World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

June 30, 2005

Code Number: **074-R**
Meeting: **133 SI - Bibliography**

Веб-индексирование[1] (Web crawling): опыт Национальной библиотеки Франции

Кристиан Луповичи
Руководитель национального библиографического
агентства Национальной библиотеки Франции
Christian.lupovici@bnf.fr

Резюме:
В настоящее время Национальная библиотека Франции (НБФ) в рамках выполнения своей миссии по отношению к обязательному экземпляру проводит эксперимент по выработке методики сбора сетевых материалов и организации долговременного хранения цифровых документов. Для достижения этой цели необходимо разобраться в сущности обязательного экземпляра и библиографической обработки интернет-ресурсов.

[1] Web crawler – индексная программа, которая строит индексы путем последовательного перехода по гиперссылкам с одной Web-страницы на другую (сленговый термин «ползатель») (Воройский Ф.С. Информатика : новый систематизированный толковый словарь-справочник. М., 2001. С.377)- *примеч. переводчика.*

1. Обязательный экземпляр в НБ Франции

1.1 История вопроса

Первые попытки принятия законодательства об обязательном экземпляре были предприняты во Франции Франсуа Первым в 1537 г. С тех пор целью всех последующих распоряжений было расширение состава обязательного экземпляра с учетом новых технических достижений.

1537: книги;

1667: офорты;

1689: гравюры;

1925: любая художественная графическая продукция;

1941: плакаты, ноты и фотографии;

1963: звукозаписи любого характера;

1975: фотодокументы или записи движущегося изображения, независимо от технологии производства;

1977: киноработы;

1992: все виды документов независимо от носителя, включая базы данных и специальные системы, если они доступны для пользователей.

Включение сетевых ресурсов в состав обязательного экземпляра должно произойти в 2005 г., поскольку сейчас этот вопрос обсуждается в Парламенте.

1.2 Миссия и состав Обязательного экземпляра

Цель обязательного экземпляра состоит в сохранении и предоставлении долговременного доступа ко всем документам, являющимся культурным наследием, и которые были изготовлены и распространены в большом количестве экземпляров. Обязательный экземпляр включает все виды ресурсов, изданных не только на французском языке на территории Франции.

Больше вопросов возникает относительно содержания и/или каналов распространения (книг, гравюр, карт, звуко- и аудиовидеозаписей), а не носителя информации: бумага, диск или он-лайн. Большинство серой (скучной, неинтересной) литературы, выявленной в интернет, будет до тех пор входить в состав регулируемого обязательного экземпляра, пока сетевые документы будут доступны для общественного пользования, и таким образом быть изданными.

Обязательный экземпляр также не означает, что собирание является процессом комплектования, т.к. последнее подразумевает политику развития фонда; это вопрос сбора материалов всех видов, что дает возможность получить полную картину знания, распространяемого в настоящее время среди французского народа, с целью сохранения его для будущих исторических исследований. Это означает, что все различные типы документов, в том числе и запрещенные работы, включены в состав обязательного экземпляра, как это и было, начиная с шестнадцатого столетия.

Как все выше сказанное применимо к сетевым ресурсам?

2. Веб индекесирование

2.1 Собирание интернет-ресурсов в качестве составной части обязательного экземпляра

Философия обязательного экземпляра по отношению к сетевым ресурсам подразумевает:

a) исключить из его состава все имена доменов, которые не имеют никакого отношения к французской культуре;
b) сфокусировать внимание на специальных французских доменных именах (.fr, .pf, .wf, .pm, .re, .tf, .ad, .yt);
c) собирать все остальное.

В первом случае невозможно узнать, что на самом деле имеет отношение к французской культуре, и придется провести работу, чтобы найти способ, с помощью которого осуществить сложный отбор, основанный на семантическом анализе содержания сайтов. Несмотря на то, что вопросы объема памяти уже не являются самой большой проблемой – количество сайтов, отвечающих требованиям обязательного экземпляра намного меньше, чем количество сайтов, которые в настоящее время уже хранится в НБ Франции – все же надо найти лучшее решение. В любом случае, та часть сети, которая будет отбираться и сохраняться, всегда будет больше.

2.2 От экспериментов к повседневной практике

2.2.1 Первые опыты

НБ Франции собирает французские веб-сайты с помощью «ползателя» HTTrack с 2001 г.[2] Во время первых опытов основное внимание было сосредоточено на нескольких операциях отбора во Франции. Надо было выяснить следующие вопросы:

a) как строить типологию веб-сайтов. И соответственно, какая должна быть периодичность индексирования;
b) решение технических проблем (форматы, глубина индексирования и настройка «ползателя»;
c) сравнение ручных и автоматических способов отбора веб-сайтов.

Эта стадия была важна для оценки автоматического отбора элементов сети, которые будут обрабатываться вручную, и выяснения вопросов использования вспомогательного инструментария в момент принятия решения.

