НАУЧНЫЙ ОБЗОР

УДК 004.9
DOI 10.17513/snt.40399

ИНСТРУМЕНТЫ АВТОМАТИЗАЦИИ ДЛЯ ОБЕСПЕЧЕНИЯ ВОСПРОИЗВОДИМОСТИ ИССЛЕДОВАНИЙ В НАУКЕ О ДАННЫХ

[1, 2]Горбунов В.И., [1]Салимов Т.А., [1]Горбань Е.В.

[1]ФГАОУ ВО «Национальный исследовательский институт ИТМО», Санкт-Петербург, e-mail: gorbunov.v93@gmail.com;
[2]ФГБОУ ВО «Санкт-Петербургский государственный университет», Санкт-Петербург

Современные междисциплинарные проекты в области науки о данных характеризуются высокой сложностью, множеством участников и необходимостью координации организационных и технических процессов. Одной из ключевых проблем в таких проектах является обеспечение воспроизводимости методов и результатов исследований. Целью работы является проведение обзора современных практик и инструментов, направленных на повышение воспроизводимости в проектах науки о данных, и их анализ с точки зрения управления исследовательским процессом. Был проведен систематический обзор из более чем 50 публикаций за 2015-2025 годы, направленный на выявление современных организационных практик и инструментов, применяемых для обеспечения воспроизводимости в проектах науки о данных из научных и прикладных публикаций, документации инструментов в науке о данных и открытых репозиториев. Из них 30 работ легли в основу данного обзора. В работе рассмотрены пять ключевых категорий решений: контроль версий кода, данных и отчетов; управление зависимостями и средами исполнения; автоматизация процессов и оркестрация пайплайнов; стандартизация хранения данных; документирование и обеспечение прозрачности. Особое внимание уделено управленческому эффекту от их применения – снижению издержек, рисков и трудозатрат на коммуникации и выполнение типовых работ. Основными ограничениями внедрения инструментов воспроизводимости в организационные процессы остаются необходимость зрелой инфраструктуры, организационных изменений и обучения персонала. Представленные выводы могут быть использованы при разработке стандартов управления исследовательскими проектами, формировании корпоративной культуры прозрачности и выборе инструментов для применения.

Ключевые слова: управление исследованиями, наука о данных, воспроизводимость, управление автоматизацией исследований, организационно-технические системы

AUTOMATION TOOLS FOR ENSURING REPRODUCIBLE RESEARCH IN DATA SCIENCE

[1, 2]Gorbunov V.I., [1]Salimov T.A., [1]Gorban E.V.

[1]ITMO University, St. Petersburg, e-mail: gorbunov.v93@gmail.com;
[2]Saint-Petersburg State University, Saint-Petersburg

Modern interdisciplinary data science projects are characterized by high complexity, multiple participants, and the need to coordinate organizational and technical processes. One of the key challenges in such projects is to ensure reproducibility of research methods and results. The aim of the work is to review modern practices and tools aimed at improving reproducibility in data science projects, and to analyze them from the point of view of managing the research process. A systematic review of more than 50 publications from 2015-2025 was conducted, aimed at identifying modern organizational practices and tools used to ensure reproducibility in data science projects from scientific and applied publications, documentation of tools in data science and open repositories. Of these, 30 papers formed the basis of this review. The paper considers five key categories of solutions: version control of code, data and reports; dependency management and execution environments; automation of processes and pipeline orchestration; standardization of data storage; documentation and transparency. Special attention is paid to the management effect of their use – reducing costs, risks and labor costs for communication and standard work. The main limitations of implementing reproducibility tools in organizational processes remain the need for mature infrastructure, organizational changes, and staff training. The presented conclusions can be used in the development of standards for the management of research projects, the formation of a corporate culture of transparency and the selection of tools for application.

Keywords: research management, data science, reproducibility, research automation management, socio-technical systems

Введение

Современные исследования в области науки о данных всё чаще рассматриваются не только как технические проекты по построению моделей, но и как сложные организационные системы, в которых задействованы многопрофильные команды, сложные вычислительные пайплайны и разнообразная инфраструктура. В таких проектах затрагивается широкий круг участников, зачастую из разных подразделений: менеджеры, владельцы продуктов, бизнес-аналитики, IT-архитекторы, специалисты по инфраструктуре, эксперты в доменной области и другие заинтересованные лица. Многообразие ролей и интересов предъявляет высокие требования к координации, прозрачности и контролю работ на всех этапах исследования. Одним из ключевых факторов успеха таких междисциплинарных проектов является воспроизводимость

применяемых методов и полученных результатов [1-3].

