Работа с обязательным экземпляром сетевых публикаций в Германии

Автор: Рената Гёмпель, Ларс Свенссон

УДК 021.843

Освещены технология Немецкой национальной библиотеки по работе с сетевыми ресурсами и использование постоянных идентификаторов для обеспечения их долговременной доступности.

Доклад на совместном заседании "Обязательное депонирование электронных документов: от законодательства к реализации, от комплектования к доступу" Секции библиографии, Союза по выработке электронных стратегий Конференции директоров национальных библиотек. Секции информационных технологий и Секции "Управление знаниями" в ходе 77-й Генеральной конференции ИФЛА (13 - 18 авг. 2011 г., Сан-Хуан, Пуэрто Рико).

Публикуется с одобрения аппарата ИФЛА.

Ключевые слова: Немецкая национальная библиотека, электронные документы, сетевые публикации, комплектование, каталогизация, индексирование, метаданные, архивирование, цитирование.

В 2006 г. вступил в действие новый закон о Немецкой национальной библиотеке (*Deutsche Nationalbibliothek* - DNB). Наиболее заметное в нём изменение по сравнению с предыдущим законом - введение положения об обязательном депонировании электронных документов: комплектование, каталогизация, индексирование и архивирование ресурсов.

DNB создала полностью автоматизированную технологию комплектования и обработки всех видов сетевых материалов. С использованием этой технологии можно обрабатывать электронные книги и диссертации, журналы и газеты; сейчас мы работаем над архивированием веб-сайтов.

Создание технологии включало в себя работу с форматами метаданных, поступающих из различных источников (например от издателей, вузовского сообщества, отдельных лиц), а также с различными форматами

данных, которые придется архивировать. Необходимое внимание при этом уделено авторским правам.

Разработанная технология предоставляет широкий набор вариантов доставки сетевых публикаций: через сетевую форму (для небольшого количества публикаций); через OAI PMH - *(Open Archive Initiative Protocol for Metadata Harvesting)* (используется, например, для издательств и университетов); через WebDAV или .ftp "в горячую папку" (предназначен в основном для коммерческих издательств).

Во всех трех случаях метаданные поставляются создателем или издателем документа и передаются в каталог без интеллектуального вмешательства. Как только ресурс передан в репозитарий и архивирован, его заглавие можно увидеть в каталоге, а саму публикацию - прочитать в читальном зале. Все вновь поступившие сетевые публикации отражаются в Немецкой национальной библиографии - в особой серии, известной как *O серия.*

Долговременная сохранность гарантируется, но пока что не все типы форматов могут сохраняться. DNB работает над расширением диапазона принимаемых к обработке метаданных и форматов документов с тем, чтобы соответствовать современным требованиям.

Для обеспечения надежного доступа и считывания сохраняемых цифровых ресурсов, мы присваиваем постоянный идентификатор всем сетевым публикациям, поступающим в библиотеку по линии обязательного экземпляра. Использование идентификатора гарантирует возможность однозначного распознавания и доступа к ресурсу, даже если он изменит свое местоположение.

Обязанностями являются сбор, обработка и архивирование всех немецких и немецкоязычных документов (начиная с 1913 г.), обеспечение их долговременной сохранности и предоставление к ним доступа широкой публике. С момента введения в действие Закона о Немецкой национальной библиотеке (от 22 июня 2006 г.) эти обязанности, или задачи, распространяются и на произведения "в нематериальной форме" - сетевые публикации.

Эта миссия не только отражена в законе, но и уточнена в соответствующих документах, где обозначено, как и когда применяется требование об обязательном экземпляре.

Сетевые публикации - это всё, что представлено и доступно в публичных сетях в текстовом и иных форматах; электронные книги и журналы, презентации, музыкальные файлы, веб-сайты и др. Описания тех электронных продуктов, которые включаются в собрание сетевых публикаций.

можно найти в пересмотренной версии Положения об обязательном экземпляре *(Legal Deposit Regulation)* и в Руководстве по формированию коллекции (http://files.d-nb.de/pdf/sammelrichtlinien.pdf; доступно только на

немецком языке).

Любой создатель сетевых публикаций (включая коммерческие организации и частных лиц) по закону обязан предоставлять свои публикации в DNB.

До настоящего времени не все участники были в состоянии передавать файлы и метаданные. Следовательно, необходимо было найти простой способ передачи документов и одновременно обеспечить для DNB возможность автоматически обрабатывать основную часть огромного массива электронных публикаций.

