы автоматизации с предоставлением управляемого доступа помогают в сохранении знаний, регламентов и других «сущностей» лабораторной работы, не требуют постоянных согласований. В самом начале этих работ встают следующие вопросы: 1) использование существующих стандартов оформления данных, разработанных профессиональным сообществом; 2) формализация или создание единого «рабочего языка» внутри коллектива; 3) развертывание, внедрение и сопровождение информационной системы и настройка прав доступа для групп пользователей.

Переход на существующие стандарты и форматы представления данных или создание собственных форматов с исчерпывающей документацией, достаточной для однозначной интерпретации значений, позволяет преодолеть проблему передачи знаний между сотрудниками внутри коллектива и вне его. Сопроводительная документация будет использована для автоматизации работы с информационной системой, например для построения модулей генерации сводных диаграмм и отчетов. Формальные схемы описания результатов научной деятельности в последнее время полезны для быстрого поиска информации и интерпретации этих файлов не только машинами, но и людьми. В качестве примеров могут служить математические модели в форматах SBML (Hucka et al., 2019), SBGN (Novère et al., 2009), поддерживаемые сообществом CO.MBINE (Schreiber et al., 2015). Отметим также подход MIRIAM для описания целостных биохимических систем (Novère et al., 2005) и формат MIAME (Brazma et al., 2001) для описания результатов секвенирования на микрочипах или РНК-последовательностей.

Когда определены стандарты представления данных, наступает этап формализации или создания единого рабочего языка и протоколов обмена внутри коллектива для упорядочения передачи знаний предметной области. Если оставить подход к оформлению данных «как удобно/как раньше», то вопрос с неоднозначными или пропущенными

знаниями в базе не будет решен, что в дальнейшем приведет к дополнительным затратам ресурсов на исправление данных на более поздних стадиях работы. В решении задачи формализации и создании единого рабочего языка могут помочь инструменты с онтологиями (Guizzardi, 2020). Онтологии являются более широким классом систем организации знаний по описанию результатов в сравнении с вышеупомянутыми формальными схемами. В системах онтологий можно устанавливать «понятия» и «отношения» между понятиями, а не строго следовать за готовой схемой, предложенной кем-то ранее. Онтологии создаются с целью описания смысловой информации и однозначной интерпретации системы понятий и процессов внутри коллектива и за его пределами. Коллективы используют как простые методы описания онтологий, такие как язык логики первого порядка, так и более сложные древовидные структуры, например OntoUML (Guizzardi et al., 2018) или схемы RDF (Gutierrez et al., 2007). Для составления онтологических связей предметной области также набирает популярность математическая теория категорий (Kuś, Skowron, 2019), призванная соединять различные области математики и предметные области друг с другом. Был реализован графический язык «онтологических журналов» (англ. Ologs, по сути описания предметной области в виде графов, где в узлах описаны объекты с определенными свойствами, а в ребрах – функции по преобразованию из одного объекта в другой) с использованием основ данной теории (Spivak, Kent, 2012). В настоящее время инструментарий и язык теории категорий не используются широко в научных публикациях и системах, однако есть работы по реализации этого языка в нейробиологии (Brown, Porter, 2003) и по математическому описанию развивающейся модели памяти (Ehresmann, Vanbremeersch, 2007).

Одним из путей формализации этапов работы лаборатории является создание метаданных – информации, описывающей сами данные (Roche et al., 2015). Формат их описания довольно свободный. Метаданные могут быть описаны/представлены в виде структурированного файла (XML или JSON) или таблиц баз данных как реляционной (Postgrespro.ru), так и документно-ориентированной структуры (MongoDB.com). Описанием может быть любая информация, например: что означают колонки в таблицах, какие используются единицы измерений, из какого организма были получены материалы, каким образом получались эти результаты. Метаданные могут дополнять системы онтологий и формальные схемы представления научных результатов для быстрого поиска нужной информации и однозначного интерпретирования результатов.

