УДК 004.414.22 doi: 10.15827/0236-235X.149.122-133 2025. Т. 38. № 1. С. 122–133

Определение принципов проектирования отечественной платформы обмена научными датасетами

К.В. Гарев [1, 2]✉

[1] Департамент информационных технологий Банка России, г. Москва, 107016, Россия
[2] Национальный исследовательский центр «Курчатовский институт» – НИИСИ, г. Москва, 117218, Россия

Ссылка для цитирования
Гарев К.В. Определение принципов проектирования отечественной платформы обмена научными датасетами. 2025. Т. 38. № 1. С. 122–133. doi: 10.15827/0236-235X.149.122-133
Информация о статье
Группа специальностей ВАК: 2.3.1
Поступила в редакцию: 01.07.2024 После доработки: 26.07.2024 Принята к публикации: 31.07.2024

Аннотация. В работе рассматриваются предпосылки и предложения по созданию отечественной платформы обмена научными датасетами в контексте глобальных тенденций, связанных с развитием открытой науки и распространением FAIR-принципов. На основе анализа зарубежных инициатив (EUDAT, EOSC, DataONE, Dryad, Zenodo) выявляются ключевые проблемы, затрудняющие эффективное использование, сохранение и повторное применение научных данных. К таким проблемам относятся отсутствие единых регламентов для описания датасетов, разрозненность инфраструктурных решений, недостаточная межплатформенная совместимость, а также сложность обеспечения воспроизводимости исследований. Особое внимание уделяется роли профессионального сообщества, важности формирования среды для обмена опытом, проведения междисциплинарных проектов и совершенствования навыков работы с большими массивами данных. При этом подчеркивается необходимость систематизации работы с научными данными и унификации требований к их сбору, хранению, обработке и представлению, что позволит повысить прозрачность исследовательских процессов. В статье обоснована целесообразность внедрения механизмов распределенного хранения, федеративной аутентификации и высокопроизводительных вычислительных ресурсов, способных удовлетворять потребности отечественного научного сообщества. Выдвигаются предложения по проектированию единой платформы обмена научными датасетами – от разработки методологических регламентов и стандартов взаимодействия с внешними системами до принципов интеграции инструментов аналитики и обеспечения надежной защиты данных.

Ключевые слова: датахабы, научные данные, научные датасеты, цифровые платформы обмена научными данными, открытая наука, FAIR, EUDAT, EOSC, DataONE, Dryad, Zenodo

Введение. В международном научном сообществе постепенно укрепляется и развивается концепция Open Science, призванная обеспечить более широкое и открытое распространение научных данных, результатов исследований и методологических наработок. Европейская комиссия подчеркивает, что открытое взаимодействие в научно-исследовательской среде стимулирует инновации и экономический рост, а также способствует решению крупных социально-экономических задач [1]. В свою очередь, Организация экономического сотрудничества и развития (OECD) отмечает, что доступность и повторное использование исследовательской информации являются важнейшими условиями повышения эффективности научного сотрудничества и ускорения научно-технического прогресса [2]. Основная идея данной концепции заключается в повышении прозрачности научных исследований и их воспроизводимости, что, как предполагается, может содействовать ускорению прогресса в различных областях знаний и укреплению междисциплинарных связей. Несмотря на то, что потенциальные эффекты от внедрения принципов Open Science выглядят многообещающими, в мировом сообществе все еще ведутся дискуссии о реальном масштабе и траекториях будущих изменений, обусловленных данной концепцией.

Неотъемлемой составляющей современных подходов к организации научных данных является концепция FAIR-принципов (*Findable*, *Accessible*, *Interoperable*, *Reusable*). Ее авторы акцентируют внимание на необходимости обеспечения удобного поиска, доступности, совместимости и возможности повторного использования данных в широком спектре научных дисциплин, что должно способствовать повышению эффективности обмена научными данными и прозрачности процесса получения исследовательских результатов [3].

Вышеупомянутые принципы реализуются путем создания различных цифровых инфраструктур, в том числе международных науч-

ных датахабов. К числу известных ресурсов, формирующих основу для открытого обмена данными и предоставляющих доступ к широкому спектру научных датасетов, относятся такие, как EUDAT (*European Data Infrastructure*), EOSC (*European Open Science Cloud*), DataONE (*Data Observation Network for Earth*), Dryad, Zenodo и др. Их создание и развитие во многом стимулируются как государственными, так и неправительственными инициативами, стремящимися упростить доступ к актуальным наборам данных и повысить эффективность научного взаимодействия.

В России данное направление находится в стадии формирования: несмотря на появление отечественных баз научных данных, до последнего времени отсутствовал системный подход к созданию универсальных платформ для обмена научными датасетами. Новая стратегия научно-технологического развития Российской Федерации декларирует необходимость разработки и внедрения таких платформ, указывая на их потенциальную значимость для обеспечения технологической независимости и повышения глобальной конкурентоспособности отечественных научных организаций. Вместе с тем реализация подобных проектов может потребовать существенного комплексного анализа потребностей различных научных направлений, обеспечения дополнительных ресурсов, а также принятия соответствующих нормативных актов.

Данная статья посвящена рассмотрению существующих инициатив и подходов в области создания научных датахабов и возможных направлений их создания и развития.

Анализ зарубежного опыта построения и использования систем управления научными данными

В мире продолжают создаваться и развиваться системы управления научными датасетами (научными датахабами), позволяющие осуществлять взаимообмен результатами экспериментов, а также однозначно ссылаться на них. Системы необходимы для консолидации и долгосрочного хранения данных, предоставления доступа к ним исследовательских сообществ.