Разработка автоматизированной технологии - поэтапная реализация

DNB начала разрабатывать технологию, цель которой - создание автоматизированного процесса передачи сетевых публикаций, их отражение в каталоге и архивирование файлов. Для этого была выбрана поэтапная процедура, позволяющая производить тестирование и совершенствование технологии в ходе реальных операций с небольшими объемами данных.

В качестве первого шага была запланирована регистрация репозитариев сетевых публикаций с учетом всех процедур и интерфейсов. Записанные данные (имя, адрес, электронная почта) могут быть показаны только самому поставщику и сотруднику DNB, работающему с сетевыми публикациями. Процесс регистрации учитывает случаи, когда сервис-провайдеры, предоставляющие электронную среду для издателей, депонируют материалы третьей стороны.

Сетевые форматы были первым вариантом, который мы использовали для получения различных типов сетевых публикаций. Новые форматы основаны на тех, которые в свое время использовались для монографий (электронных книг), и на формате, применявшемся начиная с 2006 г. для добровольного депонирования диссертаций.

Новый формат допускает простую передачу сетевых публикаций: формат для монографий может использоваться для электронных книг, сетевых диссертаций и музыкальных произведений. Однако обязательными для различных типов публикаций являются лишь очень немногие поля (наименование, дата и адрес публикации); в зависимости от типа издания могут добавляться еще какие-либо поля (например, информация о диссертации или специфические идентификаторы, такие, как ISMN - для музыкальных произведений).

Следующие форматы дают возможность передавать наименования электронных периодических изданий и сами издания. Все форматы связаны между собой: наименования регистрируются один раз, и когда издатель периодики передает последующий документ, ему показывают список

периодических изданий, и большинство полей для отражения издания заполняется автоматически.

Сетевые форматы идеальны для передачи небольшого числа публикаций, поскольку представление метаданных производится вручную. Кроме того, имеются ограничения по объему отправки отдельной единицы (50 мегабайт) и для представления через URL (500 мегабайт).

Возможность комплектования через интерфейс. Следующим шагом было формирование интерфейса, через который публикации становятся доступными организации-репозитарию и с которого материал извлекает DNB.

В этом сервисе используется протокол на основе HTTP, разработанный в рамках проекта *Open Archive Initiative - OAI-PMH.* При пользовании этим протоколом клиент или специалист, занимающийся подбором информации, запрашивает данные, отправляя HTTP-запрос GET-requests на сервер или в репозитарий. Метаданные извлекаются из сервера депонирующей организации и передаются в DNB посредством полностью автоматизированного процесса, который исключает "ручное вмешательство" со стороны любого из участников.

Затем в метаданные включается адрес передачи *(transfer URL),* с помощью которого публикация также обнаруживается автоматически. Метаданные вносятся в каталог DNB без транспортного URL, а файлы добавляются в репозитарий. (Более подробную информацию - только на немецком языке - по установке интерфейса OAI можно найти по адресу http://www.d-nb.de/netzpub/ablief/pdf/automatisierte_ablieferung.pdf.) После тестирования этот процесс осуществляется автоматически с обеих сторон и пригоден для передачи большого количества файлов.

Депонирование через "горячую папку". Этот дополнительный интерфейс создан в апреле 2011 г. Технология "горячей папки" *(hotfolder)* подходит для передачи больших объемов данных. Папка называется "горячей", поскольку каждый шаг происходящих в ней процессов контролируется другим процессом. Депонирующая организация пересылает материал в эту папку, которая непрерывно просматривается. После регистрации аккаунта депонирующей организацией документ содержится в архивном контейнере Zip-Container вместе с метаданными. Подходящие методы передачи контейнера - FTP *(File Transfer Protocol)* или WebDAV. Посредством

автоматизированной процедуры метаданные интегрируются в каталог, а файлы архивируются в репозитарии.

Технология "горячей папки" требует от поставщика активных действий по представлению публикаций и данных, и тем не менее этот интерфейс охотно используется издателями, поскольку они уже знакомы с вариантами передачи

данных (такими, как FTP).

Использование форматов данных. Сетевые публикации принимаются в том формате, в котором они созданы. Важно упомянуть, что пригодная для передачи сетевая публикация должна представлять собой независимую логическую единицу, которую можно отделить от ее среды. Она не должна быть зависимой от внутренней связи в сервере, когда значительная часть документа извлекается специально для этого случая динамическим образом из системы хранилища данных в момент взаимодействия с пользователем.