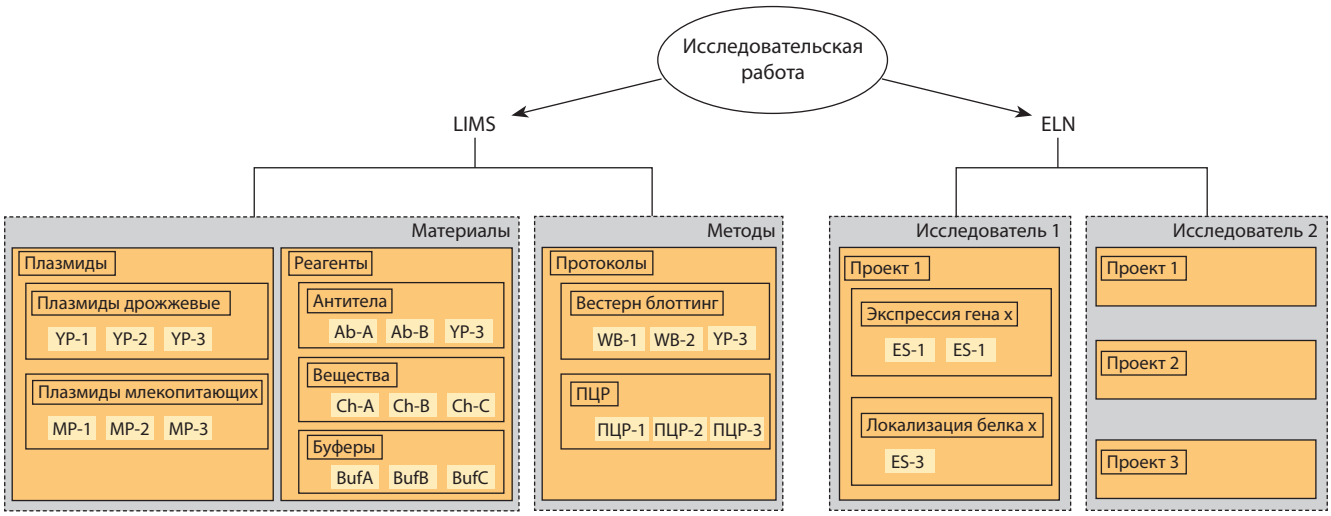
Сообщество исследователей FAIR (Wilkinson et al., 2016) предложило свой набор принципов описания данных и метаданных в задачах хранения и передачи информации как между коллективами исследователей, так и между различными программами анализа данных. Ими были сформулированы следующие четыре принципа, которыми должна обладать лабораторная информационная система:

1. Определенность (Findable) – (мета)данные уникальные и однозначно определяемые. Система должна обладать базовым механизмом чтения подробного описания и возможностью искать эти данные по ключевым полям.
2. Доступность (Accessible) – данные доступны для чтения как людям, так и компьютерам для дальнейшей работы. Достигается с помощью стандартных форматов и протоколов.
3. Интерпретируемость (Interoperable) – (мета)данные описаны в машиночитаемом виде, в удобном формате и аннотированы с помощью онтологии.
4. Повторная используемость (Reusable) – (мета)данные достаточно хорошо описаны, чтобы передавать эти данные другим людям и системам для дальнейшего анализа. Этот пункт является логичным следствием выполнения вышеупомянутых пунктов.

Далее рассмотрим программные инструменты для решения задач управления данными и автоматизации исследовательских работ.

Программные инструменты

Две концепции – LIMS и ELN (Barillari et al., 2016), которые реализовываются в программных комплексах для задач контроля выполнения исследовательских работ, приведены на рисунке.



Описание структуры данных, которые хранятся в LIMS и ELN системах.

LIMS (Laboratory Information Management System) – система управления лабораторной информацией. В ее задачи входят управление и контроль за лабораторными материалами и методами. С помощью этой системы исследователи могут осуществлять документооборот с администрацией и компаниями, создавать расписание использования инструментов, учет реактивов, объектов исследований и др.

ELN (Electronic Laboratory Notebook) – электронный лабораторный журнал. В задачи таких систем входит управление проектами, экспериментами, пользователями, исследовательскими группами, а также протоколирование (журналирование) и контроль проведения экспериментов. По сути, эти системы заменяют функции бумажных блокнотов для ведения и передачи заметок по ходу экспериментов.

