УДК 025.17:004.738.5(100)
ББК 78.375.0(0)

Опыт национальных библиотек зарубежных стран по сбору и долговременному сохранению ресурсов Интернета

Надежда Викторовна Браккер,
главный специалист Центра по проблемам информатизации сферы культуры (Центра ПИК)

Леонид Абрамович Куйбышев,
генеральный директор Центра по проблемам информатизации сферы культуры (Центра ПИК)

В статье представлен обзор опыта национальных библиотек Австралии, Германии, Дании, Китая, Литвы, Нидерландов, Новой Зеландии, Норвегии, Португалии, Соединенного Королевства, США, Финляндии, Франции, Чехии и Швеции по сбору и архивированию сетевых информационных ресурсов. Приводится краткое описание технологий сбора и сохранения ресурсов Интернета, а также правовых проблем, связанных с этими процессами.

Ключевые слова: *долговременное сохранение, архивирование Интернета, национальные библиотеки, технологии сбора и архивирования сетевых ресурсов, правовые проблемы.*

Технологии сбора и архивирования ресурсов Интернета

Сбор ресурсов Интернета для целей долговременного сохранения и предоставления доступа к ним может осуществляться автоматически с помощью программ-роботов или путем выборочного отбора, глубокого сбора и архивирования отдельных сайтов или документов.

Результатом автоматического сбора сетевых ресурсов или веб-харвестинга являются все материалы определенного сегмента сети в момент сбора данных. Веб-харвестинг осуществляют программы-роботы или веб-кроулеры, основанные на тех же принципах, что и поисковые машины. В начале процесса выполняется ручная настройка параметров сбора информации, при которой определяется, из каких доменов и с какой периодичностью собираются материалы для хранения (например, национальный домен или материалы по определенной тематике). После

окончания работы кроулера необходимы проверка и архивирование собранной информации, что требует некоторого участия человека. Небольшие трудозатраты — несомненное преимущество этого метода.

Как правило, процедура веб-харвестинга выполняется регулярно через определенные, достаточно большие промежутки времени (например, раз в полгода). Изменения, произошедшие в сети за этот период, не архивируются и полностью утрачиваются.

Качество и полнота результатов веб-харвестинга зависят от используемых роботов, которые постоянно совершенствуются. Результат работы кроулера — статические представления интернет-страниц, как правило, первого и второго уровней.

В результате веб-харвестинга образуются огромные объемы информации для хранения. Эта информация не может быть каталогизирована обычным способом, поэтому для автоматического аннотирования и структурирования разрабатываются и используются специальные программы, основанные на методах семантического веба. Недостатком харвестинга является дублирование, так как архивируются повторно размещенные материалы и зеркала сайтов, т. е. один и тот же материал может быть собран несколько раз.

Этот метод недостаточно эффективен для сбора и сохранения таких интернет-ресурсов, как газеты, потоковые видео- и аудиоресурсы, результаты работы веб-камер, интерактивные документы, цифровые материалы различных типов, хранящиеся в базах данных, интернет-ресурсы с коротким жизненным циклом.

Выборочный тематический отбор с глубоким (многоуровневым) сбором и архивированием таких материалов реализуется на основе закона об обязательном экземпляре или договоров с издателями и дает более качественный результат на небольшом сегменте сети. Сотрудничество с издателями позволяет качественно каталогизировать собранные ресурсы. Обычно используется сочетание обоих методов — полный автоматический сбор материалов каких-либо сегментов сети через определенные периоды времени и глубокое выборочное архивирование наиболее ценных ресурсов Интернета.

Правовые проблемы сбора и сохранения сетевых ресурсов

Для реализации проектов по архивированию Интернета необходимо, как минимум, достаточное долговременное финансирование и адекватная законодательная база, желательно в форме законодательства об обязательном экземпляре документов. Очевидно, что невозможно обязать всех, кто публикует свои материалы в Интернете, передавать их на депозитарное хранение, поэтому веб-харвестинг является единственным вариантом, обеспечивающим полноту охвата определенного сегмента сети (например, национального домена). А вот определение ресурсов Интернета, имеющих большое общественно-политическое, культурное и научное значение, и сбор их для долговременного хранения и доступа на основе закона об обязательном экземпляре (легальном депозите) и/или договоров с издателями вполне возможны и практикуются во многих странах.

