Вопросы проектирования цифрового архива федерального исследовательского центра

Designing the digital archive of the federal research center

Е. В. Ковязина,
Институт вычислительного моделирования
Сибирского отделения Российской академии наук,
Красноярск, Россия

Elena Kovyazina
Institute of Computational Modeling
of the Russian Academy of Sciences Siberian Branch,
Kransoyarsk, Russia

Наличие институционального репозитория в научном центре позволяет проводить качественный учет и продвижение научных публикаций, обеспечить открытость результатов научных исследований для мирового сообщества. В докладе показано, что в процессе проектирования необходимо учесть особенности организационной структуры и спектра исследовательской деятельности научного центра. А проблему содержательного наполнения репозитория целесообразно интенсифицировать с помощью импорта описательных метаданных из внешних библиографических источников.
Ключевые слова: институциональный репозиторий, открытый доступ, описательные метаданные.

The institutional repository of the scientific center allows for the qualitative registration and promotion of scientific publications, ensuring the openness of scientific research results for the world community. The author demonstrates that in the design process, it is necessary to take into account the features of the organizational structure and range of research activities of the scientific center. The repository content development can be intensified by importing descriptive metadata from external bibliographic sources.
Keywords: Institutional repository, open access, descriptive metadata.

Введение. Основной деятельностью научного центра являются фундаментальные и прикладные научные исследования, определенные государственным заданием и текущими научными проектами. Результаты таких исследований, как правило, представлены научными публикациями, полное собрание которых служит целям сохранения научного наследия, неразрывно связанного с деятельностью центра. В силу указанных обстоятельств во всех научных организациях одной из традиционных функций библиотеки является как можно более полный учет таких публикаций. А так как научная деятельность не может существовать в отрыве от научных коммуникаций, то большинство ученых заинтересованы в возможно более широком распространении и всестороннем обсуждении результатов их исследований. Таковое распространение носит характер *продвижения* научного исследования, служит повышению престижа научной организации, а также персональной известности каждого отдельного ученого. В настоящее время заинтересованность исследователей в продвижении их научных результатов проявляется в том, что они широко практикуют такие подходы как, например, дублирование текстов на английском языке для вывода публикации на международный уровень; публикацию в цифровом формате в журналах или трудах конференций, представленных в зарубежных коммерческих информационных ресурсах; выбор для опубликования высокорейтинговых изданий, отраженных в российских и международных индексах научного цитирования и т.п. Такой подход позволяет увеличить цитируемость публикации, что в свою очередь служит показателем ее качества и научной значимости.

Однако среди вышеперечисленных мер продвижения результатов исследований отсутствует то, что в современном научном сообществе обозначается термином «открытая наука» и тесно связано с открытым доступом к результатам научных исследований. Строго говоря, открытая наука предполагает не только доступ к публикациям, но и открытую доступность любых данных научных исследований. Однако предметом рассмотрения здесь являются только научные публикации, размещение которых в открытом доступе повышает их «видимость» и увеличивает, часто

многократно, цитируемость. По указанным причинам априори предполагается заинтересованность научного центра и его сотрудников в полноценном учете и продвижении их публикаций, а как следствие, в открытом доступе к ним.

Инициатива открытого доступа и присоединение к ней многочисленных университетов и научных институтов мира проявляется наличием в них «институциональных репозиториев (ИР) – электронных архивов для длительного хранения, накопления и обеспечения долговременного и надежного открытого доступа к результатам научных исследований, проводимых в учреждении» (Википедия). Институциональные репозитории получившие широкое распространение в мире, построены, как правило, на специально предназначенном для этих целей программном обеспечении, реализующем Протокол Инициативы открытых архивов для сбора метаданных (OAI-PMH), и полностью обеспечивают функции учета и продвижения публикаций научного центра. В техническом плане ИР представляет собой цифровой архив, поэтому далее эти термины будут использоваться как синонимы.

Особенности цифрового архива научного центра. Базовое определение открытого доступа, временные рамки в его регламентах и особенности российского законодательства об авторском праве, как правило, не позволяют администрации научной организации в полной мере присоединиться к Инициативе открытого доступа, а институциональный репозиторий может считаться цифровым архивом открытого доступа только условно. Определенные сомнения внушают также доводы сторонников централизованных в национальном масштабе репозиториев, а также известия об определенном противодействии издателей европейским инициативам публикации в открытом доступе результатов исследований, поддержанных государственным финансированием. Однако, российские реалии не оставляют надежды на решение проблемы открытого доступа к результатам научных исследований в национальном масштабе, хотя некоторые ее проблески и пробудил проект НОРА, в рамках которого уже опубликованы «Методические рекомендации по разработке репозиториев» [1] и ряд аналитических докладов.