Trello

Trello (https://trello.com/) является условно-бесплатным веб-сервисом по организации рабочего процесса и коммуникации. В этой системе пользователи настраивают виртуальную доску, на которой располагаются «карточки» с «заданиями». Сама доска разделена на участки, между которыми перемещаются эти карточки, демонстрируя движение по этапам работ. Чаще всего участки доски помечают статусами выполнения работ, например: «задачи в очереди», «в работе», «ждут отклика», «задача выполнена». Возможно самостоятельно создавать участки/разделы по своему сценарию, наиболее отражающему рабочий процесс коллектива. Таким образом, сотрудники и руководители могут: 1) наблюдать в режиме реального времени за прогрессом работ; 2) изменять статусы задач, добавлять к задачам комментарии; 3) связывать друг с другом задачи; 4) реагировать на ранних этапах в случаях зависших работ.

К недостаткам Trello можно отнести невозможность модифицировать функционал системы собственными модулями и ограниченную функциональность в бесплатной версии. Аналогами можно считать решения Яндекс.Трекер (https://cloud.yandex.ru/services/tracker), GutHub Projects (https://docs.github.com/en/issues/planning-and-tracking-with-projects/learning-about-projects/quickstart-for-projects) и Kanboard (https://kanboard.org/). Предложенные инструменты ориентированы на реализацию требований ELN, однако пользователи могут адаптировать их под решение задач LIMS. Они направлены на управление процессами работы коллектива. Для организации хранения и перемещения самих данных надо использовать другие инструменты.

GitHub

При совместной работе коллектива над кодами программ, документами и отчетами стоит важная задача по контролю за изменениями. Почтовые клиенты и пересылка по сети от человека к человеку плохо справляются с этой задачей, так как самим пользователям нужно контролировать актуальность версий этих документов. Также не решается задача версионирования данных и текста ввиду отсутствия системы централизации хранения файлов и фиксации их изменений. Именно эти задачи можно решить с помощью программы Git (Chacon, Straub, 2014).

Программа Git создает в локальной папке файлы репозитория, позволяющие перемещаться между изменениями в файлах. Как правило, данную систему используют программисты для одновременной работы над проектом, сравнивая и объединяя изменения кода от разных разработчиков. Проекты с открытым кодом обычно хранятся публично в серверах проекта GitHub (https://github.com). Некоторые исследовательские группы используют систему контроля версий Git для подготовки статей и диссертаций. К примеру, с их помощью писалась математическая книга по гомотопической теории типов (The Univalent Foundations Program, 2013). Над книгой работало около 20 человек, и сервис облачного хранения Dropbox не справлялся с задачей синхронизацией текста. В результате команда выпустила книгу объемом 600 страниц менее чем за полгода (https://math.andrej.com/2013/06/20/the-hott-book/).

Сам GitHub нельзя установить на локальном компьютере, однако есть аналогичные решения с возможностью установки в локальное хранилище, например GitLab (https://gitlab.com), Gogs (https://gogs.io), Gitea (https://gitea.com), GitWeb (https://git-scm.com/docs/gitweb). В рамках этих систем возможно решать задачи ELN и задачи LIMS, но пользователям придется подробно разобраться с Git.

Redmine

Redmine (https://redmine.org/) используется в качестве системы контроля проектов и распределения задач. Чаще всего главный управляющий проекта (менеджер, заведующий лабораторией и т. д.) создает набор задач и назначает ответственных исполнителей. Исполнители по мере выполнения меняют статус готовности задачи. Система автоматически отслеживает состояние задач проекта и строит сводные диаграммы, на которых видно расхождение по срокам между планом и фактическим исполнением. Также в основные функции данной системы входят:

- создание и ограничение ролей – администратор может создать несколько дополнительных ролей пользователей и установить для них правила работы в системе (чтение и/или запись «задачи», вики-страниц и т. д.);
- гибкая система по контролю ошибок – функция широко используется в сфере разработки ПО, когда тестировщики или пользователи добавляют «задачу» вида «ошибка» в систему для оповещения разработчиков;
- календарь и диаграмма Ганта. Служат для отслеживания сроков исполнения задач;
- добавление новостей по проекту с оповещением участников;
- добавление документов и файлов в систему;
- оповещение пользователей по электронной почте или RSS-ленте;
- оформление знаний для каждого проекта в формате Википедии – электронная энциклопедия/справочник в виде интернет-страниц;
- система форумов для каждого проекта – возможность публично обсудить в одном месте решение задач. Воз-

можность быстро пробежать глазами цепочки сообщений по теме;
- учет времени работы над задачами и проектом в целом;
- создание пользовательских форм и полей для дополнительного описания «задач», «проектов», «пользователей» и других сущностей в рамках данной системы.