Фрагментация инструментов, несогласованность этапов обработки данных, отсутствие стандартов в организации работы, а также неоптимальные коммуникации между управленческими и исполнительными ролями затрудняют управление исследовательской деятельностью [4]. Данные проблемы создают значительные барьеры для управляемости процессов, усложняют принятие решений и затрудняют внедрение результатов исследований в практическое использование, а также сервисную поддержку разработанных решений [5]. Современные проекты в области науки о данных требуют не только высокой точности моделирования, но и прозрачности, воспроизводимости и управляемости процессов проведения исследований в составе команды [6; 7].

Рассмотрение исследовательского проекта как организационно-технической системы, подлежащей управлению с использованием инструментов автоматизации, позволяет повысить надежность, воспроизводимость и устойчивость всего процесса получения знаний из данных [8; 9].

В условиях высокой динамики и сложности проектов с применением машинного обучения необходимость в формализации и стандартизации организационных процессов становится особенно актуальной. В таком случае воспроизводимость выступает не только инженерной, но и управленческой задачей, т.к. обеспечивает:

- прозрачность и возможность многосторонней верификации результатов – любые участники, от команды разработки до руководства организацией, могут проследить ход исследований и подтвердить достоверность выводов;
- управляемость проекта – при наличии четко описанных процедур и регламентов снижаются риски, связанные с нехваткой документированных знаний, что особенно важно при ротациях персонала и изменениях в приоритетах бизнеса;
- эффективное распределение ресурсов – формальные инструменты контроля и отчетности позволяют менеджерам отслеживать динамику исследований и вовремя принимать управленческие решения, например о перераспределении бюджета, назначении новых исполнителей и выделении вычислительных ресурсов.

Целью настоящего исследования является проведение обзора современных практик и инструментов, применяемых для обеспечения воспроизводимости в проектах науки о данных, и их анализ с точки зрения управления жизненным циклом исследовательского проекта в организационных системах.

Материалы и методы исследования

В рамках исследования был проведен систематический обзор более чем 50 публикаций за 2015-2025 годы, направленный на выявление современных организационных практик и инструментов, применяемых для обеспечения воспроизводимости в проектах науки о данных из научных и прикладных публикаций (Scopus, Web of Science, IEEE Xplore, ACM Digital Library, arXiv, PubMed), документации инструментов в науке о данных (DVC, ClearML, Airflow и др.) и открытых репозиториев (GitHub, GitLab). Из них 30 работ легли в основу данного обзора, в рамках которого воспроизводимость трактуется как управляемая характеристика исследовательской деятельности, обеспечиваемая взаимодействием организационных ролей, процессов и программных инструментов. Каждое из решений было сгруппировано по категориям обеспечения воспроизводимости и анализировалось с точки зрения влияния на эффективность организационного управления, включая старт выполнения новых задач, разделение ролей, поддержание стандартов и минимизацию рисков, связанных с человеческим фактором.

Результаты исследования и их обсуждение

В рамках анализа были выделены ключевые управленческие эффекты, отражающие зрелость и полноту поддержки воспроизводимости на уровне организационных решений:

- сокращение затрат ресурсов на типовые операции;
- сокращение затрат ресурсов на коммуникации между сотрудниками и согласование действий команды проекта;
- сокращение рисков из-за человеческого фактора при реализации задач (конфликтов версий, утрат критически важной информации, ручных ошибок).

Также были систематизированы основные категории технических решений, обеспечивающие воспроизводимость методов и результатов, применение которых влияет на перечисленные выше управленческие параметры.

1. Контроль версий и управление изменениями

Контроль версий и управление изменениями в проектах науки о данных выполняют функцию фиксирования состоя-

ния системы на каждом этапе жизненного цикла исследования, а также обеспечения возможности отката, сравнения, воспроизведения и анализа истории изменений. В условиях разработки решения несколькими командами и высокой динамики процессов контроль версий может быть рассмотрен как метод управления изменениями в проекте – позволяет сохранять историю развития проекта, обеспечивать прозрачность решений и обоснованность результатов, дает возможность проводить аудит и верификацию исследовательских действий, сокращает издержки на повторную настройку и поиск ошибок, обеспечивает передачу проекта между участниками без потери знаний.

Можно выделить 3 категории решений контроля версий и управления изменениями.

