Модель исследовательской инфраструктуры в области комьютерных наук

Калинин Н. А. , Скворцов Н. А.

Федеральный исследовательский центр «Информатика и управление»
Российской Академии Наук, Москва 119333

Аннотация В последние годы научное сообщество уделяет все более пристальное внимание вопросам, связанным с эффективным обменом исследовательскими данными. Однако предпринимаемые усилия и разрабатываемые инструменты не всегда одинаково эффективны для различных областей научного знания. В данной статье проводится анализ существующих инструментов и подходов для размещения исследовательских данных в области компьютерных наук, выделяются основные трудности при обеспечении принципов FAIR для таких данных, приводятся соображения по разработке исследовательской инфраструктуры для преодоления выявленных трудностей.

Keywords: Ontology · FAIR · Research infrastructures · Open Science

1 Введение

В наше время ученые при проведении исследований сталкиваются с огромным объемом данных, которые необходимо хранить, обрабатывать и публиковать для обеспечения воспроизводимости результатов. Однако не всегда эти данные легко доступны и могут быть использованы в других исследованиях без дополнительной работы по их подготовке. В связи с этим, в последнее время все большее внимание уделяется принципам FAIR [20], которые декларируют необходимость обеспечения обнаруживаемости, доступности, интероперабельности и пригодности для повторного использования исследовательских данных.

Принципы FAIR с момента первой публикации получили ряд расширений для конкретных предметных областей, таких как [18] и [9], и были учтены в той или иной степени практически во всех крупных сервисах, предназначенных для размещения исследовательских данных. Достигнутые результаты демонстрируют значительное влияние развитых исследовательских инфраструктур, основанных на этих принципах, как на общую эффективность исследований, так и на смежные ненаучные области [12]. К сожалению, несмотря на значительное количество предпринимаемых усилий в области обеспечения повторного исспользования научных данных с учетом принципов FAIR, в ней сохраняются существенные трудности. Они включают в себя стороности, связанные с недостатком осведомленности ученых о

важности публикации и использования FAIR-данных и сервисов, проблемы научного администрирования, проблемы стандартизации, а также технические сложности, возникающие при формировании даннных и реализации сервисов по принципам FAIR. Именно последним будет посвящена настоящая работа.

Как было показано и отмечено в нашей предыдущей обзорной работе [8], не во всех областях популяризация принципов FAIR-данных проходит одинаково хорошо. Одной из причин такой диспропорции является недостаточность поддержки потребностей предметной области в инфраструктурах, предлагаемых исследователям. В работе [5] также отмечается, что большая часть исследователей ограничивается публикацией своих данных в неспециализированных сервисах (gitlab, archive) или хорошо известных репозиториях с ограниченным функционалом (zenodo, dryad). Эта ситуация особенно ярко наблюдается в области компьютерных наук.

Цель данного исследования заключается в разработке подходов к организации исследовательской инфраструктуры в компьютерных наук. Для достижения этой цели в первом разделе работы проводится анализ уже существующих инструментов и подходов, используемых в них. Во втором разделе формулируются критерии, которым должны соответствовать исследовательские инфраструктуры в области компьютерных наук. В третьем разделе предлагаются некоторые соображения по реализации инфраструктуры, соответствующей сформулированным критериям.

Указанная цель является одним из этапов в решении актуальной научной проблемы: разработки методов построения инфраструктуры исследовательских данных, которые могли бы использоваться во всех основных предметных областях без дополнительной адаптации, и учитывали подходы к анализу потребностей предметных областей при их развитии с течением времени. Другие этапы исследования включают в себя разработку аналогичных подходов для всех основных предметных областей, систематизацию самого процесса анализа потребностей предметных областей, классификацию предложенных подходов и обобщение разработанных программных и архитектурных решений.

2 Обзор существующих инструментов

2.1 Отдельные элементы исследовательских инфраструктур

Инфраструктуры научных данных представляют собой системы предназначенные, в первую очередь, для хранениями обработки исследовательских объектов, поэтому ключевым их компонентом можно считать репозитории данных. Такие репозитории представлены и отдельным продукатми, например zenoodo[2] и dryad[19]. По мере увеличения колличества различных репозиториев данных, для обеспечения обнаруживаемости данных были созданы каталоги репозиториев и схем метаданных, такие как re3data[13] и fairsharing[15]. Данные каталоги получили свое развитие в виде сервисов

поиска научных данных, таки как EOSC portal/openAIRE [11] и CORE [1], которые обеспечивают интеграцию не только репозиториев данных (или метаданных), но и других полезны источников исследовательских данных. Работа с исследовательскьими объектами, как в момент публикации, так и при повторном использовании, обеспечивается засчет различных сервисов-посредников, например EGI cloud compute [4] и Language Resource Switchboard [22], сервисов OSF [7]. Для интеграции различных компонентов инфраструктуры также использую стандарты и каталоги метаданных и интерфейсов (API), например DCAT, RO-CRATE, OPEN-API. Далее перечисленные элементы будут расмотрены более подробно.