Цифровой архив научного центра строится на основе организационных документов и решений, принятых в свое время в Институте вычислительного моделирования Сибирского отделения РАН. Перечень организационных документов и принятый на практике процесс отбора документов в репозиторий Федерального исследовательского центра «Красноярский научный центр Сибирского отделения Российской академии наук» (ФИЦ КНЦ СО РАН) подробно приведен в [2].

Исходные информационные ресурсы и их развитие в историческом контексте. В рамках традиционных библиотечных технологий библиографические описания публикаций компоновались в базу данных трудов сотрудников, а печатные версии публикаций составляли базовую часть резервно-сохранного фонда библиотеки. По мере развития компьютерных технологий и роста цен на бумажные носители печатные версии собрания публикаций сменяли их цифровые копии. Традиционные методы работы диктовали единственный вид цифровых документов – электронные версии печатных публикаций – статичные цифровые документы, практически не отличающиеся от их печатных источников. С течением времени в библиотеках сформировались два отдельных, но взаимосвязанных электронных ресурса – база данных трудов сотрудников, формируемая в системе автоматизации библиотеки (САБ), и собрание цифровых копий печатных публикаций, связанных с библиографическими описаниями в базе ссылками. Такой характер работы не требовал какого-либо пересмотра традиционных библиотечных технологий и легко управлялся средствами САБ. По мере дальнейшего развития под влиянием преимущественно административных требований трансформировались библиографические описания документов, пополняясь данными об импакт-факторах изданий, ссылками на российские и международные индексы цитирования, библиографическими ссылками из списков литературы и т.п. Для некоторых данных не существовало однозначных полей в традиционных MARC-форматах, приходилось приспосабливать для этих целей наиболее подходящие из имеющихся, тенденциозно истолковывая инструкции ГОСТов. Менялись и цифровые документы архива: практически каждый пункт списка литературы пополнился URL-ссылкой, появились также многочисленные ссылки в тексте, указывающие на связанные с исследованиями данные, справочную информацию, научно-технические проекты, с помощью которых проводились исследования и т.д. Значительная часть документов в сегодняшнем архиве уже на

этапе формирования была предназначена для использования только в электронном виде и не предполагала печатной копии.

Взрывной характер развития информационных и Интернет технологий в последние годы существенно изменил как характер и природу самих научных публикаций, так и привычки, манеру работы, требования и ожидания пользователей библиотеки. Научные публикации мигрировали в цифровую среду, приобретя при этом новые качества, требующие новых технологий и методов работы, о чем уже в течение ряда лет писали как отечественные, так и зарубежные исследователи, например, [3-4]. В приведенных публикациях выделен ряд требований к работе с цифровыми документами, нереализуемых или реализуемых с трудом в рамках традиционных библиотечных технологий, как-то обеспечение ссылочной целостности, связывание данных, поиск с использованием онтологий, глобальная интероперабельность, контроль доступа, поддержка хранилищ и т.д. Для работы с цифровыми объектами разрабатываются новые модели информационных систем как общего характера [5], так и отражающие специфику конкретной ограниченной области использования, как например, концептуальная модель репозитория академических библиотек Индонезии, основанная на управлении знаниями [6].

Выбор программного обеспечения (ПО). Требования учета (надежность хранения) и одновременно продвижения (открытость доступа) создают некоторые технические противоречия, из-за которых программное обеспечение («коробочный» вариант) части крупных IT-компаний не пригодно к использованию для организации доступа к институциональным репозиториям. Ограниченность финансовых возможностей организаций РАН, исключающих проприетарное ПО, с одной стороны, и наличие квалифицированных специалистов в области IT, с другой стороны, позволяет остановить свой выбор на открытом ПО, удовлетворяющем следующим требованиям:

а) лицензировано как ПО с открытым исходным кодом;

б) соответствует протоколу OAI-PMH;

в) имеет большое количество установок по всему миру.

В силу наибольшего распространения в России (по данным OpenDOAR http://www.opendoar.org/countrylist.php?cContinent=Europe#Russian Federation) и мире для реализации цифрового архива научного центра был выбран DSpace (разработка и поддержка MIT Libraries and Hewlett-Packard Labs).