Для данной системы существует функция по ее разворачиванию в локальной информационной среде (вплоть до персонального компьютера). Также пользователи могут реализовывать новую функциональность через реализацию подмодулей (плагинов). К недостаткам Redmine можно отнести отсутствие доски задач по типу Trello, которая понятна и проста в использовании, а также ограниченность функциональности стандартной версии. Поэтому для полноценной работы приходится устанавливать сторонние подмодули.

На основе программного комплекса Redmine построены рабочие процессы многих коллективов в секторе информационных технологий. В 2019 г. была начата реализация проекта ENVRI-FAIR (Petzold et al., 2019) по объединению ресурсов и данных между кластером Европейской инфраструктуры экологических исследований (ENVRI) и вычислительным облаком Европейской «Открытой науки» (EOSC) с использованием Redmine (эта информация была получена из технической документации данного проекта). На основе Redmine возможно реализовать решение задач и ELN, и LIMS.

Система SEEK

Система SEEK (Wolstencroft et al., 2015) предназначена для управления, распространения и изучения математических моделей и ассоциированных данных системной биологии. SEEK организовывает информацию исследовательского проекта, включающего экспериментальные данные и результаты биоинформатической обработки в рамках структуры из трех сущностей: Исследования, Стадии, Образцы (ISA) (Rocca-Serra et al., 2010). «Исследование» раскрывает суть конкретного проекта (кто выполняет работу, какой институт, время проведения исследования). «Стадия» описывает конкретный этап исследования (экскреция ДНК или белка из ткани исследуемого организма, картирование РНК-прочтений на референсный геном и т. д.). «Образец» – единица результата выполненной работы. Также в системе можно устанавливать ассоциативную связь между образцами.

Достоинством этой системы является связывание данных между собой в рамках вышеописанной структуры с описанием коллектива исследователей, а также переформатирование метаданных в граф знаний RDF (Gutierrez et al., 2007) с помощью сервера Virtuoso (Software, 2022). Метаданные описываются в основном в табличной форме (сокр. ISA-Tab), также есть возможность использования JSON схемы. Для ручной аннотации данных разработчики SEEK предлагают программное обеспечение FightField. Поиск данных по графу RDF с помощью языка запросов SPARQL является гибким в использовании в сравнении с SQL, в котором, помимо написания правил отбора данных, от пользователя требуется вручную расписать список таблиц и способ их объединения. Проблема SQL также в том, что пользователь вынужден оптимизировать свои запросы для быстрого выполнения поиска.

Основным направлением SEEK является хранение и передача математических моделей биологических процессов. Ресурс также позволяет работать с SBML моделями и открывать их в JWS Online (Olivier, Snoep, 2004) и в COPASI (Hoops et al., 2006). Эта система в основном реализовывает требования ELN по биоинформатическим проектам, а LIMS не реализован в ней.

Система OpenBIS

В рамках работы лаборатории перед исследователями стоит задача по созданию протоколов экспериментов, следованию этим протоколам с фиксацией результатов, фиксации событий и т. д. Необходимо выстраивать результаты серии данных в рамках одного проекта, например, связывание экспериментов с различными организмами, их фенотипами, генотипами, средой развития и другими данными. OpenBIS (Bauch et al., 2011) предоставляет функционал по хранению и выстраиванию метаданных под подробное описание экспериментов, их результатов, параметров и т. д. Система OpenBIS состоит из трех модулей: сервер приложения, сервер данных и база метаданных.