1.1. Контроль версий кода

Контроль версий кода традиционно осуществляется с использованием Git. Git является стандартом в разработке программного обеспечения и обеспечивает отслеживаемость изменений, совместную работу, откат к предыдущим версиям, проведение код-ревью и автоматизацию процессов [10]. На практике применяются различные модели работы с Git – Trunk-Based Development, Git Flow, GitHub Flow и т.д. Существуют также альтернативные инструменты – mercurial, fossil и прочие, которые менее распространены и могут потребовать больше ресурсов на обучение новых сотрудников. Для автоматизации проверки кода перед коммитами существуют инструменты pre-commit, lefthook, позволяющие запускать тесты, линтеры и другие проверки до фиксации изменений в репозитории. Для стандартизации сообщений коммитов и поддержания единого стиля истории изменений, а также упрощения анализа истории проекта могут применяться инструменты Conventional Changelog, такие как commitizen, commitlint.

1.2. Контроль версий данных

Для обеспечения воспроизводимости исследований необходимо в точности фиксировать версию данных, на которых исследование было проведено [11; 12]. Выделяют три уровня версионирования данных.

1. Версионирование файлов – отслеживание изменений файловых представлений наборов данных. Работать с уровнем версионирования данных позволяет, например, DVC (Data Version Control), где метаданные версий хранятся с помощью Git, а сами данные располагаются в объектных хранилищах.

2. Версионирование на уровне хранилищ – систематизированный контроль версий объектов в системах хранения, включая управление состоянием. Широко распространённым инструментом здесь является LakeFS, который представляет собой систему контроля версий, располагающуюся поверх S3-совместимых хранилищ, и использует Git-терминологию для управления данными: ветки, коммиты, слияния.

3. Версионирование на уровне таблиц и транзакций – версионирование на уровне логических представлений, включая поддержку ACID-транзакций и временных запросов. Одним из наиболее используемых инструментов для поддержки такого версионирования является инструмент Delta Lake, который оптимизирован под эффективное хранение и управление табличными данными [13].

Выбор решения напрямую зависит от специфики данных, используемых в проекте, наличия внешних интеграций с инструментами обработки и требований к транзакционности. Данные решения позволяют обеспечить воспроизводимость исследований, исходя из конкретных версий входных наборов данных [14].

1.3. Контроль версий отчетов об исследованиях

Исследовательские документы, как правило, содержат не только текст, но и исполняемый код, визуализации экспериментов, результаты работы моделей и метаинформацию. Для создания и распространения отчетов применяются различные инструменты: генераторы отчётов (Jupyter Notebooks, Quarto и Marimo) и системы управления отчетами для структурирования и хранения результатов (ClearML Reports, MLflow Tracking, Weights & Biases Reports). Данные инструменты отличаются степенью интеграции с вычислительной средой, поддержкой воспроизводимости и масштабируемости [15].

Применение инструментов версионирования кода, данных и отчетов позволяет сократить трудозатраты на коммуникации между сотрудниками при постановке новых исследовательских задач на выбранных данных, воспроизведении результатов выполненных работ, а также минимизировать риски утраты артефактов.

2. Управление зависимостями и средами исполнения

Внешними зависимостями являются библиотеки и пакеты, используемые в проекте, а также вся сопутствующая инфраструктура, включая системные библиотеки,

драйверы, компиляторы и аппаратно зависимые компоненты. Ниже рассмотрены основные подходы, направленные на решение задачи управления зависимостями и средами исполнения.

2.1. Управление Python-зависимостями

Для управления зависимостями в Python-проектах используются менеджеры пакетов:

- Pip + requirements.txt – инструмент, в котором указываются фиксированные версии Python-зависимостей. Однако Pip выполняет функции только установки пакетов без менеджмента зависимостей, что не дает отслеживать вложенные зависимости и не может гарантировать однозначного восстановления окружения.
- Poetry, PDM – современные менеджеры пакетов, которые добавляют зависимости в pyproject-файл и однозначно фиксируют версии всех пакетов в lock-файле, что позволяет прозрачно и безопасно управлять изменениями.
- Conda, Mamba – менеджеры пакетов и окружений, которые позволяют управлять не только Python-зависимостями, но и научными и системными пакетами.
- Pixi – современный инструмент управления зависимостями, который обладает функциональностью управления виртуальными окружениями, управления Python-пакетами и научными пакетами, воспроизведения зависимостей при помощи lock файла, сборки и публикации Python и научных пакетов, использования единого файла метаданных pyproject, автоматизации выполнения команд.

2.2. Управление системными зависимостями

При реализации проектов в области науки о данных часто возникает необходимость обеспечения совместимости версий системных компонентов, таких как драйверы, компиляторы и различные библиотеки, которые находятся вне управления менеджерами Python. Особенно это актуально для фреймворков глубокого обучения (например, TensorFlow и PyTorch), где критически важно строго зафиксировать версии компонентов CUDA и cuDNN для обеспечения корректного функционирования приложений. В таких случаях применение системных менеджеров пакетов (например, apt для систем на базе Debian/Ubuntu) является необходимым для корректного управления внешними зависимостями.