Наличие подобных платформ дает возможность систематизировать подходы к управлению научными данными, что делает их доступными для поиска, легко интегрируемыми в разные системы и пригодными для повторного использования. Реализация FAIR-принципов в научных датахабах способствует формированию глобальной инфраструктуры данных. Кроме того, в международном научном сообществе принято делиться научными датасетами в открытом доступе. Этот подход позволяет обеспечить воспроизводимость результатов, подтвердить их достоверность и исключить риск фальсификации данных. Как показано в [4], открытый доступ к датасетам является важным элементом современной научной практики, поскольку предоставляет возможность независимой проверки выводов и использования данных для новых исследований. Данные, размещенные на специализированных платформах для сбора, хранения, обработки и представления научных данных, сопровождаются описанием метаданных и стандартами цитирования, что гарантирует корректное использование и ссылку на источник.

Датахабы условно можно разделить на следующие группы:

– по областям наук: специализированные, профильные, междисциплинарные;

– по назначению: хранилища данных, аналитические платформы, каталоги справочников и стандартов;

– по организации инфраструктуры: облачные, локальные, гибридные.

Рассмотрим примеры таких систем.

EUDAT (https://www.eudat.eu/) – крупномасштабная европейская инфраструктура, предназначенная для управления научными данными и предоставления исследовательским сообществам инструментов для хранения, обработки и обмена данными. Ее основная цель заключается в создании единой унифицированной платформы для обеспечения FAIR-принципов и поддержки междисциплинарных исследований. EUDAT предоставляет широкий спектр функций и технологий для управления научными данными, обеспечивая их долгосрочное хранение, обработку, доступ и интеграцию. Предлагает комплекс взаимосвязанных сервисов для управления научными данными: B2SAFE обеспечивает их репликацию, повышая надежность и длительное сохранение; B2STAGE предоставляет высокопроизводительные средства (GridFTP, HTTP) для передачи больших объемов данных на внешние вычислительные ресурсы; B2SHARE, основанный на платформе Invenio, упрощает загрузку, аннотирование и обмен исследовательскими данными с присвоением постоянных идентификаторов; B2FIND обеспечивает единый ка-

талог метаданных и расширенный поиск; B2DROP служит для хранения, синхронизации и совместного использования динамических данных; B2ACCESS поддерживает федеративную аутентификацию и авторизацию через OpenID, SAML, x.509 и OAuth 2, а B2HANDLE, работающий на основе системы Handle, позволяет регистрировать постоянные идентификаторы и гарантирует доступность и цитируемость данных. EUDAT активно применяется в различных предметных областях науки. В исследовании [5] приведен пример использования данной платформы как основы для управления научными данными в биомедицине на базе географически распределенной высокопроизводительной инфраструктуры. Проект ESCAPE Dark Matter Science Project, реализуемый в рамках инициативы EOSC Future, использует инфраструктуру EUDAT для управления и обработки данных в исследовании темной материи, включая сервис для масштабируемого хранения и обработки данных, сервис для авторизации и управления доступом, а также платформу на базе JupyterHub для проведения воспроизводимого анализа данных и обеспечения соответствия принципам FAIR [6].

EOSC (htpps://eosc.eu/) – инициатива Европейской комиссии, направленная на создание федеративной и открытой мультидисциплинарной среды, где европейские исследователи, инновационные компании и граждане могут публиковать, находить и повторно использовать данные, инструменты и сервисы для научных, инновационных и образовательных целей. Техническая реализация EOSC основана на федеративной архитектуре, которая объединяет существующие и новые научные данные и сервисы в единую интероперабельную экосистему, устраняя фрагментацию и предоставляя пользователям единую точку доступа к ресурсам. Для обеспечения взаимодействия компонентов применяются EOSC Interoperability Framework [7] с рекомендациями по технической, семантической и организационной интероперабельности, а также федеративная система аутентификации и авторизации с интеграцией через eduGAIN (https://edugain.org/), что позволяет безопасно использовать институциональные учетные данные. Через EOSC Portal исследователям доступен каталог сервисов и данных со стандартизированными метаданными, упрощающий поиск и интеграцию различных ресурсов, включая ПО и вычислительные мощности. В качестве примера использования можно привести исследование, где в рамках Большой латиноамериканской обсерватории (LAGO) была проведена масштабная интеграция симуляционного фреймворка ARTI в облачные сервисы EOSC-Synergy (проект в контексте EOSC, направленный на увеличение международной сети обмена научными данными), что позволило обеспечить централизованное хранение, обработку и моделирование больших объемов данных о космических лучах, поступающих с распределенной сети датчиков в различных регионах и на высотах до 5 500 м [8].

DataONE (https://www.dataone.org/) – сеть репозиториев данных, призванная сохранять и предоставлять доступ к данным в области наук о Земле и окружающей среде. DataONE функционирует на базе распределенной архитектуры с координирующими узлами и узлами-участниками, которые связаны стандартизированными RESTful API. Координирующие узлы отвечают за регистрацию репозиториев, управление метаданными, предоставление сервисов поиска и обнаружения данных, а узлы-участники представляют собой репозитории, чьи коллекции индексируются и реплицируются для повышения доступности и надежности. Система устойчивых идентификаторов обеспечивает однозначную идентификацию объектов данных, а метаданные описываются унифицированными схемами, что упрощает их поиск и повышает интероперабельность. Механизмы репликации между узлами-участниками гарантируют долговременное хранение данных и устойчивость к сбоям, обеспечивая широкий доступ к информации. В статье [9] авторы описывают, как DataONE предоставляет инфраструктуру для хранения, обмена и управления биологическими и экологическими данными, что способствует инновациям в этих областях, акцентируя внимание на интеграции сервисов DataONE в исследовательские процессы.

