ОТКРЫТЫЙ ДОСТУП. ОТКРЫТЫЕ АРХИВЫ ИНФОРМАЦИИ

УДК 022.42+026:62(470.311) ГПНТБ
doi: 10.33186/1027-3689-2021-12-15-28

М. В. Гончаров, К. А. Колосов
ГПНТБ России, Москва, Российская Федерация

Анализ метаданных российских репозиториев открытого доступа по научно-технической тематике с целью их использования в системе Единого Открытого архива информации ГПНТБ России

Аннотация. В ГПНТБ России разрабатывается Единый Открытый архив информации (ЕОАИ), объединяющий все электронные полнотекстовые ресурсы, создаваемые или собираемые библиотекой. При обработке поисковых запросов пользователей портал ЕОАИ предусматривает использование не только собственных источников контента, но и метаданных, собранных из других репозиториев открытого доступа, в первую очередь российских.

Цель анализа метаданных российских репозиториев открытого доступа – составление списка доступных источников, отвечающих заданным критериям: техническая доступность провайдера OAI-PMH, актуальность (регулярность) пополнения данных. В статье приведены результаты анализа, описан характер выявленных проблем. Составлен список российских репозиториев открытого доступа, метаданные из которых планируется использовать в ЕОАИ ГПНТБ России в качестве дополнительного источника информации.

Статья подготовлена в рамках государственного задания № 075-01300-20-00 «Разработка и совершенствование системы Открытого архива интегрированных информационно-библиотечных ресурсов ГПНТБ России как современной системы управления знанием в цифровой среде: на пути к Открытой науке» на 2020–2022 гг.

Ключевые слова: открытый доступ, открытые архивы информации, институциональные репозитории, метаданные, OAI-PMH

OPEN ACCESS. OPEN INFORMATION ARCHIVES

UDC 022.42+026:62(470.311) ГПНТБ
doi: 10.33186/1027-3689-2021-12-15-28

Mikhail V. Goncharov and Kirill A. Kolosov
Russian National Public Library for Science and Technology,
Moscow, Russian Federation

Analyzing metadata of Russian open access repositories in science and technology for using in RNPLS&T's system of Single Open Information Archive

Abstract. RNPLS&T has been designing and developing the Single Open Information Archive (SOIA) to integrate all digital full-text resources generated or otherwise acquired by RNPLS&T. SOIA is intended to use not only its own content sources but also metadata collected from other open access repositories, in particular Russian repositories. The goal of this analysis is to draw the list of available sources that meet the following criteria: OAI-PMH provider technical accessibility, continuity (regularity) of updating. The authors present the findings of their analysis and describe the challenges revealed. The list of Russian open access repositories is drawn for metadata intended to be used in RNPL&T's SOIA as extra source of information.

The article is prepared within the framework of the State Order № 075-01300-20-00 "Design and development of the system of Open Archive of Integrated Information Library Resources of Russian National Public Library for Science and Technology as a modern system of knowledge management in digital environment: On the way to Open Science" for the years 2020–2022.

Keywords: open access, open archive, institutional repository, metadata, OAI-PMH

Цель создаваемого в ГПНТБ России ЕОАИ – развитие пользовательских сервисов и создание базы для научных исследований и обеспечения научно-образовательной деятельности. Как отмечено в [**1**],

основными источниками формирования контента системы являются электронные документы, появившиеся в результате оцифровки изданий из фонда ГПНТБ России; объекты интеллектуальной собственности, созданные в результате её научно-исследовательской и образовательной деятельности, а также другие материалы, права на использование которых имеет библиотека. Кроме собственных источников портал ЕОАИ в качестве дополнительного источника информации предусматривает набор пользовательских сервисов, в число которых входят метаданные, собранные из других репозиториев открытого доступа (ОД), в первую очередь российских.

