ТЕКУЩИЕ ПРОБЛЕМЫ РОСТА ЭБ РГБ

Перли Б.С.
Российская Государственная библиотека, 101000, Москва,
ул.Воздвиженка, 3/5
e-mail: bsperli@rsl.ru

CURRENT PROBLEMS OF THE DIGITAL LIBRARY OF THE RUSSIAN STATE LIBRARY

Boris S. Perli
RUSSIAN STATE LIBRARY, 3/5 Vozdvishenka str., Moscow, Russia 101000
e-mail: bsperli@rsl.ru

Described are the initial conditions under which the RSL as the first library among Russian foremost libraries is engaged in the setting up of a full-text digital library. Given are the definitions the creators of the digital library go by while solving current tasks of its development as well as criteria of picking up the literature to be included into the stock of the library.
Listed are the demands made towards the premises, recommendations in regard to handling the materials chosen for digitization and towards the equipment it is done on (scanners, computers).

Indicated are the sources of completion of the digital library of the RSL and ways of grouping the electronic documents in the digital library as well as consecutive wordings of the site orel.rsl.ru and measures taken to boost it. Finally a version of the development of the digital library of the RSL as a line of advancing the publishing activity is put forward.

Путь к мудрости прост (говорю без улыбки)
Он мной до конца досконально прослежен
Сначала ошибки, и после ошибки,
И снова ошибки, ошибки, ошибки ...
Но ... реже, и реже, и реже.

В. Левин

ЭЛЕКТРОННАЯ БИБЛИОТЕКА С ТОЧКИ ЗРЕНИЯ РГБ

В настоящее время эпоха восприятия понятия «Электронная библиотека» как разной (иногда глобальной) сложности системы электронных каталогов постепенно сменяется эпохой наполнения русского Интернета полноценной информацией.

РГБ первой из гигантских библиотек страны приступила к созданию полнотекстовой электронной библиотеки. Первой не просто встрети-

лась с проблемами, с которыми встречаются или еще встретятся все, кто создает свои электронные библиотеки, но решает их в максимально усложненных условиях – при объеме фондов в 43 миллиона единиц хранения любое технологическое нововведение усложняется тысячекратно. А ,вдобавок, последние два года главное книгохранилище библиотеки (и страны) закрыто на ремонт.

Провозгласив: «РГБ делает электронную библиотеку!» мы сначала должны были решить — «А ЧТО, собственно, мы делаем?». Ведь можно электронной библиотекой называть коллекцию электронных журналов от Эльзевира, Шпрингера и т.д. Можно затратить массу сил на перевод ВСЕХ каталогов в электронную форму … Слов нет, занятия необходимые и достойные. Но мы в первую очередь — Библиотека — одна из древнейших на земле институций, призванных накапливать знания и предоставлять их желающим этим знанием воспользоваться. Поэтому основными задачами, которыми мы осмелились перед собой поставить, стали:

1. Раскрытие богатств бесценных фондов самой большой библиотеки России — Российской Государственной Библиотеки

Для решения этой задачи ЭБ предоставляет для общественного пользования свои коллекции электронных версий книг, статей, рефератов, нот, коллекций изобразительного, военного и остальных отделов РГБ. В частности собранное с участием европейских библиотек собрание произведений кирилловского шрифта XV – XVI вв.

2. Представление русскому читателю электронных версий основных произведений русской классической литературы, собранных профессионалами и любителями - ценителями русской литературы.

В соответствии с «Концепцией Электронной библиотеки Российской государственной библиотеки», принятой в 2002 г., электронная библиотека, или цифровая библиотека (ЭБ) представляет собой:

Вид информационных систем, в котором документы хранятся и могут использоваться в машиночитаемой («электронной») форме, причем программными средствами обеспечивается единый интерфейс доступа из одной точки к электронным документам, содержащим тексты и изображения. База данных ЭБ может состоять из различного вида электронных коллекций документов. Электронные издания на оптических компакт-дисках включаются в ЭБ только при условии, если Библиотека выставляет их в сети (локальной или глобальной), обеспечивая ту же систему доступа и поиска, что и к остальным документам ЭБ.

Содержание доклада посвящено преимущественно аспекту создания контента[1] ЭБ – тому из-за чего сегодня на сайт orel.rsl.ru обращаются от 2 по 4 тыс. посетителей в день.

[1] Да простят нас ревнители русского языка, но это одно из слов, приобретших с возникновением Интернета совершенно определенный смысл, отличный от привычного «content» = содержание (оглавление). Оно, скорее, соответствует русскому «наполнение», но емче его. Т.ч. еще раз, простите.

