

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ АВТОМАТИЗИРОВАННОГО СБОРА И СИСТЕМАТИЗАЦИИ НАУЧНЫХ ПУБЛИКАЦИЙ

А. О. Кондратьев^{1,}, А. С. Бондяков¹*

¹ Объединенный институт ядерных исследований, Дубна, 141980, Россия

Представлен программный комплекс, разработанный для автоматизации процессов сбора, обработки и систематизации научных публикаций в рамках создания институционального репозитория Объединенного института ядерных исследований (ОИЯИ). Описаны архитектура комплекса, ключевые компоненты и алгоритмы, обеспечивающие интеграцию данных из внешних источников (eLibrary, INSPIRE-HEP) в платформу цифрового репозитория DSpace. Особое внимание уделено разработке специализированных сервисов для решения критически важных задач: идентификации авторов, исключения дублирования публикаций и обеспечения синхронизации данных с внутренней информационной системой Института. Приведены результаты внедрения: на текущий момент в репозитории ОИЯИ систематизировано и доступно более 16 тыс. публикаций. Подчеркивается значимость разработанного решения для повышения эффективности управления научной продуктивностью и оценки интеллектуального потенциала Института.

The paper presents a software complex developed to automate the processes of collecting, processing, and systematizing scientific publications as part of creating an institutional repository for the Joint Institute for Nuclear Research (JINR). The architecture of the complex, its key components, and algorithms that ensure data integration from external sources (eLibrary, INSPIRE-HEP) into the DSpace digital repository platform are described. Particular attention is paid to the development of specialized services designed to address critical tasks: author identification, elimination of publication duplication, and ensuring data synchronization with the institute's internal information system. Implementation results are presented: currently, the JINR repository contains and provides access to more than 16 000 systematized publications. The significance of the developed solution for enhancing the efficiency of managing scientific productivity and assessing the institute's intellectual potential is emphasized.

PACS: 89.20.Ff

ВВЕДЕНИЕ

Цифровые репозитории публикаций играют ключевую роль в обеспечении доступности научных результатов и эффективного управления научной информацией. Современные институциональные репозитории позволяют интегрировать данные из доверенных источников, обеспечивать

* E-mail: kondratyev@jinr.ru

долговременное хранение и удобный доступ к информационным активам. В условиях отсутствия в ОИЯИ полноценного институционального цифрового репозитория разработка системы автоматизированного сбора и систематизации публикаций становится приоритетной задачей [1–4].

Управление авторскими данными требует точной идентификации публикаций и их авторов, особенно в контексте роста объема научных работ и числа соавторов. Автоматизированные решения позволяют минимизировать ручной ввод, исключить дублирование и повысить достоверность данных. Важными задачами являются привязка публикаций к авторам, унификация метаданных и поддержание согласованности информации между различными системами.

В рамках решения этих задач был разработан программный комплекс, обеспечивающий автоматизированный сбор, обработку и систематизацию научных публикаций с последующей интеграцией в цифровой репозиторий [5–8].