2.2.2 Веб индексирование в повседневной практике

Каковы объем и источники сбора французского обязательного экземпляра?
Весной 2004 г. с помощью сетевого «ползателя» Alexa был сделан моментальный снимок всей сети: были отобраны соответствующие домены (*.fr, .com, .net, .edu, .com, .biz, .info, .int, .pf, .ad, .coop, .name, .aero, .tf, .re, .museum, .pro, .pm, .wf, .yt*), на основе которых в дальнейшем был произведен отбор. В результате было выявлено более 2 миллиардов URL (универсальных указателей ресурсов) и 65 миллионов хост-компьютеров, объемом 27 терабайтов.
Кроме того, в конце 2004 г. и январе 2005 г. при помощи «ползателя» Heritrix был произведен отбор домена .fr. В результате было проиндексировано 500 000 общественных и 4 000 личных сайтов. Это - больше чем 118 миллионов URL и полмиллиона хост-компьютеров объемом 3 терабайта.
Такое количество данных в полном объеме невозможно ни каталогизировать, ни индексировать, если только какую-то часть. Для того, чтобы отсортировать и закаталогизировать все общественные сайты потребовалось бы 500 человеко/лет. Это в два раза больше, чем весь штат каталогизаторов подразделения национальной библиографии НБ Франции.
Вот почему необходимо комбинировать различные методы.

2.3 Обработка сетевого обязательного экземпляра

Основным методом является автоматическое индексирование. Таким способом можно собрать информацию со всей поверхности сети и частично с более глубокого уровня, это зависит от технической задачи, поставленной перед «ползателем» и от настройки программы. Но, вообще говоря, автоматическое индексирование, главным образом, охватывает поверхность сети и создает карту сети.
Второй и дополнительный метод заключается в том, чтобы сфокусировать автоматический сбор на специфических областях, например, на *.fr* домене, и нацелить «ползатель» на сбор более глубинной информации.
Два последних метода представляют собой вершину, ориентированную вниз.
Для того чтобы получить целостную картину сети, мы должны комбинировать регулярный

[2] http://bibnum.bnf.fr/conservation/aristote2004/roche.pdf

автоматический сбор материала с методом перевернутой вершины, который подразумевает сочетание коллекций, состоящих из глубинных сетевых ресурсов, депонированных производителями сайтов, и постоянно обновляемых ссылок к уже собранной сетевой поверхности. Эти материалы должны быть обработаны частично вручную, частично автоматически. Депозит имеет преимущества по срокам страховки сохранности, но это - трудоёмкий процесс, который начинается с анализа сайта, и для его осуществления требуется в среднем 5 человеко/дней.
Все собранные данные хранятся в НБ Франции в центральном репозитарии данных, предназначенном для внутреннего использования и для работы персонала. Веб-архив является частью глобального цифрового хранилища НБ Франции и программы сохранения SPAR *(Système de Préservation des ARchives),* которые были созданы в конце 2004 г.

2.4 Вопрос и метод отбора

Конечно, НБ Франции хотела бы архивировать наиболее существенные веб-сайты. Поэтому было бы интересно найти метод, с помощью которого можно было бы отбирать или осуществлять отбор среди всех уже собранных сайтов. Французский научно-исследовательский проект под названием "Watson" был осуществлен на лингвистической основе, которая используется в веб-архиве НБ Франции, предназначенном для пользователей и сотрудников. Проект продолжался с 2003 г. до 2004 г.[3] В ходе исследования был использован метод характеристики сайта на основе лингвистического анализа метаданных и фраз, имеющихся в полном тексте. С помощью этого метода были составлены рефераты, отражающие содержание сайтов. Этого недостаточно, чтобы полностью исключить участие человека, но этого достаточно, чтобы использовать его в качестве вспомогательного аппарата в процессе отбора сайтов.
В НБ Франции организована группа по отслеживанию сети, в которую вошли сотрудники, работающие в разных подразделениях Библиотеки, в основном, из обязательного экземпляра. Перед группой была поставлена задача, выяснить, как действовать по отношению к новой среде цифровых ресурсов и как наладить необходимую взаимосвязь между новыми ресурсами и традиционными коллекциями на носителях. С помощью средств доступа и методов наблюдения надо было оценить результат индексирования и качество сбора. Группа имеет доступ к средствам, которые позволяют ей контролировать и архив НБ Франции и сеть с целью выработки предложений по более глубокому индексированию определенных веб-сайтов посредством сфокусированного индексирования. Этой осенью при индексировании будут учитываться результаты работы, достигнутые группой наблюдения, и в параметры отбора будет включено ядро предложенных URL.
В любом случае, никогда нельзя будет отобрать те веб-сайты, которые точно соответствуют параметрам французского обязательного экземпляра, и очевидно, что веб-архив НБ Франции всегда будет большим. Если это будут делать все страны, то дублирование неизбежно, но это не так уж и плохо. Поскольку проблема хранения уже не является столь острой, то дублирование в разных странах явится гарантией лучшей безопасности при долговременном хранении. С другой стороны, техника моментального снимка не гарантирует полноту коллекций, а представляет типичный образец фактической культуры. Такова сущность обязательного экземпляра.
Однако, необходимость “интеллектуального «ползателя»”, способного сфокусировать свое внимание на сайтах в соответствии с заданными параметрами и способного осуществлять глубинное индексирование с высокой степенью релевантности, была признана международным консорциумом IIPC (Международный консорциум сохранения интернета). Весной этого года Британская библиотека и Национальная библиотека Франции призвали к созданию интеллектуального «ползателя».