Открытые ресурсы как объект комплектования повышают значение библиотек в системе научных коммуникаций. По мнению авторов статьи [**2**], самые востребованные виды изданий ОД – журнальные статьи, монографии и материалы конференций. Наиболее авторитетными и востребованными точками доступа к открытым ресурсам являются сайты научных организаций. Однако, как было отмечено в [**3**], в 2006 г. 21% репозиториев был недоступен для поисковых сервисов *Google*, *Yahoo* и *MSN*. Исследование 2017 г. показало, что собственные репозитории имеют лишь немногие вузы (рассматривались федеральные, научно-исследовательские, опорные университеты) и ещё меньше – обладают действительно открытыми репозиториями, доступ к которым не ограничивается сотрудниками и учащимися вуза.

В соответствии с концепцией открытых архивов институциональные репозитории ОД представляют собой публично доступные архивы научных, исследовательских и образовательных организаций, в которых члены сообщества размещают свои опубликованные и подготовленные к печати статьи и другие материалы научно-исследовательской и научно-организационной деятельности. Репозитории метаданных доступны для авторизованных приложений-сборщиков метаданных, функционирующих в соответствии с протоколом *OAI-PMH* [**4**].

Создаваемый в ГПНТБ России открытый архив выполняет функции как сервис-провайдера, так и клиента *OAI-PMH*. Сервис-провайдер предоставляет внешним приложениям-сборщикам метаданные, описывающие объекты, которые хранят на сервере хранения ресурсов ЕОАИ. Клиент *OAI-PMH* позволяет импортировать метаданные из других открытых архивов (репозиториев).

Согласно сведениям, полученным от *Registry of Open Access Repositories (ROAR)* [**5**], по состоянию на сентябрь 2021 г. в России насчитывается 67 зарегистрированных институциональных репозиториев. Они созданы и поддерживаются следующими организациями: университеты, научные институты РАН, агрегаторы («КиберЛенинка», «Национальный агрегатор открытых репозиториев»), а также отдельными библиотеками, электронными изданиями и др. В [**6**] отмечено: в 2006 г., по данным *ROAR,* в России насчитывалось 58 репозиториев (среди них значительное число институтов РАН), которые зарегистрировали и внесли данные о своих коллекциях в 2004–2006 гг. К сожалению, эти данные не обновляются годами или даже десятилетиями.

Цель анализа метаданных российских репозиториев ОД – составление списка доступных источников для сбора метаданных, который отвечает следующим критериям: техническая доступность провайдера *OAI-PMH*; актуальность (регулярность) пополнения данных.

Для проверки доступности провайдера *OAI-PMH* использовались *URL*, выбранные из строки «*OAI-PMH Interface»* реестра *ROAR* для анализируемого репозитория. Затем через браузер проверялся результат выполнения запроса *ListRecords* протокола *OAI-PMH* для схемы метаданных *oai_dc*.

Проверка актуальности (регулярности) пополнения данных анализируемого репозитория проводилась в два этапа. Сначала выбирались наиболее поздние даты размещения ресурсов (<*дата размещения*>), представленные на странице веб-интерфейса соответствующего репозитория. Затем проводилась проверка с использованием запроса протокола *OAI-PMH* вида «*verb=ListRecords&metadataPrefix=oai_dc&from=<дата размещения>*».

Результаты анализа приведены в табл. 1. В неё не были включены репозитории, по *URL* которых не удалось получить ответ. Также не анализировался репозиторий «Национальный агрегатор открытых репозиториев (НОРА)» [**7**], поскольку исследовалась доступность отдельных институциональных репозиториев. Порядок расположения репозиториев в табл. 1 соответствует последовательности в реестре *ROAR*.

Таблица 1

Результаты анализа доступности российских институциональных репозиториев, зарегистрированных в *ROAR*
(по состоянию на 01.09.2021 г.)