Сложно дать такое определение сетевой публикации, чтобы оно позволяло собирать все материалы, требующие сохранения. Сетевые публикации не являются линейными и законченными. Они включают в себя ресурсы различных форматов (текст, статические изображения, видео- и аудиофрагменты, потоковые видео и аудио, 3D объекты, виртуальную реальность и пр.), имеют сложную структуру. Не всегда возможно установить автора и издателя материала, страну происхождения, можно определить только местонахождение сервера и страну регистрации домена. Дата публикации также является расплывчатым понятием в случае динамического формирования страниц и постоянно обновляющихся сайтов. Для публикации в Интернете понятие «экземпляр» теряет свой первоначальный смысл.

Технологии развиваются очень быстро, появляются новые формы и виды сетевых изданий, и законодательство не поспевает за этими изменениями. Если ориентироваться на существующие сегодня технологии и типы сетевых публикаций, законодательство придется пересматривать слишком часто. Если закон носит более общий характер, ориентируется на завтрашние технологии, которые сложно предсказать, то невозможно точно определить все необходимые для сбора и сохранения информации процедуры, обеспечить контроль исполнения, предусмотреть штрафные санкции [1].

Некоторые страны, в которых архивирование интернет-ресурсов еще не закреплено законодательно, тем не менее, реализуют проекты по веб-харвестингу и глубокому тематическому архивированию на основе добровольного предоставления издателями материалов на хранение. Эта деятельность позволяет точнее сформулировать предложения по изменению законодательства и отработать совершенно новые для традиционных учреждений памяти стратегии и технологии сбора и сохранения цифровой информации.

Возможности доступа к архивированным материалам определяются международными и национальными законами об охране прав на интеллектуальную собственность, поэтому в большинстве стран доступ к ним ограничен или даже закрыт.

Опыт национальных библиотек зарубежных стран

Австралия. Национальная библиотека (НБ) Австралии занимается проблемами долговременного сохранения цифровой информации с 1996 года. В созданном для этой цели архиве PANDORA[1] сохраняются специально отобранные веб-ресурсы, имеющие национальное значение и достойные того, чтобы обеспечить их долговременное сохранение. В марте 2013 г. объем архива составлял 8.19 ТВ. Для автоматического сбора и описания веб-ресурсов и организации доступа к ним разработано специальное программное обеспечение PANDORAS.

Германия. НБ Германии в сотрудничестве с другими учреждениями культуры создала экспертную библиотечную сеть NESTOR[2] для разработки методик и рекомендаций по сохранению цифрового наследия и обучению работников библиотек. Экспертная сеть создает базис для архивирования, долговременного сохранения и защиты, а также доступности цифровых ресурсов и их последующего использования. Это информационный форум для обсуждения проблем архивного хранения и долговременной доступности цифровых информационных ресурсов Германии. Участники проекта обсуждают следующие темы:

- критерии оценки вызывающих доверие цифровых репозитариев;
- процедуры сертификации систем для цифровых архивов;
- принципы и критерии отбора для архивного хранения цифровых ресурсов;
- стратегии долговременного архивного хранения цифровых ресурсов;
- исследование долговременной доступности цифровых ресурсов в музейной сфере;
- концепция устойчивой формы организации экспертной сети и информационного форума;
- координация распределения ответственности по долговременному сохранению цифровых ресурсов, особенно между библиотеками, архивами и музеями.

Проект KOPAL[3] посвящен разработке цифрового информационного архива для долговременного хранения электронных документов с сохранением всех функций полноценного доступа к ним в будущем. В проекте участвуют НБ Германии, Нижнесаксонская государственная и университетская библиотека Геттингена, а также офис IBM в Германии.