КРИТЕРИИ ОТБОРА КОЛЛЕКЦИЙ

В первую очередь перед РГБ встала проблема отбора книг для включения в электронную библиотеку. С чего начинать в океане из 43 миллионов книг?!

Тут очевидные критерии вытекли, просто, из названия учреждения: «Российская Государственная библиотека». В первую очередь нас - ее сотрудников - должно интересовать то, что связано с Россией, русским языком, русской культурой. Русской государственностью. И, несомненно, Библиотечным делом, поскольку мы являемся головной организацией в стране по данной специальности. Ну и, конечно, диссертации, потому что РГБ является депозитарием диссертаций по большому количеству дисциплин. В результате отбор приоритетных коллекций документов для включения в ЭБ привело к следующим темам:

1. Россия. История и культура
    i. А. Москва. История и культура
2. Образование
    a. А. Высшее
    b. Б. Среднее
3. Библиотечное дело
4. Диссертации

Следующей проблемой стала задача постоянного пополнения коллекций. На первых порах (и эта практика продолжается, но менее интенсивно) мы широко использовали тексты из Интернета. Но оказалось, что эти тексты далеко не всегда качественны и комплектны – на 2-х десятках сайтов «Дон-Кихот» был оборван на 42 главе (из 74). А чтобы включить выявленные тексты в свою ЭБ, их все равно надо просматривать, убивать рекламу, несуразную разбивку, чуть ли не вычитывать заново. При этом, не забудьте, что далеко не все, как Мошков и Пескин, дают ссылку на то, какое именно издание оцифровывалось. Как прикажете составлять библиографическую запись!? То, что допустимо для любителя, не может себе позволить РГБ.

Но, самое главное, 45000 проиндексированных сегодня YANDEX-ом книг не дают сегодня исчерпывающей коллекции ЭД ни по одной специальности. Даже для достаточно невысоких требований средней школы. Мы сложили списки для обязательного прочтения в средней школе, гимназии и колледжах гуманитарного направления – вроде бы, в этот список должны были войти основные тексты, совершенно необходимые в Рунете. И оказалось, что в Интернете не было ни Радищева, ни программных произведений Ломоносова, ни «Корсара» Байрона …
Уж эти-то лакуны надо было восполнять в первую очередь.

Поэтому пришлось осваивать и развивать оцифровку собственными силами. Дальше речь будет идти об ЭД, сделанных нами или по нашим заказам.

Итак, выбранная экспертами и утвержденная Советом директоров книга должна поступить на СКАНИРОВАНИЕ

Но до непосредственно сканирования надо произвести экспертную оценку состояния коллекций, предназначенных для оцифровки. Меры по защите оригиналов при их сканировании определяются специалистами по сохранности нашим НИЦ КД.

Их требования:

Ценные библиотечные документы должны сканироваться в том же корпусе, где хранятся. Обязательно в присутствии хранителя.

Для сканирования необходимо охраняемое чистое помещение со столами размерами в шесть раз превышающими самый крупный сканируемый объект. Прямо скажем, нет у нас подходящего стола для карты 2,5 х 1,5 $м^2$.

Во избежание повреждения документов от облучения при сканировании, лампы, освещающие сканируемый объект, должны содержать UV-фильтр или иметь минимальные количество UV.

На рабочем место сканировщика должен быть обеспечен температурный и влажностный режим, соответствующий режиму хранения бумажных подлинников в хранилище.

Анализ оборудования для оцифровывания и особенностей хранящихся в библиотеках документов дал основания рекомендовать:

Оцифровывать на планшетном сканере: копии негативов; плоские бумажные документы в хорошем состоянии без красящих носителей; брошюры и ноты в хорошем состоянии, раскрывающиеся на 180 градусов.

Оцифровывать, используя камеры: книги; фотографии; изоиздания и издания на мелованной бумаге; все документы, которые нельзя безопасно прижимать; большеформатные документы в неудовлетворительном состоянии, переворачивать которые небезопасно; все документы, размеры которых больше планшетного сканера.

Не отбирать на оцифровывание архивные экземпляры.

По мере возможности оцифровывание производить с микрокопий.

Ограничить оцифровывание с оригинала одним разом.