№ п/п	Учредитель репозитория	Регулярность пополнения	Выявленные проблемы
1	Сибирский федеральный университет	Активно пополняется	
2	Объединённый институт ядерных исследований	Нет записей после 2018 г.	Зависает интерфейс *OAI-PMH*, нет ответа
3	Уральский федеральный университет	Активно пополняется	
4	Научная электронная библиотека «КиберЛенинка»	Активно пополняется	Ограниченный набор полей метаданных
5	Институт вулканологии и сейсмологии РАН	Активно пополняется	
6	Тверской государственный университет	Активно пополняется	
7	Южно-Уральский государственный университет	Активно пополняется	Ошибка *HTTP Status 500* при любых запросах
8	Белгородский государственный университет	Активно пополняется	
9	Ярославский государственный университет		Ошибка. Ресурс недоступен
10	Высшая школа менеджмента (СПбГУ)		Ошибка. Ресурс недоступен
11	Институт прикладной математики им. М. В. Келдыша РАН	Нет записей после 2019 г.	
12	Институт философии РАН	Нет записей после 2008 г.	
13	Институт народнохозяйственного прогнозирования РАН	Нет записей после 2017 г.	

Продолжение таблицы 1

№ п/п	Учредитель репозитория	Регулярность пополнения	Выявленные проблемы
14	Институт экономики РАН	Нет записей после 2012 г.	
15	Центр египтологических исследований РАН	Нет записей после 2007 г.	
16	Институт Европы РАН	Нет записей после 2007 г.	
17	Институт аграрных проблем РАН	Нет записей после 2007 г.	
18	Институт США и Канады РАН	Нет записей после 2014 г.	
19	Институт проблем рынка РАН	Последнее пополнение было в 2020 г.	Не открываются полные тексты
20	Отделение международных экономических и политических исследований (филиал Института экономики РАН)	Нет записей после 2012 г.	
21	Институт экономических проблем им. Г. П. Лузина Кольского научного центра РАН	Нет записей после 2008 г.	
22	Институт экономики и организации промышленного производства Сибирского отделения РАН	Нет записей после 2008 г., кроме раздела «Научные отчёты ИЭОПП СО РАН»	
23	Центральный экономико-математический институт РАН	Нет записей после 2019 г. (в 2019 г. было две публикации)	
24	Сочинский научно-исследовательский центр РАН	Нет записей после 2014 г.	
25	Институт проблем региональной экономики РАН	Нет записей после 2012 г.	
26	Вологодский научно-координационный центр ЦЭМИ РАН	Нет записей после 2007 г.	

Продолжение таблицы 1

№ п/п	Учредитель репозитория	Регулярность пополнения	Выявленные проблемы
27	Институт социально-экономических исследований Уфимского научного центра РАН	Нет записей после 2008 г.	
28	Башкирский государственный медицинский университет	Нет записей после 2010 г.	
29	Тюменский государственный университет	Нет записей после 2020 г.	
30	Уральский государственный медицинский университет	Регулярно пополняется	
31	Северо-Кавказский федеральный университет	Активно пополняется	Недоступен поиск по *OAI-PMH*, массив не индексирован
32	Институт биологии южных морей им. А. О. Ковалевского РАН	Активно пополняется	
33	Томский государственный университет	Активно пополняется	
34	Казанский федеральный университет	Активно пополняется	
35	Самарский университет	Активно пополняется	
36	Томский политехнический университет	Активно пополняется	Не найден базовый *URL* для поиска по *OAI-PMH*
37	Московский НИИ психиатрии	Активно пополняется	Не найден базовый *URL* для поиска по *OAI-PMH*
38	ОЭБ «Оренбуржья»	Активно пополняется	Отсутствуют индексы для поиска по *OAI-PMH* для документов с датой после 2018 г.

