От сканированных изображений к базе знаний. Модель и стратегия научных исследований

Дэвид Бирман[1] и ***Дженнифер Трант***[2], Архивная и музейная информатика, Торонто, Канада

Аннотация

---

[1] Дэвид Бирман является Президентом компании «Архивная и музейная информатика». Его компания консультирует по вопросам электронных записей и архивов, интеграции разноформатных систем информации по культуре и музеям. Он также является издателем-учредителем ежеквартального журнала «Архивная и музейная информатика» нидерландского издательства Kluwer Academic Publishers с 2000 года. С 1991 Дэвид Берман занимается организацией и является председателем Международных конференций по гипермедиа и интерактивным музеям (ICHIM), а с недавних пор – и конференций «Музеи и Веб», а также различными образовательными семинарами и встречами на разные темы. Бирман является автором более 125 книг и статей по вопросам управления информацией в музеях и архивах. До 1986 г. Дэвид Берман занимал должность заместителя директора Смитсониевского института по управлению информационными ресурсами и директора целевой группы по Национальным информационным системам Американской ассоциации архивистов (SAA) с 1980 по 1982 г. С 1987 по 1992 г. являлся председателем Инициативы по компьютерному обмену музейной информацией (CIMI). В 1989 г. Д.Бирман предложил к рассмотрению Принципы политики управления электронными записями, которые были приняты Консультативным комитетом ООН по координации информационных систем (ACCIS) в 1990 году; в 1995 году он предложил исходную модель коммуникаций, приемлемую для бизнеса как часть исследовательской программы по функциональным требованиям к электронным данным. Д. Бирман также работал в качестве директора по стратегии и исследованиям для Консорциума по изображениям для художественных музеев (AMICO). Более подробную биографию и публикации можно найти на сайте http://www.archimuse.com/consulting/bearman_cv.html

[2] Дженнифер Трант является партнером компании «Архивная и музейная информатика», сопредседателем конференций Музеи и Веб и ICHIM, а также работала в программном комитете конференции «Совместные электронные библиотеки» (JDL) в 2001г.; она также была членом совета директоров Комитета по медиа и технологиям Американской ассоциации музеев. Дж.Трант работала исполнительным директором Консорциума по изображениям для художественных музеев (AMICO). Была главным редактором журнала «Архивная и музейная информатика» – ежеквартального издание по культурному наследию издательства Kluwer Academic Publishers с 1997 по 2000 г. До перехода в компанию «Архивная и музейная информатика» в 1997 году Дженнифер Трант занималась развитием коллекций и стандартов в Информационной службе по культуре и искусству Королевского колледжа в Лондоне, Англия. В качестве директора по управлению информацией по искусству она занималась применением технологий в арт-галереях и музеях. Среди ее клиентов – бывший Институт информации Фонда Гетти (в то время Информационная программа по истории искусства Фонда Гетти), для которого она разрабатывала Инициативу по изображению и руководила деятельностью Проекта образовательных сайтов музеев MESL. Она также проводила дисциплинарный обзор и подготавливала отчет для целевой группы по информации по искусству (AITF) под названием «Категории описания произведений искусства для Ассоциации искусства колледжей и AHIP». Будучи специалистом по управлению информацией по искусству, Дж. Трант работала с системами документооборота в крупнейших музеях Канады, включая Национальную галерею Канады и Канадский центр архитектуры, где она разрабатывала и внедряла общие стандарты каталогизации для изданий и рисунков, фотографий и архивных коллекций. Она активно участвовала в определении стандартов музейных данных, участвуя в многочисленных комитетах и регулярно публикуя статьи и презентации по вопросам доступа и интеллектуальной интеграции. В настоящее время ее интересует использование информационных технологий и сетей связи для улучшения доступа к информации о культурном наследии и интеграции культуры в электронные библиотеки для исследования, обучения и досуга. Более подробную биографию и публикации можно найти на сайте http://www.archimuse.com/consulting/trant_cv.html

В следующем десятилетии печатное наследие мира будет оцифровано. Если национальные правительства будут играть в этом процессе активную роль, то он сможет принести значительные выгоды для развития человечества путем демократизации доступа к различным печатным материалам. Предпосылками успеха являются всеобъемлющие программы оцифровки, которые делают общедоступными изображения страниц, а также позволяют комбинировать алгоритмы оптического распознавания текстов (OCR) с декодированием содержания, заложенного в типографских традициях, представляя слова в контексте их функций в документах – в заглавиях, ссылках, подписях к иллюстрациям и т.д.