Дания. Законодательство об обязательном экземпляре в 1997 г. было распространено на электронные публикации на материальных носителях и статические сетевые документы. По мандату депозитарной библиотеки Королевская библиотека Дании начала выборочно комплектовать интернет-ресурсы с 1998 года. Электронные журналы с разрешения издателей скачивались в регистрационную систему, доступную в Интернете. По закону издатели обязаны указывать информацию об электронных продуктах, поступающих в регистрационную систему, что обеспечивает Королевской библиотеке получение информации об интернет-адресах, к которым можно обращаться для скачивания. Затем персонал собирает, проверяет, каталогизирует материалы и передает их на архивный сервер.

Параллельно с этим процессом Королевская библиотека, Государственная библиотека Орхуса и Центр исследований Интернета Университета Орхуса организовали сетевой архив[4] для тестирования различных архивных стратегий и содействия комплектованию материалов научных исследований. На основе результатов тестирова-

ния выработаны рекомендации, ставшие основой новой редакции закона об обязательном экземпляре 2005 года. Городская и университетская библиотека совместно с Королевской библиотекой комплектуют и сохраняют датский сегмент Интернета. Собранные материалы хранятся в общем архиве, который для большей безопасности расположен в обоих учреждениях. К архиву нет общего доступа, он открыт только для исследовательских целей с разрешения датского Агентства по охране данных. Разработаны рекомендации по применению нового закона об обязательном экземпляре для библиотек, интернет-издателей и контент-провайдеров.

Китай. Для сохранения китайского интернет-наследия НБ Китая[5] с 2003 г. реализует проект по сбору и сохранению веб-информации (WICP). Проект China Events отбирает, индексирует и публикует сайты, которые связаны с историческими событиями, имеющими большое значение для Китая. Проект направлен на защиту наиболее ценного культурного наследия страны и его долговременное сохранение для будущих поколений.

Литва. Литовская НБ им. Мартинаса Мажвидаса в соответствии с измененным в 1996 г. законом об обязательном экземпляре собирает и хранит обязательный экземпляр документов, в том числе электронных. Библиотека архивирует документы из национального домена .lt, другие документы, важные для страны, коммерческие и правительственные электронные документы. Документы, распространяемые на материальных носителях, комплектуются и обрабатываются точно так же, как традиционные, а материалы Интернета собираются методом харвестинга в виде статических страниц. Динамические информационные ресурсы и контент общественно значимых баз данных собираются для депозитарного хранения на основе договоров с издателями.

В 2002 г. для хранения сетевых материалов создан Архив электронных ресурсов[6] как подсистема Литовской интегрированной библиотечной информационной системы (LIBIS) и является депозитарной системой для хранения сетевых публикаций, основанной на модели NEDLIB. Библиотека собирает сетевые ресурсы своего национального домена один раз в полгода, одновременно проясняя вопрос о правах для каждого ресурса и снабжая его метаданными в стандарте Dublin Core.

Нидерланды. Разработка системы для добровольного депонирования электронных публикаций началась в 1995 году. Нидерланды — одна из немногих стран мира, где нет законодательства об обязательном экземпляре документов, и Королевская библиотека Нидерландов собирает все документы на основе договоров с издателями. После нескольких лет экспериментов в 2000 г. Королевская библиотека совместно с IBM разработала депозитарный архив для хранения электронных публикаций, основанный на стандарте OAIS.

В результате с 2002 г. в Нидерландах функционирует e-Depot[7] — официальный архив таких крупнейших издательских домов, как Kluwer и Elsevier Science, а также других членов Голландской ассоциации издателей.

С 2006 г. Королевская библиотека занимается веб-архивированием и долговременным сохранением специально отобранных сетевых ресурсов на основе договоров с издателями.

Новая Зеландия. Одной из первых стран, в которых обязательный экземпляр был распространен на все цифровые материалы, в том числе интернет-ресурсы открытого доступа, включая блоги, вики и пр., стала Новая Зеландия. В соответствии с принятым в 2003 г. законом НБ Новой Зеландии получила мандат на сбор и сохранение всех электронных публикаций и интернет-ресурсов страны. Это позволило основать Национальный архив цифрового наследия (NDHA[8]) для обеспечения бессрочного хранения цифровой информации о Новой Зеландии.