2.3. Управление средой исполнения

Использование технологий контейнеризации (Docker, Podman, Dev Containers в VS Code) позволяет создавать изолированные и воспроизводимые окружения, в которых могут фиксироваться как версии Python-пакетов, так и системных компонентов. Такой комплексный подход способствует снижению вероятности конфликтов версий и обеспечивает стабильность работы программного обеспечения [16; 17].

Для хранения моделей, данных и других артефактов используются как специализированные хранилища (GitLab Registry, Nexus), так и объектные хранилища (S3). Такие системы обеспечивают доступ к артефактам и их интеграцию с CI/CD пайплайнами.

Важной практикой является использование Feature Store – централизованных систем хранения признаков, которые используются повторно в различных моделях, что обеспечивает согласованность признаков на всех этапах жизненного цикла модели [18].

Стандартизация файловой структуры хранения данных и артефактов может достигаться через шаблонные репозитории (Cookiecutter, Copier). Данные инструменты задают общую структуру проекта, включая папки для сырого и обработанного набора данных, моделей, конфигураций и логов экспериментов. Единообразие рабочих окружений обеспечивает согласованность операций чтения и записи в различных средах.

Применение инструментов управления зависимостями и средами исполнения в процессе разработки и эксплуатации решений позволяет уменьшить затраты на коммуникации между сотрудниками при постановке задач и воспроизведении результатов выполненных работ (ревью результатов). Таким образом можно обеспечить стабильность и воспроизводимость вычислительных окружений, снизить риски несовместимости при передаче проекта между участниками. Внедрение таких инструментов позволяет централизованно управлять ресурсами и средами, поддерживая согласованность результатов на протяжении всего жизненного цикла проекта.

3. Оркестрация и автоматизация исследовательских процессов

Автоматизация процессов в исследовательских проектах позволяет стандартизировать и формализовать этапы обработки данных, гарантируя повторяемость результатов при повторном запуске [19; 20]. Основной механизм – построение детерминированных пайплайнов обработки данных с использованием workflow-менеджеров [21].

Среди общих принципов автоматизации можно выделить:

- явное описание зависимостей между этапами обработки и анализа – каждый шаг представляется как независимая единица работы с чётко определёнными входами, выходами и средой;
- поддержку повторного использования и кеширования, если входные данные и конфигурации не изменились;
- отделение логики обработки от окружения исполнения – пайплайн должен быть независим от локальной конфигурации машины;
- интеграцию с системами логирования и мониторинга для отслеживания состояния выполнения шагов и фиксации ошибок.

Для реализации автоматизации существует широкий спектр инструментов:

- пайплайны обработки данных могут быть описаны с помощью DSL-языков или YAML/JSON-конфигураций, что упрощает их чтение и поддержку. Примеры: Snakemake, CWL, Nextflow;
- оркестраторы задач и DAG-системы, такие как Apache Airflow или Luigi, предоставляют расширенные возможности для построения сложных графов зависимостей с возможностью планирования задач, мониторинга и масштабирования;
- системы автоматизации общего назначения (например, GNU Make, CMake, Just) позволяют автоматизировать не только компиляцию программного обеспечения, но и шаги предобработки данных или запуска экспериментов.

Современные системы централизованного управления конфигурациями, такие как Hydra, OmegaConf и Pydantic, позволяют динамически компонировать и валидировать настройки экспериментов. Они дают возможность объединять настройки из различных источников, валидировать и типизировать параметры, а также переопределять параметры без изменения базовых конфигурационных файлов.

Для организации тестирования в изолированных окружениях может использоваться инструмент Tox, который помогает обеспечить стабильность и воспроизводимость экспериментов за счет тестирования в разных виртуальных окружениях.

Применение данных решений позволяет выстраивать четкую ролевую систему с разграничением зон ответственности между исполнителями (исследовательские пайплайны, продукционные пайплайны для промышленной эксплуатации), строить воспроизводимые и масштабируемые конвейеры обработки данных, снимая нагрузку с исполнителей и снижая вероятность ошибок, связанных с ручными действиями [22; 23].

4. Стандартизация хранения данных

Одним из ключевых аспектов обеспечения воспроизводимости исследований является стандартизация хранения данных. В условиях многоэтапных вычислительных пайплайнов и командной разработки становится критически важным единообразие в обращении с промежуточными и итоговыми наборами данных.