2.2 Репозитории научных данных

Самая многочисленная группа решений в области открытой науки и исследовательских данных представлена различными репозиториями данных. Репозитории можно разделить по характеру хранимых в них данных на репозитории общего назначения и репозитории, адаптированные под конкретную предметную область. Также репозитории отличаются по типу организации; можно выделить крупные централизованные репозитории, часто создаваемые на уровне государств или крупных научных ассоциаций, и репозитории на уровне исследовательских институтов или отдельных групп.

Поиск репозиториев для настоящего исследования проводился на основе каталогов репозиториев: re3data.org, fairsharing.org. Рассматривались репозитории общего назначения, так как крупные репозитории с эксклюзивной поддержкой для данных в области компьютерных наук отсутствуют. Репозитории для анализа выбирались на основе объема размещенной в них информации и популярности.

Zenodo[1]. Zenodo - это крупный междисциплинарный репозиторий, позволяющий хранить и распространять исследовательские данные в произвольном формате. Объекты в Zenodo хранятся вместе с уникальным идентификатором DOI и метаданными, соответствующими критериям DataCite [17]. Метаданные отделены от данных и всегда доступны. Данные связаны с разместившими их пользователями. Доступ к данным предоставляется с помощью REST-API, предусмотрены возможности ограниченного доступа. Также осуществляется рецензирование на базе "сообществ".

В Zenodo отсутствует специфическая поддержка для разных предметных областей. Однако следует отметить интеграцию с GitHub и большое количество размещенного программного кода (примерно 100 000 записей по состоянию на май 2023 года).

Dryad[2]. Репозиторий Dryad управляется консорциумом исследовательских институтов и поддерживает публикацию произвольных данных в аналогичной Zenodo манере. Для хранения данных используется протокол SWORD,

[1] https://zenodo.org/
[2] https://datadryad.org

данные доступны по API и хранятся вместе с метаданными. Поиск осуществляется после извлечения метаданных из хранилища с помощью компонента HARVEST по протоколу OAI-PMH. Поддерживаются идентификаторы DOI и ORCID. Главная особенность Dryad это процесс курирования данных, который обеспечивает качество метаданных и доступность публикаций.

Science Data Bank[3]. Это крупный репозиторий, поддерживаемый китайскими институтами. Он отличается значительным объемом депонированных данных и поддерживает идентификаторы DOI и CSTR. Доступны для публикации различные типы данных, включая наборы данных, изображения и таблицы, слайды, программный код. В репозитории присутствует 36 тысяч записей с типом "данные программного кода". Метаданные доступны по открытой лицензии и могут быть загружены в нескольких форматах. У репозитория отсутствует специфическая поддержка для программного кода, однако он сфокусирован на управлении цитируемостью данных и подсчете статистики для оценки результатов научных исследований.

EUDAT B2SHARE[4]. B2SHARE поддерживает описания метаданных через схему метаданных EUDAT, регистрирует DOI для наборов данных и Handle PIDs для объектов данных, поддерживает многоверсионность данных, собирается инструментами B2FIND и OpenAIRE Explorer, разрешает прямую загрузку из B2DROP, доступен через веб-интерфейс и API для поддержки автоматических процессов публикации, поддерживает предметные сообщества, позволяет определять расширения метаданных, правила доступа и рабочие процессы для публикации данных и позволяет аннотировать исследовательские данные через расширение B2NOTE. В репозитории отсутствует специфическая поддержка для программного кода, но существует возможность расширения метаданных. B2SHARE является частью EOSC и направлен на использование в качестве компонента инфраструктуры, поэтому обладает ограниченными возможностями по поиску, анализу и публикации данных. Интересной особенностью B2SHARE является поддержка пяти различных типов лицензий.