Национальная стратегия, которая лежит в основе NDHA, не делает различий между контентом, созданным авторизованными организациями, и контентом, созданным гражданами. Хранящиеся в NDHA цифровые материалы поступают в НБ из четырех источников: обязательные экземпляры, интернет-поиск, безвозмездно предоставленные и оцифрованные материалы. В том, что касается обязательных экземпляров, издатели, выпускающие электронные книги, используют возможности предоставления материалов в Интернете в онлайновом режиме или же отправляют материалы на традиционных носителях (например, CD или DVD) в Бюро обязательных экземпляров. Программа НБ по поиску в Интернете предусматривает отбор сайтов с помощью программного обеспечения Web Curator Tool. Неопубликованные материалы по цифровому наследию обычно поступают в НБ от дарителей таким же образом, как и публикации. Еще одним крупным источником цифровых материалов являются внутренние программы по оцифровке звуковых, аудиовизуальных, изобразительных и печатных материалов. Пользование этими ресурсами основано на политике НБ Новой Зеландии по сбору материалов [2].

Архив NDHA размещен в киберпространстве как область для сохранения и преобразования данных и обмена ими. В качестве государственного ресурса NDHA обеспечивает возможность хранения материалов НБ на трех официальных языках Новой Зеландии: английском, маори и новозеландском языке жестов. Как хранилище цифрового наследия архив обеспечивает сохранность сайтов, цифровых изображений, CD, DVD и других цифровых материалов из постоянно об-

новляемого собрания НБ (несмотря на их техническое старение) и может предоставлять ученым, студентам и читателям библиотеки доступ к этим материалам в настоящее время и в будущем.

Проект выполняется в сотрудничестве с компаниями Ex Libris Group и Sun Microsystems, которые занимаются разработкой программного и технического обеспечения хранения цифровых материалов [2].

Норвегия. НБ Норвегии начала заниматься долговременным сохранением цифровой информации с 2001 г. в рамках проекта Paradigma.

В 1989 г. был введен в действие измененный закон об обязательном экземпляре документов, который охватывает все виды документов на всех возможных носителях: бумажные документы, микроформы, фотографии, комбинированные документы, фонограммы, фильмы, видео, цифровые документы, радиопрограммы. Закон относится как к документам, созданным в Норвегии, так и к зарубежным изданиям, произведенным для норвежских издателей или адаптированным для Норвегии.

На основе закона об обязательном экземпляре и рекомендаций проекта Paradigma НБ Норвегии начала харвестинг всего национального домена .no. В декабре 2002 — январе 2003 г. были собраны информационные ресурсы 3,1 млн сайтов, а в августе 2003 г. — уже 4,1 млн сайтов.

Помимо ресурсов национального домена собираются относящиеся к Норвегии документы из доменов .com, .org, .net.

В 2003 г. НБ начала ежедневный сбор и архивирование статических страниц 65 онлайн-газет, а с 2005 г. — полную выгрузку баз данных всех периодических интернет-изданий норвежского домена.

Норвежский закон об обязательном экземпляре устанавливает, что документы, собранные в соответствии с ним, могут предоставляться в пользование в целях исследовательской деятельности и документоведения, т. е. закон ограничивает доступ к документам для всех пользователей. Предоставление доступа к документу определяется целями, в которых он будет использован. Например, к архиву сетевых документов могут получить доступ исследователи, преподаватели, студенты и некоторые другие группы пользователей со специфическими интересами или хобби (например, генеалогия). Доступ может предоставляться только с компьютеров, находящихся в библиотеке[9].

Португалия. Национальный фонд научной информатики (FCCN[10]) занимается сбором информационных ресурсов португальского национального домена .pt и долговременным сохранением онлайн-документов в веб-архиве Португалии[11]. Архив предоставляет открытый доступ более чем к 130 млн страниц, архивированных в 1998—2009 годах. FCCN — некоммерческая организация, которая объединяет учреждения науки и образования Португалии и ведет регистр информационных ресурсов верхнего уровня в домене .pt.

Веб-архив Португалии создавался на основе результатов работы исследовательской группы Университета Лиссабона, занимавшейся вопросами веб-архивирования с 2001 года. Группа разработала прототип веб-архива Tomba и кроулер, который собирал данные португальского сегмента Интернета в 2002—2006 годах.