Внутренняя структура цифрового объекта DSpace объединяет метаданные, цифровой контент и отношения с другими объектами, что вполне соответствует концептуальной модели информационных систем, работающих с цифровыми объектами. Наиболее значимо для интеграции объектов репозитория в мировое информационное пространство существование уникальных идентификаторов (URI) цифровых объектов, битовых потоков, коллекций и сообществ, позволяющих их однозначно идентифицировать. Установленный в ФИЦ КНЦ СО РАН DSpace использует СУБД Postgres SQL и веб-интерфейс Apache Tomcat. Программное обеспечение развернуто на виртуальном сервере ИВМ СО РАН, а в качестве тестового информационного массива используется база трудов сотрудников института и архив научных публикаций в цифровом виде, формируемые уже в течение многих лет.

Проектирование и тестовая эксплуатация. Обширный опыт использования DSpace в мире позволяет детально исследовать практику функционирования цифровых репозиториев и уже в период проектирования и тестовой эксплуатации, адаптировать и модифицировать программное обеспечение в соответствие с нуждами научного центра. Пошаговый переход с DSpace v5.3 до текущей v6.2 позволил убедиться в том, что программное обеспечение оперативно меняется вслед за развитием технологий, вынуждая пользователей к повышению квалификации и постоянному освоению нового. Документация содержит не только техническое описание, но и краткое пояснение теоретических основ используемых технологий. С момента запуска в тестовую эксплуатацию исследованы и протестированы следующие актуальные направления работы, предполагающие предварительные проектные действия, детальное исследование возможностей, встроенных в систему, и инструментов адаптации программного обеспечения:

Базовые установки.

1. URI (handle). Уникальные идентификаторы формируются встроенным handle-сервером DSpace с заданным незарегистрированным префиксом. Для формирования зарегистрированных уникальных идентификаторов требуется платная подписка на службу регистрации префиксов. По умолчанию в систему встроена служба CNRI Handle System.
2. В качестве формата метаданных принято DC с квалификаторами, проходит тестирование использование дополнительных полей метаданных, предназначенных для наукометрических параметров, принятых в БД трудов сотрудников.
3. Аутентификация пользователей производится по паролю и IP-адресам, не исключается и возможность анонимного входа.
4. Наряду с встроенной в систему статистикой Apache Solr, протестирована интеграция с Google Analitics.
5. Настроено отслеживание контрольных сумм в хранилище с помощью назначенного задания.

Ограничения доступа. Большая часть представленных в репозитории документов являются служебными произведениями и не допускают передачи имущественных прав. Однако добавленная стоимость издательств в результате рецензирования, корректуры, правки и верстки вынуждает к компромиссу с издателями, регулируемому с помощью эмбарго – временных ограничений, налагаемых издателями на публикацию произведения на сайте автора или его места работы. Наиболее полный перечень продолжительности таких ограничений представлен на сайте проекта SHERPA/ROMEO, посвященного открытому доступу (http://www.sherpa.ac.uk/romeo/PDFandIR.php?la=en).

1. Требования к доступу в цифровом архиве зависят от правового статуса документа, регулируются программными средствами, и могут быть следующими:
2. Полностью открытый доступ для служебных произведений с истекшим периодом эмбарго. В случае наличия соавторов-сотрудников сторонних организаций предполагается наличие авторского договора с ними. Поле даты окончания эмбарго в этом случае пустое.
3. Отсроченный открытый доступ для документов, находящихся под действием эмбарго, либо ожидающих договора/лицензии соавторов. Описание таких документов должно содержать информацию об эмбарго и сроках его действия, либо формулировку иных причин временного ограничения доступа. Устанавливается дата окончания эмбарго и\или причина ограничения доступа.
4. Закрытый доступ для произведений, вышедших в издательствах, условия публикации в которых вообще не предполагают открытого доступа. Если институциональный репозиторий будет зарегистрирован в соответствующих службах учета репозиториев открытого доступа, то, по-видимому, даже описательные метаданные таких документов должны быть недоступны анонимным пользователям. Документ помечается как приватный.

Электронные документы, содержащие государственную тайну, не подлежат учету в репозитории, хранятся и содержатся в соответствующих подразделениях и условиях.

Электронные документы, определяемые пп.2-3, предполагают «темное» размещение в репозитории, пока разрешение на открытый доступ не может быть получено, то есть в цифровой архив такие документы занесены, но «не видны» никому, кроме администратора (приватные документы).