Все эти рекомендации нами выполняются по мере возможности – всегда не хватает средств, сотрудников, площадей. Но самое главное при сканировании редкого документа – принять решение – для нас важнее оставить книгу непереснятой и ждать, пока она погаснет от естественного старения, или, все-таки снять с нее копию (фото, электронную …), чтобы хоть что-то осталось потомкам?!? Самое страшное, что доводилось сканировать, это не книги кирилловского шрифта XV в., а издания 20-х гг. XX в., напечатанные на Дальнем Востоке на бумаге из опилок. Переворачива-

ешь страницу и краешек ее остается на пальцах. А ведь книги еще и погибают от биологических факторов, какие-то от сырости … Вот и принимай решение …

Начинали мы сканировать книги, как и большинство неофитов, на планшетных сканерах. Сейчас на вооружении сканеры А4 и А3 разных марок. Измерения разрушения опытных образцов бумаги на установке искусственного старения после сканирования, достаточно убедительно показывают, что облучение ультрафиолетом и светом на планшетных сканерах даже меньше, чем на бесконтактных. Наверно, это связано с тем, что освещенная полоса проходит через участок страницы достаточно быстро.

Сканируем мы, в основном от 300 до 600 dpi, а когда выставляем изображения в Интернет, переписываем их на 72 – 144 dpi.

Сейчас сканерный парк пополняется бесконтактными Book-Scaner`ами. Первыми из них в РГБ появились BooKYE – стоимость $18000, разрешение 300 dpi, позволяют сканировать в Line Art и Grey. Скорость сканирования до 300 в час. Стол плоский. Прижимных приспособлений нет. Слабое место – лампы. За год меняем 2 – 3 комплекта. Стоимость замены ламп с работой до $250.

Сейчас появляются Минолты стоимостью $28000, разрешение 600 dpi, позволяющие сканировать в Grey. Скорость сканирования до 200 экспонирований в час. В среднем — 150. Стол плавающий, благодаря чему книгу можно при съемке не полностью раскрывать. Прижимных приспособлений также нет. В работе стабильны. Коммерческая фирма ООО «Электронная книга», отсканировавшая в РГБ до 40000 книг, сделала это, в основном, на Минолтах. Понятно, что это были, в основном, черно-белые тексты. Но для сканирования цветных вклеек фирма была вынуждена включить в технологию цветные планшетники.

Вот уже полтора года в РГБ (как и РНБ) работают поставленные Библиотекой Конгресса Phase One – для бесконтактного сканирования в цвете. Стоимость их от $50000.

Сканирование производится с разрешением до 900 dpi, для размещения в рабочем поле оптической системы плотно прошитых книг предусмотрена специальная люлька. Сканирование (в зависимости от разрешения и размера кадра) занимает от 1,5 до 7,5 мин. Иногда мы сканируем в цвете изображения, обычно сканируемые в GREY. Изображение получается глубже, лучше передается фактура носителя (бумаги, ткани). Если просвечивают знаки оборотной стороны, под лист подкладывается черная бумага. Сейчас в РГБ работает второй Phase One, уже принадлежащий нам.

В настоящее время вступает в строй роторная линия – для сканирования листовых материалов. Линия устанавливается в Химках для массового сканирования диссертаций. Последние расшиваются, листы пропускаются через сканер, после чего диссертации вновь переплетаются.

При массовой оцифровке документов естественно встает вопрос о соотношении между числом распознанных книг и предоставляемых читателям в картинках. Наглядным примером может служить та же ООО «Электронная книга». Начинали они с того, что считали себя в силах полностью распознать все, что сканируют. При 10 сканерах посадили до сотни распознавальшиков – редакторов. Но сколько ни усугубляй нормы, больше чем 5000 знаков корректор за час не вычитает. И отставание вычитанных книг от отсканированных стало настолько разительным, что в результате фирма перешла на изготовление т.н. «быстрых книг»: изображения (tiff), полученные от Минолт, проверяются, обрезаются (если надо, поворачиваются), переводятся в JPG, а оттуда в PDF. Каждый такой файл снабжается титульным листом, выполненным по ГОСТ … и БЗ. В общем, даже изготовление «быстрой книги» занимает от 3 до 6 час.

В результате за прошлый год в отчете ООО «Электронная книга» значилось – отсканировано 1500000 стр., распознано 150 книг.

Проблема обработки отсканированных материалов заставила проанализировать расход времени на каждую операцию с полученным изображением.

КОРРЕКТИРОВКА ПОДХОДА К ВЫБОРУ КОМПЬЮТЕРОВ:

Если изначально (5-6 лет назад) требования к ним были, преимущественно, количественные – их не хватало, то сегодня нам требуются, в основном, компьютеры - графические станции. Потому что файл – карты объемом до 2,5 Гб на ширпотребовских компьютерах, которых большинство в Библиотеке, можно обрабатывать по часу на каждую операцию, не говоря уже просто о перекачке с компьютера на компьютер.
Сегодня у нас несколько машин с конфигурацией:

Системная плата со слотом шины 66 Mhz, 64-бит PCI/ процессор Intel Pentium 4, 1600 Mhz, кэш 512 Kb/ DIMM 1 Gb/ SCSI HDD 40 Gb/ FDD 3.5”/ Video 32 Mb Matrox/ Case ATX/ Optical Mouse/ SCSI CD-ROM 40x/ SCSI CD-Recorder 20x/10x/40x internal/ двухканальный SCSI-контроллер Ultra 160/ сетевая карта PCI 10/100.
Монитор Iiyama Vision Master Pro 452, 19 –21 дюйм.