Dryad (https://datadryad.org/) – международный открытый репозиторий, предоставляющий исследователям платформу для публикации и сохранения данных, лежащих в основе научных и медицинских публикаций. Dryad использует открытую платформу Stash (на базе Ruby on Rails) для управления процессом публикации научных данных, обеспечивая загрузку, описание и распространение наборов данных в различных форматах с применением адаптированной схемы метаданных Dublin Core. Долговременное хранение данных осуществляется в сотрудничестве с Merritt, репозиторием Калифорнийской цифровой библиотеки, который гарантирует надежное битовое хранение и управ-

ляет копиями файлов, размещенными у нескольких облачных провайдеров в двух географических регионах. Каждый опубликованный набор получает DOI от DataCite для постоянного доступа и цитирования, а поддержка Schema.org повышает видимость данных в поисковых системах. Команда кураторов Dryad проверяет загружаемые материалы на соответствие стандартам качества и этики, а партнерство с Zenodo, установленное в 2019 году, связывает опубликованное в Dryad исследовательское ПО с репозиторием CERN. Аутентификация через ORCID способствует точному учету авторства и вклада ученых [10].

Zenodo (https://github.com/zenodo/zenodo) – универсальный открытый репозиторий, разрабатываемый в рамках программы OpenAIRE и управляемый CERN, предоставляет исследователям возможности по хранению и распространению статей, наборов данных, ПО и других цифровых объектов. Техническая инфраструктура платформы основана на ПО Invenio, созданном в CERN, что обеспечивает масштабируемость и поддержку множества форматов данных. Размещаясь в высокопроизводительном вычислительном центре CERN в Швейцарии, Zenodo гарантирует надежное долгосрочное хранение материалов; каждому загружаемому объекту присваивается DOI через DataCite, способствуя его обнаружению и корректному цитированию. Гибкая политика лицензирования и интеграция с GitHub позволяют автоматически сохранять версии ПО и обеспечивать их цитируемость, а запуск репозитория в 2013 году сделал его широко востребованной платформой для международного научного сообщества.

В качестве иллюстрации одного из примеров использования можно привести долгосрочное археологическое исследование в Южной Румынии [11]. Платформа Zenodo была интегрирована в рабочий процесс для хранения и обмена данными, полученными с помощью дронов, включая ортофотоснимки, цифровые модели рельефа и 3D-модели археологических памятников. Использование Zenodo позволило обеспечить долговременное хранение больших объемов пространственных данных, предоставление уникальных DOI для удобного цитирования и соблюдение принципов открытого доступа.

Рассмотренные международные платформы демонстрируют ряд общих принципов.

Во-первых, каждая из них делает ставку на использование устойчивых идентификаторов для обеспечения цитируемости и глобальной прослеживаемости данных (DOI, Handle), что упрощает их поиск и интеграцию.

Во-вторых, ключевым аспектом во всех системах выступает работа с качественными и стандартизированными метаданными (Dublin Core, Schema.org и др.), обеспечивающими высокую интероперабельность между различными областями исследований.

В-третьих, данные платформы ориентированы на соблюдение принципов FAIR.

Повсеместно применяется федеративный подход к аутентификации (SAML, OpenID, eduGAIN и т.д.), что повышает безопасность и упрощает управление доступом. Немаловажны и механизмы репликации и распределенного хранения (например, интеграция с облачными провайдерами и национальными репозиториями), которые обеспечивают устойчивость к сбоям и долговременное сохранение данных.

Таким образом, совокупность этих характеристик закладывает основу для эффективного функционирования датахабов.

Отечественные решения для работы с научными данными

По данным портала «Научно-технологическая инфраструктура Российской Федерации», в России функционируют более 400 уникальных научных установок, в рамках использования которых так или иначе могут генерироваться научные датасеты. Использование физических научных установок увеличивает издержки, связанные с закупкой расходных материалов и амортизацией инфраструктуры, а также влияет на время проведения исследования за счет физической очереди. Практика переиспользования научных датасетов, напротив, позволяет произвести необходимые вычисления зачастую без необходимости получения физического доступа к научной установке.

В России отсутствуют централизованные датахабы, аналогичные зарубежным. При этом существуют репозитории научных данных, созданные отдельными институтами. Например, Институт космических исследований РАН предоставляет доступ к данным через Центр коллективного пользования «ИКИ-Мониторинг», который обеспечивает хранение и обработку спутниковых наблюдений для изучения и мониторинга окружающей среды [12].

Одним из примечательных является проект «Инфраструктура научно-исследовательских данных» (ИНИД) (https://data.rcsi.science/about/),

разработанный Центром перспективных управленческих решений. Изначально проект ИНИД был создан для содействия цифровой трансформации науки и совершенствования взаимодействия между учеными и органами государственной власти. Однако реализованный функционал позволяет рассматривать данную платформу как сервисную основу для развития отечественного междисциплинарного научного датахаба на базе Российского центра научной информации, являющегося текущим оператором ИНИД.

Иная ситуация с платформами, направленными на работу не с научными датасетами, а с научно-технической информацией, являющейся ключевым ресурсом для организации и управления научной и инновационной деятельностью. Она включает в себя сведения о научных исследованиях, технологических разработках, патентах, публикациях, конференциях и других аспектах, отражающих динамику научно-технического прогресса. Эффективное управление этим видом информации является критически важным для обеспечения ее доступности, анализа и использования в процессе принятия решений. В России функционируют несколько таких систем, предназначенных для поддержки научных исследований, технологического развития и обеспечения обмена информацией.

Единая государственная информационная система учета результатов научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения – централизованная платформа для учета данных о научных исследованиях и разработках в России, охватывающая все этапы проекта от запуска до итоговых результатов. Для каждой работы формируется регистрационная карта с основными сведениями, а исполнители обязаны предоставлять информацию об итогах и созданных интеллектуальных продуктах, полученных за счет средств федерального бюджета. Такой подход обеспечивает наличие единого ресурса для анализа, планирования и мониторинга научной деятельности в стране.