Окончание таблицы 1

№ п/п	Учредитель репозитория	Регулярность пополнения	Выявленные проблемы
39	Российский государственный профессионально-педагогический университет	Активно пополняется	
40	Журнал «Вестник Кузбасского государственного технического университета»	Регулярно пополняется	Не найден базовый *URL* для поиска по *OAI-PMH*
41	Уральский государственный педагогический университет	Активно пополняется	
42	Алтайский государственный университет	Активно пополняется	Недоступен поиск по *OAI-PMH*, массив не индексирован
43	Свердловская областная универсальная научная библиотека им. В. Г. Белинского	Активно пополняется	
44	Санкт-Петербургский государственный университет	Активно пополняется	
45	Институт философии СПбГУ	Нет записей после 2020 г.	
46	Уральский государственный лесотехнический университет	Активно пополняется	
47	Удмуртский государственный университет	Активно пополняется	

По результатам анализа институциональных репозиториев на предмет их доступности и регулярности пополнения можно сделать следующие выводы:

из 68 институциональных репозиториев, зарегистрированных в *ROAR*, по состоянию на сентябрь 2021 г. фактически доступны 47 репозиториев ОД;

доступ по протоколу *OAI-PMH* по техническим причинам не поддерживается у 9 репозиториев;

регулярно пополняются 24 репозитория, наиболее интенсивно добавляется контент в 22 репозиториях;

наибольшие проблемы с пополнением контента выявлены в репозиториях научных организаций РАН, использующих домен *socionet.ru*.

Дополнительно были выборочно проанализированы метаданные институциональных репозиториев на предмет их полноты и возможности проведения тематического поиска, в результате которого получены следующие результаты:

хотя большинство репозиториев поддерживают схемы метаданных *Dublin Core (DC)* и *MARC*, записи в формате *MARC* содержат те же данные, что и в формате *DC*, меняется лишь формат их представления. Это связано с тем, что базовым форматом метаданных для открытых архивов является именно *DC* [**8**];

наиболее полно (по числу использованных элементов метаданных и по объёму данных) представлены метаданные в репозиториях университетов. Они включают значительное количество элементов <*subject*> и <*description*>, которые могут быть использованы при тематическом поиске;

в записях репозитория Тверского госуниверситета (http://eprints.tversu.ru) в элементе <*subject*> каждого метаописания содержится индекс УДК, что значительно повышает точность тематического поиска. Кроме того, в названии коллекций также присутствуют индексы УДК, что может быть использовано при выборе источника импорта метаданных по заданной тематике;

в записях репозитория Института вулканологии и сейсмологии ДВО РАН (http://repo.kscnet.ru) в элементе <*subject*> каждого метаописания, а также в названии коллекций содержится индекс ГРНТИ, что также улучшает возможность проведения точного тематического поиска;

в записях репозитория Научной электронной библиотеки «КиберЛенинка» (https://cyberleninka.ru/oai) элементы <*subject*> и <*description*> не представлены, что существенно ограничивает возможность использования метаданных этого репозитория при тематическом поиске, поскольку и названия части коллекций в нём не позволяют классифицировать тематику (например, «Вестник науки и образования», «Научный журнал», «Современные инновации»).

По результатам анализа составлен список наиболее перспективных российских репозиториев ОД по научно-технической тематике с целью их использования в системе ЕОАИ ГПНТБ России на первом этапе опытной эксплуатации в качестве дополнительного источника информации при обработке пользовательских запросов (см. табл. 2).

Таблица 2

Список российских репозиториев ОД, метаданные из которых планируется использовать в ЕОАИ ГПНТБ России в качестве дополнительного источника информации

Название репозитория	Базовый *URL* провайдера данных	Число документов (по состоянию на 01.09.2021)
Архив электронных ресурсов Сибирского федерального университета	http://elib.sfu-kras.ru/oai/request	79 045
Электронный научный архив Уральского федерального университета	http://elar.urfu.ru/oai/request	89 229
Репозиторий Тверского государственного университета	http://eprints.tversu.ru/cgi/oai2	Нет данных
Репозиторий Белгородского государственного университета	http://dspace.bsu.edu.ru/oai/request	42 767
Репозиторий Казанского федерального университета	http://dspace.kpfu.ru/oai	84 076
Репозиторий Самарского университета	http://repo.ssau.ru/oai/request	33 791
Архив ОД Санкт-Петербургского государственного университета	https://dspace.spbu.ru/oai/request	31 142

