УДК 001:37.0:004
DOI 10.20913/2618-7515-2019-3-23-28

ЦИФРОВОЙ РЕПОЗИТОРИЙ В ИНФОРМАЦИОННЫХ НАУЧНО-ОБРАЗОВАТЕЛЬНЫХ СИСТЕМАХ

DIGITAL REPOSITORY FOR RESEARCH AND EDUCATION INFORMATION SYSTEMS

© ***Федотова Ольга Анатольевна***
научный сотрудник лаборатории по развитию электронных ресурсов, Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук (ГПНТБ СО РАН), Новосибирск, Россия, fedotovao@gpntbsib.ru

© ***Федотов Анатолий Михайлович***
доктор физико-математических наук, член-корреспондент РАН, главный научный сотрудник, Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН), Новосибирск, Россия, fedotov@sbras.ru

© ***Жижимов Олег Львович***
доктор технических наук, ведущий научный сотрудник, Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН), Новосибирск, Россия, zhizhim@mail.ru

© ***Самбетбаева Мадина Аралбаевна***
аспирант, Новосибирский государственный университет (НГУ), Новосибирск, Россия, madina_jgtu@mail.ru

Fedotova Olga Anatolyevna
Researcher of the Laboratory for Electronic Resources Development, State Public Scientific Technological Library of the Siberian Branch of the Russian Academy of Sciences (SPSTL SB RAS), Novosibirsk, Russia, fedotovao@gpntbsib.ru

Fedotov Anatoliy Michailovich
Doctor of Physics-Mathematic Sciences, Chief Researcher, Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (ICT SB RAS), Novosibirsk, Russia, fedotov@sbras.ru

Zhizhimov Oleg Lvovich
Doctor of Technical Sciences, Leading Researcher, Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (ICT SB RAS), Novosibirsk, Russia, zhizhim@mail.ru

Sambetbayeva Madina Aralbayevna
Post-graduate student, Novosibirsk State University (NSU), Novosibirsk, Russia, madina_jgtu@mail.ru

Статья посвящена обзору наиболее популярных систем поддержки цифровых репозиториев и их информационной модели. Обосновывается выбор системы DSpace для хранилища данных научно-образовательной информационной системы.

Ключевые слова: *информационная система, электронная библиотека, цифровой репозиторий, хранилище данных, эталонная модель OAIS, DSpace, EPrints, Fedora, GreenStone, CDC Invenio*

The article is devoted to the overview of the most popular digital repository support systems and their information model. The choice of DSpace for the data warehouse of the scientific and educational information system is justified.

Keywords: *information system, electronic library, digital repository, data warehouse, the reference model OAIS, the protocol OAI-PMH, DSpace, EPrints, Fedora, GreenStone, CDC Invenio*

Работы по созданию информационных систем поддержки научных исследований, интегрирующих информационные ресурсы, ведутся в Институте вычислительных систем (ИВТ) СО РАН с 1996 г. В результате сложилось понимание того, что информационная система для поддержки научных исследований должна основываться на использовании концепции электронных (цифровых) библиотек. В рамках нашего подхода цифровые библиотеки рассматриваются как отдельная конкретная технология работы с цифровой информацией, образующая класс информационных систем (ИС), предназначенных для управления информационными ресурсами (ИР).

Под термином электронная библиотека (ЭБ) в данной работе будем понимать систему управления структурированными каталогизированными коллекциями разнородных электронных (цифровых) объектов (ресурсов).

ЭБ как система управления ИР, в отличие от печатных изданий, микрофильмов и других носителей, не только обеспечивает многосторонний поиск и навигацию по рубрикаторам (словарям-классификаторам, управляемым словарям), но и непосредственно предоставляет пользователю конкретный найденный ресурс (публикацию, документ, фотографию, описание факта и др.), а также дополнительные сведения о нем: например, географическую привязку, информацию об авторах или фактах, библиографию, перечень организаций, имеющих отношение к ресурсу, и т. д. [1].

Основными целями, стоящими перед ЭБ, являются: организация хранения информации и обеспечение доступа к ней, управление ИР, предоставление результатов научных исследований мировому сообществу, предотвращение утраты ценных научных и культурных коллекций для будущих поколений, повышение эффективности научных исследований и обучения, а также создание новых технологий научных исследований и эффективного инструментария для их проведения.

Основные задачи, решаемые ЭБ, – это управление и интеграция ИР, включая поддержку унифицированного доступа к ним, а также эффективная навигация.