Harvard Dataverse[5]. Один из самых крупных репозиториев, основанных на популярном решении с открытым исходным кодом Dataverse. Он поддерживает как интеграцию с другими репозиториями Dataverse, так и размещение собственных материалов. Для идентификации используется DOI, а метаданные соответствуют требованиям DataCite. Репозиторий обеспечивает доступность метаданных и включает ссылки на страницу с ними при цитировании. Поддерживает предметные онтологии, такие как DDI. Возможно размещение программного кода, но отсутствует специфическая поддержка. Данные публикуются в виде коллекций Dataverse, которые представляют собой контейнеры для данных и выбранных пользователем метаданных наборов. Для обеспечения машиночитаемости предоставляется API, а для повторного использования - возможность просмотра метрик цитиро-

[3] https://www.scidb.cn
[4] https://b2share.eudat.eu/
[5] https://dataverse.harvard.edu/

вания. Этот репозиторий можно считать образцовым для всех решений на основе Dataverse.

Figshare[6]. Репозиторий, ориентированный на хранение объектов-приложений к публикациям в научных журналах. Данные в репозитории размещаются с идентификатором DOI и сопровождаются изменяемым набором метаданных, зависящим от журнала. Для поиска используются внешние поисковые решения. Все метаданные доступны по открытым лицензиям, поддерживается выгрузка в различных форматах. Объекты могут сопровождаться ссылками на рецензируемые публикации, что позволяет проверить их происхождение. Значительной особенностью Figshare является поддержка интеграций во внешние процессы рецензирования. Несмотря на теоретическую возможность использования наборов метаданных для публикаций в области компьютерных наук, на практике они отсутствуют.

MathRepo[7]. Несмотря на то, что данный репозиторий не является крупным или популярным, его включение в обзор обусловлено демонстрацией наличия специфической поддержки рассматриваемой области в небольших предметных репозиториях. MathRepo специализируется на исследованиях в области компьютерной алгебры MathPro и включает в себя поддержку размещения моделей в виде специального программного кода, а также сохранение его вывода в виде ноутбуков.

В заключение можно сказать, что в крупных междисциплинарных репозиториях отсутствуют механизмы специфической поддержки предметных областей. Эта задача перекладывается на компоненты более высокого уровня или на плечи предметных сообществ. Для области компьютерных наук это означает фактическое отсутствие такой поддержки, за исключением поиска по метаданным типов файлов, соответствующих программному коду. Большая часть репозиториев является централизованной (с одним главным провайдером хранения данных) и представляет стандартный минимальный набор метаданных, необходимых для цитирования, использует идентификатор DOI и поддерживает разделение метаданных и данных с точки зрения доступа. Первые доступны всегда, для вторых поддерживается ограничение доступа. Под возможностями повторного использования обычно понимается программная доступность метаданных. Репозитории часто интегрированы как между собой, так и с внешними источниками, такими как GitHub.

2.3 Каталоги репозиториев и сервисы поиска научных данных

Как было показано в предыдущем разделе, существует значительное количество как крупных, так и более мелких репозиториев данных. Для обеспечения обнаруживаемости данных в этих репозиториях были созданы каталоги репозиториев, которые позже превратились в порталы научных данных. Наиболее известными каталогами являются упомянутые ранее re3data.org и fairsharing.org.

[6] https://figshare.com/
[7] https://mathrepo.mis.mpg.de/

Поскольку каталоги не всегда обладают удобным интерфейсом для поиска, появились сервисы-агрегаторы и связанные с ними порталы научных данных. Примером таких решений являются сервисы openAIRE[8] и связанный с ним портал EOSC[9], а также портал CORE[10].

Каталоги репозиториев re3data.org и fairsharing.org организованы сходным образом. Для различных репозиториев и источников создается запись, которая дополняется метаданными. Основные метаданные включают принадлежность к предметной области, стране и организации, тип хранимых данных. Re3data.org обладает более широким перечнем допустимых метаданных. А особенностью fairsharing является хранение записей не только о самих репозиториях, но также и об отраслевых стандартах, политиках. Каталоги репозиториев представляют ценность как источник данных для исследований и поисковых сервисов.

В Европейском облаке октрытой науки EOSC, отличающейся федеративной структурой, особенно интересны сервисы интеграции и поиска, предоставляемые в рамках инициативы openAIRE. Интеграция осуществляется через сервис OpenAIRE Connect.