Следующая задача – составление списков коллекций для каждого репозитория, из которых будут собираться метаданные. Как было отмечено выше, ЕОАИ ГПНТБ России не ставит целью собирать или агрегировать все доступные метаданные открытых архивов. Такую задачу успешно решают агрегаторы, например «Национальный агрегатор открытых репозиториев (НОРА)» [**7**]. В нашем проекте метаданные

собираются для поддержки дополнительных сервисов, связанных с обработкой поисковых запросов пользователей, а также для информирования пользователей о новых публикациях по отслеживаемой тематике. Соответственно, провайдер данных ЕОАИ ГПНТБ России не будет передавать собранные метаданные внешним клиентам *OAI-PMH*. Экспортироваться будут только собственные метаданные.

В дальнейшем также планируется интеграция ЕОАИ с действующей в ГПНТБ России системой унифицированного сводного каталога научных библиотек и библиотек образовательных организаций (ИС ЭКБСОН). ИС ЭКБСОН содержит более 60 млн библиографических записей, которые затем дедублицируются и сводятся в единую сводную запись. Интеграция с точки зрения как поиска, так и сервисов, существенно повысит возможности качественного обслуживания пользователей как ЕОАИ, так и ИС ЭКБСОН.

Разработка сервисов для поддержки поиска информации через единую точку доступа является одной из целей создания ЕОАИ ГПНТБ России. Решение этой задачи должно способствовать развитию технологии обмена данными между информационными системами библиотек, научно-информационных подразделений научных организаций, университетов, общеобразовательных учреждений.

СПИСОК ИСТОЧНИКОВ

1. **Шрайберг Я. Л.** О разработке концепции Открытого архива информации ГПНТБ России / Я. Л. Шрайберг, М. В. Гончаров, К. А. Колосов // Науч. и техн. б-ки. – 2020. – № 12. – С. 45–58.

2. **Лакизо И. Г.** Ресурсы открытого доступа как объект формирования фондов академических библиотек (Опыт ГПНТБ СО РАН) / И. Г. Лакизо, Н. И. Подкорытова, Л. В. Босина // Там же. – 2019. – № 5. – С. 78–93.

3. **Чехович Ю. В.** Открытый доступ и проблема качества квалификационных работ [Электронный ресурс] / Ю. В. Чехович, М. А. Суворова // Научное издание международного уровня – 2018: ред. политика, открытый доступ, науч. коммуникации : материалы

7-й Междунар. науч.-практ. конф., Москва, 24–27 апр. 2018 г. – Москва, 2019. – С. 163–169. – doi: 10.24069/konf-24-27-04-2018.29.

4. **The Open** Archives Initiative Protocol for Metadata Harvesting [Электронный ресурс]. – URL: https://www.openarchives.org/OAI/openarchivesprotocol.html (дата обращения: 01.09.2021).

5. **Registry** of Open Access Repositories [Электронный ресурс]. – URL: http://roar.eprints.org/ (дата обращения: 01.09.2021).

6. **Земсков А. И.** Открытый доступ: роль библиотек / А. И. Земсков // Науч. и техн. б-ки. – 2016. – № 6. – С. 41–61.

7. **Скалабан А. В**. НОРА – национальный агрегатор открытых репозиториев российских университетов. Текущее состояние и перспективы развития [Электронный ресурс]. – URL: https://scholar.google.com/scholar?hl=ru&as_sdt=0,5&q=НОРА–национальный+агрегатор+открытых+репозиториев+российских+университетов.+Текущее+состояние+и+перспективы+развития&btnG= (дата обращения: 01.09.2021).

8. **Открытый** доступ: история, современное состояние и путь к открытой науке : моногр. / Вахрушев М. В., Гончаров М. В., Засурский И. И. [и др.] ; под общ. и науч. ред. д-ра техн. наук, проф., чл.-кор. Рос. акад. образования Я. Л. Шрайберга. – Санкт-Петербург [и др.] : Лань, 2020. – 165, [**1**] с. : ил. – ISBN 978-5-8114-5034-3.