Под интеграцией ИР понимается их объединение в целях использования (с помощью удобных и унифицированных пользовательских интерфейсов) разнородной информации с сохранением ее свойств, особенностей представления и пользовательских возможностей манипулирования с ней. При этом объединение ресурсов не обязательно должно осуществляться физически, оно может быть виртуальным, главное – оно должно обеспечивать пользователю восприятие доступной информации как единого информационного пространства. В частности, такие системы обеспечивают работу с гетерогенными наборами и базами данных или системами баз данных, обеспечивая пользователю эффективность информационных поисков независимо от особенностей конкретных систем хранения ресурсов, к которым осуществляется доступ.

Под эффективной навигацией в ИС понимается возможность для пользователя находить интересующую его информацию с наибольшей полнотой и точностью при наименьших затратах усилий во всем доступном информационном пространстве.

Исходя из целей ЭБ можно сформулировать следующие функциональные требования к ее модели [2]:

- надежное долговременное и защищенное от исчезновения хранение информации;
- актуальность, полнота, достоверность происхождения документов;
- историчность информации;
- географическая привязка информации;
- наличие большого числа словарей-классификаторов (справочников), для обеспечения идентификации и классификации ресурсов;
- поддержка неоднородных и слабо структурированных информационных ресурсов;
- поддержка взаимосвязей информационных ресурсов;
- наличие интеллектуальных служб обслуживания запросов пользователя;
- поддержка требований интероперабельности как на программном, так и на семантическом уровне;
- предоставление информации пользователю в виде, выбранном пользователем;
- наличие интеллектуальных служб обеспечения запросов пользователя;
- наличие программных интерфейсов для поддержки аналитической работы пользователя с помощью программных приложений;
- поддержка работы с внешними источниками.

Наиболее важным выводом из вышесказанного является то, что информационная модель ЭБ должна быть многоуровневой и состоять как минимум из следующих компонент: хранилище данных (репозиторий), сервер метаданных, сервер приложений (диспетчер), словари-справочники (рис. 1) [3].

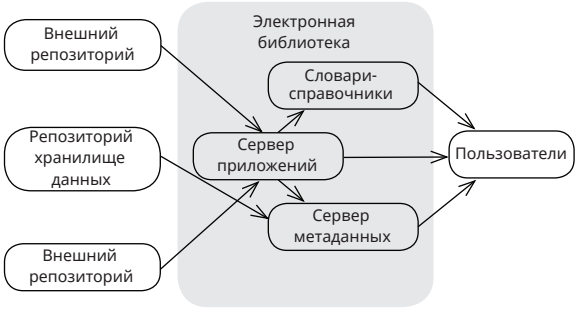

Рис. 1. Архитектура цифровой библиотеки

Репозиторий – это независимая система долговременного хранения и доступа к разнородным цифровым объектам. Цифровой репозиторий является одним из важнейших компонентов распределенной системы и предназначен только для обеспечения «функции» долговременного хранения ИР. Могут измениться система, интерфейсы и сервисы, но ИР, несущий информацию не изменяется, поэтому должен храниться вечно и независимо. Тем самым функция хранения данных отделена и не зависит от других функций и сервисов системы.

Сервер метаданных должен обеспечить работу с метаданными – каталогизацию всех ИР в соответствии с общепринятыми международными стандартами.

Сервер приложений обеспечивает сервисы, необходимые для формирования ИР с использованием и без использования диалоговых пользовательских интерфейсов. Сервисы позволяют использовать метаданные других ИС в диалоговом режиме и пакетных режимах. Их функциональность должна обеспечивать поиск и извлечение метаданных из других систем, конвертирование полученных метаданных в схемы и структуры локальной системы.

Справочники (управляемые словари, ключевые термины) – это особый вид метаданных, отражающих наиболее существенные свойства информационного объекта и имеющие наиболее важное значение с точки зрения ЭБ. Специфика словарей определятся терминологией конкретной предметной области. Необходимо рассматривать различные типы ключевых терминов (ключевые термины в стандартном понимании; ключевые термины, описывающие персону; ключевые термины, описывающие организацию; ключевые термины, описывающие временные периоды; ключевые термины, описывающие географические понятия). Это тематические словари-классификаторы, тезаурусы, рубрикаторы, описания предметной области и классификаторы документов.

Описывается технология создания и поддержки ИР с успехом была использована в научно-образовательной сфере на примере разработки ЭБ научной школы А. А. Ляпунова, а также в виде ЭБ учебных пособий по курсам «Современные проблемы информатики и вычислительной техники», «Вычислительные системы», «Информатика» и другим (http://fedotov.nsu.ru/lecture.php).

