УДК 026.06+930.25:004
DOI 10.33186/978-5-85638-221-0-2019-15-19

Цифровой архив: обмен данными с системой автоматизации библиотеки

Digital repository: Data exchange with library automation system

Е. В. Ковязина
Институт вычислительного моделирования СО РАН,
Красноярск, Россия

Elena V. Kovyazina
Institute of Computational Modeling, Russian Academy of Sciences Siberian Branch,
Krasnoyarsk, Russia

Движение мирового сообщества по пути развития инфраструктуры открытой науки инициировало массовое внедрение технологий открытых архивов в повседневную работу библиотек научно-исследовательских организаций. Сбор и формирование данных в таких системах проводит, как правило, библиотека, поэтому немаловажной задачей является интеграция и обмен данными цифрового репозитория с системой автоматизации библиотек. В докладе представлен опыт заимствования данных и возможные пути решения этой задачи.
Ключевые слова: цифровые репозитории, обмен данными, DSpace.

The advance of the international community on the way to open science has triggered the mass implementation of digital repositories into the libraries of academic organizations. These repositories are designed to perform the role of open archives. As a rule, it is the library that acquires and generates data in such systems, therefore the integration and exchange of archive data with the library automation system makes a primary goal. The experience in data sharing and possible solutions are discussed.
Keywords: Digital repository, data exchange, DSpace.

Введение. Продвижение человечества к цифровому обществу, инициировало кардинальные изменения в принципах и подходах к науке и образованию. Повсеместное распространение Интернет, развитие цифровой техники и информационных технологий сформировали условия для широкого и открытого распространения знаний. В международном сообществе активно обсуждаются принципы и технологии осуществления перехода к открытой науке и открытому образованию [1-4]. Создаются и активно развиваются новые подходы к работе с информацией, трансформируется понятийный аппарат, формируется терминология, направленная на работу с цифровыми ресурсами.

Мировое библиотечное сообщество активно включено в этот процесс. Библиотеки, претендуют на ведущие роли в формировании интеллектуального наполнения глобальной паутины, организации и упорядочении Интернет-пространства, оказании разнообразных услуг обеспечения релевантности поиска, оценки значимости ресурсов и документов. Библиотеки научных и научно-образовательных организаций активно участвуют в оценке публикационной активности, способствуют продвижению научных трудов и иных данных научных исследований, зачастую берут на себя функции издателей [1], финансируют публикации открытого доступа, оказывают услуги по идентификации ученых и их трудов в Интернет-пространстве, проводят собственные наукометрические и библиометрические исследования.

Одним из важных аспектов продвижения по этому пути является открытый доступ к результатам научных исследований. При этом результаты трактуются в широком смысле этого понятия – предполагается, что доступ будет открыт не только к текстам – публикациям, отчетам, научным заметкам и т.п., но и к разнородным научным данным, включая динамически изменяемые. Такой подход подразумевает постепенное эволюционное слияние технологий институциональных репозиториев и центров обработки данных. Ряд авторов, например [4], отмечают уже имеющееся развитие функциональности цифровых архивов в направлении поддержки функций, присущих ранее хранилищам данных.