При этом постепенно мы пришли к соотношению: на 1 сканер – 3-4 компьютера для обработки. С соответствующим числом сотрудников.

СЕГОДНЯШНИЕ ИСТОЧНИКИ ПОПОЛНЕНИЯ ЭБ РГБ

Нам постоянно передают изготовленные ими ЭД (электронные документы) ООО «Электронная книга» и «Русский курьер». «Электронная книга» на 1 августа 2002 г. передала 5200 ЭД — электронных версий книг. «Русский курьер» — около 3000 ЭД — в основном, кусочков статей, книг, диссертаций. Т.к. РГБ по договорам имеет право выставлять передаваемые ей ЭД только в Интранете, то диски, переданные этими коммерческими организациями, установлены в Jukebox`e, доступ к которому открыт с нескольких компьютеров в зоне обслуживания.

Число книг кирилловской печати, выставляемых по программе Память России - достигло в этом году 76. Совместно с РНБ мы готовим полную электронную версию «Ведомостей». По совместному с Библиотекой Конгресса проекту «Встречи на границах» – мы за прошлый год сделали 42 книги. 6 из них мы выставили на своем сайте — на большее не хватило сил. Ничего, через год-два мы их все увидим на сайте www.loc.gov — у них больше сил, денег, но даже они все равно отстают в обработке от обильных поступлений их России.

Продолжается совместная работа с обществом «Мемориал» - нами сделано 30 книг, а выставлено около 70. Интересные метаморфозы претерпел Проект Образование.

Мы заключили 5 договоров с ВУЗАМИ России

На нашем сайте выложены результаты сотрудничества с двумя из них:

Сотрудничество с МИРЭА

Сотрудничество с МИСИ

С нашей точки зрения проект был, безусловно, выгоден всем. Суть его в том, что по отбору ЛЮБОЙ кафедры мы БЕСПЛАТНО сканируем первые 10 книг, а нераспознанные пакеты изображений отдаем вузам. Они (руками студентов) распознают монографии и отдают нам WORD`овский файл. После чего мы сканируем для них одиннадцатую книгу. Современные монографии, которые по закону об авторском праве никто не имеет право выставлять в Интернете, таким образом, попадают на экраны компьютерных классов в стенах библиотек вузов и РГБ.

И, удивительное совпадение, — мы сравнили список отказов по Библиотеке за последние годы с заказами 3-х вузов — они процентов на 60 совпали. Мы по этому списку очень медленно, но движемся, а вот ни одного вуза, который бы стабильно продолжал бы с нами сотрудничество, так и не нашлось. Контакты прекращались после того, как первый диск с 10 книгами передавался нашим контрагентам.

СТРУКТУРА ЭБ РГБ

Организации библиотеки осложняется тем, что по сути дела мы создаем несколько ЭБ —

Первую — знакомую многим присутствующим, — orel.rsl.ru. В нее включено 3000 описанных выше произведений, а также 62 произведения *кирилловского шрифта XV —- первой четверти XVI века* и 14 — *кирилловского шрифта второй четверти XVI века.*

Входят передаваемые нам из отделов и издательств РГБ журналы и диссертации.

Вторая — это библиотека книг, доступных читателям ТОЛЬКО в стенах библиотеки — переданное «Электронной книгой» и «Русским курьером», а также оцифрованное нами, по тем или иным соображениям выставленное для читателей только в Интранете.

Третья — библиотека только для сотрудников. Это архивы, справочные материалы, полуфабрикаты будущих работ.

Подобная структура требует программных решений и определенного устройства Интранета Библиотеки. Из-за огромных объемов потоков РГБ, перманентного ремонта и необходимости поддержания системы при этом в работоспособном (и растущем) состоянии приходится многие решения принимать временными.

Особенность программного обеспечения, сегодня используемого в РГБ, состоит в том, что поиск по каталогу осуществляет программа «Алеф», а полнотекстовая библиотека построена на программе «Библиотека 2000», т.к. «Алеф» в чистом виде не предусматривает пристегивание текстов к БЗ (библиографической записи).