OpenAIRE Connect предоставляет инфраструктуру для сбора, хранения и обмена метаданными, связанными с научными результатами и деятельностью. Этот сервис облегчает сбор и стандартизированное хранение метаданных, что способствует повышению доступности научных результатов. Ключевым сервисом с точки зрения обеспечения обнаруживаемости является OpenAIRE Explore. Это сервис поиска, который позволяет пользователям искать научные результаты в различных источниках, включая институциональные репозитории, архивы данных и издательства. Пользователи могут выполнять поиск по ключевым словам, авторам, организациям и другим критериям, а также фильтровать результаты поиска по дате публикации, типу документа и языку. OpenAIRE Explore использует рекомендации OpenAIRE для метаданных, обеспечивая последовательность и стандартизацию метаданных от разных поставщиков данных. Визуализация агрегированных метаданных доступна с помощью сервиса openAIRE Graph, который позволяет представить общий ландшафт агрегированных научных данных.

Другим примером решения этой задачи является платформа CORE, предоставляющая сервисы для поиска, доступа и анализа научных статей и других исследовательских результатов, размещенных в открытых репозиториях и архивах по всему миру. Поиск осуществляется через сервис CORE Discovery, а также доступен сервис CORE Recommender для рекомендации статей на основе алгоритмов машинного обучения.

С точки зрения области компьютерных наук, сервисы CORE и OpenAIRE практически не предоставляют специфических возможностей для поиска данных. Поиск осуществляется по метаданным, предоставляемым самими

[8] https://openaire.eu/
[9] https://eosc-portal.eu/
[10] https://core.ac.uk

репозиториями, которые обычно ограничиваются информацией о форматах исследования и происхождении (страна, автор, область знания).

2.4 Сервисы-посредники в исследовательских инфраструктурах

Как отмечено в [16], посредники открытых данных предоставляют специализированные ресурсы и возможности для использования открытых данных. В современных исследовательских инфраструктурах эта функциональность поручается сервисам-посредникам. Именно эти сервисы обеспечивают агентам дополнительные возможности по обработке, анализу, визуализации и, в конечном итоге, обеспечивают повторное использование данных.

Для целей данной работы были использованы сервисы из каталога EOSC[11], как самого широкого из доступных. Из всех сервисов, представленных на портале, были выбраны следующие, как наиболее интересные:

1. EGI Notebooks: Облачный сервис для работы с данными с помощью платформы Jupyter Notebook.

2. DODAS: Сервис для интерактивного анализа, основанный на платформах Spark и Jupyter Notebook для интерактивного анализа, а также HTCondor для пакетной обработки данных.

3. TSD: Сервис для работы с приватными данными, предоставляющий безопасное окружение.

4. VD-Maps: Сервис для визуализации карт на основе наборов данных.

5. Enrichment API: Сервис для автоматического создания метаданных, использующий закрытое программное обеспечение expert.ai для извлечения метаданных на основе семантической сети.

6. PSI Remote Desktop: Сервис предоставления удаленного рабочего стола.

7. AI4GEO: Сервис для анализа данных с множеством возможностей, включая Jupyter, IDE и открытые AI-фреймворки.

8. AMNESIA: Сервис для анонимизации конфиденциальных данных с возможностью управления процессом.

9. EGI Cloud Container Compute: Сервис для запуска контейнеризованных приложений.

10. HOSTKEY GPU GRANT PROGRAM: Программа предоставления вычислительных мощностей графических ускорителей.

11. EGI Workload Manager: Сервис для организации распределенных вычислений.

12. EGI Online Storage, EGI Data Transfer: Сервисы для хранения и гарантированной передачи данных.

13. Nuvla.io SAAS: Сервис для предоставления облачных технологий широкого спектра от Nuvla.io.

14. robotbenchmark: Веб-сервис с облачным программным обеспечением для 3D-моделирования роботов, который предоставляет задачи по робототехнике исследователям и студентам.

[11] https://search.marketplace.eosc-portal.eu

15. IFCA-CSIC Cloud Infrastructure: Сервис для предоставления облачной инфраструктуры на основе OpenStack.

16. Infrastructure Manager: Сервис для управления вычислительной инфраструктурой, включая виртуальные машины и облачные решения, с поддержкой процессов CI/CD на основе Ansible.

17. DEEP Training Facility: Сервис, предоставляющий инструменты для создания моделей машинного обучения и искусственного интеллекта в распределенных инфраструктурах.

Таким образом, можно выделить следующие основные направления, в которых разработаны сервисы-посредники: виртуальные платформы для исследований на основе Jupyter Notebook, организация облачных вычислений и других услуг, предоставляемых облачными провайдерами (хранение на S3, передача данных), сервисы для работы с приватными данными. Особый интерес представляют сервисы Enrichment API, решающие проблему автоматизированного обогащения метаданных, и сервис DEEP, предоставляющий доступ к известным моделям и алгоритмам в области искусственного интеллекта и машинного обучения.