Как отмечено в [3], Россия находится в начальной стадии процесса «цифровой трансформации науки и образования». Отсутствуют инфраструктура, законодательная база для открытости доступа к научным ресурсам, необходимые компетенции сотрудников, вовлеченных в этот процесс и т.д. Однако, определенные усилия по представлению результатов научных исследований в публичном Интернет пространстве предпринимаются. В качестве результатов на первых парах рассматриваются, как правило, научные публикации. Библиотеки научных и образовательных учреждений имеют богатый опыт учета публикаций. Формируются базы трудов сотрудников институтов и с их помощью проводится оценка публикационной активности организации. Базы данных пополняются и администрируются средствами системы автоматизации библиотек (САБ) и содержат библиографические описания публикаций, дополненные наукометрическими компонентами. Библиографические описания трудов сотрудников представлены в Интернет с помощью веб-шлюза, обеспечивающего поиск и отображение его результатов на экране компьютера, а также вызов полного текста публикации, если таковой имеется, по ссылке на место его хранения. Безусловно, такой способ доступа к результатам исследований лучше, чем полное его отсутствие. Однако, для продвижения к открытой науке любой цифровой объект (для библиотеки актуален документ) должен обладать рядом качеств, многие из которых не обеспечиваются системой автоматизации библиотеки, а требуют привлечения технологий цифрового репозитория. Перечисление и описание указанных качеств, а также обоснование целесообразности ведения цифрового архива библиотекой научно-образовательной организации приводятся, например, в [1,5], а здесь хотелось бы подробнее остановиться на особенностях, используемых в цифровых архивах и САБ схем метаданных документов, определяющих возможность взаимного обмена. Сомнения, связанные с предполагаемыми проблемами обмена данными цифрового архива и используемой в библиотеке САБ являются одной из ряда практических причин, препятствующих широкому распространению репозиториев в научных библиотеках России, наряду с почти нулевой практикой работы с UNIX-подобными операционными системами, отсутствием русскоязычной документации и программистов. Связано это с тем, что, как правило, метаданные цифровых репозиториев хранятся в формате Дублинского ядра с квалификаторами (QDC). Изменение его на один из распространенных в России форматов нежелательно, так как приведет к утрате интероперабельности и затруднит обмен данными с другими репозиториями. Различие форматов хранения служит препятствием для обмена данными с САБ, преодоление которого могло бы позволить быстрее и эффективнее формировать контент цифрового архива. Пути решения задачи обмена данными изучались на примере САБ ИРБИС64 и DSpace v.6.2, работающего под управлением ОС Linux. Основное направление исследования состоит в выявлении наиболее эффективных и малозатратных по времени методов пополнения цифрового репозитория данными из внешних источников. Особенно актуальным является импорт данных из САБ ИРБИС64. Предпочтительными представляются методы, не требующие высокой программистской квалификации, и ограниченные минимальной корректировкой базовых настроек и программного кода. Для форматной совместимости полей описательных метаданных были использованы расширения схем данных, хранимых в DSpace, в частности расширенная схема данных, эксплуатируемая в репозитории ИВТ СО РАН [6].

Репозиторий, развернутый на основе DSpace, обладает двумя типами пользовательских интерфейсов:

1. Интерфейс на основе технологий Java – JSPUI.
2. Интерфейс на основе xml-технологий – XMLUI.

JSPUI – более ранняя разработка, обладающая в настоящее время более широким набором функций, часть из которых постепенно отмирает из-за появления новых, более эффективных информационных технологий. XMLUI – разработка более поздняя, активно развивающая и обрастающая все новыми функциями. Программная платформа DSpace располагает большим набором средств и методов обмена данными, работающих как в обоих интерфейсах, так и в каком-либо одном. Часть из них ограничивается только работой с метаданными документов, другая включает как метаданные, так и полный текст в виде файла. При рассмотрении методов пополнения цифрового репозитория из внешних источников, одним из которых является ИРБИС64, методы рассматривались в порядке популярности и наличия опыта работы с ними.

Обмен метаданными в формате csv. Самый распространенный из методов обмена метаданными – загрузка и выгрузка в формате csv – работает в обоих интерфейсах. Базовый DSpace выгружает данные в формате квалифицированного дублинского ядра (QDC). Данный метод неплохо подходит для выгрузки данных из DSpace с дальнейшим импортом их в ИРБИС64 и по реализации мало отличается от хорошо известного [7] метода заимствования метаданных из международных индексов научного цитирования Web of Science и Scopus с последующей переброской их в ИРБИС64. Файл читается с помощью редактора ISO-файлов, формируется таблица соответствия полей, с помощью которой данные размещаются в ИРБИС. Необходимо точно определить соответствие полей записей, что бывает затруднительно в ситуации отсутствия однозначности между полями MARC-записи и QDC. Как следствие, полученные записи нуждаются в дальнейшей глобальной корректировке для приведения вида полей к принятому стандарту. Еще одной дополнительной сложностью применения метода является отсутствие задокументированной возможности выгрузки данных в формате csv из ИРБИС64, что делает обмен данными исключительно однонаправленным.