Установление основного формата хранения данных позволяет упорядочить пайплайны и упростить обмен результатами исследований (например, Parquet, Arrow, CSV). В случае работы со структурированными данными оптимально использование реляционных СУБД для обеспечения единого формата хранения и доступа к данным. При работе с неструктурированными данными возможно использование объектных хранилищ.

Ключевым элементом воспроизводимости является обеспечение возможности установить связь между данными и порождающим их этапом вычислений [24]. Для этого применяются подходы по описанию и версионированию артефактов, включая хранение всех промежуточных результатов и их привязку к этапам исследований; в частности, используются системы каталогизации, такие как DataHub, Amundsen и др., обеспечивающие формирование единой карты зависимости между данными и пайплайнами, которые их порождают [25].

Применение инструментов стандартизации хранения данных позволяет сократить коммуникации, т.к. исследователи работают с данными в одном формате и знают, где и как искать нужные данные, как осуществлять их преобразование, где искать существующие наработки для их переиспользования, а также минимизировать риски, связанные с утерей промежуточных артефактов для получения итоговых моделей и аналитических выводов.

5. Документирование и прозрачность

Документирование процесса и результатов исследований – необходимое условие для обеспечения внутренней прозрачности в команде и внешней воспроизводимости результатов [26; 27]. Современные инструменты позволяют формализовать отчёты, связывая текст с исполняемым кодом и визуализациями. Существуют два распространенных подхода – использование статических отчетных систем (например, Quarto, Sphinx, MkDocs) и динамических платформ (например, ClearML, MLflow). Первый под-

ход ориентирован на генерацию научных отчетов и статической документации с воспроизводимыми блоками кода. Второй обеспечивает онлайн-отслеживание экспериментов, автоматическую фиксацию метрик, артефактов и визуализаций.

Обеспечение прозрачности требует доступности результатов для заинтересованных лиц и достигается путем интеграции отчетов, артефактов и данных в корпоративную инфраструктуру: Confluence, wiki-системы, BI-системы и серверы с отчетами. Наличие централизованного ресурса позволяет отслеживать ход исследований и проводить аудит.

Высокое качество кода критично для интерпретируемости и надёжности. Инструменты проверки качества: статическая проверка типов (mypy, pyright), линтеры (Flake8, pylint, ruff), форматтеры (black, isort, ruff) и системы централизованных конфигураций (nitpick), обеспечивают соответствие стилевым и техническим стандартам в проекте.

Унификация проектной структуры через шаблонные репозитории обеспечивает единый подход к ведению документации, кодовому стилю, структуре отчетности и организации вычислений. Это снижает когнитивную нагрузку при переходе между проектами и способствует формированию устойчивой инженерной культуры в командах.

Решения для документирования повышают прозрачность процессов и упрощают аудит как внутри команды, так и для внешних заинтересованных сторон (например, руководства или контролирующих органов), что уменьшает длительность взаимодействия между сотрудниками, например при рецензировании и воспроизведении результатов.

Перечисленные категории инструментов автоматизации обеспечивают реализацию основных управленческих функций: планирования, контроля, документирования и коммуникации. При этом управление инструментами автоматизации становится частью системы управления исследовательской инфраструктурой, поэтому важно рассматривать их не как отдельные изолированные решения, а как компоненты единой организационной системы управления проектами, что обусловливает необходимость их интеграции на организационном уровне. Формализация процессов посредством использования пайплайнов, конфигурационных файлов и шаблонных репозиториев обеспечивает возможность совместной и параллельной работы сотрудников с различными компетенциями, предоставляя доступ к общим стандартам и унифицированным методикам проверки результатов. Это также позволяет на организационном уровне формально определить роли и зоны ответственности между специалистами, такими как инженеры данных, исследователи данных, ML-инженеры и аналитики данных.

Формальная фиксация результатов исследований предоставляет руководству объективную информацию о прогрессе и качестве выполняемых задач. Такие данные являются основой для принятия решений по приоритизации проектов, перераспределению ресурсов и дополнительному финансированию перспективных направлений. Кроме того, ретроспективные отчёты, фиксирующие всю историю исследований, упрощают стратегическое планирование, позволяя менеджменту анализировать динамику развития и точнее прогнозировать сроки достижения целевых показателей. В результате повышается обоснованность управленческих решений и доверие к итоговым результатам проектов.