В настоящее время в дополнение ко всем разновидностям формата PDF (PDF/A и все другие типы PDF) могут также автоматически архивироваться документы в форматах EPUB и HTML. Все форматы данных должны передаваться без кодировки, чтобы гарантировать возможность долговременного обращения к документам с минимальными усилиями.

Использование форматов метаданных. При использовании метода сетевой формы метаданные вводит поставщик документов. Обязательного стандарта на этот процесс нет - должны лишь выполняться требования, относящиеся к самому формату. Метаданные (из полей, заполненных поставщиком) извлекаются и передаются в соответствующие поля внутреннего формата библиотечного каталога.

При применении двух других методов депонирования метаданные для реализации автоматизированной технологии должны предоставляться в точном и заранее согласованном стандарте. В настоящее время приемлемым форматом является ONIX (для книг). MARC-XML или XMetaDissPlus; со временем будут добавлены и другие форматы.

Для максимального упрощения процесса депонирования установлены минимальные требования к элементам метаданных. (С деталями можно ознакомиться по адресу http://www.d-nb.de/netzpub/ablief/pdf/metadaten_kernset_definitionen.pdf; только на немецком языке.)

Как обрабатываются сетевые публикации?

Метаданные интегрируются в каталог в том виде, в каком они поступили. Когда метаданные загружены, ссылок к авторитетным файлам не существует, заполнено только несколько обязательных полей. Однако со-

держание полей не проверяется, поскольку невозможно обработать вручную огромное количество записей. В рамках проекта, выполняемого DNB. разработан сценарий, который улучшает содержание записей посредством автоматизированной описательной каталогизации и предметной индексации:

представленные в записях имена (автор, редактор, переводчик и др.) проверяются по авторитетным файлам персональных имен и сопоставляются с авторитетными записями;

новая запись о сетевой публикации проверяется на наличие параллельного печатного издания и, если таковое существует, связывается с ним, и на этой стадии информация именно из печатного издания вносится в запись о сетевой публикации. Сетевая публикация автоматически индексируется - проставляются индексы Десятичной классификации Дьюи и предметные рубрики.

Метаданные по представленным или архивированным сетевым публикациям доступны бесплатно для просмотра в каталоге DNB. Однако управление правами на цифровые документы представляет определенные сложности. В процессе передачи документа организация-поставщик может указать, какие права на распространение документов передаются DNB.

Диапазон прав достаточно широк - от доступа только в читальных залах библиотеки, доступа через Интернет для зарегистрированных пользователей (они могут работать с документом, не выходя из дома) и до глобального неограниченного доступа для любого пользователя. Коммерческие издатели обычно предлагают платный доступ к документам, находящимся на их веб-сайте, и ограничивают доступ в библиотеке: только в её помещении. В тех случаях, когда документ доступен только в читальных залах, пользователь не имеет права делать электронные копии документов или пересылать их по электронной почте. (Информацию, касающуюся законодательных основ авторского права, можно найти по адресу http://undesrecht.juris.de/dnbg/index.html и http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:EN:HTML)

Немецкая национальная библиография серии О *(O-series).* Сразу же после того как документ депонирован, его метаданные включаются в каталог библиотеки. Затем все поступившие сетевые публикации отражаются (анонсируются) в списке, который называется *серия О* (онлайновые публикации) Немецкой национальной библиографии. Один раз в месяц список поступивших записей компилируется, и этот файл может быть бесплатно выгружен. Записи доступны в обменных форматах MAB и MARC21. В отличие от других частей Немецкой национальной библиографии, этот файл не распространяется в формате PDF.

Возможность цитирования и архивирования. Все электронные публикации, депонированные в DNB. не только индексируются в системе Национальной библиографии - им также присваивается так называемый постоянный идентификатор *(Persistent Identifier - PI).*

Этот уникальный идентификатор используется во всем мире для объектов,

которым можно присвоить адрес, таким, как документы, изображения, звукозаписи, анимация или описание метаданных. В случае обращения через PI. активируется инструмент, известный как определитель *(resolver),* связывающий имя и адрес цифрового объекта. Задачей определителя является укрепление связи между именем и адресом. При смене адреса данные в определителе обновляются так, что всегда можно выйти на нужный адрес. Таким образом, ссылка остается стабильной. Разделение идентификации объекта и его местоположения посредством определенной системы символов - это фундаментальный принцип PI.

DNB ввела PI в ходе работы над проектом Эпикур *(Epicur project 2002 - 2005);* для PI в качестве идентификаторов используется единый список имен (*Uniform resource Names, urn*) из списка имен Национальной библиографии *(National Bibliography Number).* Список имен разработан специально для идентификации библиографических ресурсов. Для обеспечения возможности децентрализованного администрирования имена географически структурированы; DNB "отвечает" за Германию и. следовательно, её часть пространства имен обозначается urn:nbn:de.