Как уже было отмечено выше, цифровой репозиторий является одним из важнейших компонентов ИС. Для организации системы долговременного хранения ИР международной организацией по стандартизации (ISO) при участии NASA предложен стандарт ISO-14721:2012 RM OAIS (Reference Model for an Open Archival Information System) [4]. Эталонная модель стандарта OAIS основана на расширенной схеме данных Dublin Core [5] с квалификаторами. Эта модель была использована многими организациями для разработки наборов метаданных и создания крупных хранилищ цифровых объектов. В более широком смысле цифровой репозиторий – это некоторая ИС, функционирующая совместно с хранилищем цифровых объектов и предоставляющая сервис как по управлению этими объектами, так и по организации доступа к ним. В последнем случае роль цифрового репозитория может выполнять практически любая система управления контентом (CMS, Content Management System), наделенная функциями работы с цифровыми объектами. Функциональность цифровых репозиториев зависит большей частью от функциональности используемого программного обеспечения (ПО).

В этом классе ПО существует достаточно большое разнообразие, причем не только среди проприетарного ПО, но и среди свободно распространяемого. Согласно данным сайта OpenDOAR (The Directory of Open Access Repositories) (http://www.opendoar.org), большинство открытых репозиториев основаны на эталонной модели RM OAIS и созданы на свободно распространяемом ПО.

На основе RM OAIS создана концепция «институционального репозитория» (IR, Institutional Repositories) как системы долговременного хранения, накопления информации и обеспечения надежного доступа к цифровым объектам, представляющим собой результат интеллектуальной деятельности научного или образовательного учреждения.

Институциональные репозитории связаны с вопросами цифровой интероперабельности, инициативой открытых архивов OAI (Open Archives Initiative), протоколом для сбора

метаданных OAI-PMH (Protocol for Metadata Harvesting), а также с понятием ЭБ, то есть с функциями сбора, хранения, классификации, каталогизации ресурсов (данных) и обеспечения доступа к цифровому контенту. Процесс интеграции цифрового репозитория в ИС основан на модели агрегирования и распространения метаданных. Применение этой модели закреплено в протоколе OAI-PMH [6], который поддерживается большинством систем, предназначенных для хранения ИР.

В настоящий момент в мире насчитывается более десятка систем поддержки цифровых хранилищ (институциональных репозиториев). Наиболее популярные из них: DSpace (http://www.dspace.org) (47,5% установок), EPrints (http://www.eprints.org) (13,8% установок), Fedora (http://www.fedoracommons.org), Greenstone, CDC Invenio.

DSpace – это самое популярное в академической среде ПО для создания архива электронных ресурсов (институционального цифрового репозитория) [7]. DSpace обеспечивает платформу для долгосрочного хранения цифровых материалов (изображения, медиафайлы, документы в различных форматах и т. п.), используемых в академических исследованиях. Платформа DSpace разрабатывалась совместно компанией Hewlett-Packard и библиотеками MIT (Massachusetts Institute of Technology). Движение Scholarly Communication оказало влияние на развитие DSpace, вследствие чего конфигурация по умолчанию направлена на поддержку научных публикаций [7, 8].

Для базовой организации данных в DSpace зафиксирована определенная модель данных, основанная на схеме Dublin Core [5] и ее расширениях. Система хранит (конвертирует) и индексирует метаданные в разнообразных форматах (DIM, MODS, METS, QDC, XOAI, MARC, RUSMARC, МЕКОФ, ORE и др.). Список форматов может быть расширен добавлением новых, в том числе собственной генерации, конвертеров. Поддерживается резервное копирование контента и система LOCKSS-compliant для организации надежного хранилища данных. Система хранит информацию о пользователях системы и поддерживает авторизацию и разграничивает доступ к содержимому репозитория. Кроме того, такие функции, как депонирование и редакторская проверка, привязаны к пользователям. DSpace работает со стандартными для библиотечной сферы протоколами OAI-PMH, OpenURL (ANSI/NISO Z39.88-2004 (R2010) – The OpenURL Framework for Context-Sensitive Services. National Information Standards Organization) и SWORD (http://swordapp.org/sword-v2/sword-v2-specifications).

С 2009 г. DSpace поддерживается сообществом DURASpace, которое образовано путем слияния двух проектов цифровых репозиториев DSpace Foundation и Fedora Commons.

Fedora (Flexible Extensible Digital Object Repository Architecture) – репозиторий разработан исследователями из Корнельского университета в 1997 г. в качестве платформы для хранения, управления и доступа к цифровому контенту (цифровым объектам) [9]. Открытое (лицензия Apache License, Version 2.0) Java приложение поддерживает резервное копирование контента и программу LOCKSS-compliant для обеспечения надежного хранения ресурсов. Fedora определяет набор абстракций для выражения цифровых объектов, отношения между цифровыми объектами, их связи и поведение. В отличие от DSpace, Fedora больше подходит для хранения произвольных цифровых объектов, например, ПО.