3 Поддержка компьютерных наук в исследовательских инфраструктурах

3.1 Прицнипы FAIR

В статье [20] описаны 15 принципов FAIR. Эти принципы разбиты на 4 категории, входящие в акроним: обнаруживаемость (F), доступность (A), интероперабельность (I) и возможность повторного использования (R). В качестве обозначения для принципов используются буква соответсвующая категория и номер принципа. С момента публикации указанной основопалагающей работы в 2016 году принципы FIAR стали основной для оценки качества публикуемых научных данных и средств для их публикации. Несмотря на то, что они не представляют собой строгих инструкций, они задают направление, следование которому, как показала практика и многочисленные работы, обеспечивает более эффективный научный обмен. В этом разделе рассмотрен вопрос полноты воплощения этих принципов для исследовательских данных в существующихи исследовательских инфраструктурах в области компьютерных наук.

3.2 Обеспечение доступности

Большая часть существующих инструментов частично соответвует принципа доступности.

(Мета)данные извлекаются по их идентификатору с использованием стандартизированного протокола (A1). Важно отметить, что для настоящей доступности требуется открытость не только протокола, но и открытое описание программного интерфейса.

Протокол допускает процедуру аутентификации и авторизации при необходимости (A1.2). Вопрос разграничения доступа (как с точки зрения лицензирования, так и с точки зрения конфиденциальности), в большинстве случаев, решается путем обеспечения полной открытости метаданных и ограниченного набора параметров для ограничения доступа к данным. Именно такое решение наблюдается во всех рассмотренных в предыдущем разделе репозиториях. Как показано в [3], существует потребность и в разграничении доступа к метаданным, особенно если эти метаданные автоматически извлечены из исследовательского объекта. Особенность области компьютерных наук, в данном случае, заключается в необходимости поддерживать различные типы ограничений доступа к элементам одного исследовательского артефакта, например, при совместном использовании компонентов с закрытым и открытым исходным кодом.

Метаданные доступны, даже если данные больше недоступны (A2). Во всех рассмотренных репозиториях принцип A2 выполняется на концептуальном уровнею, но важно помнить, что соответветсвие ему неразрывно связана с вопросом обеспечения технической надежности хранения данных. Примером гарантии этой надежности является использование хранилища CERN для Zenodo. Альтернативой дорогим высоконадежным централизованным системам могут являться децентрализованные, распределенные системы. Их возможности в рассмотренных репозиториях не используются.

3.3 Обеспечение обнаруживаемости

Как показано в предыдущем разделе, существующие ресурсы обеспечивают некоторый уровень соответствия принципам обноаруживаемости, в том числе, для области компьютерных наук. (Мета)данным присваивается глобальный уникальный и постоянный идентификатор (F1) Метаданные четко и недвусмысленно включают идентификатор данных, которые они описывают (F3) Данные регистрируются или индексируются на ресурсе с возможностью поиска (F4).

Таким образом, дополнительных усилий требует только выполнения принципа. Данные описаны богатыми метаданными (F2) Данные описываются с использованием обширных метаданных. Существует два подхода к решению этого вопроса. Первый подход, отмеченный, в частности, в статье [6], состоит в разработке более полных стандартов метаданных для предметных областей. Являясь основным подходом на протяжении длительного времени, он, тем не менее, обладает ограниченной эффективностью, так как требует проведения серьезной дополнительной работы со стороны исследователя, публикующего свои результаты. Второй подход представлен, в частности, сервисом Enrichment API в EOSC, и представляет собой дополнение вручную сформированных стандартных метаданных метаданными, извлеченными автоматически. Примером метаданных, извлекаемых автоматически для исследований в области компьютерных наук, являются схемы наборов данных, используемые библиотеки, внешнее программное обеспечение, характе-

ристики программного кода, такие как используемые алгоритмы, шаблоны проектирования и так далее. Совмещение этих двух подходов, которое позволит использовать отраслевые стандарты и освободит исследователей от тяжелой и неочевидной работы, должно быть обеспечено инфраструктурой.