- Сервер приложения является точкой доступа для пользователей. Модуль реализовывает доступ к программному комплексу через графический пользовательский интерфейс, а также по HTTP протоколу (для OpenBIS предоставлены библиотеки на языках программирования Python, Java и Matlab для взаимодействия по сети). Для добавления новых функций (например, хранение данных по масс-спектрометрии) OpenBIS предоставляет систему модулей, каждый из которых должен быть реализован на языке программирования Jython. Этот модуль разделяет полномочия среди пользователей (чтение данных, чтение/запись данных).
- Сервер данных выполняет работу по организации хранения первичных данных на дисковых накопителях.
- База метаданных представляет собой систему управления базой данных (СУБД) PostgreSQL. Этот модуль связывает данные в проектах, хранит метаданные, указывает на данные из сервера данных, обеспечивает задачи поиска в данных.
- Возможность ссылок к данным на внешних ресурсах (модуль BigDataLink). Метаданные сохраняются в базе метаданных, при этом исходная информация не хранится на сервере данных, а остается на сторонних ресурсах. Эта функция используется в случае работы с файлами большого размера.
- Расширение функционала с помощью библиотек на Java, Python, JavaScript, Matlab для взаимодействия с системой OpenBIS (получение/загрузка данных, поиск метаданных). Эти библиотеки используют аппаратный интерфейс REST API сервиса OpenBIS; таким образом, можно реализовать модули для взаимодействия с системой на других языках программирования. Может использоваться для реализации автоматизированных вычислений с привлечением хранимых данных из системы OpenBIS.

- Структура хранения данных является иерархической и организована следующим образом: область (space), проект (project), эксперимент/коллекция (experiment/collection), Объект/Образец (Object/Sample), данные (Data Set).
- Для связи объектов и данных друг с другом существует метод по установлению связей «предок–потомок», т. е. система может создавать граф объектов и данных.
- Импорт/экспорт данных в табличном виде.
- Реализация дополнительного функционала самой системы с помощью системы модулей.
- Система выполняет аудит каждого изменения в своих базах данных.
- Семантическое аннотирование данных – описание результатов в удобном и интерпретируемом формате. Для описания семантики используется RDF схема (Gutierrez et al., 2007).
- Интеграция с системой SEEK.

Система OpenBIS хорошо себя зарекомендовала для первичного хранения биологической информации, полученной в ходе экспериментов. В работе (Friedrich et al., 2015) была реализована система по добавлению и учету экспериментальных данных по различным тканям организмов при применении разных препаратов. На первом уровне системы хранения описывается объект исследования (например, определенная мышь в лаборатории, которой ввели конкретный препарат). На втором уровне – определенная биологическая ткань, которую извлекли из объекта. На третьем уровне – последовательности (нуклеотидные или белковые), полученные из исследуемой ткани объекта. Система основывается на требованиях LIMS и ELN, является образцовой их реализацией.

Galaxy

Выше были описаны в основном системы для контроля лабораторных данных, однако для биоинформатических лабораторий задачи стоят точно такие же: контроль за потоком данных, воспроизводимость вычислений, доступ к данным и их сохранение в сервере. Для решения подобных задач была реализована система Galaxy (Galaxy Community, 2022). Galaxy состоит из следующих модулей: 1) сервер с программным и графическим интерфейсом; 2) рабочие процессы, которые и запускают аналитические конвейеры по запросу пользователей. Пользователи могут запускать самостоятельным образом установленные в сервере программы и там же хранить свои данные (последовательности, аннотации, список белков и т. д.).

Доступна реализация вычислительных конвейеров в виде графа, где в вершинах обозначены программы с настроенными параметрами, а связаны они между собой ребрами, которые обозначают направление движения данных от выхода одной программы ко входу другой. Также