Ядро репозитория Fedora предоставляет набор веб-сервисов с четко определенными API. Кроме того, Fedora предоставляет широкий спектр вспомогательных сервисов и приложений, включая поиск, поддержку протоколов OAI-PMH и SWORD, обмен сообщениями, управление клиентами и многое другое.

Метаданные представлены в формате Dublin Core. Метаданные могут быть преобразованы в форматы METS, FOXML, Atom. Обеспечивается поддержка RDF.

EPrints – это вторая в академическом мире по популярности после DSpace система, которая используется для формирования и управления открытыми архивами и предназначена для поддержания институциональных репозиториев открытого доступа и создания архивов научных исследований с большим разнообразием ИР (научные статьи, отчеты, диссертации, монографии, учебно-методические пособия, материалы конференций, данные результатов экспериментов и наблюдений и т. п.). EPrints был разработан при университете Саутгемптона [10].

Модель данных EPrints отличается от DSpace, использующего строгую иерархическую систему организации данных, которая позволяет отразить структуру организации тем, что все записи эквивалентны и являются одноуровневыми.

Открытые архивы, созданные в среде EPrints, поддерживают протоколы обмена метаданными OAI-PMH, SWORD, которые обеспечивают глобальные услуги доступа и поиска.

Greenstone (http://www.greenstone.org) – свободно распространяемое ПО (выпускается под лицензией GNU General Public License) для создания и поддержания институциональных репозиториев открытого доступа (цифровых онлайн библиотек). Оно разработано в рамках Проекта новозеландской цифровой библиотеки при Университете Вайкато в сотрудничестве с ЮНЕСКО и Неправительственной организацией гуманитарной информации.

Основная схема данных Dublin Core с квалификаторами, основные форматы документов HTML и MS Word. Для программного доступа к ресурсам имеется собственный API, отличный от стандартных протоколов доступа к цифровым репозиториям, таких как OAI-PMH или SRU / SRW (http://www.loc.gov/standards/sru) (Search/Retrieve via URL / Search/Retrieve Web service), однако есть поддержка протокола Z39.50 [11]. Для организации просмотра материала предусмотрено использование внутренних классификаторов [12].

CDS Invenio (http://invenio-software.org) (ЦЕРН / CERN (http://cds.cern.ch), Швейцария) – интегрированная электронная библиотечная система, которая представляет собой набор приложений для построения и управления автономным сервером ЭБ. ПО бесплатное, распространяемое под лицензией GNU General Public License. Технология, предлагаемая данным продуктом, покрывает все аспекты поддержки ЭБ, поддерживает протокол OAI-PMH, использует формат MARC21 как основной библиографический стандарт. CDS Invenio является комплексным решением управления репозитариями документов средних и больших объемов.

Посредством CDS Invenio создан и поддерживается архив публикаций сервера документов CERN (CERN Document Server). В CERN CDS Invenio управляет более чем 500 коллекциями данных, состоящих из более чем 800 000 библиографических записей и 350 000 полнотекстовых документов, покрывая препринты, статьи, книги, журналы, фотографии, видеоматериалы и др. Помимо CERN, CDS Invenio в настоящее время инсталлирована и используется в 14 научных и образовательных учреждениях мира.

В качестве репозитория для разработанной нами ИС была выбрана система DSpace. Выбор обусловлен тем, что она является самой популярной в мире (рис. 2) и уже эксплуатируется в ИВТ СО РАН более 10 лет. Кроме того, система DSpace имеет ряд привлекательных возможностей, знание которых позволяет существенно повысить функциональность и интероперабельность ИС, использующих это ПО.

Для более полного соответствия локальным требованиям в базовую систему DSpace внесены многочисленные модификации (расширение схем данных, номенклатуры обменных форматов, возможность работы с географической информацией, авторитетный контроль и пр.).

В модернизированной системе доступ к данным репозитория возможен не только через web-интерфейсы DSpace, но и по протоколам OAI-PMH, SOLR (http://lucene.apache.org/solr), SRW / SRU, Z39.50. При этом поддержка SRW/SRU и Z39.50 обеспечивается связью DSpace с системой ZooSPACE [13]. Возможность доступа к репозиторию (поиск и извлечение данных) в обход графических веб-интерфейсов, которые являются неотъемлемой частью любой CMS, существенно расширяет его функциональные возможности, поскольку позволяет использовать содержимое репозитория другими приложениями и интегрировать его в информационное пространство.