ПРИНЦИПЫ ПОСТРОЕНИЯ И ОФОРМЛЕНИЯ САЙТА

А. Первый вариант – был построен на обращениях к людям Сети, заявках на дальнейшую деятельность и отработках приоритетов.

В. Второй – третий варианты — при сформулированном скелете сайта были перегружены графическими материалами. Но здесь было отработано требование: от первого экрана до книги (добираясь любым способом) не более 2-х экранов

С. Сегодняшний – максимально лаконичный, с привязкой оформления к сайту РГБ.

Обращаем вниманием на постоянное присутствие во всех итерациях сайта, кроме необходимого с точки зрения профессионалов окна поиска, присутствующего на всех любительских сайтах алфавитного указателя. Мы считали необходимым такое дублирование, чтобы посетители сайта orel.rsl.ru, работающие на клавиатурах, не имеющих кириллицы, имели возможность выбрать необходимую книгу.

В следующей модификации сайта вместо алфавитного указателя появится виртуальная клавиатура, которая позволит пользователю в Гонолулу, Аделаиде и Бангкоке набрать на местной клавиатуре слово «Гоголь», и наслаждаться прекрасной русской прозой.

КАК РАСКРУТИТЬ САЙТ?

Для увеличения скорости загрузки мы перешли к многоэтапной загрузке длинных страниц.

Уменьшили и число картинок, и размер фона для сокращения времени общей загрузки сайта.

Зарегистрировали сайт на всех значительных ресурсах Рунета. Несколько изменили политику индексирования сайта более чем на 30 поисковых системах и рейтингах.

Пытаемся создать простую и понятную навигацию по сайту.
Ставим скрытые счетчики. Они не вносят изменений в дизайн сайта, но при участии в рейтингах полезны - дают много посетителей.

ЭЛЕКТРОННАЯ БИБЛИОТЕКА РГБ КАК ЭЛЕКТРОННОЕ ИЗДАТЕЛЬСТВО

Время показало, что мы не слишком ошибаемся в выборе книг, включаемых в ЭБ РГБ. Это хорошо видно на TOP-листе, выставленных нами произведений. На первом месте, конечно соперничают Некрасов с Фрейдом (студенты со школьниками). А вот дальше – изумительные детские стихи Ренаты Мухи, и «Статистика Москвы» Гастева, 1841 г., о которой мы долго думали – распознавать – не распознавать. Как? Три года назад у нас не было Finereader`a со шрифтами Palantino Linotape и Arial Unicode (берем мы их из комплекта WINDOW XP).

Постепенно поднимаются по рейтингу выставленные у нас современные учебники для средней школы. С каждым автором мы заключили договор.

В результате накопления электронных документов мы столкнулись с необходимостью и как-то уже научились обрабатывать, каталогизировать, хранить и создавать сотни CD – дисков с ЭД.

Но, кроме этого, накопленный объем позволил нам самим комплектовать CD-диски, на которые есть социальный или гуманитарный заказ. Это совсем, вроде бы, не библиотечная деятельность. Но в наш маркетинговый век мы все занимаемся тем, на что есть спрос. Кто-то, зарабатывая деньги, а другие во имя престижа своей Библиотеки.

В настоящее время нами подготовлен к тиражированию диск «РГБ-школе». На этом диске-хрестоматии записаны, практически, все литературные произведения, используемые в средней школе. Сейчас мы попол-

няем его, в основном, за счет программ специализированных колледжей и гимназий различных профилей.

Накапливается и материал для диска «Библиотечное дело в России». Уже распознано более 30 книг – «Трудов БВЛ» за много лет. Среди них книги, возвращаемые современному читателю, авторы которых были репрессированы, а единичные экземпляры сохранились благодаря мужеству библиотечных работников. В течение 2002 г. РГБ выложит в Сеть до 100 книг по этой тематике. Желающие смогут прочесть их не только на сайте, но и приобретя диск. Мы считаем, что эта работа будет весьма полезна для библиотечного сообщества.

Уже сейчас можно было бы на основании уже сделанного и намеченного к оцифровке прогнозировать издание подразделениями РГБ компакт-дисков на темы античной литературы, по культурологии, подлинников по истории Москвы.

Ведь кто лучше сотрудников библиотек лучше произведет поиск и подбор книг по данной тематике?! Ведь культура каталогов создавалась столетиями, и системно-предметный поиск не заменить поиском по ключевым словам!!!

А современная техника в современной библиотеке дает возможность любому пользователю стать не только читателем нужных ему книг, но и обладателем ЭД – их копий, вплоть до уникальных творений человечества.

Возможно, что именно это — издание дисков на любой вкус и является одним из главных направлений развития библиотечного дела в стране.