Поисковая платформа Роспатента – цифровая система для доступа к патентной информации, позволяющая осуществлять поиск и анализ сведений об объектах интеллектуальной собственности (патенты, товарные знаки, полезные модели) по различным параметрам, включая текст, изображения и прочее. Кроме пользовательского интерфейса, платформа предлагает документированное API, возвращающее структурированные данные, что облегчает интеграцию в аналитические задачи и разработку специализированных приложений.

eLIBRARY.RU – одна из крупнейших российских научных электронных библиотек с более чем 50 млн публикаций (статьи, монографии, диссертации, материалы конференций) и интеграцией с РИНЦ. Она обеспечивает доступ к ресурсам для проведения исследований, анализа публикационной активности и мониторинга научной деятельности.

Анализ международного опыта в сфере создания и развития инфраструктуры научных датахабов, а также рассмотрение существующих российских систем работы с научно-технической информацией позволяют выявить ряд проблем.

Во-первых, неоднородность форматов и стандартов метаданных затрудняет интеграцию и обработку крупных и разнородных массивов данных, а недостаток специализированной инфраструктуры не обеспечивает на должном уровне хранение и анализ больших объемов информации.

Во-вторых, отсутствие единого подхода к интероперабельности и стандартам взаимодействия между различными платформами порождает сложности доступа к данным и их повторного использования.

В-третьих, проблемы воспроизводимости научных результатов, а также неоднородная стандартизация правил по формированию метаданных отрицательно сказываются на возможности воспроизводимости исследований.

Кроме того, барьеры доступа к данным, связанные с авторским правом и правовыми ограничениями, этические аспекты использования данных и необходимость защиты конфиденциальности и интеллектуальной собственности также требуют учета при формировании эффективной экосистемы.

Важно отметить, что российские информационные системы в большей степени ориентированы на учет результатов научной деятельности, публикационную активность и патентную статистику, тогда как готовых решений, специализирующихся именно на совместном использовании структурированных научных датасетов, на сегодняшний день существенно меньше. Международная практика показывает, что успешное формирование платформ обмена научными данными подразумевает наличие целостной инфраструктуры, единой системы идентификаторов и аутентификации пользователей, а также строго прописанных регламентов обмена информацией.

Общие принципы построения платформы обмена научными датасетами

Формирование единой междисциплинарной платформы для обмена научными датасетами в России предполагает комплексный подход, охватывающий несколько ключевых принципов (см. рисунок):

– создание и поддержка среды для обмена знаниями, совместных междисциплинарных исследований и обучения;

– разработка четкой методологии, включающей регламенты описания и оценки качества датасетов, а также механизмов защиты интеллектуальных прав;

– выбор и адаптация инструментов, позволяющих эффективно формировать, обрабатывать и распространять данные, включая решение задач по загрузке, выгрузке и визуализации.

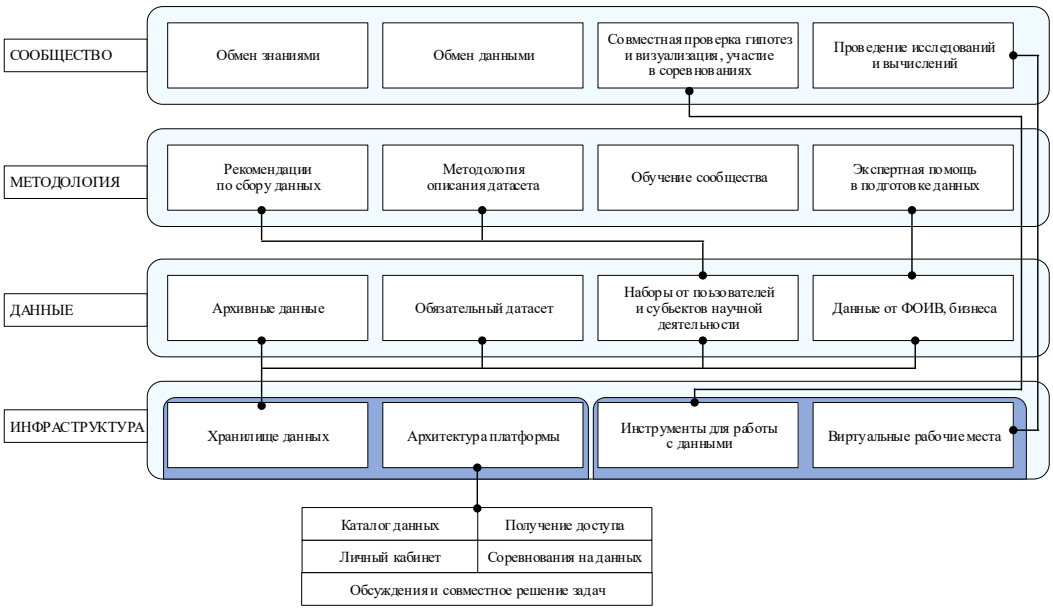
Еще одним определяющим фактором становится организация самих данных – создание единого каталога с детальным классификатором и реализация FAIR-принципов доступности и совместимости. Устойчивой основой для перечисленных элементов должна выступать соответствующая инфраструктура, отвечающая требованиям быстродействия, безопасности и интероперабельности.

Сообщество – это не только носитель культурных и научных традиций, но и движущая сила для непрерывного развития платформы обмена датасетами. Международные проекты показывают важность комплексной поддержки взаимодействия между учеными. Так, в рамках EOSC созданы отдельные тематические рабочие группы, занимающиеся формированием рекомендаций и обменом опытом в области обработки, документирования и распространения данных. Аналогично платформа DataONE развивает программу обучения работе с научными данными, выпуская подробные руководства и проводя сетевые семинары, что способствует повышению цифровой грамотности и стимулирует сотрудничество участников. В экосистеме Kaggle (https://www.kaggle.com/) особую роль играют регулярные соревнования и форумы, где исследователи и инженеры обмениваются лучшими методами анализа и обработки датасетов. В результате создается культура кооперации, ускоряющая внедрение новых решений.