3.4 Обеспечение интероперабельности

Принципы (мета)данные используют формальный, доступный и широко применимый язык для представления знаний (I1), и (мета)данные используют словари (I2) концептуально учитываются в существующих решениях. Однако, как показано в [14], степень соответствия им зависит от качества представлений метаданных и их словарей соответственно. Для конкретных предметных областей это означает, что выбранный язык должен удовлетворять требованиям этой области и что выбраны или построены словари, предоставляющие корректную и общепринятую интерпретацию терминов. В рассмотренных репозиториях и сервисах экосистемы EOSC использование словарей уровня инфраструктуры ограничено междисциплинарными стандартами.

Несколько хуже обеспечивается соответствие принципу (мета)данные содержат квалифицированные ссылки на другие (мета)данные (I3). Наличие квалифицированных ссылок означает интеграцию с другими инфраструктурами и в минимальной степени представлено во всех инфраструктурах. Во всех рассмотренных репозиториях используются ссылки на другие репозитории (обычно на Zenodo или GitHub), DOI, а также указание провайдеров сервисов. Наиболее развитой реализацией этой интеграции является сервис OpenAIRE. Таким образом, описанная интеграция осуществляется путем внесения данных в каталоги данных, подключения сервисов поиска, использования общепринятых идентификаторов и базовых наборов метаданных. Кроме того, как показано в [10], это требование также означает необходимость наличия квалифицированных ссылок на программные зависимости. Поскольку программные зависимости являются специфичным типом данных, в рассмотренных инфраструктурах вопрос о наличии квалифицированных ссылок на них не решен.

3.5 Обеспечение воспроизводимости и повторного использования

(Мета)данные богато описаны с множеством точных и актуальных атрибутов (R1). Реализация этого принципа неразрывно связана с реализацией принципа F2, за исключением того, что точно и актуально должны быть описаны не только атрибуты для поиска, но и атрибуты, необходимые для воспроизводимости и повторного использования исследовательского объекта. Актуальность данных и метаданных предполагает реализацию сервисов для поддержки процессов сопровождения данных, а также сервисов анализа метаданных на предмет их точности и актуальности.

(Мета)данные выпущены с четкой и доступной лицензией на использование данных(R1.1) Функционал лицензирования широко представлен в существующих инфраструктурах и реализован во всех рассмотренных репозиториях. Единственной специфичной для области компьютерных наук особенностью, как отмечалось выше, является необходимость поддерживать отдельные лицензии для отдельных компонентов одного и того же исследовательского объекта (например, одна лицензия для кода, другая для данных и моделей, третья для внешних зависимостей).

(Мета)данные связаны с подробной историей происхождения.(R1.1) Доступные решения в некоторой мере обеспечивают учет этого принципа. Предоставляются метаданные об авторах, организация, и в некоторых случаях предоставляется также сервис рецензирования публикуемых материалов. Для публикуемого программного кода актуальным является перенос понятий происхождения и авторства с целого объекта на его составные части. Программные средства контроля версий позволяют определить авторство с точностью до строки, однако ни одно из рассмотренных решений не использует эту особенность. То же самое относится и к реализации многоверсионности в существующих решениях, которая может быть расширена для типов исследовательских объектов, позволяющих такое расширение.

(Мета)данные соответствуют стандартам, релевантным для соответствующей области.(R1.3) Реализация этого принципа обеспечивается наличием предметных сообществ в рассмотренных решениях, что особенно заметно по федеративной структуре EOSC. Однако стандарты для области компьютерных наук тесно связаны со стандартами, принятыми вне научной среды, и должны подвергаться регулярному пересмотру по мере их развития. Как показано в [10], для программного обеспечения этому требованию должны удовлетворять не только метаданные, но и документация.

Стоит отметить, что для многих областей потребности для успешного повторного использования не ограничиваются приведенными принципами. В частности, в [21] показана исключительная важность обеспечения технических возможностей повторного использования. Это означает, что должна существовать возможность программного исполнения, вычисления, запуска и обработки для всех рассматриваемых артефактов. Такие возможности успешно реализованы, например, в сервисах сопровождения облачных вычислений, таких как EGI, nulva.io, infrastructure manager и др. Перспективным вариантом развития этих сервисов является их более плотная интеграция с другими компонентами инфраструктуры, чтобы анализ исследовательских артефактов мог быть осуществлен исследователем полностью внутри инфраструктуры, без необходимости дополнительной обработки данных или использования сторонних вычислительных и аналитических ресурсов. Разумеется, инфраструктура не может удовлетворить все потребности в воспроизведении исследования, но она должна поддерживать указанные требования, как минимум, для самых популярных типов исследовательских артефактов.