Соединенное Королевство. Программа архивирования Интернета Британской библиотеки обеспечивает долговременное сохранение веб-ресурсов по образованию и культуре в домене .uk и предоставляет доступ к ним. Цели программы: создать всеобъемлющий веб-архив как часть цифровой коллекции Британской библиотеки; обеспечить долговременное сохранение архива с возможностями доступа к нему в будущем; подготовить персонал и организовать все

процессы и системы, необходимые для легального депонирования веб-ресурсов.

С 2004 г. Британская библиотека с разрешения создателей архивирует веб-сайты по социальной истории и культурному наследию страны в соответствии со специально разработанной политикой комплектования[12]. Отобранные сайты доступны через веб-архив, содержащий регулярно обновляемые статические представления 5 тыс. сайтов. Архив предоставляет возможности полнотекстового поиска, поиска по названию, по предметным рубрикам и по интернет-адресу (URL). Например, за октябрь 2010 г. было заархивировано 30 тыс. 885 страниц 8 тыс. сайтов, объем информации составил 7,46 Тб.

Статические представления страниц веб-сайтов собираются с помощью программного обеспечения Web Curator Tool (WCT), разработанного НБ Новой Зеландии и Британской библиотекой в рамках Международного консорциума сохранения Интернета. Это открытое программное обеспечение, свободно распространяемое на основе публичной лицензии Apache, внедрено также в НБ Норвегии.

WCT предоставляет веб-архивистам средства для управления следующими процессами:

• авторизация харвестинга (получение разрешения собирать веб-материалы и предоставлять доступ к ним);

• отбор сайтов, определение объемов и составление графика (что будет собрано, каким образом, когда и как часто);

• описание (добавление метаданных);

• харвестинг (скачивание материалов в назначенное время с помощью кроулера Heritrix);

• контроль качества (проверка соответствия результатов харвестинга на соответствие заданию и корректировка мелких ошибок).

WCT работает как браузер. Программное обеспечение идет по ссылкам внутри сайта и собирает все доступные файлы, которые находит. WCT может собирать динамические сайты, разработанные с помощью технологий PHP или ASP, но не может собирать контент баз данных, так называемый «глубокий веб», например библиотечные каталоги. Используется разработанный интернет-архивом кроулер Heritrix, который настраивается таким образом, чтобы минимизировать его влияние на собираемые сайты.

В последние годы Британская библиотека играет ведущую роль в международных консорциумах по разработке технологий веб-архивирования. Библиотека участвует в работе группы национальных библиотек по совершенствованию Heritrix, в декабре 2009 г. была выпущена версия 3.0.

Британская библиотека стала одним из основателей Международного консорциума сохранения Интернета, в котором участвуют национальные библиотеки и другие заинтересованные организации, обмениваясь опытом и продвигая использование общих стандартов и технологий. Кроме того, Британская библиотека возглавляет Консорциум веб-архивирования Соединенного Королевства (UKWAC[13]). В него также входят Объединенный комитет информационных систем, Национальная библиотека Уэльса и компания Wellcome Trust. Для создания специализированных коллекций Британская библиотека сотрудничает и с другими учреждениями.

Британская библиотека, Национальные библиотеки Шотландии и Уэльса, Бодлианская библиотека в Оксфорде, библиотека Кембриджского университета и библиотека Тринити-колледжа в Дублине, которым предоставляются обязательные экземпляры изданий, с 2013 г. имеют право собирать и хранить все, что публикуется в Сети в Великобритании. По оценкам специалистов, около миллиарда страниц в год будут доступны для исследований. В архив войдут данные из социальных сетей и 4,8 млн веб-сайтов, включая журналы, книги, научную периодику, а также альтернативные источники литературы, новостей и критики. Осуществлению проекта предшествовали 10 лет планирования[14].

США. В декабре 1994 г. Комиссия США по сохранению и доступу к информации и Группа научных библиотек организовали Рабочую группу по цифровому архивированию. В 1995 г. Рабочая группа подготовила проект отчета «Долговременное сохранение цифровой информации», который активно обсуждался всеми заинтересованными сторонами (в окончательной редакции опубликован в 1996 г.[15]). В отчете определены ключевые проблемы (организационные, технологические, правовые, экономические и пр.), которые необходимо разрешить для обеспечения долговременного доступа к электронным цифровым материалам, и предложены пути решения этих проблем.