Сравнение программных решений

Название системы	Основная сфера работы	Уровни иерархии	Использование метаданных	LIMS	ELN	Используемые средства разработки	Развертывания
Trello	Организация задач в виде заметок на доске (Kanban стиль)	Проект Стадии Задача		–/+	+	Невозможно установить в локальной среде	Не нуждается в развертывании, для локального развертывания требуются другие инструменты (например, Kanboard)
Git	Версионирование текстовых файлов	Свободное	Файлы изменений и дерево «коммитов»	–/+	+/–	Само приложение git, GitHub нельзя установить локально	Для эффективной работы требуются навыки работы с git. Также стоит решить, использовать сторонний сервис (например, GitHub) или разворачивать локальный сервер (GitLab, Gitea)
Redmine	Организация работы по проектам (используется в IT)	Проекты Задачи	Сервер PostgreSQL Можно добавлять пользовательские поля для описания	+/–	+	Ruby, PostgreSQL	Требуется развертывание системы и базы данных
OpenBIS	Управление лабораторией (LIMS) и проектами (ELN)	Проекты Эксперименты Образцы Набор данных	Сервер PostgreSQL Пользовательские поля	+	+	Java, PostgreSQL	Требуется развертывание системы и базы данных
SEEK	Управление данными и моделями системной биологии (ELN)	ISA стандарт: Исследование Стадия Образец	Схемы RDF Пользовательские поля на уровне «Образец»	–	+	Ruby, MySQL, Virtuoso	Требуется развертывание системы и базы данных
Galaxy	Воспроизводимость вычислительных экспериментов/протоколов	Отсутствует, есть связь между данными	База метаданных PostgreSQL	–	+	Python, PostgreSQL	Требуется развертывание системы и базы данных, также требуется настройка кластера

эти процессы могут запускать программы на удаленном сервере или кластере, а обмен файлами выполнять через общую файловую систему. Воспроизводимость вычислительных программ достигается с помощью системы окружений Conda (Yan Y., Yan J., 2018), когда для каждой программы создается свое независимое окружение (набор библиотек, программ и модулей на Python\R строго определенных версий). Может также использоваться система легковесной виртуализации Docker (Rad et al., 2017), в рамках которой программа запускается в «виртуальной» и «легковесной» операционной системе семейства Linux. Galaxy является FAIR-подобной системой (Hiltemann et al., 2023). По сути, Galaxy реализовывает ELN систему требований, но в области биоинформатических конвейеров, т. е. не является полноценной ELN. LIMS не реализован в полной мере, есть лишь многопользовательский вход и ограничение на хранение результатов вычислений.

Заключение

В настоящей работе было рассмотрено ограниченное множество информационных решений в сфере организации проектной деятельности лабораторий, работающих в области биологии. Краткие характеристики систем описаны в таблице. Такие решения, как OpenBIS, SEEK и Galaxy, были созданы специально для сопровождения научных работ, тогда как Trello и Redmine являются системами управления проектами более общих категорий, хотя и могут использоваться в работе научных групп. Программный комплекс Git может быть рассмотрен крупными коллективами как инструмент для обмена и версионирования программного кода, данных, текстов статей, монографий и других научных текстов. Следует отметить, что Git не предназначен для хранения бинарных файлов (в частности, файлов в формате DOCX, PDF и др.), так как учитывает лишь изменения текстовых файлов. Более подходящие форматы для такого использования Git – это Markdown и LaTeX.

Перед внедрением тех или иных инструментов важно понимать цели их внедрения. Исходя из целей сформулировать требования, определить набор задач, которые должна решать система, а также отслеживать применение рекомендаций конкретными исполнителями. Учитывая сложность перечисленных процессов, можно рекомендовать начинать с внедрения открытых форматов и стандартов по представлению и передаче биологических данных, предложенных и развиваемых научным сообществом. Использование систем документооборота общего назначения в лаборатории позволит получить опыт эксплуатации, что, в свою очередь, поможет определить форматы данных, протоколы работы и программные продукты, необходимые для работы лаборатории, и исходя из этого принимать решение о масштабировании автоматизации работы с данными, включая создание структур онтологий, метаданных, схем хранения, сценариев работы программных систем.

Список литературы / References

Barillari C., Ottoz D.S.M., Fuentes-Serna J.M., Ramakrishnan C., Rinn B., Rudolf F. openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics.* 2016;32(4):638-640. DOI 10.1093/bioinformatics/btv606

Bauch A., Adamczyk I., Buczek P., Elmer F.J., Enimanev K., Glyzewski P., Kohler M., Pylak T., Quandt A., Ramakrishnan C., Beisel C., Malmström L., Aebersold R., Rinn B. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*. 2011;12:468. DOI 10.1186/1471-2105-12-468

Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., Ball C.A., Causton H.C., Gaasterland T., Glenisson P., Holstege F.C., Kim I.F., Markowitz V., Matese J.C., Parkinson H., Robinson A., Sarkans U., Schulze-Kremer S., Stewart J., Taylor R., Vilo J., Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. 2001;29(4):365-371. DOI 10.1038/ng1201-365.

