БИБЛИОГРАФИЧЕСКИЕ БАЗЫ ДАННЫХ, СОДЕРЖАЩИЕ ССЫЛКИ НА ПОЛНЫЙ ТЕКСТ ДОКУМЕНТА

Е. В. Ковязина
Институт вычислительного моделирования СО РАН, Красноярск, Россия
e-mail: ted@icm.krasn.ru

This contribution describes specifics of the design and maintenance of the bibliographic databases which contain references to the full texts in the form of URL or file name. The experience of creation of such databases at ICM SB RAS library is presented. Software aimed the automation of the bibliographic description based on a full text as well as integration of these descriptions to the IRBIS system is described.

Наиболее востребованной формой информационного обслуживания читателей библиотеки является предоставление им полного текста необходимого документа в электронном виде. Предварительно нужно отыскать такой документ среди других, имеющихся на сервере библиотеки или в сети Интернет. Необходимым условием эффективности такого поиска является наличие стандартных описаний таких документов, собранных в единую базу данных или несколько баз данных, разделенных по тематическому признаку. Вследствие этого актуальным направлением работы научной библиотеки является формирование библиографических баз данных, в каждой записи которых содержится ссылка на полный текст документа. В публикациях часто принято называть такие базы данных полнотекстовыми, хотя они не являются таковыми в полном смысле этого слова, так как обычно в них не предусмотрен поиск по всему тексту документа. Для обеспечения многоаспектного поиска по документу в таких базах данных требуется описание его содержания в ключевых словах, предметных рубриках и т. п.

В библиотеках научно-исследовательских учреждений или вузов такие базы данных могут быть условно разделены на две группы в зависимости от того, кем и где опубликована работа:

— базы данных, производимые организацией (труды сотрудников, публикации организации);

— базы данных, потребляемые организацией (тематические базы данных, формируемые в соответствии с направлениями исследований организации или ее отдельных подразделений).

Полные тексты документов при этом хранятся в виде файлов различных типов в локальной сети организации или представляют собой URL-ссылки на ресурсы, находящиеся в сети Интернет. Заметим, что если локальная сеть построена по технологии Ethernet, то

документы, хранящиеся на серверах локальной сети, не отличаются доступом от ресурсов Интернет.

Возможности серверов корпоративной сети Красноярского научного центра СО РАН позволяют хранить достаточно большое количество информации. В Институте вычислительного моделирования СО РАН дисковое пространство предоставляется в пользование сотрудникам научных подразделений и служит для хранения информации как собственного производства, так и найденной в Интернет. Существенную часть этой информации составляют научные статьи и книги. В течение ряда лет эта информация просто накапливалась без сортировки и использовалась каждым сотрудником индивидуально по мере необходимости. В результате был стихийно сформирован обширный банк полезной информации, которая не могла далее быть эффективно используема вследствие большого объема и отсутствия средств поиска. Администрацией институтов была инициирована работа по сортировке имеющихся публикаций, в ходе которой объем информации был существенно сокращен вследствие устранения дублирования и удаления утративших актуальность публикаций. Из получившегося объема документов были выделены ресурсы собственного производства (статьи, монографии), которые были внесены в базу данных трудов сотрудников института и публикаций института. Оставшиеся документы были разделены по тематике направлений исследования института и переданы в библиотеку для формирования тематических баз данных.

Для ведения библиографических баз данных в институте применяется АБИС ИРБИС, где для привязки файла, содержащего полный текст документа, используется повторяющее поле, в котором документ описывается как файл или URL. Для базы данных безразлично, в каком формате хранится такой файл. Напротив, для публикации ресурсов в сети Интернет необходимым является требование, чтобы формат, в котором хранится файл, поддерживался большинством браузеров Интернет, хотя следует отметить, что современные технологии определяют возможность понимания любых форматов, если на компьютере имеется соответствующее программное обеспечение. Таким образом, для обеспечения совместимости документы, предназначенные для обработки, представлены ограниченным набором форматов, а именно HTML, PDF и PostScript — для текстов, JPEG и GIF — для графики.

