Управление технологией наполнения электронной библиотеки «Научное наследие России»

Н.Е. Каленов
Библиотека по естественным наукам РАН
nek@benran.ru

Аннотация

В докладе рассматриваются организационная структура управления проектом создания ЭБ «Научное наследие России», основные решения по организации процесса распределенного наполнения метаданными ЭБ, контролю технологической дисциплины, получению динамической статистической информации о ходе реализации проекта.

Начиная с 2007 г., в рамках целевой научной программы РАН разрабатывается интегрированная Электронная библиотека «Научное наследие России» (ЭБ ННР), представленная в открытом доступе в Интернет (http://e-heritage.ru) [2–4]. Основной целью создания Библиотеки является информирование пользователей о выдающихся ученых, внесших вклад в развитие российской науки, и их научных достижениях.

В ЭБ ННР отражаются биографические данные об ученых, наиболее значительные их публикации (библиография и отсканированные полные тексты), архивная и музейная информация, относящаяся к деятельности ученого.

В основу функционирования ЭБ ННР положен принцип распределенной подготовки данных с централизованной редакторской обработкой, загрузкой и поддержкой контента.

Головным исполнителем работ по созданию ЭБ является Межведомственный суперкомпьютерный центр (МСЦ) РАН, разработчиками технологии и программного обеспечения – МСЦ РАН, Вычислительный центр им. А.А. Дородницына (ВЦ) РАН, Библиотека по естественным наукам (БЕН) РАН.

В число участников проекта, обеспечивающих подготовки контента для ЭБ ННР, входит достаточно большое количество академических организаций – центральные библиотеки (БАН, БЕН, ИНИОН, ЦНБ УрО РАН), использующие как собственные фонды, так и фонды своих отделений в академических институтах; Архив РАН со своими филиалами Государственный геологический музей им. В. И. Вернадского (ГГМ) РАН и ряд институтов Москвы и Санкт-петербурга.

Наряду с поставщиками контента организационная структура ЭБ как действующей автоматизированной системы включает Совет Системы, в состав которого входят авторитетные представители организаций – участников; административную группу, редакторскую группу, группу технической поддержки.

Совет Системы определяет принципы формирования контента и предоставления его пользователям, основные направления развития ЭБ, решает вопросы привлечения к созданию Библиотеки новых организаций.

Административная группа, представляющая собой временный научный коллектив в составе МСЦ РАН, осуществляет общее руководство технологическими процессами наполнения ЭБ, рассматривает планы и отчеты участников проекта.

Редакторская группа, состоящая из штатных сотрудников МСЦ РАН, принимает окончательное решение о включении в ЭБ тех или иных изданий, осуществляет «выходной контроль» подготовленной информации, включающий проверку правильности метаданных и качества (постранично) каждого оцифрованного издания, подготавливает прошедшие контроль издания для загрузки в программную оболочку демонстрационной части ЭБ, доступную пользователям по адресу http://e-heritage.ru.

Группа технической поддержки, включающая сотрудников ВЦ РАН и БЕН РАН, обеспечивает сохранность информации, работоспособность программных и технических средств, поддерживающих все элементы ЭБ.

В задачи поставщиков контента входит отбор материалов в соответствии с установленными принципами, формирование метаданных о принятых к включению в ЭБ объектах (персоналии, издания, архивные документы, музейные предметы, фотографии, мультимедийные материалы), оцифровка изданий и обработка информации в соответствии с правилами системы (для ЭБ ННР принято решение, согласно которому отсканированный текст не распознается, за

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

исключением оглавления, по которому обеспечивается навигация внутри издания), передача обработанных материалов в редакторскую группу.

Принцип распределенной подготовки данных требует разработки специальных средств, обеспечивающих исключение возможности дублирования материалов (когда разные организации сканируют одни и те же издания или вводят данные об одном и том же ученом), достаточно жесткий контроль соблюдения «технологической дисциплины» и оперативное исправление ошибок.

Специфика технологии создания ЭБ ННР еще более обостряет эти проблемы. Это обусловлено тем, что, во-первых, в качестве участников, формирующих контент, выступают не только разнородные организации, но и представители различных регионов (в настоящее время - Москва, Санкт-Петербург, Екатеринбург, Борок), а, во-вторых, тем, что подготовка информации для отражения в ЭБ включает ряд растянутых во времени процессов, в которые вовлечены различные службы.

Технологические процессы формирования контента ЭБ ННР на примере подготовки изданий (в предположении, что данные об ученом уже введены в систему) представлены в табл. 1.