4 Разработка исследовательской инфраструктуры

4.1 Ключевые задачи системы

На основе проведенного исследования были сформулированы следующие требования к исследовательской инфраструктуре в области компьютерных наук, исходя из того, что она должна способствовать максимально полному соответствию размещаемых в ней научных данных принципам FAIR:

1. Как было показано в разделах 3.2 и 3.3, существующие решения не предоставляют отдельных возможностей для размещения в них специфических для рассматриваемой предметной области типов данных. Кроме того, существующие подходы к регистрации различных версий одного исследовательского артефакта не соответствуют общепринятым подходам для программного кода. Таким образом, для устранения указанных недостатков инфраструктура должна обеспечивать хранение и версионность популярных типов данных, включая, но не ограничиваясь, следующими: программным кодом, моделями машинного обучения, ноутбуками, данными, представленными в виде дампов популярных типов баз данных, электронными таблицами, файлами JSON. Список популярных типов исследовательских артефактов был составлен на основе анализа датасета DBLP [12].

2. Для обеспечения интероперабельности, как показано в разделах 3.3 и 3.4, необходимы полные и понятные метаданные. Составление таких метаданных представляет собой существенную сложность для исследователей, поэтому инфраструктура должна предоставлять соответствующие сервисы: с одной стороны, популярные словари и схемы метаданных с учетом предметной области, а с другой стороны, сервисы по их автоматизированному извлечению. Важно отметить, что такого рода сервисы в современных инфраструктурах практически не представлены.

3. Как было отмечено в разделе 3.5, примечательной особенностью области компьютерных наук является то, что большая часть исследовательских артефактов представляет собой компьютерные программы, в том или ином виде. Повторное использование таких артефактов означает, что должна обеспечиваться возможность их исполнения. Инфраструктура должна реализовывать эту возможность.

4. Обеспечение одновременно конфиденциальности и доступности - нетривиальная задача. Которая, фактически, не решается в существующих инфраструктурах, как было показано в разделе 3.1. Решением этой дилеммы может стать поддержка многоуровневых метаданных на уровне инфраструктуры.

5. Одной из проблем существующих инфраструктур, как показано в разделе 3.4, является их многообразие, которое существенно затрудняет интеграцию между собой. Для того чтобы поддерживать корректные и полные ссылки на другие исследовательские артефакты, инфраструктура должна быть интегрирована с другими инфраструктурами и продуктами как в области исследовательских данных, так и в области компьютерных наук.

[12] https://arxiv.org/pdf/2204.13384.pdf

4.2 Основные компоненты

Для решения сформулированных задач необходима реализация следующих высокоуровневых составляющих:

1. Подсистема хранения исследовательских артефактов. Подсистема должна обеспечивать хранение перечисленных выше типов исследовательских данных, а также связанных с ними метаданных. Кроме того, указанная система должна предоставлять интерфейс доступа, соответствующий отраслевым стандартам для репозиториев - git или svn. Также, подсистема хранения должна поддерживать процессы, гарантирующие достаточность базовых метаданных при публикации. Наконец, связь артефактов между собой может обеспечиваться на двух уровнях: логическом - на уровне метаданных и физическом - путем размещения вместо использованного объекта ссылки на него.

2. Подсистема запуска исполнимых артефактов. Данная подсистема должна обеспечивать возможность запустить те артефакты, для которых это возможно. Для того чтобы это стало возможным, такие артефакты должны быть представлены в стандартизованном виде или иметь соответствующие описания. Кроме того, должна обеспечиваться возможность запустить исполнение на внешних вычислительных ресурсах. Для реализации этой подсистемы могут быть использованы технологии контейнеризации и оркестрации (docker, kubernetes), а также фреймворки для декларативного описания инфраструктуры (terraform, ansible).

3. Подсистема управления метаданными. Как было описано выше, она должна обеспечивать несколько связанных с управлением метаданными процессов. Во-первых, предоставлять сервис-конструктор метаданных, включающий в себя онтологии высокоуровневых терминов, стандартные описания, а также рекомендации по составлению собственных наборов метаданных. Во-вторых, сервис по автоматизированному извлечению метаданных. А в третьих, поддерживать несколько уровней метаданных, чтобы доступ к метаданным разных уровней можно было разграничивать.

4. Подсистема интеграции с внешними системами. Эта подсистема должна реализовывать возможность получения данных из других инфраструктур и прочих источников.