---

[3] José Coch, Julien Masanès**. -** Language engineering techniques for web archiving. In : 4th International Web Archiving Workshop (IWAW04), 16 septembre 2004, Bath, UK - http://www.iwaw.net/04/index.html (visited 30/05/2005).

2.5 Новая среда, новый документ

Имея дело с сетью, мы должны оставить традиционный подход к обработке документа и двигаться по направлению к новому способу управления ресурсами.
Веб-архив имеет новые параметры:
Должны быть функции навигации, которые обновляются изнутри. Ни один из документов не является автономным ресурсом. Документ сам по себе преобразуется. В модели НБ Франции документальной единицей является сам веб-сайт (в модели коллекции).

2.6 Лингвистические методы анализа ресурса

Для поиска необходимой информации, которая постоянно пополняет хранилище, и сотрудникам и читателям Библиотеки необходим соответствующий инструментарий. Возможно, им потребуется точная информация по отдельным темам, также возможно им необходимо будет обработать эту информацию, используя статические или лингвистические методы.
Проект Watson был ориентирован на отбор соответствующих ресурсов определенного домена и осуществление некоторого лингвистического анализа этих ресурсов.
Дать пользователю представление о содержании документа можно с помощью информации о структуре, ключевых словах, именах и местонахождении и т.д., имеющихся непосредственно в документах, или через навигационный реферат.
В ходе проекта были выработаны способы, с помощью которых пользователь может осуществлять тематические подборки, навигацию и семантический анализ содержания сайтов. Эксперименты были проведены на основе французского архива.

3. Новая предполагаемая каталогизация

3.1 Каталогизация метаданных

Традиционно метаданные используются для описания, идентификации физических объектов с целью осуществления поиска. Такая каталогизация достаточно трудоемкая работа и осуществляется в ручном режиме, но т.к. это связано с физическими объектами, то избежать этого нельзя.
В цифровой среде ситуация другая:
Цифровые объекты готовы (или должны быть готовы) к информационной обработке. Они уже содержат в себе необходимые данные, управлять которыми в системе можно посредством имеющихся технических и юридически закрепленных метаданных.
Когда есть текст, то его содержание может быть проиндексировано, классифицировано без любого человеческого вмешательства (на этой стадии). Таким образом, раскрывается структура и семантическое содержание информации.
Если в сети описательные метаданные представлены плохо, то это сбалансируется эффектом сети. В веб-архиве цифровые объекты не являются изолированными документами. Они связаны с другими ресурсами точными ссылками, содержанием и описанием содержания, которые включают средства поиска – или традиционные или совершенно новые.
Это не означает, что нам больше не нужны описательные метаданные или ключевые слова. Но библиотека не лучшее место для создания такой информации, лучше передать эту задачу еще на стадии создания; эти данные не так важны, как в традиционной среде.

3.2 Другие виды метаданных

Появляются новые проблемы, такие как, управление правами, гарантирование подлинности, не только в юридическом плане, но также для гарантированного воспроизводства цифровых ресурсов в больших количествах и моментальный их ввод в обращение. Форматы (а один и тот же документ может быть реализован в разных форматах) являются одной из наиболее важных проблем, поскольку серверная или документальная информация не достаточно надежны. Необходимо определить формат собранных документов.
В рамках деятельности национальной библиотеки также очень важно распространить метаданные на информацию по сохранности по отношению к веб-архиву.

3.3 Будущее каталогизации

Проблемы каталогизации в сетевом контексте, в конечном счете, связаны с разработкой спецификаций для созданных цифровых документов и анализом ресурсов в процессе сбора (модель OAIS), включая инструкции для «ползателя» и взаимодействие с сетевыми производителями.

4. Какой может быть сетевая национальная библиография?

Во французском законе об обязательном экземпляре (даже новом) говорится, что национальная библиография существует на основе депонированного документа.
Но на что в данном случае похожа национальная библиография?
Поскольку цель национальной библиографии заключается в рекламировании новых общедоступных публикаций, мы можем предусмотреть издание на веб-сайте НБ Франции списка новых (не модифицированных) сайтов, используя автоматическую сортировку и "каталогизационную" обработку на стадии архивирования (модель OAIS). Результатом будет список URL с именным указателем, предметным указателем (насколько это позволит сделать лингвистический анализ, близко к тому, что было сделано в проекте Watson).