Связывание сканированных страниц с библиографическими метаданными и использование оптического распознавания текстов – распространенный метод получения дополнительной информации при сканировании книг. Но для извлечения полезного знания, содержащегося в типографских традициях (печать и представление страницы) требуются дальнейшие исследования, чтобы их можно было использовать при декодировании электронных версий печатных книг. В данной работе исследуются некоторые вопросы кодирования информации, заключенной в печатных традициях, и то, как получающиеся в итоге базы знаний и семантический анализ могут использоваться для получения обогащенного культурного контента. Рекомендуемые национальные стратегии могут превратить электронные версии печатных текстов во взаимосвязанные базы знаний и предоставить для всех доступ к печатному наследию, сохраняя его разнообразие.

Предпосылки

На протяжении последних двух десятилетий правительства и частный сектор финансировали ряд небольших выборочных проектов по оцифровке книг и газет (http://ota.ahds.ac.uk/). Исходя из стратегии, разработанной для микрофильмов, эти проекты созывали научные консультационные комитеты, предлагая им выбрать книги, подлежащие оцифровке. Результатами их работы были списки, содержащие несколько сотен или несколько тысяч названий каждый. Всего в рамках таких проектов было отсканировано несколько сот тысяч томов. Этот небольшой объем работ стал причиной возникновения «Проекта миллиона книг» в рамках университета Карнеги Меллон, хотя в ходе этого проекта на июнь 2004 года было отсканировано лишь 50 000 книг (Carnegie Mellon University Libraries & Troll, 2005). В 2004 году динамика сканирования изменилась. Это случилось после того, как Google (http://print.google.com) объявил, что он будет сотрудничать с Университетом штата Мичиган, Гарвардским университетом, Стэнфордским университетом, Публичной библиотекой Нью-Йорка и библиотеками Оксфордского университета в проекте, включающем в себя десятки миллионов книг, который сделает поисковую машину Google более привлекательной (Google, 2005).

Возможность широкого (возможно, даже всеобъемлющего) коммерческого доступа к печатному наследию заставила национальные библиотеки пересмотреть свою политику в свете этой новой потенциальной конкуренции. Обзор затрат по проектам оцифровки показал, что значительная часть средств выделяется в данный момент на отбор книг для оцифровки, т.е. на дискриминационный процесс, по своей сути сокращающий объем работ (http://www.ninch.org/forum/price.resources.html). Изучение Интернет-архива (http://www.archive.org) показывает, что можно значительно сократить затраты на оцифровку, сделав их меньшими, чем покупка в среднем одной копии книги. Это

меняет природу оцифровки и открывает возможность цифрового доступа к печатным коллекциям национальных библиотек (Internet Archive, 2005).

С небольшими затратами на массовую оцифровку национальные библиотеки могут начать планировать крупные проекты оцифровки в общенациональных масштабах. Канадский библиотекарь и архивист Иэн Вилсон возглавляет деятельность в этом направлении, как он сам отметил в заключительном пленарном обращении на Конференции Museums and the WEB – 20053. Параллельные действия планируются также в нескольких европейских странах. Скорость распространения этого подхода такова, что сегодня представляется возможным заявить, что мировая печатная литература будет полностью отсканирована в течение десятилетия (за исключением последних 2-3%, на оцифровку которых могут потребоваться значительные усилия).

Масштабная оцифровка влечет за собой несколько важных последствий. Во-первых, после ее завершения каждый, кто пользуется библиотекой всего интернета, сможет (в принципе) иметь доступ ко всей литературе, независимо от степени удаленности сообщества, в котором он живет. Это означает более полный доступ, чем тот, который имеет на сегодняшний момент какой бы то ни было пользователь, невзирая на его статус или доступ к крупнейшим библиотекам. Во-вторых, это означает, что категория книг с правом издания, но уже распроданных и не переиздающихся4 – а такие книги составляют более трети всемирной литературы – исчезнет, тем не менее, любая книга будет доступна в цифровом формате. В результате с держателями авторских прав необходимо будет заключить новый социальный контракт на разрешение электронного доступа к большей части литературы прошлого века. Вероятной моделью решения этого вопроса об интеллектуальной собственности может стать дальнейшее развитие понятия «права публичного заимствования»5, что обеспечит создателям культурного контента по всему миру компенсацию, основанную на использовании их материалов, а не на предполагаемой ценности информации.