Следуя данному принципу, функционально необходимо реализовать:

– тематические рабочие группы (виртуальные исследовательские сообщества); создание специализированных секторов внутри платформы, где ученые по смежным дисциплинам могут общаться, совместно формировать библиотеки рабочих инструментов, обсуждать кейсы и получать экспертизу от более опытных коллег;

– форумы и систему «вопрос–ответ»; аналог Kaggle Discussions, где участники задают практические вопросы по обработке датасетов, об-



Общие принципы построения платформы обмена научными данными

General principles for building a scientific data exchange platform

ращаются за технической или методической поддержкой и получают обратную связь от членов сообщества с разным уровнем компетенций;

– регулярные конкурсы и хакатоны; организация соревнований по анализу предоставленных датасетов в различных предметных областях;

– инструменты для совместной работы; предоставление возможности вести коллективную разработку внутри платформы (совместное редактирование метаданных, настройка рабочих процессов для обработки данных и т.п.) с сохранением версионности и историй изменений.

Преимуществом такой модели станет формирование сплоченного научного сообщества, где обмен датасетами сочетается с постоянной профессиональной поддержкой и обучением. В результате появится платформа, в которой исследователи не просто загружают и скачивают данные, а ведут активный диалог, совместно прорабатывают методологии анализа и решают прикладные научные проблемы.

Наличие методологии является фундаментальным принципом, определяющим структуру, стандарты и регламенты работы с данными в рамках платформы обмена научными датасетами. Международный опыт показывает, что выстраивание четких методологических основ позволяет обеспечить прозрачность, воспроизводимость и высокое качество научных исследований. Подобного рода подходы охватывают вопросы корректной структуры данных, единых правил описания и цитирования, а также правовой и этической сторон работы с научными результатами.

При разработке методологической основы необходимо учесть следующие основные аспекты.

- Регламенты описания датасетов (управление метаданными), фиксирующие:

– разработку единых руководств по созданию и ведению метаданных с учетом принятых в мире стандартов (Dublin Core, ISO 19115 для геоданных, Darwin Core для биологических данных и т.д.);

– интеграцию принципов FAIR в качестве ориентира для формирования структуры метаданных и описания данных;

– многоуровневое описание: от базовых полей (автор, название, ключевые слова) до расширенных (условия сбора, протокол экспериментов, единицы измерения), что повышает воспроизводимость научных исследований.

- Стандарты и процедуры оценки качества датасетов, включающие:

– установку критериев проверки достоверности, целостности и репрезентативности данных; например, Dryad и Zenodo внедряют базовый кураторский надзор для проверки формата, точности описания и соответствия лицензионным требованиям;

– разработку формализованных методик для оценки качества метаданных и самих наборов данных (ошибки пропущенных значений, корректность классификаций), а также определение ответственных за подобные проверки (кураторы, эксперты);

– использование показателей для мониторинга качества и востребованности датасетов: число загрузок, цитирований, перепубликаций, что помогает формировать репутацию и рейтинг отдельных наборов данных.

- Порядок обмена научными данными и защита авторских прав, предусматривающие:

– создание типовых лицензионных соглашений для регулирования использования данных и защиты авторских прав;

– описание процедур ознакомления и согласия всех участников научной деятельности с установленными правилами лицензирования, этическими нормами, а также механизмов контроля за соблюдением прав собственности;

– меры, направленные на соблюдение законов в сфере персональных данных и конфиденциальной информации.

- Юридические и этические регламенты, обеспечивающие:

– учет законодательства в области интеллектуальной собственности, авторских прав и защиты персональных данных;

– включение в методологические документы положений об этических аспектах: согласование протоколов исследования, анонимизация данных, обеспечение прозрачности исследования для участников экспериментов;

– регулярное обновление методических материалов с учетом изменения законодательства и появления новых международных и национальных актов.

Таким образом, методологический принцип формирует основу, на которой строится вся деятельность платформы от создания датасетов и их аннотирования до правового регулирования и обеспечения воспроизводимости результатов. Адаптируя проверенные мировые практики к отечественным условиям, можно разработать единый свод регламентов и процедур, отвечающих потребностям российских науч-

ных организаций. Эти методологические решения позволят наладить прозрачный и эффективный обмен научными данными.

Данные занимают центральное место в любой платформе обмена научными датасетами и определяют качество проводимых исследований и возможности для междисциплинарного сотрудничества. Международные проекты показывают, что единая точка доступа к разнородным наборам данных значительно упрощает их поиск, интеграцию и повторное использование. Особое внимание уделяется созданию подробных каталогов, каждый датасет сопровождается подробными метаданными, что позволяет отслеживать происхождение, качество и условия применения научного материала. Подход в формате единого окна позволит исследователям находить соответствующие наборы данных из различных направлений и оперативно выявлять возможности для междисциплинарных проектов.

Для эффективной работы каталога необходим определенный порядок наполнения платформы.

- Зафиксировать требования к загружаемым датасетам:

– определить минимальный набор метаданных, соответствующий как международным стандартам, так и требованиям отечественных исследовательских сообществ;

– назначить уникальный идентификатор (например DOI) для каждого набора данных, что упростит цитирование и отслеживание переиспользования;

– обеспечить автоматизированную проверку загружаемого материала на корректность форматов (CSV, NetCDF, XML и т. д.) и целостность файлов, а также валидацию ключевых полей метаданных.