СТРУКТУРА ПРОГРАММНОГО КОМПЛЕКСА

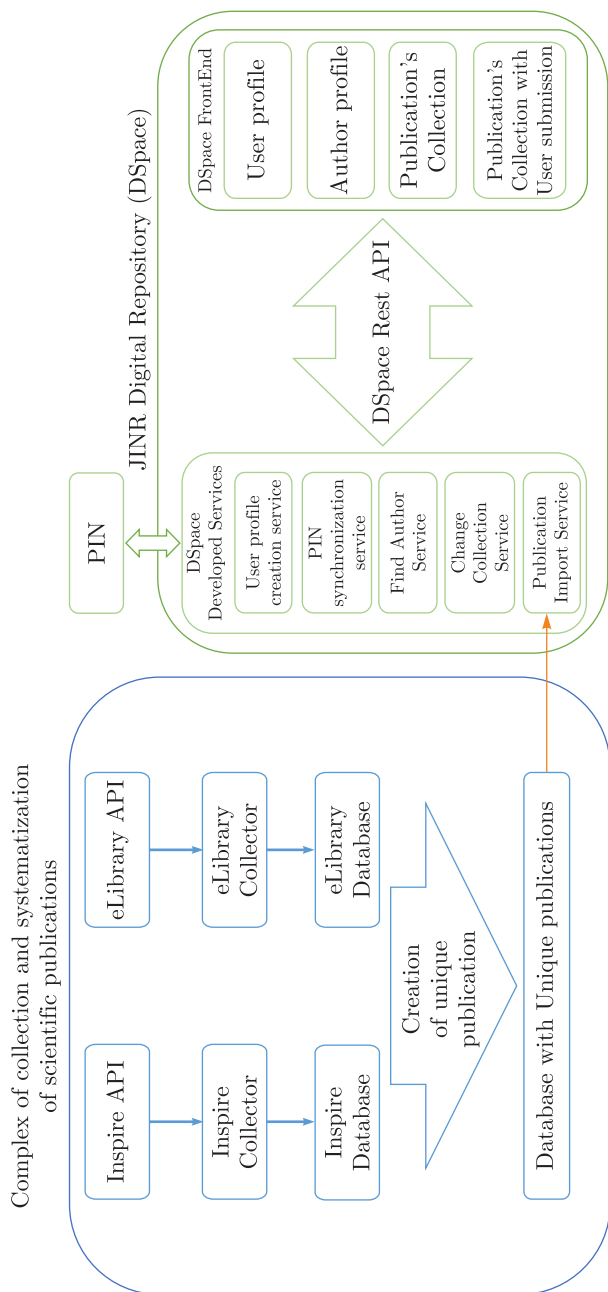
В качестве платформы цифрового репозитория выбрана система DSpace — открытая платформа для управления научным контентом, поддерживающая гибкую работу с метаданными, настройку прав доступа и масштабируемость. Преимущества DSpace включают активное сообщество разработчиков, поддержку различных форматов и расширяемую архитектуру. Однако система требует квалифицированного администрирования и дополнительной функциональности для автоматизации процессов сбора и сопоставления данных [9, 10].

Для расширения возможностей репозитория разработан программный комплекс, включающий следующие компоненты:

- «сборщики» — алгоритмы, осуществляющие регулярный сбор метаданных публикаций из внешних источников (eLibrary, INSPIRE-HEP) через API и сохранение их в промежуточные базы данных;
- модули систематизации преобразуют и объединяют данные из различных источников, исключая дублирование и насыщая метаданные;
- центральная база данных хранит обобщенные и нормализованные записи о публикациях;
- алгоритмы актуализации обеспечивают коррекцию и обновление информации на основе новых данных.

Комплекс дополняется рядом специализированных сервисов, обеспечивающих взаимодействие с репозиторием и информационной системой института (ИСИ).

Сервис импорта публикаций в репозиторий извлекает данные из центральной базы комплекса, проверяет их на дублирование с уже загруженными в DSpace публикациями и при необходимости обновляет метаданные. Для новых записей осуществляется автоматическая привязка



Общая схема работы комплекса и разработанных сервисов

авторов, аффилированных с ОИЯИ, с использованием идентификаторов (ORCID, eLibrary ID) и нормализованных вариантов написания ФИО. После обработки публикации импортируются в репозиторий, а их метаданные синхронизируются с базой данных ИСИ.

Сервис смены коллекции публикаций предназначен для обработки публикаций, внесенных вручную. Он проверяет такие записи на дублирование с автоматически загруженными. При обнаружении совпадения публикация направляется в архив для ручной проверки модератором. В противном случае она перемещается в основную коллекцию. Метаданные также передаются в ИСИ, что обеспечивает целостность данных.

Сервис синхронизации с ИСИ извлекает актуальные данные о сотрудниках (ФИО, должность, e-mail, подразделение) и сопоставляет их с профилями в DSpace. При выявлении расхождений или появлении новых сотрудников система генерирует нормализованные варианты написания имен и формирует список для обновления в репозитории. Это обеспечивает согласованность персональных данных и поддерживает актуальность профилей.

Сервис создания профиля сотрудника формирует профили пользователей DSpace на основе данных ИСИ. Профиль сотрудника включает SSO-логин, e-mail и пароль — данные, необходимые для аутентификации. Профиль автора, в свою очередь, содержит научную информацию: ФИО, должность, список публикаций. Связь между ними устанавливается через e-mail. При создании профиля проверяются уникальность логина и e-mail, что предотвращает дублирование.