Для конверсии постоянных идентификаторов в адреса доступа DNB использует определитель, расположенный по адресу http://nbn-resolving.org/. Этот определитель использует не только имена из списка, выделенного для Германии (urn:nbn:de), но также и из списков для Швейцарии и Австрии (urn:nbn:ch; urn:nbn:at). Кроме того, имеется возможность переправить запрос о постоянных идентификаторах, с которыми DNB не работает. Определитель пересылает запросы на urn:nbns Чешской Республики. Финляндии. Венгрии. Нидерландов. Норвегии и Швеции к соответствующим национальным определителям, а также формулирует запросы о постоянных идентификаторах DOI (*Digital Object Identifier* идентификатор цифрового объекта). Handle и Ark (*Archival Resource Key* - ключ архивных ресурсов), находящихся у различных сервис-провайдеров. Этот определитель разработан как часть системы *EuropeanaConnect* и входит в инфраструктуру *Europeana* (http://europeana.eu).

Определитель DNB содержит не только постоянные идентификаторы для депонирования электронных документов. Любой создатель цифрового документа может присвоить постоянные связи к созданным объектам. Например, в мае 2011 г. этой услугой пользовались более 400 учреждений,

которые зарегистрировали около 5 млн. urn. В день поступает примерно 3 500 запросов. Количество посещений все еще относительно невелико, но будет расти по мере расширения использования постоянных идентификаторов.

DNB архивирует все депонированные электронные объекты в репозитарий, который предоставляет доступ к рабочим копиям документов: это может быть копия оригинальной публикации или (особенно, когда речь идет об оцифрованных объектах) некое производное. Когда сетевая публикация депонируется в репозитарий, автоматически генерируется набор данных для

обеспечения долговременной сохранности, который передается в отдельную архивную систему, контролирующую долговременную сохранность объекта.

Комбинация процессов обеспечения сохранности и формирования системы постоянных идентификаторов обладает рядом преимуществ. С помощью PI ученые могут делать ссылки к цифровым объектам и при этом быть уверенными, что эти ссылки и в долгосрочной перспективе останутся стабильными и другие исследователи смогут обращаться к цитированным ресурсам и уточнять информацию без дополнительных усилий.

Издатели или другие поставщики документов могут запросить PI у DNB с тем, чтобы включить его в документ до публикации, так что PI будет доступен и в печатной версии документа. Использование определителя позволяет также издателям или иным хранителям данных содержать тождественные объекты в различных местах и затем запрограммировать определитель, указав, какой адрес имеет высший приоритет. Например, издатель может использовать один и тот же urn для обозначения документа, доступного как в сетевом магазине издателя, так и в библиотечном репозитарий. Когда будет задействован определитель, наивысший приоритет получит адрес объекта в сетевом магазине, и пользователя направят к издателю (для совершения покупки). Если же издатель по какой-либо причине выйдет из бизнеса или просто не будет сохранять активность ссылки, то откроется адрес электронного архива библиотеки.

К проблеме долговременного обеспечения сохранности электронных документов следует относиться очень серьезно, гарантированная защита данных от изменений или помех в течение как можно более длительного времени - весьма существенный фактор.

В настоящее время DNB поддерживает электронный архив, который основан на программном обеспечении DIAS, созданном в рамках проекта KOPAL совместно DNB, библиотекой Университета Геттингена и компанией IBM Germany. Уже на этой стадии стало очевидно, что классическое

программное обеспечение долговременного архивирования, подобное DIAS, необходимо будет заменить программами следующего поколения с многоуровневой инфраструктурой архивирования (может быть, даже в международном масштабе), в котором документы депонировались бы во многих связанных между собой архивах.

DNB участвует в проекте Европейского Союза "Устойчивый доступ к культурному наследию посредством мультивалентного архивирования" *(SHAMAN, Sustaining Heritage Access through Multivalent Archiving),* направленном на разработку технических и концептуальных основ нового поколения, создание связанных между собой систем и обеспечение связи между архивами долговременного хранения с помощью GRID - технологий с тем, чтобы решить сложную и трудоемкую проблему сохранности

электронных документов.

Для того чтобы удовлетворить будущие потребности в гибкой и масштабируемой инфраструктуре информационных технологий, DNB осуществляет модернизацию своего центра обработки данных с целью увеличить емкости хранилища электронного репозитария и архивов долговременного хранения, а также подключить дополнительные сервисы.

