Доставка документов как средство пополнения электронных библиотек России

Барышева Ольга Владимировна,
кандидат филологических наук по специальности «Информатика»,
ведущий инженер Отдела автоматизации Российской национальной библиотеки (Санкт-Петербург)

Цели и задачи доставки документов и электронных библиотек

Сегодня можно с уверенностью констатировать, что работа служб доставки документов в России стала реальностью. Несмотря на некоторые еще имеющиеся проблемы финансового и правового характера библиотеки, центры научной информации и коммерческие фирмы принимают и выполняют заказы, создавая и пересылая друг другу электронные копии документов. Услуги по доставке документов широко рекламируются в Интернет, реально становясь фактом сетевого взаимодействия.

Соответственно, в каждой организации, оказывающей услуги по электронной доставке документов, хранится определенное (пока еще незначительное) количество копий выполненных заказов в виде файлов разного формата. Многие службы предпочитают удалять сделанные копии, и главной причиной этого является отсутствие ответа на вопрос: что с ними делать дальше?, а не недостаток памяти для их хранения. Действительно, этот вопрос можно было считать правомерным до появления такого сетевого феномена как электронные библиотеки.

В основу концепции создания электронных библиотек положен принцип своевременного обеспечения пользователей необходимой информацией по скоростным каналам связи через телекоммуникационные сети. Не менее важным являются и принципы отбора материала, т.е. наполнения электронных библиотек. Документы в электронных библиотеках должны быть источником информацией качественной, верифицированной, публикуемой в авторитетных источниках (например, в научных изданиях, имеющей аннотации и рефераты). Безусловно, этим не может и не должен исчерпываться их профиль, т.к. в традиционных библиотеках есть и массовая литература, и желтая пресса, и т.п. Существующие крупнейшие электронные библиотеки можно разделить на: частные (типа библиотеки Мошкова), государственные (например, ОРЕЛ) и коммерческие (например, Национальная электронная библиотека). Профильными для большинства являются произведения художественной (и не очень) литературы, научные журналы и информационные источники. Большинство электронных библиотек имеют культурно-развлекательный и научно-образовательный уклон.

Что касается доставки документов, то тематический профиль заказываемых копий определяется недостаточно четко, но некоторые тенденции все же прослеживаются. Так, по данным Британской библиотеки[1] тематика 69% заказов - естественные и технические науки, 19% - общественная литература, 9% - гуманитарные науки, 3% - не определяется.

Из этого можно сделать вывод, что в области научной литературы профили работ служб доставки документов и электронных библиотек совпадают, как совпадает и цель их работы – быстро предоставить нужную информацию.

Хранение электронных документов

Еще одной общей чертой электронных библиотек и служб доставки является то, что документы для них создаются одними и теми же средствами – они не появляются как электронные, а копируются с традиционных в электронной форме с помощью различных технических средств – рулонных, планшетных или листовых сканеров, цифровых аппаратов и камер. Строго говоря, один из видов доставки документов – предварительное сканирование – тоже может быть назван электронной библиотекой.

Каждая электронная библиотека с технической точки зрения есть набор файлов и программ для их интерпретации. И для нее (как и при доставке документов) могут изготавливаться файлы, имеющие разные характеристики: качество изображений или текстов, форматы, кодировки. Также и наборы документов могут быть представлены как файловая система или база данных. Тут нет и не может быть никакого регламента – все определяется в каждом конкретном случае, исходя из принципа целесообразности. То же относится и к протоколам обмена данными – кто-то предпочитает ftp, кто-то http или Z39.50. Все эти технические подробности абсолютно не должны влиять на доступ для конечного пользователя.

Механизм доступа к электронным документам

Доступ к документу через службу доставки достаточно прост: есть модуль заказа с возможностью выбора наиболее предпочтительного формата копии и способа ее передачи. После этого пользователь может получить нужный ему документ, например, по электронной почте.

Доступ к документу электронной библиотеки еще проще, поскольку он уже создан, достаточно просто нажать кнопку, и на экране появится требуемая копия.

Но это все применимо только в случаях, когда заказчик / пользователь / читатель точно знает все характеристики документа, т.е. имеет его исчерпывающее описание. Согласитесь, это бывает крайне редко. Соответственно, кроме хранилища информации и модуля доступа, т.е. какого-либо сетевого интерфейса, необходим еще и механизм поиска. Тут есть практически всего два варианта: поиск по каким-то единообразным описаниям либо по полным текстам документов. В ряде случаев полнотекстовый поиск просто невозможен: например, если речь идет о графике или текстовые файлы хранятся на сервере в виде нераспознанных изображений. Именно поэтому основной задачей и для поисковых систем Интернет, и для электронных библиотек стало создание разнообразных схем для обработки электронных документов (их принято называть метаданными). Приведем несколько примеров:

