Архив российских научных онлайновых журналов

©Шварцман М. Е.

Российская государственная библиотека
shvar@rsl.ru

Аннотация

В этой статье излагаются основные проблемы, возникшие при создании архива российских научных онлайновых журналов и методы их решения, как предполагаемые, так и уже реализованные. В приведенном анализе существующих онлайновых журналов основное внимание уделено юридическим и техническим проблемам.

1. Вступление

Периодические издания вообще и журналы в частности содержат наиболее актуальную информацию, оперативно сообщают читателям о новых научных исследованиях и полученных результатах, событиях научной и культурной жизни. Поэтому научные библиотеки считают приоритетным комплектование своих фондов журналами. В последние годы, когда резко сократилось финансирование комплектования библиотек, они, осознавая важность периодики, тратят на ее подписку в среднем более половины всех средств, выделяемых на комплектование.

В этих условиях для библиотек и индивидуальных пользователей журнальной периодики особенно важно знать о новых формах существования журналов, а именно об электронных журналах, размещенных в сети Интернет.

За рубежом, как правило, она размещается на специализированных сайтах: тематических порталах и службах информирования о периодике, таких как:
Ulrich's International Periodicals Directory (http://www.ulrichsweb.com),
Publist (http://www.publist.com),
Periodicals (http://periodicals.net) и других.

Обратившись к специализированным сайтам, пользователь сразу получает важнейшую информацию: имеется ли интересующий его журнал в Интернете, в каком объеме там представлены его материалы, за какой период и на каких условиях (бесплатно или за плату). К сожалению, в Рунете подобных служб, предоставляющих весь комплекс информации по российским журналам, пока нет. Поэтому коллектив сотрудников Российской государственной библиотеки решил взяться за разработку портала российских научных журналов, размещенных в Интернете. Так возник проект "Создание архива российских научно-технических полнотекстовых журналов, опубликованных в Интернет", получивший поддержку РФФИ, грант 04-07-90056-в.

Главным компонентом портала стала общедоступная база данных структурированных описаний российских научных журналов, доступных через Интернет. Описания отражают важнейшие характеристики журналов: название, ISSN, тематику журнала, издающую организацию, наличие в открытом доступе аннотаций (рефератов) и/или полных текстов статей, включение статей из журнала в различные базы данных и другие.

Обратившись к зарубежному опыту, мы увидели, что готового решения для подобного архива нет. Существует богатый опыт архивирования всего Интернета или его региональной части, также в некоторых странах архивируют отдельные сайты, отобранные вручную[1,2]. Наиболее похожим проектом является «Архив шведских журналов» http://www.kb.se/Nbp/el-perE.htm. В открытом доступе опубликован каталог шведских онлайновых журналов, а в закрытом архиве (закрытость определяется законами об авторском праве) хранятся копии журналов. Несмотря на довольно тесные контакты со шведскими коллегами, мы не смогли воспользоваться их опытом, поскольку они имеют строго определенную периодичность скачивания для каждого журнала и скачивают его обновления, в сотрудничестве с редакцией журнала. В программе деятельности Международного консорциума по сохранности Интернет ресурсов (International Internet Preservation Consortium) [3] предполагалось создание «интеллектуального» робота, позволяющего самостоятельно определять необходимую частоту скачивания сайтов, однако у них на сайте нет обновлений с 20 июля 2004 года.

Среди российских проектов можно отметить www.eLibrary.ru, создатели которого тоже создают полнотекстовый архив российских журналов. Эту

Труды 7ой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - RCDL'2005, Ярославль, Россия, 2005.

задачу они предполагают решить путем тесного сотрудничества с редакциями журналов и использованием специального программного обеспечения для подготовки XML размеченного текста журнала. Такой текст впоследствии можно будет как загружать в электронную библиотеку, так и выпускать в печатном виде. В отличие от предыдущих проектов, мы же хотели в условиях минимизации ручного труда, архивировать многочисленные онлайновые журналы, автоматически отслеживая происходящие изменения и отбрасывая лишнюю информацию с сайтов, на которых они размещены. В том числе, нам бы хотелось охватить также ряд журналов, не имеющих постоянной редколлегии, выпускаемых группами энтузиастов, с которыми трудно налаживать договорные отношения, но, которые тоже являются частью нашей российской науки и культуры.

Проект выполняется с 2004 года. К настоящему времени выполнены следующие работы по проекту:

2. Разработка критериев отбора журналов и методики их описания

Считается журналом сетевой ресурс, представляющий содержательную информацию (от содержаний до полных текстов) о печатном журнале, регулярно выходящем не менее года. Сетевой ресурс, не являющийся электронной версией печатного журнала, считается журналом при выполнении следующих условий:

- ресурс имеет определенную тематическую направленность;
- материалы объединяются в номера (выпуски); выпуски могут выходить нерегулярно, однако не реже раза в год;
- если ресурс имеет фиксированную периодичность, она должна составлять не менее недели, причем каждый выпуск должен содержать материалы, значительные по объему и содержанию

Разработан формат описания электронных журналов, основанный на международном стандарте Dublin Core Metadata Element Set (DC MES) и методика заполнения полей формата описания.

Сформирован тестовый массив описаний журналов, присутствующих в Интернете, отражающий все разнообразие их представления: от кратких описаний до больших массивов полных текстов.

3. Изучена структура онлайновых журналов

Поскольку одна из главных задач проекта – разработка максимально автоматизированной технологии архивирования, мы исследовали возможность создания системы автоматического скачивания новых выпусков по мере их поступления на основе прогнозирования имен файлов для новых выпусков. Большинство журналов (примерно 50% из просмотренных) организуют архив журнала одним из следующих способов:

1. На одной странице (архив журнала) размещаются ссылки на все опубликованные номера журналов. По имени каждого файла (или пути к нему) можно определить год выпуска, № тома или выпуска и т.п.