В 2000 г. Библиотека Конгресса начала реализацию Национальной программы по цифровой информационной инфраструктуре и стратегии сохранения. В рамках ее реализации предполагалось установление сотрудничества с федеральными учреждениями и институтами для развития национальной стратегии цифрового сохранения. В итоге Программа объединила более 170 организаций, предоставив доступ к богатейшей национальной цифровой коллекции, что стало ярким примером неоспоримого преимущества сотрудничества и партнерских отношений.

В 2010 г. Библиотека Конгресса инициировала создание Национального цифрового альянса (NDSA). Эта совместная инициатива правительственных учреждений, образовательных институтов, некоммерческих организаций и предприятий направлена на сохранение рассредоточенных национальных цифровых коллекций и предостав-

ление доступа к ним. NDSA станет продолжением Национальной программы по цифровой информационной инфраструктуре и стратегии сохранения.

Учитывая предыдущие достижения Программы, NDSA ставит следующие цели: развитие стандартов и практик, направленных на сохранение цифрового наследия; объединение усилий с экспертами по выявлению наиболее важных категорий цифровой информации, срочно нуждающихся в сохранении; разработка мер по объединению контента в национальную коллекцию[16].

Проект веб-архивирования Библиотеки Конгресса США (LCWA[17]) — часть деятельности библиотеки по сохранению веб-материалов для будущих поколений исследователей. Проект уделяет особое внимание информации, связанной с выборами в Соединенных Штатах, а также относящейся к членам Конгресса США.

Библиотека Конгресса США формирует тематические коллекции архивированных веб-сайтов, отобранных кураторами и специалистами, и оценивает различные аспекты веб-архивирования: стратегии отбора веб-сайтов; создание метаданных, необходимых для обеспечения доступа конечным пользователям; пакетирование метаданных; поисковые системы. Каждый сайт описывается в каталоге с использованием стандарта MODS[18], необходимого для доступа к коллекциям. Кроме того, для генерального каталога Библиотеки Конгресса, доступного в Интернете, каждая коллекция веб-сайтов описывается в стандарте MARC, так что веб-коллекции описаны в общем каталоге вместе с другими материалами. Ведутся эксперименты с пакетами METS для описаний в стандарте MODS, иконками изображений, облаками тэгов и метаданными для долговременного сохранения PREMIS[19]. Создание производных вариантов сайтов (деривативов), описание, поиск, пакетирование, использование метаданных для эффективного поиска сайтов, созданных на разных естественных языках и языках программирования — это составные части постоянно ведущихся исследований в области эффективного веб-архивирования.

Финляндия. С 1997 г. для сбора, регистрации и долговременного хранения интернет-публикаций в Финляндии функционирует архивная система EVA[20] — объединенный проект библиотек, издателей и экспертного сообщества страны, который координирует библиотека Университета Хельсинки (НБ Финляндии). Собираются публично доступные статические веб-страницы финского домена. Цели проекта: формулирование критериев отбора электронных документов, создание депозитарной системы для издателей и разработка надежной системы долговременного сохранения информации. С 2002 г. система периодически собирает статические образы страниц домена .fi. Кроме того, собираются тематические материалы (например, связанные с выборами). Налажено сотрудничество с порталом финского информационного центра, который передает в EVA адреса финских серверов, расположенных в других доменах.

Изменения в законы об обязательном экземпляре и копирайте, которые разрабатывались параллельно, вступили в силу в 2005 году. НБ получила право собирать цифровые материалы онлайн и офлайн, а также радио- и телепрограммы.

Франция. Решение о распространении французского закона об обязательном экземпляре на сетевые материалы принято парламентом в 2006 году. Закон дает полномочия НБ Франции и Национальному институту аудиовизуальной информации (INA), ответственному за сохранение радио- и телепрограмм, автоматически собирать сетевые материалы, а также требовать материалы у издательств, если автоматический харвестинг невозможен, и предоставлять доступ к архиву.