Интеграция инструментов контроля версий и оркестрации в корпоративные информационные системы обеспечивает прослеживаемость (traceability) ключевых артефактов исследований, что приносит следующие преимущества [28]:

- возможность демонстрации аудиторским органам или партнёрам исходных данных и параметров, использованных для получения итоговых метрик;
- снижение рисков, связанных с концентрацией критически важной информации у отдельных сотрудников, благодаря формализации и документированию процессов, что позволяет воспроизводить исследования любому члену команды или внешнему подрядчику;
- уменьшение влияния человеческого фактора и вероятности ошибок, связанных с ручными операциями, благодаря автоматизации пайплайнов, использованию CI/CD-практик для ML, автоматическим тестированиям и валидациям;
- повышение прозрачности процессов, способствующее взвешенному управлению рисками и сроками реализации проектов, а также оперативному вмешательству при задержках выполнения отдельных этапов.

Однако ключевым ограничением внедрения инструментов воспроизводимости является необходимость организационных изменений и поддержания новой культуры прозрачности [29; 30]. Внедрение дополнительных процедур документирования и контроля качества может восприниматься

сотрудниками как усложнение рабочих процессов. Эффективность автоматизации зависит от уровня зрелости инфраструктуры и компетенций команды, и при недостаточной поддержке руководства даже современные технические решения могут оказаться неэффективными. Дополнительным барьером является необходимость обучения сотрудников работе с новыми инструментами. При большом количестве сотрудников или частых ротациях персонала передача знаний осложняется, что может приводить к временному снижению производительности до тех пор, пока культура воспроизводимости не станет стандартной практикой.

Перспективным направлением развития организационных практик является интеграция показателей воспроизводимости в систему ключевых показателей эффективности (KPI) исследовательских коллективов и организаций. Например, целесообразно учитывать время развертывания окружения для новых сотрудников, долю полноценно документированных исследований и частоту успешного воспроизведения пайплайнов. Разработка формальных стандартов по аналогии с ISO/IEC для проектов в области анализа данных способствовала бы упрощению процедур аудита и сертификации решений, а также повышению доверия к результатам со стороны заказчиков. Такие стандарты должны включать требования к хранению метаданных, процедурам валидации и документирования, методикам управления качеством данных, а также определению организационных ролей.

Заключение

Проведенный анализ современных практик и инструментов, применяемых для обеспечения воспроизводимости в проектах, связанных с наукой о данных, демонстрирует, что достижение прозрачности, управляемости и эффективности исследовательской деятельности требует интеграции широкого спектра специализированных решений. В связи с этим критически важно не только знание актуальных технологий, но и способность к их обоснованному выбору и координированному применению в рамках единой исследовательской инфраструктуры.

Современные организационно-технические решения позволяют стандартизировать и автоматизировать ключевые аспекты исследовательской работы: обработку данных, управление зависимостями и версиями артефактов, оркестрацию вычислений и документирование процессов. Их применение снижает влияние человеческого фактора, усиливает командную координацию и повышает устойчивость проектов к изменениям в составе участников и конфигурации инфраструктуры и позволяет снизить трудозатраты на выполнение типовых работ и коммуникации.

Эффективность внедрения таких решений во многом определяется готовностью организаций к институциональным изменениям. Необходимы формализация процессов, развитие культуры прозрачности и повышение зрелости инфраструктуры. Основные барьеры носят организационный характер: нехватка компетенций, сопротивление изменениям и отсутствие системной поддержки со стороны управленческих структур.

Возможными направлениями развития являются интеграция существующих инструментов воспроизводимости в единую систему, включающую ключевые показатели эффективности исследовательских коллективов, разработка формальных отраслевых стандартов и методик сертификации проектов науки о данных, а также новых информационных технологий для решения задач управления.

Список литературы

1. Goodman S.N., Fanelli D., Ioannidis J.P.A. What does research reproducibility mean? // Science Translational Medicine. 2016. Vol. 8. P. 341. DOI: 10.1126/scitranslmed.aaf5027.

2. Baker M. 1,500 scientists lift the lid on reproducibility // Nature. 2016. Vol. 533. P. 452-454. URL: https://www.nature.com/articles/533452a (дата обращения: 26.02.2025). DOI: 10.1038/533452a.

3. Munafò M.R., Nosek B.A., Bishop D.V.M., Button K.S., Chambers C.D., Percie du Sert N., Simonsohn U., Wagenmakers E.-J., Ware J.J. Ioannidis J.P.A. A Manifesto for Reproducible Science // Nature Human Behaviour. 2017. Vol. 1, Is. 1. URL: https://www.nature.com/articles/s41562-016-0021 (дата обращения: 26.02.2025). DOI: 10.1038/s41562-016-0021.