Механизм библио-трансформации. Традиционным средством заимствования данных из внешних источников в интерфейсе JSPUI DSpace является механизм библио-трансформации (BTE). Библио-трансформация производится в два шага: шаг заимствования данных в формате источника и шаг их преобразования. Сам BTE представляет собой платформу Java, и состоит из программных API для фильтрации и изменения записей, которые извлекаются из различных типов источников данных (например, баз данных, файлов, устаревших источников данных), а также для вывода их в соответствующие стандартные форматы визуализации (например, файлы базы данных, txt, xml, Excel). Структура включает независимые абстрактные модули, которые выполняются отдельно, предлагая во многих случаях альтернативный выбор пользователю в зависимости от входного набора данных, рабочего процесса преобразования, которое необходимо выполнить, и формата вывода, который необходимо создать. Стандартная версия BTE предлагает несколько предопределенных загрузчиков данных (в настоящее время поддерживаются arXiv, PubMed, CrossRef, CiNii), а также выходные генераторы для базовых библиографических форматов. Поиск во внешних источниках производится с помощью, соответствующих загрузчику идентификаторов публикаций из простой пользовательской формы. В настоящее время поддерживается четыре идентификатора (DOI, PubMed ID, arXiv ID и NAID (идентификатор CiNii)) [8], что позволяет без труда заимствовать описания большинства зарубежных публикаций, а также отечественных, имеющих DOI CrossRef. Остается «подцепить» к описаниям файл документа и объект DSpace практически полностью сформирован [9]. Механизм библио-трансформации хорош для пополнения архива из стандартных внешних источников, а для формирования собственного загрузчика данных требует знаний и опыта программирования Java. От источника заимствования требуется наличие хорошо задокументированных внешних API, которые в ИРБИС64 отсутствуют. К тому же разработчики DSpace объявили, что указанный механизм не будет поддерживаться последующими (после версии 6) версиями программной платформы.

Импорт с помощью Simple Archive Format. Базовой концепцией простого формата архивирования DSpace, предназначенного для обмена данными между однотипными репозиториями, является создание архива, который представляет собой каталог, содержащий один подкаталог для каждого цифрового объекта. Каждый каталог объектов содержит файл для описательных метаданных объекта и файлы, составляющие этот объект.

```
archive_directory/
  item_000/
      dublin_core.xml – метаданные QDC для полей,
                                        определенных схемой DC
      metadata_[prefix].xml – метаданные в других схемах, где
                [prefix] – имя схемы, зарегистрированное в
                                          реестре метаданных
      contents – текстовый файл, содержащий строки имен файлов,
                                      по одному имени в строке
      collections – текстовый файл, содержащий дескрипторы
```

```
                    коллекций, которым будет принадлежать объект.
                    Опционально, дескрипторов может быть несколько,
                    указываются по одному в строке. Коллекция в
                  первой строке будет коллекцией-владельцем объекта
        file_1.doc - файлы, добавляемые как битовые потоки в
                                                цифровой объект
        file_2.pdf
     item_001/
        dublin_core.xml
        contents
        file_1.png
  ...
```

Импортирование данных с помощью простого формата архивирования удобно тем, что позволяет переносить все метаданные объекта, включая поля расширенной схемы данных. Если аккуратно прописать все поля метаданных из ИРБИС64 в xml-структуру DSpace, то полученный файл позволяет перенести в репозиторий все необходимые метаданные, а также файлы полных текстов единым пакетом. Кроме того, такой вариант импорта позволяет перенести сразу несколько записей (максимальное их количество регулируется DSpace параметрически). Импорт работает как в JSPUI, так и в XMLUI.

Система прямого импорта из внешних источников. Для полноты информации об исследуемых методах следует упомянуть о системе прямого импорта из внешних источников, основанной на таблицах соответствия XSLT и работающей в интерфейсе XMLUI DSpace. На текущий момент эксплуатации DSpace работа преимущественно с интерфейсом JSPUI не позволила с достаточной полнотой произвести тестирование этого метода, но работа в этом направлении продолжается.