- 8 июня 1994 года Федеральным комитетом географических данных США[2] (Federal Geographic Data Committee) разработаны правила приведения метаданных, получившие название Стандарт метаданных FGDC. Он содержит 334 различных элемента, 119 из которых существуют только для того, чтобы содержать другие элементы, что необходимо для описания связей между другими элементами;
- В 1996 году техническим комитетом ИСО[3] 211 начата разработка Стандарта метаданных ИСО (проект 15046-15);
- Кроме того, существуют так называемые HTML или HTTP метаданные, поскольку об их возможном наличии упоминается в спецификациях (RFC 1866 и 2616 соответственно). Речь идет обо всем известных тэгах <meta>. Например, “http-equiv” имеет до 50 атрибутов. Реально же используются около 35, причем более 20 менее, чем в 1% случаев. Подобной схемой пользуется большинство поисковых систем Интернет, а также, в частности, электронные библиотеки, входящие в проект Compulib[4], для поиска через AltaVista.

Помимо этого, существует международное инициативное движение по метаданным, под эгидой которого создается модель описания электронных ресурсов, названная Dublin Core – Дублинское ядро. На его основе уже ведется 13 проектов, а сам Dublin Core переведен на более, чем 20 языков мира. Основу его составляют 15 элементов, определенные в RFC 2413[5].

Имя элемента	Идентификатор	Дефиниция - Определение	Комментарии
Название	Title	Имя, данное ресурсу	Обычно *название* – это имя, под которым ресурс известен
Созда-тель	Creator	Лицо (лица), несущее первичную ответственность за создание и содержание ресурса	Примеры *создателя* включают персону, организацию или службу. Обычно имя создателя следует использовать для индикации объекта описания
Предмет и ключевые слова	Subject	Предметная область, определяющая содержание ресурса	Обычно *предмет* выражается с помощью *ключевых слов*, ключевых фраз или кодов классификаций, которые описывают тематическую принадлежность ресурса.
Описание	Description	Сообщение о содержании ресурса	*Описание* может включать (но не ограничивается): реферат, оглавление, ссылки на графическое представление содержания или простое текстовое изложение содержания
Издатель	Publisher	Лицо (лица), несущее ответственность за ввод ресурса в обращение	Примеры *издателя* включают персону, организацию или службу. Обычно имя издателя следует использовать для индикации объекта описания

Соиспол-нитель	Contribu tor	Лицо (лица), несущее ответственност ь за содействие в создании содержания ресурса	Примеры *соисполнителя* включают персону, организацию или службу. Обычно имя соисполнителя следует использовать для индикации объекта описания
Дата	Date	Дата, связанная с событием в жизненном цикле ресурса	Обычно дата ассоциируется с созданием или доступностью ресурса. Рекомендуемое для практического использования при кодировке значение *даты* определено в профиле ИСО 8601 и поддерживает формат ГГГГ-ММ-ДД
Тип ресурса	Type	Свойство или жанр содержания ресурса	*Тип* включает термины, включающие общие категории, функции, жанры или объединенные уровни содержания. Для практического использования рекомендуется выбирать значение из контролируемого словаря (e.g. DCT[4]). Для описания физического или цифрового представления ресурса используется элемент *формат*
Формат	Format	Физическое или цифровое представление ресурса	Обычно *формат* может включать тип копии (медиатип) или величину ресурса. *Формат* может использоваться для определения технического и программного обеспечения или другого оборудования, необходимого для отображения или управления ресурсом. Примеры величины включают размер и продолжительность. Для практического использования рекомендуется выбирать значение из контролируемого словаря (e.g. MIME[5])
Идентифик атор ресурса	Identi fier	Однозначная ссылка на ресурс в пределах данного контекста	Для практического использования рекомендуется идентифицировать ресурс посредством строки или числа, соответствующего формальной идентификационной системе (URI[6], URL[7], DOI[8], ISBN[9])
Источник	Source	Ссылка на тот ресурс, из которого извлечен настоящий	Настоящий ресурс может быть извлечен из *источника* целиком или частично. Для практического использования рекомендуется идентифицировать ресурс посредством строки или числа, соответствующего формальной идентификационной системе

Язык	Language	Язык интеллекту-ального содержания ресурса.	Для практического использования рекомендуется значение элемента *язык,* определяемое RFC 1766, включающим двухбуквенные коды языков (взятые из стандарта ИСО 639), со следующими факультативно двухбуквенными кодами стран (взятыми из стандарта ИСО 3166[10]). Например, «en» - для английского, «fr» - для французского, «en-uk» - для английского, используемого в Великобритании.
Отноше-ние	Relation	Ссылка на родственные ресурсы	Для практического использования рекомендуется идентифицировать ресурс посредством строки или числа, соответствующего формальной идентификационной системе
Охват	Coverage	Протяжен-ность и границы содержания ресурса	*Охват* обыкновенно включает пространственное местонахождение (название местности или географические координаты), временной промежуток (временная метка, дата или диапазон дат) или юрисдикцию (такую, как названное административное подразделение). На практике рекомендуется выбирать значение из контролируемого словаря (например, Тезауруса географических названий), т.е. целесообразнее использовать названия местностей и периодов времени вместо цифровых идентификаторов (таких, как системы координат или диапазоны дат)
Правовое регулирова ние	Rights	Информация о правах по ограничению доступа и охране ресурса	Обычно элемент *права* содержит положение о правовых нормах, регулирующих функционирование ресурса, или ссылку на службу, предоставляющую эту информацию. Правовая информация обычно включает сведения о правах на интеллектуальную собственность, авторском праве и других имущественных правах. Отсутствие элемента *права* не может являться основанием для каких-либо предположений о правовом статусе относительно ресурса