4. Stodden V. The data science life cycle: a disciplined approach to advancing data science as a science // Communications of the ACM. 2020. Vol. 63, Is. 7. P. 58-66. DOI: 10.1145/3360646.

5. Gundersen O., Kjensmo S. State of the Art: Reproducibility in Artificial Intelligence // Proceedings – AAAI 2018 at New Orleans. Thirty-Second AAAI Conference on Artificial Intelligence 2018. 2018. URL: https://aaai.org/papers/11503-state-of-the-art-reproducibility-in-artificial-intelligence/ (дата обращения: 26.02.2025). DOI: 10.1609/aaai.v32i1.11503.

6. Gundersen O.E., Coakley K., Kirkpatrick C., Gil Y. Sources of Irreproducibility in Machine Learning: A Review // arXiv:2204.07610v2[cs.LG]. 2023. DOI: 10.48550/arXiv.2204.07610.

7. Waller L.A., Miller G.W. More than Manuscripts: Reproducibility, Rigor, and Research Productivity in the Big Data Era // Toxicological Sciences. 2016. Vol. 149, Is. 2. P. 275–276. URL: https://academic.oup.com/toxsci/article-abstract/149/2/275/2461691 (дата обращения: 26.02.2025). DOI: 10.1093/toxsci/kfv330.

8. Liu J., Carlson J., Pasek J., Puchala B., Rao A., Jagadish H.V. Promoting and Enabling Reproducible Data Science Through a Reproducibility Challenge // Harvard Data Science Review. 2022. Vol. 4, Is. 3. URL: https://hdsr.mitpress.mit.edu/pub/mlconlea (дата обращения: 26.02.2025). DOI: 10.1162/99608f92.9624ea51.

9. Hernandez J.A., Colom M. Repeatability, Reproducibility, Replicability, Reusability (4R) in Journals' Policies and Software/Data Management in Scientific Publications: A Survey, Discussion, and Perspectives // arXiv preprint arXiv:2312.11028v1. 2023. DOI: 10.48550/arXiv.2312.11028.

10. Chen K.Y., Toro-Moreno M., Subramaniam A.R. GitHub is an effective platform for collaborative and reproducible laboratory research // arXiv preprint arXiv:2408.09344v2. 2025. URL: https://arxiv.org/abs/2408.09344v2 (дата обращения: 26.02.2025).

11. Idowu S., Osman O., Strüber D., Berger T. Machine learning experiment management tools: a mixed-methods empirical study // Empirical Software Engineering. 2024. Vol. 29, Is.4. DOI: 10.1007/s10664-024-10444-w.

12. Klump J., Wyborn L., Wu M., Martin J., Downs R.R., Asmi A. Versioning data is about more than revisions: A conceptual framework and proposed principles // Data Science Journal. 2021. Vol. 20, Is. 1. P. 12. DOI: 10.5334/dsj-2021-012.

13. Armbrust M., Das T., Sun L., Yavuz B., Zhu S., Murthy M., Torres J., van Hovell H., Ionescu A., Łuszczak A., Świtakowski M., Szafrański M., Li X., Ueshin T., Mokhtar M., Boncz P., Ghodsi A., Paranjpye S., Senster P., Xin R., Zaharia M. Delta Lake: high-performance ACID table storage over cloud object stores // Proc VLDB Endow. 2020. Vol. 13, Is. 12. P. 3411-3424. DOI: 10.14778/3415478.3415560.

14. Semmelrock H., Ross-Hellauer T., Kopeinik S., Theiler D., Haberl A., Thalmann S., Kowald D. Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers // arXiv preprint arXiv:2406.14325. 2024. DOI: 10.48550/arXiv.2406.14325.

15. Arpteg A., Brinne B., Crnkovic-Friis L., Bosch J. Software Engineering Challenges of Deep Learning // 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Prague, Czech Republic. 2018. P. 50-59. URL: https://ieeexplore.ieee.org/document/8498185 (дата обращения: 26.02.2025). DOI: 10.1109/SEAA.2018.00018.

16. Nüst D., Sochat V., Marwick B., Eglen S.J., Head T., Hirst T., Evans B.D. Ten simple rules for writing Dockerfiles for reproducible data science // PLOS Computational Biology. 2020. Vol. 16, Is. 11. P. 1-24. DOI: 10.1371/journal.pcbi.1008316.

17. Chirigati F., Rampin R., Shasha D., Freire J. ReproZip: Computational Reproducibility with Ease // Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data, ACM. 2016. P. 2085-2088. DOI: 10.1145/2882903.2899401.