- Определить этапы и процедуры добавления новых наборов данных:

– первичная загрузка: пользователь (исследователь или организация) регистрирует датасет, указывая основные характеристики, и прикрепляет файлы; на этом этапе может быть применена автоматизированная проверка структуры данных и соответствия базовым правилам метаданных;

– кураторская проверка: специально назначенные эксперты (или кураторы) оценивают релевантность, правильность описания и соответствие установленным стандартам, включая возможные этические и правовые аспекты, в случае выявления несоответствий предоставляются рекомендации по доработке;

– публикация в каталоге: после успешного прохождения проверки датасет становится доступным для поиска и скачивания; ему присваивается уникальный постоянный идентификатор, формируется страница описания, при необходимости указываются условия лицензирования.

- Обеспечить долгосрочное хранение и управление версиями датасетов за счет:

– механизмов сохранения версий, отражающих изменения набора данных (добавление новых наблюдений, корректировки или уточнения), а также архивации старых версий в целях воспроизводимости экспериментов;

– средств интеграции с распределенными системами хранения (облачные репозитории, государственные ЦОД) для обеспечения устойчивости и резервирования данных.

- Обеспечить контроль качества и актуализацию информации:

– введение рейтинговой системы для оценки датасетов сообществом (по уровню полноты, надежности, востребованности);

– систематические обзоры качества с целью проверки востребованности и актуальности датасетов, а также определения устаревших или дублирующихся наборов, которые могут быть упразднены или объединены с другими.

- Принимать во внимание специфику отечественной науки и междисциплинарную ориентированность, для чего требуется:

– учитывать потребности различных российских научных организаций, а также специфику предметных областей;

– предусмотреть возможность интеграции с уже существующими базами данных и реестрами для упрощения импортирования и синхронизации научных данных;

– создать механизмы ссылки и совместного отображения метаданных (cross-linking), чтобы исследователи могли прослеживать связь между статьями, патентами и первичными данными.

Инфраструктура является ключевым фактором, определяющим устойчивость, производительность и безопасность работы платформы обмена научными датасетами. Международные проекты наглядно демонстрируют важность комплексного подхода к организации вычислительных ресурсов, сетевых взаимодействий и механизмов хранения, без которых невозможно обеспечить надлежащий уровень обслуживания. Например, в EUDAT применяется распределенная архитектура, основанная на координирующих узлах и узлах-участниках,

что гарантирует репликацию и надежное хранение данных в нескольких центрах обработки [13]. Подобная модель дает высокую устойчивость к сбоям, повышает пропускную способность и сокращает риски единой точки отказа.

Целесообразно учитывать уже имеющиеся национальные центры обработки данных и мощные исследовательские сети, в частности, проект НИКС (Национальная исследовательская компьютерная сеть) как возможную основу для построения единой платформы.

Приведем основные аспекты требований к инфраструктуре.

- Репликация и распределенное хранение:

– размещение основных узлов (центров обработки данных) в нескольких регионах с учетом географического фактора и доступности высокоскоростных сетей; это обеспечивает минимизацию задержек и резервирование в случае отключений или аварий;

– использование модулей репликации и кластерного хранения, позволяющих гарантировать целостность и восстановление данных при сбоях оборудования.

- Высокопроизводительные вычислительные мощности:

– интеграция с отечественными суперкомпьютерными центрами и кластерами, что позволит проводить сложные расчеты и анализ больших объемов данных непосредственно в рамках платформы; опыт EOSC показывает эффективность такого подхода, когда пользователи могут занимать вычислительные ресурсы из федеративной экосистемы на основе сервисов B2STAGE или аналогичных решений;

– возможность быстрого масштабирования вычислительных и сетевых ресурсов под пиковые нагрузки, например, во время проведения хакатонов или крупных научных проектов.

- Интероперабельность и открытые стандарты:

– использование принципов совместимости на уровне протоколов и интерфейсов, что упрощает подключение к международным сервисам и репозиториям;

– реализация федеративной аутентификации и авторизации пользователей, повышающая удобство использования учетных записей и сокращающая необходимость многократной регистрации.

- Безопасность и защита данных:

– введение многоуровневой системы контроля доступа, позволяющей разграничивать права пользователей в соответствии с их ролью (исследователь, куратор, администратор);

– шифрование данных при передаче и хранении с учетом требований российского законодательства и международных норм по защите интеллектуальной собственности;

– регулярные аудиты и тесты на проникновение для мониторинга надежности инфраструктуры.

- Высокая доступность и скорость отклика:

– организация резервных каналов связи между основными хранилищами и вычислительными центрами; при сбое в одном сегменте трафик должен автоматически переключаться на другой;

– применение механизмов кеширования для ускорения доступа к часто запрашиваемым датасетам.

- Системная интеграция и сервисный уровень (SLA):

– предоставление четкого уровня сервиса, в котором фиксируются показатели доступности и сроков реакции на инциденты;

– наличие круглосуточной службы технической поддержки и мониторинга для оперативного решения проблем с доступом, передачей или обработкой данных;

– интеграция с уже существующими отечественными платформами учета научных результатов и внешними репозиториями через стандартные API, обеспечивающая единый информационный контур.

Заключение

Таким образом, при проектировании инфраструктуры платформы стоит принимать во внимание принцип устойчивого масштабируемого и безопасного распределенного хранения данных, а также предусмотреть инструменты для межплатформенной интеграции и высокопроизводительной обработки. Опыт международных инициатив показывает, что без комплексного решения вопросов распределенного хранения, сетевой связности, федеративного управления доступом и резервирования ресурсов невозможно создать эффективный сервис для обмена научными датасетами.

Формирование единой платформы обмена научными датасетами требует целостного продуманного подхода, основанного на консолидации международного опыта, адаптации мировых стандартов и учете специфики отечественной научной инфраструктуры. Успешные зарубежные примеры демонстрируют важность единой среды взаимодействия специалистов,

детальной методологической базы, современных инструментов управления данными и соответствующей инфраструктурной поддержки. В российском контексте особую роль при этом играет объединение усилий различных исследовательских групп, вузов, отраслевых институтов и технологических партнеров, что позволит стимулировать междисциплинарные исследования и повысить качество научных результатов.