Каждый элемент определяется с помощью набора из 10 атрибутов по стандарту ИСО ISO/IEC 11179[13] для описания элементов данных.

Трудно сказать наверняка, какой набор метаданных лучше (если их вообще можно сравнивать), какой будет больше использоваться, и какой будет наиболее эффективным при поиске. На сегодняшний день Dublin Core представляется наиболее перспективным, ибо приложим практически ко всем видам электронных документов и доступен для интерпретации как машине, так и человеку, и, кроме того, интернационален. Более того, возможность и необходимость создания профилей для метаданных (а в подобном качестве может легко выступать Dublin Core) декларируется в спецификации языка HTML в версии 4.01, рекомендованной W3 Консорциумом 24 декабря 1999 года[14].

В любом случае электронными документами без описаний практически нельзя пользоваться, т.к. их невозможно найти. А документы для доставки обязательно снабжаются описаниями, соответственно, они могут легко стать частью электронной библиотеки. Задача только в том, чтобы в пределах одной электронной библиотеки схема метаданных совпадала, вне зависимости от источника пополнения библиотеки, формата и места хранения.

Проблемы использования электронных документов

Мы не будем касаться основных проблем функционирования электронных библиотек – экономической и авторского права, а остановимся лишь на взаимодействии самих документов.

Главная проблема – совмещение традиционных и электронных документов, особенно, если речь идет о библиотеках. Создание сквозных связей между документами (линкование) на основе взаимного цитирования, что давно уже используется Институтом научной информации в Филадельфии (ISI)[15], дает возможность ссылаться на описания традиционных (бумажных) документов, если они описаны в одном из машиночитаемых форматов. Пока что подобный аналог в нашей стране пытается создать лишь Институт научной информации по общественным наукам, хотя было бы неплохо применить эту концепцию к построению электронных библиотек, включающих в себя собственно электронные сетевые документы, электронные копии, создаваемые издательствами, службами доставки или службами сканирования, описания фондов традиционных библиотек. Тогда вопрос: должны ли традиционные библиотеки в новых условиях комплектоваться изданиями и документами или информацией и содержанием? Не будет столь острым.

Наше мнение таково, что библиотеки и службы доставки документов не должны уничтожать сделанные электронные копии. Необходимо лишь, чтобы они снабжались метаданными по одной из существующих схем. С другой стороны, необязательно, чтобы каждая традиционная библиотека создавала свою электронную, просто объединяя единой схемой метаданных разные, путь маленькие базы данных, библиотеки России смогут стать полноправными участниками процесса пополнения электронных библиотек.

---

[1] British library facts and figures <http://www.bl.uk/>;
[2] FGDC standards <http://www.fgdc.gov/standards/standards.html>;
[3] ISO catalogue <http://www.iso.ch/infoe/catinfo.html>;
[4] Compulib <http://www.citycat.ru/compulib/#Kluch>;
[5] Dublin Core Metadata for Resource discovery <http://www.ietf.org/rfc/rfc2413.txt>;

[6] DCT - List of Resource Types: Dublin Core Draft Working Group Report. <http://purl.org/DC/documents/wd-typelist.htm>;
[7] MIME - Internet Media Types. <http://www.isi.edu/in-notes/iana/assignments/media-types/media-types>;
[8] URI, URL - Naming and Addressing: URIs, URLs, ... <http://www.w3.org/Addressing/>;
URI - Uniform Resource Identifiers: Generic Syntax, Internet Draft Standard <http://www.ics.uci.edu/pub/ietf/uri/rfc2396.txt>;
[9] URL - Uniform Resource Locator Specification <http://www.w3.org/Addressing/URL/Overview.html>;
[10] DOI – The Digital Object Identifier <http://www.doi.org/>;
[11] ISBN – International Standard Book Numbering <http://www.reedref.com/standards/>;
[12] ISO 3166 2-letter country codes <http://www.w3.org/International/O-misc-iso3166.html>;
[13] ISO 11179 - Specification and Standardization of Data Elements, Parts 1-6. <ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/11179/>;
[14] HTML 4.01 Specification <http://www.w3.org/TR/html4/cover.html#minitoc>;
[15] ISI Web of Science <http://www.isinet.com/products/citation/wos.html>;