18. Martínez J. de la R., Buso F., Kouzoupis A., Ormenisan A.A., Niazi S., Bzhalava D., Mak K., Jouffrey V., Ronström M., Cunningham R., Zangis R., Mukhedkar D., Khazanchi A., Vlassov V., Dowling J. The Hopsworks Feature Store for Machine Learning // SIGMOD/PODS '24 – Companion of the 2024 International Conference on Management of Data, Santiago AA, Chile. 2024. P. 135-147. DOI: 10.1145/3626246.3653389.

19. Sculley D., Holt G., Golovin D., Davydov E., Phillips T., Ebner D., Chaudhary V., Young M., Crespo J.-F., Dennison D. Hidden Technical Debt in Machine Learning Systems // Proceedings of the 29th International Conference on Neural Information Processing Systems 2015. Vol. 2. P. 2503-2511. DOI: 10.5555/2969442.2969519.

20. Sculley D., Holt G., Golovin D., Davydov E., Phillips T., Ebner D., Chaudhary V., Young M. Machine Learning: The High Interest Credit Card of Technical Debt // SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop). 2014. URL: https://research.google/pubs/machine-learning-the-high-interest-credit-card-of-technical-debt/ (дата обращения: 26.02.2025).

21. Di Tommaso P., Chatzou M., Floden E.W., Barja P.P., Palumbo E., Notredame C. Nextflow enables reproducible computational workflows // Nature Biotechnology. 2017. Vol. 35, Is. 4. P. 316-319. URL: https://www.nature.com/articles/nbt.3820 (дата обращения: 26.02.2025). DOI: 10.1038/nbt.3820.

22. Zaharia M.A., Chen A., Davidson A., Ghodsi A., Hong S.A., Konwinski A., Murching S., Nykodym T., Ogilvie P., Parkhe M., Xie F., Zumar C. Accelerating the Machine Learning Lifecycle with MLflow // IEEE Data Eng Bull. 2018. Vol. 41, P. 39-45. URL: https://api.semanticscholar.org/CorpusID:83459546 (дата обращения: 26.02.2025).

23. Yasmin J., Wang J., Tian Y., Adams B. An Empirical Study of Developers' Challenges in Implementing Workflows as Code: A Case Study on Apache Airflow // Journal of Systems and Software. 2024. Vol. 219. DOI: 10.48550/arXiv.2406.00180.

24. Freire J., Koop D., Santos E., Silva C. Provenance for Computational Tasks: A Survey // Computing in Science and Engineering. 2008. V. 10, Is. 3, P. 11-21. URL: https://ieeexplore.ieee.org/document/4488060 (дата обращения: 26.02.2025). DOI: 10.1109/MCSE.2008.79.

25. Subramaniam P., Ma Y., Li C., Mohanty I., Fernandez R.C. Comprehensive and comprehensible data catalogs: The what, who, where, when, why, and how of metadata management // arXiv preprint arXiv:2103.07532. 2021. DOI: 10.48550/arXiv.2103.07532.

26. Wilson G., Bryan J., Cranston K., Kitzes J., Nederbragt L., Teal T.K. Good enough practices in scientific computing // PLOS Computational Biology. 2017. Vol.13, Is.6. P.1-20. DOI: 10.1371/journal.pcbi.1005510.

27. Pimentel J.F., Murta L., Braganholo L., Freire J. A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks // Proceedings – 2019 IEEE/ACM 16th International Conference on Mining Software Repositories, Montreal, QC, Canada. 2019. P. 507-517. URL: https://ieeexplore.ieee.org/document/8816763 (дата обращения: 26.02.2025). DOI: 10.1109/MSR.2019.00077.

28. Amershi S., Begel A., Bird C., DeLine R., Gall H., Kamar E., Nagappan N., Nushi B., Zimmermann T. Software Engineering for Machine Learning: A Case Study // IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), Montreal, QC, Canada. 2019. P. 291-300. URL: https://ieeexplore.ieee.org/document/8804457 (дата обращения: 26.02.2025). DOI: 10.1109/ICSE-SEIP.2019.00042.

29. Salama Kh., Kazmierczak, J., Schut D. Practitioners Guide to MLOps: A Framework for Continuous Delivery and Automation of Machine Learning // Google Cloud White Paper. 2021. 37 p. URL: https://services.google.com/fh/files/misc/practitioners_guide_to_mlops_whitepaper.pdf (дата обращения: 26.02.2025).

30. Treveil M., Omont N., Stenac C., Lefevre K., Phan D., Zentici J., Lavoillotte A., Miyazaki M., Heidmann L. Introducing MLOps: How to Scale Machine Learning in the Enterprise. O'Reilly Media, 2020. 183 p.