Таблица 1

Номер процесса	Наименование процесса	Организационная структура ЭБ
1	Библиографический поиск изданий конкретного автора	Поставщик контента (библиограф)
2	Формирование предложений для ввода изданий в ЭБ, передача их в редакторскую группу	Поставщик контента (библиограф)
3	Рассмотрение предложений, информирование Поставщика об их принятии или отклонении	Редакционная группа
4	Заказ и получение из фондов библиотеки принятого к сканированию издания	Поставщик контента (библиограф)
5	Формирование развернутых метаданных об издании	Поставщик контента (библиограф)
6	Сканирование издания и обработка изображений по принятой в ЭБ ННР технологии	Поставщик контента (технолог)
7	Обработка оглавления издания. связывание его с отсканированным текстом, передача материалов в редакторскую группу	Поставщик контента (технолог)
8	Контроль качества метаданных и отсканированных страниц, информирование Поставщика об ошибках или принятии издания	Редакционная группа
9	[Исправление обнаруженных ошибок]	Поставщик контента (библиограф, технолог)
10	Формирование издания для отражения в ЭБ по принятой технологии; загрузка метаданных, оглавления и полного текста издания в ЭБ	Группа технической поддержки

Для реализации перечисленных процессов применительно ко всем объектам, отражаемым в ЭБ ННР, а также для управления процессами децентрализованной подготовки данных специалистами БЕН РАН разработана специальная система, основой которой является настраиваемый программный комплекс SCIRUS [5, 6], обеспечивающий ввод, поиск и экспорт данных, описывающих взаимосвязанные сущности, характеризуемые различным набором полей (детальное описание последней версии комплекса приведено в докладе М.М.Якшина, представленном на данной конференции).

Реализация SCIRUS, настроенная на диспетчеризацию технологии создания ЭБ (http://meta.e-heritage.ru/), поддерживает технологическую базу данных (ТБД), отражающую свойства 5 взаимосвязанных сущностей, среди которых: «публикация / документ», описываемая 24 полями метаданных; «персона», описываемая 11 полями; «организация», описываемая 6 полями; «источник» (метаданные издания на сводном уровне – сборник, журнал и т.п., в который может входить публикация), описываемый 6 полями; «файл», описываемый 2 полями (название и URL).

Функции SCIRUS, связанные с настройкой системы, вводом и корректировкой данных доступны только авторизованным пользователям в зависимости от их прав. Настроенных администратором. Однако для поиска и просмотра данных система открыта для пользователей, входящих без пароля под именем guest.

Поиск в системе возможен по любым полям всех сущностей и их булевым комбинациям.

При формировании запроса пользователь задает сущность, которая должна быть выдана в результате обработки запроса. Результатом поиска является заданная сущность, все связанные с которой сущности удовлетворяют сформулированному запросу. Например, можно получить список

публикаций, авторы которых родились в 18 веке, или список ученых в области математики, которые публиковали свои работы на французском языке (естественно. речь идет о материалах, введенных в ЭБ). Система обеспечивает навигацию по связанным записям. Например, получив в результате обработки запроса список публикаций, можно выбрать автора одной из них и получить список всех его публикаций; аналогично от списка авторов можно перейти к списку их работ.

Результаты поиска могут быть отсортированы по любому полю выдаваемой сущности.

Метаданные персоны содержат «идентифицирующие» сведения об ученом (фамилия, имя, отчество, годы жизни и т.п.), перечень научных направлений, в которых он работал, а также его развернутую биографию.

Метаданные публикации включают, наряду с элементами библиографического описания, поля, характеризующие этапы технологической обработки издания – текстовые поля «Отсутствующие страницы» и «Комментарии к проблеме» и списковые поля «Статус записи» и «Статус приемки». Значения и функции последних двух полей используются для задач диспетчеризации технологических процессов, они достаточно подробно описываются в докладе, представленном М.М.Якшиным.

Система поддержки технологии подготовки метаданных устроена так, что каждый участник может, войдя в нее по своему паролю, осуществлять поиск и просмотр всей информации, введенной в систему; вводить новые данные и редактировать старые, но только введенные под его именем. При поиске может быть задано ограничение — выбирать записи, введенные данным пользователем (в меню предлагается список имен, зарегистрированных в системе). Пользователь, наделенный правами администратора, может редактировать информацию, введенную любым участником системы.

Распределенная технология подготовки данных для ЭБ организована следующим образом (рис. 1).

Руководствуясь согласованными подходами к принципам наполнения ЭБ, каждая организация-участник Программы, определяет издания из своих фондов, которые она считает целесообразным включить в ЭБ. После этого зарегистрированный представитель этой организации входит в систему диспетчеризации и проверяет, не зарегистрирована ли уже в ней данная публикация и ее автор(ы). Если в системе публикация отсутствует, он вводит в предлагаемый шаблон ее метаданные и сведения об авторе (если автор не был введен ранее). При этом поле шаблона “Текущий статус” принимает значение “предложено к оцифровке”.