Изучение опыта работы зарубежных университетов с ограничениями доступа по публикациям в DSpace выявило сходное деление документов. Интересен опыт разработки библиотекарями технологии отбора документов в цифровой архив, представленный в статье [7]. В ней технология названа workflow (рабочий процесс), хотя, по сути, очень напоминает технологическую карту в привычной российским библиотекарям терминологии. Процесс включает правила сбора и стимулирования авторов к сдаче публикаций в репозиторий, последующую проверку по справочникам SHERPA/ROMEO с дифференциацией настроек эмбарго по цвету, формы административного принуждения авторов и т.д.

Заимствование метаданных.

Длительный опыт работы с трудами сотрудников в САБ ИРБИС и корректно сформированные и полные библиографические описания делают очень заманчивой перспективу заимствования метаданных из САБ. DSpace предоставляет множество способов заимствования, применимых

прежде всего для сторонних систем, поддерживающих глобальную интероперабельность. К сожалению, для нашей САБ они не применимы, поэтому тестировались:

1. Экспорт-импорт через csv. Выявились проблемы с кодировками, существенно усложнившие процесс.
2. Выгрузка данных из САБ в формате DC. Потребовалась существенная корректировка и доработка таблиц соответствия, до сих пор не до конца завершенная.
3. Обмен данными через систему ZOOSPACE.

В целом, все три способа в конечном итоге привели к положительным результатам. Однако для промышленного использования требуется дальнейшая детальная корректировка.

Репликация данных

Хотя цифровой архив научного центра представляет собой систему сообществ и коллекций, повторяющую структуру организации, принятый при проектировании принцип «один документ – один объект – один уникальный идентификатор» требует устранения дублирования объектов в разных коллекциях. Дублирование может возникать в случае соавторства из различных подразделений. В этом случае реальный объект существует только в одной коллекции, а в коллекциях соответствующих подразделений присутствует в виде ссылки. Теоретически возможна коллекция, состоящая только из ссылок. Требует отдельного рассмотрения проблема интерпретации статистики таких коллекций, особенно при сравнительном анализе публикационной активности подразделений.

Самоархивирование, свободное лицензирование и корректировка данных

В случае отсутствия публикации в базе трудов автор имеет возможность известить библиотеку о ней, введя вручную предварительные данные в специально предназначенную для этих целей коллекцию с обязательным приложением полного текста публикации или ссылки на него. После рассмотрения введенных данных, их проверки и корректировки объект перемещается в соответствующую коллекцию.

В DSpace интегрировано свободное лицензирование Creative Common (CC). Система требует указание типа лицензии CC на этапе ввода метаданных. Связь с сервисом корректно работает, однако требует изучения регламент применения лицензирования, и этот вопрос требует дальнейшей работы.

Выявленные авторами ошибки и неточности в данных корректируются с помощью системы обратной связи, но допустима и самостоятельная корректировка квалифицированными пользователями, получившими на нее индивидуальные права

Перспективы развития. Для оценки перспектив развития цифрового архива изучалось использование DSpace зарубежными библиотеками по публикациям Web of Science и Scopus. Публикации содержат обширный опыт работы с цифровыми архивами, который позволяет избежать повторения ошибок, недочетов и выявить направления дальнейшего развития цифрового архива. Интересное исследование библиотекарей Сиднейского технологического университета [8] заинтересовало очевидным сходством проблем и ситуации. Результаты исследования будут использованы при моделировании и будущей корректировке пользовательского веб-интерфейса. Анализ причин и деталей плохой применимости MARC-форматов для описания цифровых объектов в [4, 9], а также определение направлений трансформации описаний, позволит произвести доработку формата метаданных с учетом данных рекомендаций.

Обзор исследований по автоматизации извлечения данных из библиографических списков и контента цифрового документа [10] и констатация существенного прогресса в этой области, позволяет надеяться на пополнение метаданных списками библиографии с последующим связыванием содержащихся в них данных. Связанные данные – это наиболее актуальный раздел работы с цифровыми документами, активно развивающийся и перспективный, что подтверждает большое количество публикаций на эту тему, например, [9], и выход методических указаний по работе с ними Американской библиотечной ассоциации [11].