На уровне организации и методологии необходимо обеспечить единые регламенты описания, хранения и документирования данных, а также формализовать порядок обмена ими в соответствии с принципами FAIR. Внедрение систем идентификации и единых протоколов аутентификации дает возможность поддерживать прозрачность и воспроизводимость исследований, а также защищать авторские права и интеллектуальную собственность. Рациональная структура данных, основанная на классификаторах предметных областей и междисциплинарных рубрикаторах, позволит создавать хорошо организованный научный датахаб, доступный широкому кругу пользователей.

При проектировании инфраструктуры платформы следует ориентироваться на устойчивое масштабируемое и безопасное распределенное хранение данных, предусмотрев механизмы межплатформенной интеграции и высокопроизводительной обработки больших объемов информации. Международный опыт наглядно показывает, что без комплексного решения вопросов сетевой связности, федеративного управления доступом и резервирования ресурсов невозможно создать эффективный сервис для обмена научными датасетами. Адаптируя эти подходы к отечественным условиям и опираясь на уже существующие сети, например НИКС, российская платформа сможет обеспечить непрерывный и надежный доступ к данным, а также их оперативную обработку, соответствующую требованиям современных научных исследований.

Таким образом, комплексное сочетание функциональных сервисов, регламентов взаимодействия и механизмов поддержки сообщества позволит сформировать полноценную экосистему открытой науки, способствующую повышению конкурентоспособности российской науки в глобальном пространстве. Внедрение описанных решений создаст условия для более тесного междисциплинарного сотрудничества, простимулирует инновационную активность и обеспечит доступ исследовательских групп к данным, необходимым для проведения передовых исследований и технологических разработок.

Список литературы

1. European Commission: DG Research and Innovation. Open innovation, open science, open to the world: A vision for Europe. Publications Office of the European Union, 2015. URL: https://data.europa.eu/doi/10.2777/061652 (дата обращения: 12.06.2024).

2. Making open science a reality. OECD Science, Technology and Industry Policy Papers, 2015, no. 25. URL: https://www.oecd.org/en/publications/making-open-science-a-reality_5jrs2f963zs1-en.html (дата обращения: 12.06.2024). doi: 10.1787/5jrs2f963zs1-en.

3. Wilkinson M., Dumontier M., Aalbersberg I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 2016, vol. 3, art. 160018. doi: 10.1038/sdata.2016.18.

4. Tenopir C., Rice N.M., Allard S., Baird L., Borycz J. et al. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLoS One, 2020, vol. 15, no. 3, art. e0229003. doi: 10.1371/journal.pone.0229003.

5. Pribec I., Hachinger S., Hayek M., Pringle G.J. et al. Efficient and reliable data management for biomedical applications. In: MIMB, 2024, vol. 2716, pp. 383-403. doi: 10.1007/978-1-0716-3449-3_18.

6. Bhattacharjee P., Bird I., Calore F., Doglioni C., Eckner Ch. et al. ESCAPE dark matter science project for EOSC future project (WP6.3). ARPHA Preprints, 2023. doi: 10.3897/arphapreprints.e116673.

7. Scardaci D.O., Sciacca E., Hériché J.-K., Van De Sanden M., Klaas W. et al. A landscape overview of the EOSC interoperability framework – capabilities and gaps. Zenodo, 2023. URL: https://zenodo.org/records/8399710 (дата обращения: 12.06.2024).

8. Rubio-Montero A.J., Pagán-Muñoz R., Mayo-García R., Pardo-Diaz A., Sidelnik I., Asorey H. The EOSC-Synergy cloud services implementation for the Latin American Giant Observatory (LAGO). Proc. ICRC2021, 2021, vol. 395. doi: 10.22323/1.395.0261.

9. Michener W.K., Allard S., Budden A., Cook R.B., Douglass K. Participatory design of DataONE–Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics, 2012, vol. 11, pp. 5–15. doi: 10.1016/j.ecoinf.2011.08.007.

10. Isard M., Budiu M., Yu Y., Birrell A., Fetterly D. Dryad: Distributed data-parallel programs from sequential building blocks. ACM SIGOPS Operating Systems Review, 2007, vol. 41, no. 3, pp. 59–72. doi: 10.1145/1272998.1273005.

11. Stal C., Covataru C., Müller J., Parnic V. et al. Supporting long-term archaeological research in southern romania chalcolithic sites using multi-platform UAV mapping. Drones, 2022, vol. 6, no. 10, art. 277. doi: 10.3390/drones6100277.

12. Лупян Е.А., Прошин А.А., Бурцев М.А., Кашницкий А.В., Балашов И.В. и др. Опыт эксплуатации и развития центра коллективного пользования системами архивации, обработки и анализа спутниковых данных (ЦКП «ИКИ-Мониторинг») // Современные проблемы дистанционного зондирования Земли из космоса. 2019. Т. 16. № 3. С. 151–170. doi: 10.21046/2070-7401-2019-16-3-151-170.

13. Lecarpentier D., Wittenburg P., Elbers W., Michelini A., Kanso R., Coveney P., Baxter R. EUDAT: A new cross-disciplinary data infrastructure for science. IJDC, 2013, vol. 8, no. 1, pp. 279–287. doi: 10.2218/ijdc.v8i1.260.