Сервис поиска авторов автоматизирует привязку авторов к уже импортированным публикациям. В текущей реализации идентификаторы авторов (ORCID, eLibrary ID) добавляются вручную, что обуславливает необходимость автоматизации. Сервис ищет совпадения по ФИО, аффилиации и идентификаторам, обновляя профили в DSpace и синхронизируя изменения с ИСИ.

Общая схема взаимодействия компонентов представлена на рисунке.

Сборщики получают данные из внешних источников и сохраняют их в промежуточные базы. Затем публикации сравниваются между собой для выявления дубликатов и объединяются в централизованную базу. При импорте в DSpace выполняется дополнительная проверка на дублирование и насыщение метаданных. Синхронизация с ИСИ обеспечивает согласованность данных в реальном времени.

ЗАКЛЮЧЕНИЕ

Разработанный программный комплекс решает задачу автоматизированного сбора, систематизации и интеграции научных публикаций в институциональный репозиторий ОИЯИ. Реализованные алгоритмы и сервисы обеспечивают высокую степень автоматизации процессов: исключение дублирования, точную привязку авторов, синхронизацию с внутрен-

ними информационными системами и поддержание целостности данных. Использование идентификаторов авторов и нормализованных форм ФИО повышает точность сопоставления. Применение DSpace в качестве базовой платформы позволяет эффективно управлять научным контентом при минимальных затратах на лицензирование. На момент завершения разработки в репозитории содержится более 16 000 публикаций, обработанных и загруженных с помощью комплекса. Результаты подтверждают эффективность предложенного подхода для создания и поддержания институционального цифрового репозитория в условиях крупного научного центра. Перспективы дальнейшего развития включают интеграцию с дополнительными источниками данных, расширение функций анализа научной продуктивности и внедрение механизмов машинного обучения для повышения точности идентификации авторов.

СПИСОК ЛИТЕРАТУРЫ

1. *Filozova I. A. et al.* Installation and Performance Evaluation of the DSpace System // Modern Information Technol. IT-Education. 2023. V. 19, No. 3; <https://doi.org/10.25559/SITITO.019.202303.581-587>.
2. *Bondyakov A. S., Kondratyev A. O.* Architecture of a Digital Repository for Publications Based on the DSpace Platform // Modern Information Technol. IT-Education. 2024. V. 20, No. 3; <https://doi.org/10.25559/SITITO.020.202403.602-608>.
3. *Kondratyev A., Bondyakov A.* Automated Collection and Systematization of Publications // Open Systems. DBMS. 2024. No. 1. P. 49–51; <https://www.osp.ru/os/2024/01/13058272>.
4. *Kondratyev A., Bondyakov A.* Author Data Management Services // Open Systems. DBMS. 2025. No. 1.
5. *Zaikina T., Filozova I., Shestakova G. et al.* JDS-JOIN2 Repository as a Workspace for Scientific Output // Phys. Part. Nucl. Lett. 2022. V. 19. P. 583–585.
6. *Dobiasch M., Oppl S., Stöckl M. et al.* Pegasos: A Framework for the Creation of Direct Mobile Coaching Feedback Systems // J. Multimodal User Interfaces. 2024. V. 18. P. 1–19.
7. *Okon R., Eleberi E., Uka K.* A Web Based Digital Repository for Scholarly Publication // J. Softw. Eng. Appl. 2020. V. 13. P. 67–75.
8. *Neubauer F., Bredl P., Xu M. et al.* MetaConfigurator: A User-Friendly Tool for Editing Structured Data Files // Datenbank Spektrum. 2024. V. 24. P. 161–169.
9. *Itiola C., Iwasokun G., Adetooto J.* Development of an Online Repository for Academic Research Works in FUTA // Intern. J. Sustain. Manag. Inform. Technol. 2021. V. 7, No. 1. P. 22–26.
10. *Nyaichyai L., Luitel G., Maharjan R.* Experiences of Library Professionals on DSpace Installation // Nepal J. Multidiscip. Res. 2021. V. 4, No. 2. P. 93–105.