Заключение. Наличие институционального репозитория, формируемого в технологиях архивов открытого доступа позволяет проводить качественный учет и продвижение научных публикаций, повысить показатели цитирования, провести многоаспектный анализ публикационной активности и научных связей авторов, выделяя круг профессионального взаимодействия, обеспечить

открытость результатов научных исследований для мирового сообщества, пополняя копилку знаний человечества. В процессе проектирования цифрового архива необходимо учесть особенности организационной структуры и спектра исследовательской деятельности научного центра, удачное отображение которых в цифровом архиве позволит в перспективе успешно внедрить технологии связывания данных, а также интегрировать данные в корпоративные, региональные или национальные проекты открытых архивов, в соответствие с общемировыми тенденциями развития технологий. Импорт описательных метаданных научных публикаций из имеющихся в научном центре или внешних библиографических баз данных позволит эффективно решить проблему содержательного наполнения цифрового архива. Освоение новых принципов и подходов, технологических решений, систем и сервисов, используемых мировым библиотечным сообществом, повлечет за собой новое осмысление работы с информацией, ее будущие перспективы, и роль библиотекарей в этом процессе.

Список литературы

1. Методические рекомендации по разработке репозиториев / под ред. М. Е. Шварцмана. – М.: Ваше цифровое издательство, 2018. – 34 с. – Режим доступа: https://openrepository.ru/images/docs/Metod_Schwarzman.pdf.
2. Ковязина Е. В. Открытый архив в научном центре: особенности формирования // Распределенные информационно-вычислительные ресурсы. Наука – цифровой экономике (DICR-2017). Труды XVI всероссийской конференции. Институт вычислительных технологий СО РАН. – 2017. – С. 434-440. – Режим доступа: http://elib.ict.nsc.ru/jspui/bitstream/ICT/1467/115/paper56.pdf.
3. Жижимов О. Л., Мазов Н. А., Федотов А. М. Некоторые заметки об эволюции цифровых репозитариев традиционных библиотек к полнофункциональным электронным библиотекам // Вестник Владивостокского государственного университета экономики и сервиса. Территория новых возможностей. – 2010. – Т.3. – № 7. – С. 55-63. – ISSN 2073-3984.
4. Clobridge Abby. Libraries in Transition: From book collections & union catalogues to open access & digital repositories // ProInflow : Časopis pro informační vědy. – 2011. – № 2. – pp. 121-132.
5. Федотов А. М., Барахнин В. Б., Жижимов О. Л., Федотова О. А. Модель информационной системы для поддержки научно-педагогической деятельности // Вестник НГУ. Сер.: Информационные технологии. – 2014. – Т.12. – № 1. – С. 89-101. – ISSN 1818-7900. – Режим доступа: https://cyberleninka.ru/article/v/model-informatsionnoy-sistemy-dlya-podderzhki-nauchno-pedagogicheskoy-deyatelnosti.
6. Ida Farida, Jann Hidajat Tjakraatmadja, Aries Firman, Sulistyo Basuki. A conceptual model of Open Access Institutional Repository in Indonesia academic libraries: Viewed from knowledge management perspective // Library Management. – Vol. 36. – Issue: 1/2. – pp. 168-181. – Режим доступа: https://doi.org/10.1108/LM-03-2014-0038.
7. Hazzard J., Towery S. Workflow Development for an Institutional Repository in an Emerging Research Institution // Journal of Librarianship and Scholarly Communication. – 2017. – 5 (General Issue). – eP2166. – Режим доступа: https://doi.org/10.7710/2162-3309.2166.
8. NarayanM Bhuva, Luca Edward. Issues and challenges in researchers' adoption of open access and institutional repositories: a contextual study of a university repository // Information Research. – 2016. – Vol.22. – № 4. – Режим доступа: http://www.informationr.net/ir/22-4/rails/rails1608.html.
9. Gonzales Brighid M. Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record // Information Technology and Libraries. – 2014. – декабрь. – Режим доступа: https://ejournals.bc.edu/ojs/index.php/ital/article/view/5631.
10. Hatop Götz. Extraction, analysis and publication of bibliographical references within an institutional repository / Götz Hatop // Library Hi Tech. – Vol. 34. – № 2. – pp. .Режим доступа: http://dx.doi.org/10.1108/LHT-01-2016-0003.
11. Mitchell, Eric T. Library Linked Data: Early Activity & Development / Eric T. Mitchell // Library Technology Reports. – 2016. – Vol.52. – № 1. – P. 18-23. – Режим доступа: https://journals.ala.org/ltr/issue/download/534/290.