Software & Systems doi: 10.15827/0236-235X.149.122-133 2025, 38(1), pp. 122–133

Design principles of a domestic platform for scientific dataset exchange

Kirill V. Garev [1, 2]✉

[1] Information Technology Department of the Bank of Russia, Moscow, 107016, Russian Federation
[2] National Research Centre "Kurchatov Institute", Moscow, 117218, Russian Federation

For citation
Garev, K.V. (2025) 'Design principles of a domestic platform for scientific dataset exchange', *Software & Systems*, 38(1), pp. 122–133 (in Russ.). doi: 10.15827/0236-235X.149.122-133
Article info
Received: 01.07.2024 After revision: 26.07.2024 Accepted: 31.07.2024

Abstract. The paper considers prerequisites and proposals for establishing a domestic platform for a scientific dataset exchange in terms of global trends related to the development of open science and spread of FAIR-principles. The authors analyze foreign initiatives (EUDAT, EOSC, DataONE, Dryad, Zenodo) to identify key problems that hinder effective use, preservation and reuse of scientific data. These problems include the lack of uniform regulations for dataset description, disparate infrastructure solutions, insufficient cross-platform interoperability, and difficulty in ensuring reproducibility of research. The authors focus on the role of the professional community, the importance of creating an environment for sharing experience, conducting interdisciplinary projects, and improving skills in working with large datasets. The paper emphasizes the need to systematize the work with scientific data and to unify the requirements for their collection, storage, processing and presentation. This will increase the transparency of research processes. The paper substantiates the feasibility of implementing distributed storage mechanisms, federated authentication and high-performance computing resources capable of meeting the needs of the domestic scientific community. Finally, it outlines proposals for designing a unified platform for scientific dataset exchange: from developing methodological regulations and standards for interaction with external systems to principles of integrating analytical tools and ensuring reliable data protection.
Keywords: data hubs, scientific data, scientific datasets, digital platforms for scientific data exchange, open science, FAIR, EUDAT, EOSC, DataONE, Dryad, Zenodo

References

1. European Commission: DG Research and Innovation (2015) 'Open innovation, open science, open to the world: A vision for Europe', *Publications Office of the European Union*, avilable at: https://data.europa.eu/doi/10.2777/061652 (accessed June 12, 2024).

2. (2015) 'Making open science a reality', *OECD Science, Technology and Industry Policy Papers*, (25), avilable at: https://www.oecd.org/en/publications/making-open-science-a-reality_5jrs2f963zs1-en.html (accessed June 12, 2024). doi: 10.1787/5jrs2f963zs1-en.

3. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3, art. 160018. doi: 10.1038/sdata.2016.18.

4. Tenopir, C., Rice, N.M., Allard, S., Baird, L., Borycz, J. et al. (2020) 'Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide', *PLoS One*, 15(3), art. e0229003. doi: 10.1371/journal.pone.0229003.

5. Pribec, I., Hachinger, S., Hayek, M., Pringle, G.J. et al. (2024) 'Efficient and reliable data management for biomedical applications', in *MIMB*, 2716, pp. 383–403. doi: 10.1007/978-1-0716-3449-3_18.

6. Bhattacharjee, P., Bird, I., Calore, F., Doglioni, C., Eckner, Ch. et al. (2023) 'ESCAPE dark matter science project for EOSC future project (WP6.3)', *ARPHA Preprints*. doi: 10.3897/arphapreprints.e116673.

7. Scardaci, D.O., Sciacca, E., Hériché, J.-K., Van De Sanden, M., Klaas, W. et al. (2023) 'A landscape overview of the EOSC interoperability framework – capabilities and gaps', *Zenodo*, available at: https://zenodo.org/records/8399710 (accessed June 12, 2024).

8. Rubio-Montero, A.J., Pagán-Muñoz, R., Mayo-García, R., Pardo-Diaz, A., Sidelnik, I., Asorey, H. (2021) 'The EOSC-Synergy cloud services implementation for the Latin American Giant Observatory (LAGO)', *Proc. ICRC2021*, 395. doi: 10.22323/1.395.0261.

9. Michener, W.K., Allard, S., Budden, A., Cook, R.B., Douglass, K. (2012) 'Participatory design of DataONE–Enabling cyberinfrastructure for the biological and environmental sciences', *Ecological Informatics*, 11, pp. 5–15. doi: 10.1016/j.ecoinf.2011.08.007.

10. Isard, M., Budiu, M., Yu, Y., Birrell, A., Fetterly, D. (2007) 'Dryad: Distributed data-parallel programs from sequential building blocks', *ACM SIGOPS Operating Systems Review*, 41(3), pp. 59–72. doi: 10.1145/1272998.1273005.

11. Stal, C., Covataru, C., Müller, J., Parnic, V. et al. (2022) 'Supporting long-term archaeological research in southern romania chalcolithic sites using multi-platform UAV mapping', *Drones*, 6(10), art. 277. doi: 10.3390/drones6100277.

12. Loupian, E.A., Proshin, A.A., Bourtsev, M.A., Kashnitskii, A.V., Balashov, I.V. et al. (2019). 'Experience of development and operation of the IKI-Monitoring center for collective use of systems for archiving, processing and analyzing satellite data', *Current Problems in Remote Sensing of the Earth from Space*, 16(3), pp. 151–170 (in Russ.). doi: 10.21046/2070-7401-2019-16-3-151-170.

13. Lecarpentier, D., Wittenburg, P., Elbers, W., Michelini, A., Kanso, R., Coveney, P., Baxter, R. (2013) 'EUDAT: A new cross-disciplinary data infrastructure for science', *IJDC*, 8(1), pp. 279–287. doi: 10.2218/ijdc.v8i1.260.

Авторы
Гарев Кирилл Валерьевич [1, 2],
руководитель направления,
аспирант, kv@garev.ru

[1] Департамент информационных технологий
Банка России, г. Москва, 107016, Россия
[2] Национальный исследовательский центр
«Курчатовский институт» – НИИСИ,
г. Москва, 117218, Россия

Authors
Kirill V. Garev [1, 2], Team Lead,
Postgraduate Student,
kv@garev.ru

[1] Information Technology Department of the Bank
of Russia, Moscow, 107016, Russian Federation
[2] National Research Centre
"Kurchatov Institute",
Moscow, 117218, Russian Federation