Так как количество документов достаточно велико, наиболее актуально уменьшение трудозатрат при формировании баз данных. Снизить эти затраты могла бы автоматизация описания электронных документов. В процессе решения этой задачи были выделены три последовательных этапа автоматизации:

— распознавание в тексте документа отдельных элементов библиографического описания;

— конвертирование выделенных элементов в формат хранения АБИС;

— редактирование и дополнение полученных автоматически описаний средствами АБИС.

Редактирование и дополнение описаний производились вручную уже после конвертирования их в АБИС и технологически не представляли проблем. На этапе конвертирования в результате работы программы из выделенных элементов описания формировалась текстовая строка в формате ИРБИС, которую затем импортировали в систему из текста. Элементы описания готовились в текстовом виде и разносились по полям вне зависимости от того, из какого файла они были получены. Следовательно, на данном этапе программное обеспечение было едино для файлов различных форматов.

С наибольшим количеством проблем пришлось столкнуться на этапе распознавания

и выделения элементов. Во-первых, форматы файлов принципиально различны, следовательно, для обработки файлов различных форматов требовались и абсолютно различающиеся программы. При определении возможности автоматизации были визуально исследованы имеющиеся документы с целью выяснения признаков, по которым можно выделить отдельные элементы библиографического описания. Для исследования были выбраны документы в форматах HTML и PDF как наиболее часто используемые. Выяснилось, что для распознавания требуется не только указание формата, но и тип документа — является ли он книгой или статьей из журнала или сборника, так как эти два типа документов имеют различный стиль оформления. Был проведен анализ наиболее типичного оформления документов для каждого типа и последующего выделения элементов библиографического описания по шрифту, местоположению в тексте или контексту. При визуальном исследовании выяснилось, что выделение по внешнему виду возможно (причем оно не всегда однозначно определено) только для следующих элементов:

— заглавие;

— авторы (разделялись по запятой);

— описание источника статьи (журнал или сборник): заглавие, том, номер, год;

— ISBN или ISSN (путем поиска соответствующего вхождения);

— URL или имя файла и путь;

— ключевые слова, если они выделены соответствующим словом (keywords или ключевые слова);

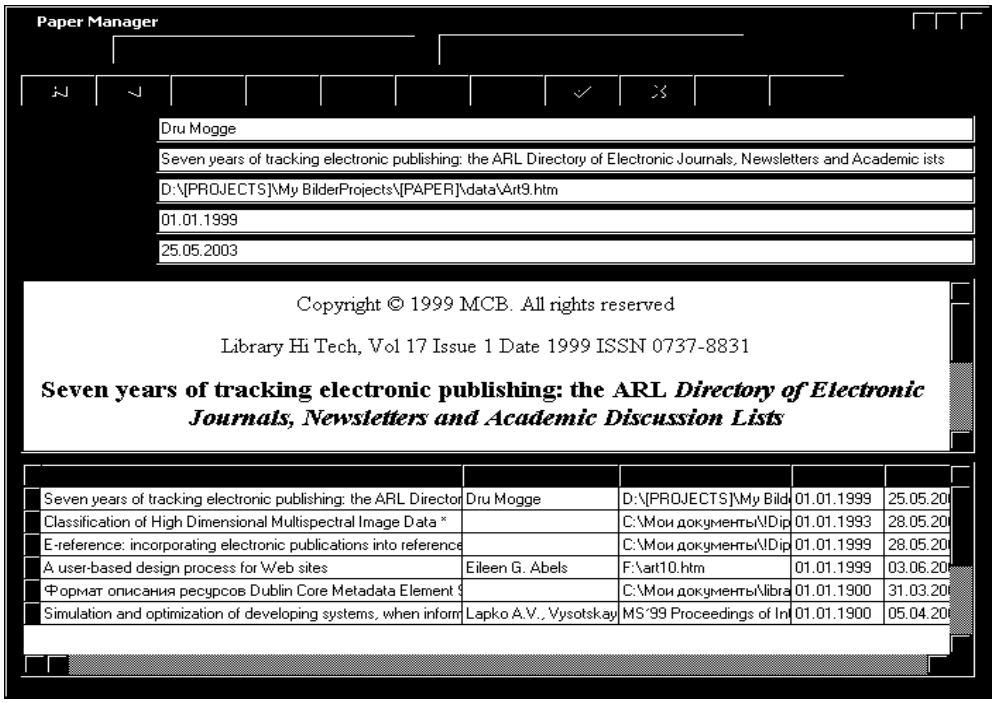
— аннотация, если она размещена после заглавия.