2. Все ссылки на номера журнала находятся на одной странице, при этом все файлы с оглавлениями журнала находятся в одной папке, а полные тексты статей находятся в другой папке.

3. Для каждого года или для каждого выпуска журнала создается новая папка. Полные тексты статей (если таковые имеются) обычно размещаются в этой же папке, но бывают случаи, когда они хранятся отдельно от оглавления журнала (в других папках).

4. В некоторых журналах новые номера добавляются в виде картинок (как правило, обложек журнал), с которых делается ссылка на нужный файл.

Рассмотренные выше способы организации архивов удобны для архивирования т.к. легко определить название файла или папки, которые должны появиться при выпуске следующего номера журнала. Но даже при такой структуре архива возникают непредвиденные ситуации: журнал может выходить с разной периодичностью, иногда выходят сдвоенные номера журналов, помимо архива номеров, журнал может содержать и другие материалы, встречаются журналы, публикующие специальные выпуски.

4. Программное обеспечение для архива

Основными проблемами, которые нам предстояло решить, были: как скачивать и как потом хранить. После изучения рекомендаций OSI по выбору программного обеспечения для создания институтских репозитариев были определены основные требования к нашей системе (поддержка OAI-MH и DC), и выбрана система GREENSTONE (http://www.greenstone.org). Эта система распространяется с открытыми кодами, и, благодаря этому, нам удалось доработать отдельные модули для решения наших задач

В процессе эксплуатации ПО Greenstone оказалось, что возможности пакета не позволяют создавать полностью функциональные архивы сайтов. Причинами этого являются нерациональное использование программы скачки сайтов wget и некорректная поддержка кодировок русского языка. Мы выделили процесс скачивания сайтов в отдельный программный модуль. Этот модуль был написан на языке PHP и использовал более совершенную версию программы wget 1.9.1. Основной задачей модуля было скачивание сайта и подготовка его для дальнейшей обработки пакетом Greenstone. В про-

цессе скачивания HTML-документы дополняются информацией (метаданными) как о самом процессе скачивания (например, дата скачивания документа), так и самом документе (кодировка документа, язык документа). Эти метаданные извлекаются из заголовков протокола HTTP, выдававшихся сервером. При скачивании автоматически удаляются рекламные баннеры и другая информация, не имеющая отношения к журналу. Изложенные в предыдущей части варианты структуры журналов были учтены при создании скачивающего модуля и структуре базы метаданных. Для полного и неизбыточного скачивания приходится задавать разрешение на скачивание внешних ссылок заданной глубины.

Помещение метаданных о языке и кодировке HTML-документа в сам документ позволило решить проблему с поддержкой кодировок русского языка пакетом Greenstone.

В целом, работа с пакетом Greenstone производит, благоприятное впечатление, среди достоинств пакета можно указать его цену (он бесплатный), поддержку платформ Windows и Linux, наличие подробной документации на русском языке, наличие оперативной и бесплатной технической поддержки со стороны разработчиков в специальных списках рассылки, протоколов Z39.50 и OAI.

5. Юридические проблемы

Приступая к созданию архива, мы ставили перед собой задачу выполнения всех требований законодательства по авторскому праву, но как показал опыт, это оказалось довольно сложно. Причем основная сложность в довольно сильной правовой неграмотности большинства российских авторов и издателей и более того в нежелании что то менять в сложившейся ситуации. Нами был разработан типовой договор на разрешение скачивания журнала и размещение его в открытом доступе в архиве в Российской государственной библиотеке. Этот договор и письмо с нашими предложениями мы разослали в 100 журналов. Большинство ответивших искренне недоумевало, зачем нужен договор, если их журнал, находящийся в открытом доступе, можно скачать и так. Выразившие же согласие подписать такой договор, как оказалось, не имели права этого делать, поскольку ни в одном из журналов не заключалось авторских договоров. Лишь в немногих журналах на сайте была надпись про то, что разрешается перепечатка. Многие журналы, как оказалось, издаются неформальным объединением, не имеющим юридического лица и непонятно с кем нужно заключать договор в таком случае. Все эти проблемы пока находятся в стадии решения, и, надеюсь, в ближайшее время мы их решим.

Литература:

1. Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004, Bath, UK, September 12-17, 2004. Proceeding./Editors: Rachel Heery, Liz Lyon ISBN: 3-540-23013-0

2. 4th International Web Archiving Workshop (IWAW04), http://www.iwaw.net/

3. International Internet Preservation Consortium, http://www.netpreserve.org/

Archiving of the Russian scientific and technical on-line magazines

Libraries and individual users of the journal periodical press it is especially important to know about new forms of existence of magazines, namely about the electronic magazines placed on the Internet. So the project «Creation of archive of the Russian scientific and technical magazines published on the Internet" was conceived at the Russian State Library.

The basis of the portal is the database of Russian scientific on-line magazines descriptions. We developed the selection criteria for magazines, the format of their description on the basis of international standard Dublin Core Metadata Element Set (DC MES) and the technique of filling the fields of the description format. We distinguished the basic types of the on-line journal sites structure to maintain the maximum of automatic work of the software. We learnt the OSI recommendations for choosing the software for institute repositories and developed the requirements to the system. We chose the system GREENSTONE which is free of charge and open source. It allowed us to complete some modules. It was rather difficult to meet all the requirements of the legislation under the copyright. These problems are now being solved.