5 Выводы и перспективы развития

В ходе данной исследовательской работы был проведен краткий обзор состояния исследовательских инфраструктур, включая область компьютерных наук. В результате этого обзора были сформулированы задачи, решение которых необходимо для устранения недостатков существующих исследовательских инфраструктур, которые замедляют их внедрение в конкретных предметных областях. На данном этапе анализ проведен в одной области. Были выделены компоненты, реализация которых обеспечит решение поставленных задач. Для некоторых компонентов предложены технологии, с

помощью которых они могут быть реализованы. Для полноценной экспериментальной проверки разработанной модели планируется создание полноценного работоспособного прототипа описанных компонентов.

Нашей глобальной целью является разработка междисциплинарных решений для публикации исследовательских данных. Для достижения этой цели не только реализуется прототип описанной архитектуры, но также предполагается ее расширение с учетом требований и проблем других научных областей.

Список литературы

1. Aggregating the world's open access research papers (May 2023), https://www.core.ak.uk/
2. Research.shared (May 2023), https://www.zenodo.org/
3. Brewster, C., Nouwt, B., Raaijmakers, S., Verhoosel, J.: Ontology-based access control for fair data. Data Intelligence **2**(1-2), 66–77 (2020)
4. Fernández-del Castillo, E., Scardaci, D., García, Á.L.: The egi federated cloud e-infrastructure. Procedia Computer Science **68**, 196–205 (2015)
5. Commission, E., for Research, D.G., Innovation: European Research Data Landscape : final report. Publications Office of the European Union (2022). https://doi.org/doi/10.2777/3648
6. Farnel, S., Shiri, A.: Metadata for research data: current practices and trends. In: International conference on Dublin core and metadata applications. pp. 74–82 (2014)
7. Foster, E.D., Deardorff, A.: Open science framework (osf). Journal of the Medical Library Association: JMLA **105**(2), 203 (2017)
8. Kalinin, N., Skvortsov, N.: Difficulties of fair principles implementation in cross-domain research infrastructures. Lobachevskii Journal of Mathematics **44**(1), 147–156 (2023)
9. Lamprecht, A.L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P.A., et al.: Towards fair principles for research software. Data Science **3**(1), 37–59 (2020)
10. Lamprecht, A.L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P.A., et al.: Towards fair principles for research software. Data Science **3**(1), 37–59 (2020)
11. Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., Principe, P., Artini, M., Becker, A., De Bonis, M., et al.: The openaire research graph data model. Zenodo (2019)
12. Martínez-García, A., Alvarez-Romero, C., Román-Villarán, E., Bernabeu-Wittel, M., Parra-Calderón, C.L.: Fair principles to improve the impact on health research management outcomes. Heliyon (2023)
13. Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.J., Gundlach, J., Schirmbacher, P., Dierolf, U.: Making research data repositories visible: the re3data. org registry. PloS one **8**(11), e78080 (2013)
14. Quarati, A., Raffaghelli, J.E.: Do researchers use open research data? exploring the relationships between usage trends and metadata quality across scientific disciplines from the figshare case. Journal of Information Science **48**(4), 423–448 (2022)

15. Sansone, S.A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M.: Fairsharing as a community approach to standards, repositories and policies. Nature biotechnology **37**(4), 358–367 (2019)
16. Shaharudin, A., Van Loenen, B., Janssen, M.: Towards a common definition of open data intermediaries. Digital Government: Research and Practice (2022)
17. Starr, J., Gastl, A.: iscitedby: A metadata scheme for datacite. D-lib magazine: a monthly magazine about innovation and research in digital libraries **17**(1/2) (2011)
18. Th, M., MartinGillian, M., Ugur, S., van OmmenGertJan, B., et al.: Enhancing reuse of data and biological material in medical research: from fair to fair-health. Biopreservation and biobanking (2018)
19. White, H., Carrier, S., Thompson, A., Greenberg, J., Scherle, R.: The dryad data repository: A singapore framework metadata architecture in a dspace environment. In: Dublin core conference. pp. 157–162 (2008)
20. 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., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)
21. Wolf, M., Logan, J., Mehta, K., Jacobson, D., Cashman, M., Walker, A.M., Eisenhauer, G., Widener, P., Cliff, A.: Reusability first: Toward fair workflows. In: 2021 IEEE International Conference on Cluster Computing (CLUSTER). pp. 444–455. IEEE (2021)
22. Zinn, C.: The language resource switchboard. Computational Linguistics **44**(4), 631–639 (2018)