Представитель редакторской группы периодически входит в систему диспетчеризации и получает список документов, имеющих статус “Предложено к оцифровке”. Редакторская группа принимает решение по каждому из них о целесообразности ввода в ЭБ. Если документ подлежит оцифровке, значение поля “Текущий статус” меняется на “зарегистрировано”, и в запись автоматически вводится номер данного документа, под которым он будет введен в ЭБ. В дальнейшем этот номер (поле «Номер МСЦ») изменению не подлежит. Если по какой-либо причине документ оцифровывать нецелесообразно, значение поля “Текущий статус” меняется на “Оцифровке не подлежит”.

Представитель организации, формирующей контент ЭБ, входит в систему диспетчеризации и выбирает свои записи, имеющие текущий статус “зарегистрировано”. После подбора изданий и отправки на оцифровку их текущий статус меняется — в это поле вводится значение “в работе”. После завершения процесса оцифровки статус записей меняется на “оцифровано”, после передачи в редакторскую группу МСЦ — на “сдано”.

Таким образом, в каждый момент времени административная группа ЭБ может получить сведения, сколько и каких изданий находится в работе, сколько и кем оцифровано и т.п.

Технологическое поле «Статус приемки» заполняется сотрудниками редакторской группы, работающими с данным изданием. При получении электронной копии издания в это поле вводится значение «принято к работе». Если замечаний по материалу нет, он передается в техническую группу, и значение поля меняется на «принято к загрузке на сайт», после загрузки поле принимает значение «опубликовано». Если в материале обнаружены ошибки, редактор присваивает полю «Статус приемки» значение «обнаружены проблемы», заполняет поля «Отсутствующие страницы» и «Комментарии к проблеме» и направляет исполнителю соответствующее сообщение. Исполнитель исправляет ошибки, соответственно меняя значение поля «Статус приемки», после чего редактор передает издание для загрузки в ЭБ.

Таким образом контролируются ход и сроки исправления ошибок.

Загрузка метаданных о персонах и публикациях из технологической системы на сервер ЭБ осуществляется автоматически при загрузке электронного издания. Синхронизация данных осуществляется по значению поля «Номер МСЦ», при этом используется специальный формат, базирующийся на XML и RDFS, принятый для системы ЕНИП РАН [1]. Экспорт данных в этом формате возможен и путем вызова опции “Экспорт в формате ВЦ РАН” непосредственно со страницы результатов поиска данных системы SCIRUS. В этом случае перед вызовом опции необходимо отметить записи, подлежащие выгрузке.

Для наглядности схема выполнения технологических процессов создания и включения публикации в ЭБ представлена на рис. 1.