Brown R., Porter T. Category Theory and Higher Dimensional Algebra: potential descriptive tools in neuroscience. *arXiv*. 2003. DOI 10.48550/arXiv.math/0306223

Chacon S., Straub B. Pro Git. Kaliforniya: Apress Berkli, 2014. DOI 10.1007/978-1-4842-0076-6

Ehresmann A., Vanbremeersch J. Memory Evolutive Systems: Hierarchy, Emergence, Cognition. Elsevier Science, 2007.

Friedrich A., Kenar E., Kohlbacher O., Nahnsen S. Intuitive web-based experimental design for high-throughput biomedical data. *BioMed Res. Int.* 2015;2015:958302. DOI 10.1155/2015/958302

Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res*. 2022;50(W1):W345-W351. DOI 10.1093/nar/gkac247

Guizzardi G. Ontology, ontologies and the "I" of FAIR. *Data Intell*. 2020;2(1-2):181-191. DOI 10.1162/dint_a_00040

Guizzardi G., Fonseca C.M., Benevides A.B., Almeida J.P.A., Porello D., Sales T.P. Endurant Types in Ontology-Driven Conceptual Modeling: Towards OntoUML 2.0. In: Conceptual Modeling – 37th International Conference, Xi'an, China, October 22–25, 2018. Proceedings. Berlin: Springer, 2018;136-150. DOI 10.1007/978-3-030-00847-5_12

Gutierrez C., Hurtado C.A., Vaisman A. Introducing time into RDF. *IEEE Trans. Knowl. Data Eng*. 2007;19(2):207-218. DOI 10.1109/TKDE.2007.34

Hiltemann S., Rasche H., Gladman S., Hotz H.-R., Larivière D., Blankenberg D., Jagtap P.D., Wollmann T., Bretaudeau A., Goué N., Griffin T.J., Royaux C., Bras Y.L., Mehta S., Syme A., Coppens F., Droesbeke B., Soranzo N., Bacon W., Psomopoulos F., Gallardo-Alba C., Davis J., Föll M.C., Fahrner M., Doyle M.A., Serrano-Solano B., Fouilloux A.C., van Heusden P., Maier W., Clements D., Heyl F., Network G.T., Grüning B., Batut B. Galaxy Training: a powerful framework for teaching! *PLoS Comput. Biol*. 2023;19(1):e1010752. DOI 10.1371/journal.pcbi.1010752

Hoops S., Sahle S., Gauges R., Lee C., Pahle J., Simus N., Singhal M., Xu L., Mendes P., Kummer U. COPASI – a COmplex PAthway SImulator. *Bioinformatics*. 2006;22(24):3067-3074. DOI 10.1093/bioinformatics/btl485

Hucka M., Bergmann F.T., Chaouiya C., Dräger A., Hoops S., Keating S.M., König M., Le Novère N., Myers C.J., Olivier B.G., Sahle S., Schaff J.C., Sheriff R., Smith L.P., Waltemath D., Wilkinson D.J., Zhang F. The Systems Biology Markup Language (SBML): language specification for Level 3 Version 2 Core Release 2. *J. Integr. Bioinform*. 2019;16(2):20190021. DOI 10.1515/jib-2019-0021

Kuś M., Skowron B. (Eds.) Category Theory in Physics, Mathematics, and Philosophy, Springer Proceedings in Physics. Cham: Springer, 2019. DOI 10.1007/978-3-030-30896-4

MongoDB: The Developer Data Platform [WWW Document], n.d. MongoDB. URL https://www.mongodb.com (accessed 9.19.23)

Novère N.L., Finney A., Hucka M., Bhalla U.S., Campagne F., Collado-Vides J., Crampin E.J., Halstead M., Klipp E., Mendes P., Nielsen P., Sauro H., Shapiro B., Snoep J.L., Spence H.D., Wanner B.L. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol*. 2005;23(12):1509-1515. DOI 10.1038/nbt1156