При просмотре кодов HTML определились два различных вида документов, в соответствии с которыми и производилась их обработка:

— документы, не имеющие содержательных метаданных;

— документы с метаданными.

В документах первого вида распознавание производилось по тегам или контексту. При этом работа осложнялась различиями в оформлении статей, для которого не существует никаких оговоренных последовательностей написания отдельных частей, входящих в библиографическую запись. Оформление электронных статей обычно повторяет оформление соответствующего печатного издания, если оно есть. Как следствие, выделенные элементы данных часто не соответствуют действительности. К счастью, хорошим тоном становится снабжение электронных страниц метаданными, определяемыми в HTML-кодах тегом META, что позволяет отнести их ко второму типу. Следует отметить, что руководствоваться только тегом META нельзя, поскольку такие команды часто формируются различными инструментами моделирования страниц и содержат только сведения о самом таком инструменте. Необходима проверка того, что данный тег содержит необходимые пары имя — значение, а именно параметр name должен содержать значения author, description, language_of_resource, originator, subject и т. п., определяемые языком HTML. Тогда параметр content содержит значение соответствующего элемента библиографического описания. Продвинутые электронные издания, предоставляющие статьи в html-виде, снабжают их метаданными в формате Dublin Core Metadata Element Set. Характерным признаком этого формата является то, что все значения параметра name начинаются с “DC.”, а далее следует имя элемента, определенное стандартом. Инструкция на русском языке по формированию метаданных электронного документа (краткая форма) находится на сайте Российской государственной библиотеки по адресу: http://www.rsl.ru/dc/Instr_s.htm и содержит перечень имен элементов, используемых в метаданных, и пояснения к ним. Наличие метаданных позволяет значительно расширить количество элементов библиографи-



Пример части документа на языке HTML, содержащий теги метаданных.

ческого описания и получить в результате более полную запись, не прибегая к анализу текста документа. Пример части документа на языке HTML, содержащий теги метаданных, приведен на рисунке.

Формат PDF не содержит метаданных. Однако принятой в Интернет формой хранения таких документов является описание документа на странице, содержащей затем ссылку на файл PDF. Если такая страница имеется, то, как правило, метаданные содержатся в ней и могут быть извлечены теми же средствами, что и из HTML-файла. Текст статьи в формате PDF хранится в закодированном виде, поэтому при работе с этим форматом статья сначала копировалась в буфер обмена из Acrobat Reader, а затем по тексту из буфера производилось распознавание. Возникающие в дальнейшем проблемы не отличаются от таковых для HTML-файлов без метаданных.

Для обработки была выбрана тестовая совокупность статей, хранящихся на сервере библиотеки ИВМ СО РАН, в количестве 204 документов. Метаданные имелись в 12 статьях, все они были корректно обработаны и преобразованы в записи ИРБИС. В статьях без метаданных вся совокупность признаков была выделена только из 62 статей. В оставшихся статьях отчетливо выделялось только заглавие и URL, иногда авторы. Из этого следует сделать вывод, что только около одной трети имеющихся в Интернет статей поддается автоматизации описания. Следовательно, эффективность применения такого программного обеспечения низка, хотя существуют еще возможности дальнейшего улучшения программы и повышения качества полученных описаний. Все же следует отметить полезность использования данной программы, так как кроме автоматического выделения элементов описания она позволяет выделять элементы вручную в собственном окне редактирования и разносить их по полям описания, что несколько ускоряет работу с документом.

Поступила в редакцию 18 марта 2005 г.