Обеспечение долговременной доступности означает не только сохранение документов в неизменном формате, но и готовность взять на себя обязательства и в будущем предоставлять возможность их использовать. Сегодня электронные архивы хранят только массивы цифровых данных, однако специалисты DNB хорошо понимают, что нужно найти пути обеспечения возможности использовать документы с операционными системами и программами следующих поколений.

В рамках проекта Европейского Союза KEEP *(Keeping Emulation Environments Portable),* DNB и его другие участники изучают, каким образом может быть достигнуто адекватное представление статических и динамических объектов различного типа: тексты, звук, изображение, мультимедийные документы, веб-сайты, базы данных, видеоигры и т.п.

Цель проекта - создание гарантированной долговременной доступности ресурсов посредством разработки гибких инструментов доступа и хранения. Когда будут созданы такие инструменты и выполнены все необходимые действия по сохранности документов в долговременных архивах, с помощью других средств будут обеспечиваться синхронизация документа в электронном репозитарии и предоставление пользователю доступа к актуализированной версии.

Проблемы и вопросы

Быстрые изменения в технической среде - это серьёзная проблема, затрудняющая сбор сетевых публикаций: из-за использования всё новых и новых технологий постоянно изменяются не только формат публикаций, но и варианты передачи цифровых данных.

Совсем недавно способность хранить, изменять и обрабатывать тексты на компьютере считалась достижением технического прогресса. Вскоре появилось множество новых форматов презентаций, а также возможность включать в документ мультимедийные средства и создавать такие документы, в которых не существует распознаваемой его финальной версии, которую можно было бы сравнить с печатным вариантом. Поскольку миссия DNB заключается в комплектовании фондов, в том числе - в формировании архивов электронных документов и обеспечении доступа к ним, мы нуждаемся в полноценных решениях, позволяющих преодолеть ограничения, вызванные разнообразием

используемых форматов.

Новые форматы для чтения ставят перед библиотекой дополнительные проблемы: зачастую форматы, в которых хранятся файлы, могут открываться только с помощью соответствующего устройства и защищены кодировкой, блокирующей широкий доступ или архивирование.

Помимо решения вопросов, возникающих в связи с применением новых технологий, нужно работать и над проблемой метаданных. В последние несколько лет (в ходе дискуссий с издателями) стало ясно: поскольку библиотеки обычно обмениваются метаданными по соответствующим стандартам, совместное с издателями использование метаданных для них не будет простым. Издатели и иные поставщики, которые продают информационные продукты, имеют свой взгляд на метаданные: для них это, скорее, средство для ведения бизнеса. Конечно, у библиотек, особенно у тех, которые обязаны собирать национальные публикации, иные цели и задачи. Столь разные подходы к метаданным могут затруднить гармонизацию, и свободный обмен данными можно ожидать только среди участников, имеющих единые цели и использующих общепринятые стандарты, построенные на едином понимании метаданных (например между библиотеками или между национальными книготорговыми организациями, где создание метаданных считается приоритетной деятельностью).

Следующие шаги

Помимо сетевых монографий, комплектуются сетевые периодические издания и газеты. Начиная с мая 2010 г. продвигается проект DNB, в рамках которого комплектуются и отражаются в каталогах 300 ежедневных газет (по мере поступления); они архивируются в репозитарии с помощью сервис-провайдера.

В настоящее время усилия DNB направлены на комплектование электронных журналов. Как известно, в них используются индивидуальные форматы метаданных, а это означает, что передача метаданных либо усложнена, либо невозможна. Следовательно, нужно искать альтернативные методы для каталогизации журналов.

В последние годы немецкие библиотеки ведут крупномасштабные работы по оцифровке печатных книг и периодики. При условии, что эти документы общедоступны, они подпадают под законодательство об обязательном экземпляре. Чтобы улучшить процесс архивирования. DNB стремится везде, где это возможно, получить мастер-файлы, которые позволят сохранять информацию без потерь. DNB обсуждает с другими библиотеками и государственными учреждениями, имеющими опыт оцифровки, пути совместного решения проблемы.

В настоящее время мы не архивируем веб-сайты. Но в рамках реализуемого

проекта DNB разрабатывает также технологию сбора, архивирования и индексирования веб-сайтов.

В наши задачи входит расширение способов доставки сетевых ресурсов, для того чтобы открыть возможность большему числу издателей стать репозитариями, выбрав, по крайней мере, один интерфейс. Сегодня подготавливаются соответствующее сопоставления, и они будут постоянно обновляться.