Поддержка протоколов OAI-PMH и OAI-ORE (Open Archives Initiative – Object Reuse and Exchange) (http://www.openarchives.org/ore/1.0/toc) позволяет разрабатывать собственные ИС, взаимодействующие с хранилищем данных, построенном на основе DSpace. В качестве примера приведем электронные версии журнала «Вестник НГУ. Серия информационные технологии» (http://jit.nsu.ru/index.php?+ru_RU) (рис. 2) и «Дайджеста прессы по проблемам российской науки» (http://www.prometeus.nsc.ru/science/digest) (рис. 3).

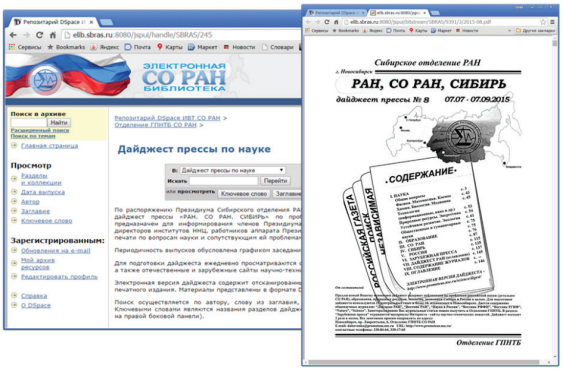

Рис. 2. Вывод из DSpace страниц журнала

Рис. 3. Дайджест прессы по проблемам российской науки

В заключение следует заметить, что свободно распространяемое программное обеспечение DSpace позволяет не только создавать и эксплуатировать цифровые репозитории с разнородным контентом, но и обеспечивает созданным репозиториям эффективные механизмы интеграции с другими ИС. Последнее делает DSpace полноправной компонентой распределенной гетерогенной информационной системы.

Список литературы

1. Эволюция информационных систем: от web-сайтов до систем управления информационными ресурсами / Ю. И. Шокин [и др.] // Вестн. Новосиб. гос. ун-т. Сер.: Информ. технологии. 2015. Т. 13, № 1. С. 117–134.

2. *Федотова О. А.* Требования к информационной модели электронной библиотеки по научному наследию // Zbornik radova konferencije MIT 2013. Beograd, 2014. С. 141–149.

3. Модель информационной системы для поддержки научно-педагогической деятельности / А. М. Федотов [и др.] // Вестн. Новосиб. гос. ун-т. Сер.: Информ. технологии. 2014. Т. 12, № 1. С. 89–101.

4. ISO-14721 Reference model for an Open archival information system (OAIS), draft recommended standard, CCSDS650.0-P-1.1.2012. URL: https://public.ccsds.org/pubs/650x0m2.pdf (дата обращения: 15.08.2018).

5. DCMI – Dublin core metadata initiative. URL: http://www.dublincore.org/ (дата обращения: 15.08.2018).

6. The open archives initiative protocol for metadata harvesting: protocol version 2.0 of 2002.06.14. 2004. URL: http://www.openarchives.org (дата обращения: 15.08.2018).

7. DSpace: an open source solution for accessing, managing and preserving scholarly works // MIT Libr. 2007. URL: http://www.dspace.org (дата обращения: 15.08.2018).

8. *Кудим К. А., Проскудина Г. Ю., Резниченко В. А.* Создание научных электронных библиотек с помощью системы DSpace // Проблемы программирования. 2007. № 3. С. 49–60.

9. Introducing Pergamos. A Fedora-based DL system utilizing digital object prototypes / G. Pyrounakis // In Research and Advanced Technology for Digital Libraries. 2006. P. 500–503.

10. A study on the Open Source Digital Library Software: Special Reference to DSpace, EPrints and Greenstone / S. Tramboo [et. al.] // Intern. J. of Computer Applications. 2012. № 59 (16). P. 1–9.

11. ANSI/NISO Z39.50–2003. Information retrieval (Z39.50): Application service definition and protocol specification. Bethesda : NISO Press. 2002.

12. *Резниченко В. А., Проскудина Г. Ю., Овдий О. М.* Создание цифровой библиотеки коллекций периодических изданий на основе Greenstone // Электрон. б-ки. 2005. 8. Вып. 6. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part6 (дата обращения: 15.08.2018).

13. *Жижимов О. Л., Федотов А. М., Шокин Ю. И.* Платформа ZooSPACE – организация доступа к разнородным распределенным ресурсам // Электрон. б-ки. 2014. Т. 17. № 2.