Еще до внесения изменений в законодательство НБ Франции с 2000 г. разрабатывала комбинированную методику, включающую в себя автоматический широкомасштабный харвестинг несколько раз в год; более частый сбор 10% автоматически отобранных сайтов; глубокое архивирование сайтов, которые нельзя собрать автоматически; сбор тематических коллекций, связанных с определенными событиями.

Осознавая необходимость международного сотрудничества по проблемам веб-архивирования, НБ Франции инициировала Международный семинар по веб-архивированию (IWAW[21]) и активно участвовала в создании Международного консорциума сохранения Интернета.

Чехия. Проект создания архива веб-ресурсов Чешской республики[22] реализуется НБ[23] при сотрудничестве с Моравской библиотекой и Институтом информатики Университета Масарика с 2000 года. Первоначально средства на реализацию проекта выделяло Министерство культуры, но затем он развивался почти исключительно благодаря грантовому финансированию.

Цель проекта — сохранение культурного веб-наследия Чехии (например, веб-ресурсы Богемии, Чешская национальная библиография). Используются как технологии автоматического сбора всех национальных веб-ресурсов, так и выборочный сбор тематических коллекций.

В Чешской республике законодательная база еще недостаточно эффективна. Закон об обязательном экземпляре не распространяется на сетевые ресурсы, хотя соответствующие изменения находятся в стадии подготовки. Закон об авторском праве с июня 2006 г. приведен в соответствие с Директивой ЕС 2001/29/ЕС, т. е. весь архив можно предоставлять в открытый доступ в помещениях библиотеки. Кроме того, в отношении наиболее важных интернет-ресурсов заключаются договора с издателями, которые дают библиотеке право делать эти архивы доступными в сети.

Сбор интернет-ресурсов — это автоматизированный процесс, осуществляемый программными комплексами, которые обеспечивают сбор, индексирование и сохранение данных в соответствии с заранее установленными параметрами. Большая часть программного обеспечения — это открытые программные системы (например, Heritrix), разрабатываемые Международным консорциумом сохранения Интернета, остальные программные средства разрабатываются самими участниками проекта. Собранные файлы и метаданные сохраняются в стандартном архивном формате, который поддерживается Консорциумом. Информация хранится на сервере, а также в резервной системе RAID. Объем данных, собранных с сентября 2001 г., составляет 15,5 Тб. Отдельный сервер используется для доступа к тем ресурсам, которые подпадают под действие соглашений с издателями. Полнотекстовое индексирование реализуется системой с открытым кодом Nutch, для доступа используются системы Nutchwax и WERA.

Для описания и идентификации ресурсов используются международные стандарты (MARC 21, Dublin Core, ISSN и URN), а для архивирования — стандарт ARC. Записи регистрируются в Чешской национальной библиографии.

Швеция. Королевская библиотека Швеции с 1996 г. изучала вопросы сбора и долговременного сохранения статических и динамических сетевых документов Швеции в рамках проекта Kulturarw3[24].

Швеция была первой страной, которая занималась исследованиями технологии харвестинга для архивирования сетевых информационных ресурсов. Первый харвестинг в 1997 г. собрал данные национального домена .se, а в следующие годы собирался важный для Швеции веб-контент из других доменов. Королевская библиотека получила мандат на сбор сетевых документов Швеции в 2002 г. и с тех пор собирает данные 2—3 раза в год.

Краткие выводы

- Национальные библиотеки многих стран мира уже более 10 лет занимаются проблемами архивирования сетевых ресурсов, создавая архивы, которые комплектуются путем сочетания технологий веб-харвестинга и глубокого тематического архивирования;
- законодательство некоторых стран допускает веб-харвестинг и/или глубокое тематическое архивирование сетевых ресурсов; если эти процессы не обеспечены законодательно, национальные библиотеки занимаются долговременным сохранением сетевых ресурсов на основе договоров с издателями и при ограничении доступа к архивам;
- архивирование и долговременное сохранение сетевых ресурсов требуют серьезных научных исследований и технологических разработок, в которых ключевую роль играет международное сотрудничество;
- в результате международного сотрудничества разработаны стандарты, схемы метаданных и открытое программное обеспечение для долговременного сохранения сетевой информации, которые необходимо использовать и в России.