Novère N.L., Hucka M., Mi H., Moodie S., Schreiber F., Sorokin A., Demir E., Wegner K., Aladjem M.I., Wimalaratne S.M., Bergman F.T., Gauges R., Ghazal P., Kawaji H., Li L., Matsuoka Y., Villéger A., Boyd S.E., Calzone L., Courtot M., Dogrusoz U., Freeman T.C., Funahashi A., Ghosh S., Jouraku A., Kim S., Kolpakov F., Luna A., Sahle S., Schmidt E., Watterson S., Wu G., Goryanin I., Kell D.B., Sander C., Sauro H., Snoep J.L., Kohn K., Kitano H. The Systems Biology Graphical Notation. *Nat. Biotechnol*. 2009;27(8):735-741. DOI 10.1038/nbt.1558

Olivier B.G., Snoep J.L. Web-based kinetic modelling using JWS Online. *Bioinformatics*. 2004;20(13):2143-2144. DOI 10.1093/bioinformatics/bth200

Petzold A., Asmi A., Vermeulen A., Pappalardo G., Bailo D., Schaap D., Glaves H.M., Bundke U., Zhao Z. ENVRI-FAIR-interoperable environmental FAIR data and services for society, innovation and research. In: 15th International Conference on eScience (eScience), San Diego, CA, USA, 2019. IEEE, 2019;277-280. DOI 10.1109/eScience.2019.00038

PostgreSQL: the world's most advanced open source database [WWW Document], n.d. URL https://www.postgresql.org/

Rad B.B., Bhatti H.J., Ahmadi M. An introduction to Docker and analysis of its performance. *Int. J. Comput. Sci. Netw. Secur.* 2017;17(3):228-235

Rocca-Serra P., Brandizi M., Maguire E., Sklyar N., Taylor C., Begley K., Field D., Harris S., Hide W., Hofmann O., Neumann S., Sterk P., Tong W., Sansone S.-A. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*. 2010;26(18):2354-2356. DOI 10.1093/bioinformatics/btq415

Roche D.G., Kruuk L.E.B., Lanfear R., Binning S.A. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol*. 2015;13(11):e1002295. DOI 10.1371/journal.pbio.1002295

Schreiber F., Bader G.D., Golebiewski M., Hucka M., Kormeier B., Novère N.L., Myers C., Nickerson D., Sommer B., Waltemath D., Weise S. Specifications of standards in systems and synthetic biology. *J. Integr. Bioinform*. 2015;12(2):1-3. DOI 10.1515/jib-2015-258

Software OpenLink. Virtuoso Open-Source Edition: Building. 2022. URL https://github.com/openlink/virtuoso-opensource

Spivak D.I., Kent R.E. Ologs: a categorical framework for knowledge representation. *PLoS One*. 2012;7(1):e24274. DOI 10.1371/journal.pone.0024274

The Univalent Foundations Program. Homotopy Type Theory: Univalent Foundations of Mathematics. Princeton, NJ: Institute for Advanced Study, 2013

Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.W., da Silva Santos L.B., Bourne P.E., … van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*. 2016;3:160018. DOI 10.1038/sdata.2016.18

Wolstencroft K., Owen S., Krebs O., Nguyen Q., Stanford N.J., Golebiewski M., Weidemann A., Bittkowski M., An L., Shockley D., Snoep J.L., Mueller W., Goble C. SEEK: a systems biology data and model management platform. *BMC Syst. Biol*. 2015;9:33. DOI 10.1186/s12918-015-0174-y

Yan Y., Yan J. Hands-On Data Science with Anaconda: Utilize the right mix of tools to create high-performance data science applications. Packt Publishing Ltd., 2018

Zeeberg B.R., Riss J., Kane D.W., Bussey K.J., Uchio E., Linehan W.M., Barrett J.C., Weinstein J.N. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*. 2004;5:80. DOI 10.1186/1471-2105-5-80

ORCID ID
A.M. Mukhin orcid.org/0000-0002-1102-0934
F.V. Kazantsev orcid.org/0000-0002-5711-7539
S.A. Lashin orcid.org/0000-0003-3138-381X

Благодарности. Работа выполнена при поддержке Курчатовского геномного центра Института цитологии и генетики СО РАН (№ 075-15-2019-1662).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 13.07.2023. После доработки 28.09.2023. Принята к публикации 29.09.2023.