REFERENCES

1. **Shrayberg Ya. L.** O razrabotke kontseptsii Otkrytogo arhiva informatsii GPNTB Rossii / Ya. L. Shrayberg, M. V. Goncharov, K. A. Kolosov // Nauch. i tehn. b-ki. – 2020. – № 12. – S. 45–58.

2. **Lakizo I. G.** Resursy otkrytogo dostupa kak obekt formirovaniya fondov akademicheskih bibliotek (Opyt GPNTB SO RAN) / I. G. Lakizo, N. I. Podkorytova, L. V. Bosina // Tam zhe. – 2019. – № 5. – S. 78–93.

3. **Chehovich Yu. V.** Otkrytyy dostup i problema kachestva kvalifikatsionnyh rabot [Elektronnyy resurs] / Yu. V. Chehovich, M. A. Suvorova // Nauchnoe izdanie mezhdunarodnogo urovnya – 2018: red. politika, otkrytyy dostup, nauch. kommunikatsii : materialy 7-y Mezhdunar. nauch.-prakt. konf., Moskva, 24–27 apr. 2018 g. – Moskva, 2019. – S. 163–169. – doi: 10.24069/konf-24-27-04-2018.29.

4. **The Open** Archives Initiative Protocol for Metadata Harvesting [Elektronnyy resurs]. – URL: https://www.openarchives.org/OAI/openarchivesprotocol.html (data obrashcheniya: 01.09.2021).

5. **Registry of** Open Access Repositories. [Elektronnyy resurs]. – URL: http://roar.eprints.org/ (data obrashcheniya: 01.09.2021).

6. **Zemskov A. I.** Otkrytyy dostup: rol bibliotek / A. I. Zemskov // Nauch. i tehn. b-ki. – 2016. – № 6. – S. 41–61.

7. **Skalaban A. V.** NORA – natsionalnyy agregator otkrytyh repozitoriev rossiyskih universitetov. Tekushchee sostoyanie i perspektivy razvitiya [Elektronnyy resurs]**.** – URL: https://scholar.google.com/scholar?hl=ru&as_sdt=0,5&q=HOPA–natsionalnyy+agregator+otkrytyh+repozitoriev+rossiyskih+universitetov.+Tekushchee+sostoyanie+i+perspektivy+razvitiya &btnG= (data obrashcheniya: 01.09.2021).

8. **Otkrytyy** dostup: istoriya, sovremennoe sostoyanie i put k otkrytoy nauke : monogr. / Vahrushev M. V., Goncharov M. V., Zasurskiy I. I. [i dr.] ; pod obshch. i nauch. red. d-ra tehn. nauk, prof., chl.-kor. Ros. akad. obrazovaniya Ya. L. Shrayberga. – Sankt-Peterburg [i dr.] : Lan, 2020. – 165, [1] s. : il. – ISBN 978-5-8114-5034-3.

Информация об авторах / Information about the authors

Гончаров Михаил Владимирович – канд. техн. наук, доцент, ведущий научный сотрудник, руководитель группы перспективных исследований и аналитического прогнозирования ГПНТБ России, доцент Московского государственного лингвистического университета, Москва, Российская Федерация

goncharov@gpntb.ru

Mikhail V. Goncharov – Cand. Sc. (Engineering), Associate Professor, Leading Researcher, Head, Group for Perspective Research and Analytic Forecasting, Russian National Public Library for Science and Technology; Associate Professor, Moscow State Linguistic University, Moscow, Russian Federation

goncharov@gpntb.ru

Колосов Кирилл Анатольевич – канд. техн. наук, ведущий научный сотрудник ГПНТБ России, доцент Московского государственного лингвистического университета, Москва, Российская Федерация

kolosov@gpntb.ru

Kirill A. Kolosov – Cand. Sc. (Engineering), Leading Researcher, Russian National Public Library for Science and Technology; Associate Professor, Moscow State Linguistic University, Moscow, Russian Federation

kolosov@gpntb.ru