Примечания

1 URL: http://pandora.nla.gov.au/ (на англ. яз.).
2 URL: http://www.langzeitarchivierung.de/Subsites/nestor/DE/Home/home_node.htm (на нем. яз.).
3 URL: http://kopal.langzeitarchivierung.de/ (на англ. и нем. яз.).
4 URL: http://netarchive.dk/index-en.php (на англ. и дат. яз.).
5 URL: http://www.nlc.gov.cn/newen/ (на англ. и кит. яз.).

[6] URL: http://www.inforum.cz/archiv/inforum2003/prispevky/Jodelis_Remigijus.pdf (на англ. яз.).
[7] URL: http://www.kb.nl/en/expertise/e-depot-and-digital-preservation/publications-and-links/the-e-depot-and-digital-preservation (на англ. яз.).
[8] URL: www.natlib.govt.nz/ (на англ. яз.).
[9] URL: http://archive.ifla.org/IV/ifla71/papers/151r_trans-Rustad.pdf (на рус. яз.).
[10] URL: http://www.fccn.pt/en/ (на португал. и англ. яз.).
[11] URL: http://www.arquivo.pt/?l=en (на португал. и англ. яз.).
[12] URL: http://www.bl.uk/aboutus/stratpolprog/coldevpol/index.html (на англ. яз.).
[13] URL: http://www.webarchive.org.uk/ukwa/ (на англ. яз.).
[14] http://www.prlib.ru/news/Pages/Item.aspx?itemid=7127 (на русск. яз.)
[15] URL: http://www.oclc.org/research/activities/past/rlg/digpresstudy/final-report.pdf (на англ. яз.).
[16] URL: http://www.prlib.ru/news/Pages/Item.aspx?itemid=1715 (на рус. яз.).
[17] URL: http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html (на англ. яз.)
[18] URL: http://www.loc.gov/standards/mods/ (на англ. яз.)
[19] URL: http://www.loc.gov/standards/premis/ (на англ. яз.)
[20] URL: http://web.archive.org/web/20041010005510/www.lib.helsinki.fi/eva/english.html (на англ. и фин. яз.).
[21] URL: http://www.iwaw.net/ (на англ. яз.).
[22] URL: http://en.webarchiv.cz/ (на англ. яз.).
[23] URL: http://www.nkp.cz/_en/index.php3 (на англ. яз.).
[24] URL: http://www.kb.se/english/find/internet/websites/ (на англ. яз.).

Список источников

1. *Анжелаки Дж.* Обязательное хранение цифровых материалов в государствах — членах Европейского Союза : обзор законодательства и правоприменительной практики // Библиотеки в правовом пространстве. Современные проблемы : сб. статей. — СПб. : Рос. нац. б-ка, 2008. — С. 124—168.
2. *Мохи Д.Г.* Национальный архив цифрового наследия // Языковое и культурное разнообразие в киберпространстве : сб. материалов Междунар. конф. (Якутск, 2—4 июля 2008 г.) / сост. Е.И. Кузьмин, Е.В. Плыс. — МЦБС, 2010. — С. 398—404.
3. *Филозова И.А.* Открытые архивы научной информации [Электронный ресурс] // Системный анализ в науке и образовании. — 2010. — № 1. — С. 70—75. — Режим доступа: http://www.sanse.ru/archive/15

Анонс

ПРЕДСТОЯЩИЕ КОНГРЕССЫ ИФЛА

• **Всемирный библиотечный и информационный конгресс — 79-я Генеральная конференция и Ассамблея ИФЛА** на тему: **«Библиотеки будущего: безграничные возможности»** (Future Libraries: Infinite Possibilities) состоится **17—23 августа 2013 г.** в Сингапуре.

Определены регионы предстоящих в 2014—2019 гг. конгрессов ИФЛА:

- **2014 — г. Лион, Франция**
- **2015 — Африка**
- **2016 — Северная Америка**
- **2017 — Европа**
- **2018 — Латинская Америка или страны Карибского бассейна**
- **2019 — Европа**