Заключение. Реализация и успешная эксплуатация цифрового репозитория, аккумулирующего научные публикации, принадлежащие научной организации на правах служебного произведения, является хорошей практикой на пути цифровой трансформации и первым шагом в направлении открытой науки. 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. Однако, эксплуатация репозитория в режиме дифференцированного доступа позволит в будущем обсуждать и согласовывать с администрацией вопросы лицензирования документов и коллекций, разделения прав между авторами, издателями и научной организацией, а также вопросы объединения данных в корпоративных проектах. Разрешение вопросов обмена данными между цифровым архивом и САБ позволит научным библиотекам пополнять архив, использовать в работе его возможности и, в конечном итоге, включиться в процесс цифровой трансформации науки, подобно тому, как это делают зарубежные библиотеки.

Список источников

1. **Ayris P., Ignat T.** Defining the role of libraries in the Open Science landscape: a reflection on current European practice // Open Information Science. 2018. № 2. pp. 1-22. URL: https://doi.org/10.1515/opis-2018-0001 (дата обращения 21.10.2019).
2. **Clobridge A.** Libraries in Transition: From book collections & union catalogues to open access & digital repositories // ProInflow: Časopis pro informační vědy. 2011. № 2. pp. 121-132. URL: http://knihovna.fss.muni.cz/caslin2011/soubory/clobridge-p.pdf (дата обращения 21.10.2019).
3. **Качан Д.А., Богатко А.В., Богатко И.Н., Енин С.В., Кулаженко В.Г., Лазарев В.С., Лис П.А., Скалабан А.В., Юрик И.В.** Интеграция информационных ресурсов открытого доступа для обеспечения научно-образовательного процесса в учреждениях высшего образования. // Открытое образование. 2018. 22(4). С. 53-63. URL: https://doi.org/10.21686/1818-4243-2018-4-53-63 (дата обращения 21.10.2019).
4. **Castelli D.,Manghi P., Thanos C.** A vision towards Science Communication Infrastructures // International Journal on Digital Libraries. 2013. V. 13. Issue 3-4. P. 155-169. URL: https://doi.org/10.1007/s00799-013-0106-7 (дата обращения 21.10.2019).

5. **Ковязина Е.В.** Открытый архив: импорт данных // Информатизация образования и методика электронного обучения: материалы II международной научной конференции, Красноярск, СФУ, 24 - 26 сентября 2018 г. Красноярск: СФУ, 2018. Ч.1. стр. 333-337. URL: https://elibrary.ru/item.asp?id=35595574 (дата обращения 31.10.2019).
6. **Шокин Ю.И., Жижимов О.Л., Федотов А.М.** Информационные системы ИВТ СО РАН // XVI Российская конференция «Распределенные информационно - вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017): Труды XVI Всероссийской конференции (4-7 декабря 2017 г., г.Новосибирск). Новосибирск: ИВТ СО РАН, 2017. С.11-18. URL: http://elib.ict.nsc.ru/jspui/bitstream/ICT/1467/5/paper01.pdf (дата обращения 20.10.2019).
7. **Баженов С. Р., Рогозникова О.А., Данилин М.В.** Интеграция базы данных публикаций организации с индексами научного цитирования: реализация средствами САБ ИРБИС64 // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса : материалы международной конф. М.: ГПНТБ России, 2015. URL: http://elis.gpntb.ru/node/683 (дата обращения 31.08.2019).
8. **Документация** по DSPACE 6.x. URL: https://wiki.duraspace.org/display/DSDOC6x/DSpace+6.x+ Documentation (дата обращения 31.08.2019).
9. **Hidalgo Y.,Ortiz E., Febles J.P.** A Method for Integrating Bibliographic Data from OAI-PMH Data Providers // IEEE Latin America Transactions. 2017. V. 15, № 9. P. 1695-1698. https://ieeexplore.ieee.org/document/ 8015075 (дата обращения 20.10.2019).