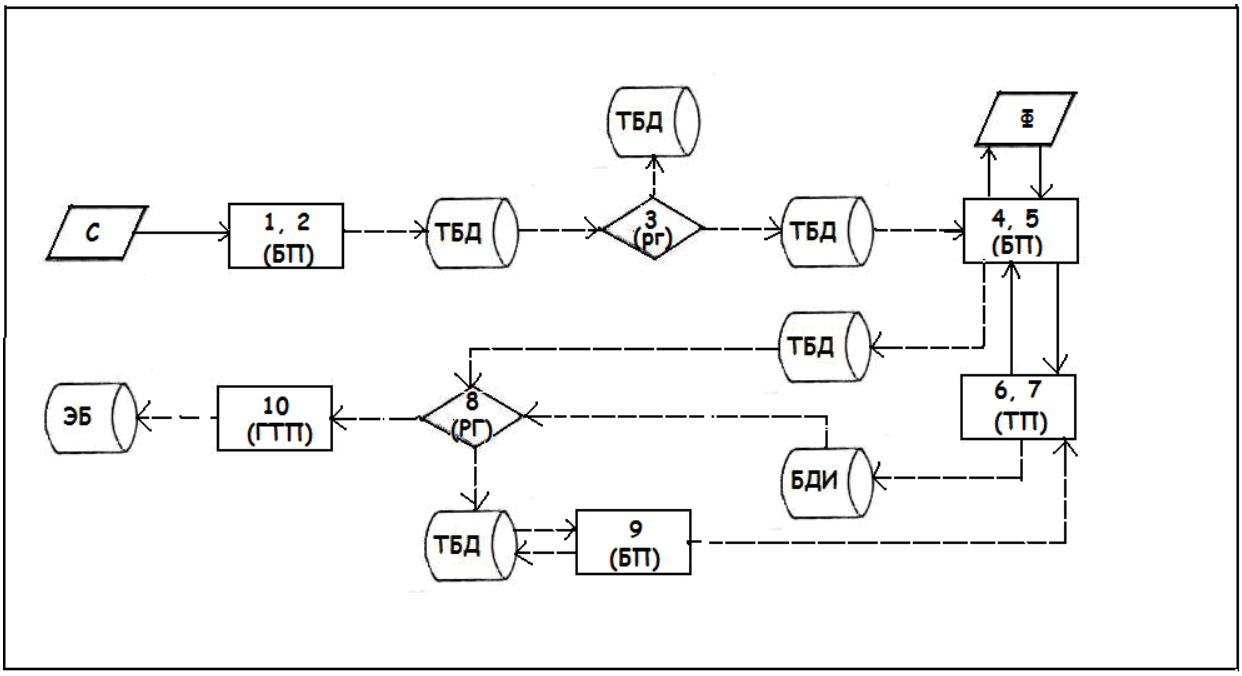

Рис. 1

Здесь приняты следующие обозначения.

Цифры внутри блоков соответствуют номерам процессов, приведенных в таблице 1.

«С» – внешние справочно-информационные ресурсы, которыми пользуются поставщики контента при поиске изданий (каталоги библиотек, библиографические издания и т.п.);

«Ф» – фонды библиотек, где хранятся печатные издания;

«БП» – библиографы поставщика;

«ТБД» – технологическая база данных;

«РГ» – редакторская группа»;

«ТП» – технологи поставщика, обеспечивающие сканирование и обработку изображений;

«БДИ» – база данных изображений (отсканированных страниц);

«ГТП» – группа технической поддержки.

Сплошные линии означают перемещение материальных объектов (издания, включаемые в ЭБ, библиографические материалы, выписки из каталогов), пунктирные – ввод данных в компьютер.

Наиболее узкое место в технологии (и это можно заметить на схеме) связано с возможностью обнаружения на этапе 8 ошибок, связанных со сканированием (в первую очередь, пропущенные страницы). В этом случае приходится снова заказывать книгу (возвращаться к этапу 4) в фондах библиотеки и направлять ее на сканирование. Эта проблема встает достаточно остро в случае, когда сканируются издания, находящиеся в удаленных хранилищах, что характерно для БЕН РАН, фонды которой распределены по 60 отделам в академических организациях московского региона. Поэтому сканирование изданий (процесс 6) требует очень тщательного контроля и должно выполняться высококвалифицированными специалистами.

В настоящее время в технологическую базу данных ЭБ «Научное наследие России» загружены данные о более чем 5000 ученых и более чем 18000 их публикаций. По многим ученым в ЭБ введены оцифрованные портреты, архивные материалы. В ЭБ также введены описания ряда, коллекций ГГМ им. В.И. Вернадского РАН, метаданные отдельных входящих в них экспонатов, ссылки на ученых, связанных с этими экспонатами.

Основное направление развития проекта связано с модификацией пользовательского интерфейса ЭБ, интеграцией с другими отечественными и зарубежными электронными библиотеками, включением в контент мультимедиа.

Литература

[1] Бездушный А.А., Бездушный А.Н., Серебряков В.А., Филиппов В.И. Интеграция метаданных Единого Научного Информационного Пространства РАН // Материалы Международной : Монография / ВЦ РАН. – М., 2005. – С. 238.

[2] Каленов Н.Е., Савин Г.И., Сотников А.Н. Электронная библиотека «Научное наследие России» // Информационные ресурсы России. – 2009. – № 2. – С. 19–20.

[3] Каленов Н.Е., Савин Г.И., Серебряков В.А., Сотников А.Н. Принципы построения и формирования электронной библиотеки «Научное наследие России» // Программные продукты и системы. – 2012. – № 4. – С. 30–40.

[4] Каленов Н.Е., Сотников А.Н., Ильина И.Н. Архивная информация в электронной библиотеке «Научное наследие России» // Фундаментальная наука: проблемы изучения, сохранения и реставрации документального наследия научной

конференции / отв. ред. В.Ю. Афиани. – М.: Архив РАН, 2013. – С. 25–35.

[5] Сенько А.М. Информационная система SciRus: принципы построения и перспективы развития // Научный серис в сети ИНТЕРНЕТ: технологии параллельного программирования. Всероссийская науч. конф., Новороссийск, 18–23 сент. 2006. – М, 2006. – С. 58–59.

[6] Якшин М.М. WEB-интерфейс системы «Наука России» // Современные технологии в информационном обеспечении науки: Сб. науч. тр. / под ред. Н.Е.Каленова. – М., 2003. – С. 47–52.

Management of the Technology for Filling the Digital Library Entitled “The Scientific Heritage of Russia”

Nikolay E. Kalenov

The main decisions on the management of the Digital Library project “Scientific Heritage of Russia”, on the processes of filling of distributed metadata for the DL, technological discipline of control of the DL filling are considered.