Широкомасштабная оцифровка создает сырье – сканированные изображения книжных страниц – для ряда дополнительных услуг, которые расширят границы возможного на сегодняшний день, а именно ограниченный поиск библиографических метаданных и просмотр страниц книги онлайн. Сканирование и распознавание текстов всех книг за десятилетие – похвальная цель, но нам нужно стремиться к большему, а именно к поиску смысла во всем этом массиве информации. Нам нужно использовать семантический анализ для создания аннотаций и указателей, ссылок и аллюзий, а также стимулировать развитие национальных языков на новом уровне, недоступном ранее. В-четвертых, у нас

[3] www.archimuse.com/mw2005/ (прим. редактора).
[4] Being in copyright but out-of-print (прим. редактора).
[5] Public lending right состоит в том, что когда в библиотеке (или другом пункте общественного доступа к информации) читатель (пользователь) обращается к какому-либо изданию (или его цифровой копии), то держатель авторских прав на это издание получает определенное вознаграждение. Соответствующие расходы библиотеки (или другого пункта общественного доступа) возмещаются государством (прим. редактора).

есть возможность сделать мировое наследие печатной литературы универсальной базой знаний путем исследования смысла, заложенного в печатных текстах, обнаружения и добавления ссылок на цитаты, оглавления, определения из справочников, фактов, содержащихся в графиках и таблицах, и многого другого. Подобная база знаний даст исследователям возможность изучать значимую часть текста, делать школьные издания, вскрывать историю мысли и влияний, и в конечном итоге – через знания, содержащиеся в многоязычных словарях и тезаурусах – способствовать преодолению языковых барьеров и созданию всеобщего культурного наследия человечества.

Давайте изучим эту четвертую составляющую подробнее, потому что в настоящее время этот вопрос еще не нашел широкого понимания.

Преобразование сканированных изображений

Технология печати рукописного текста привела к демократизации доступа к письму. За последние четыре столетия художники-оформители и издатели разработали традиции представления текста на странице для того, чтобы способствовать эффективному чтению, в частности, путем выделения слов с другим контекстным значением, представляя их в другом виде, чем обычные. Сама страница делится на текстовые блоки, которые на вид отличаются один от другого, так как их окружает больше или меньше белого пространства (пробелов) или сдвиг по сравнению с другими текстовыми блоками. Вверху или внизу страницы могут быть отдельные зоны текста, они могут располагаться и в центре страницы или колонки. В этих текстовых блоках определенные слова или группы слов могут быть напечатаны другим видом или размером шрифта, отличающимся от всего остального текста. Так, они могут находиться чуть выше или ниже линии обычного шрифта, выделяться курсивом или подчеркиваться. Все эти подсказки помогают людям читать: это задача, которую мы выполняем на уровне восприятия изображения, до того, как «читаем» слова. Эти подсказки позволяют нам мгновенно узнать (исходя из нашего опыта чтения), является ли текст, на который мы смотрим, романом, письмом, научной статьей, словарем, банковской выпиской или текстом другого жанра. Самые образованные люди могут находить различия между сотнями общих жанров и десятками специализированных профессиональных жанров. Каждый из этих жанров имеет свое оформление и типографские подсказки; знание, которое они содержат, может подразумеваться или выявляться путем изучения и понимания этих подсказок, даже независимо от того, понимаем ли мы сам язык написанного.

Некоторые крайне структурированные жанры, такие как справочники, финансовые отчеты и т.д., позволяют нам «читать» содержимое печатного текста, как если бы это была схема данных. Знаки препинания, пробелы, рисунок шрифта и графические формы букв – все это может показать читателю, что они имеют дело с элементами сравнительно однородной структуры. В повествовательных жанрах иерархическая структура, подобная схеме языка XML или SGML, отображается менее частыми типографскими подсказками, показывающими места разделения секций печатного текста, а также областей с другими характеристиками, например,

разделов, имеющих высокую значимость для содержания, представленных автором/издателем или служащих иллюстрациями важнейших положений. Некоторые разделы печатного текста могут иметь особое оформление, например, оглавление, предисловие, введение, благодарность, алфавитный указатель, ссылки. Такая важная информация, как заголовки иллюстраций, оси графиков, заголовки и подзаголовки, списки и их соотнесение с характеристиками описания сигнализируются в тексте через типографский дизайн.

В настоящий момент в процессе оцифровки мы буквально «выкидываем» всю эту информацию, фокусируясь практически исключительно на переводе печатных изображений в «символы», закодированные в формате ASCII или Unicode. В результате, если мы продолжим недооценивать эту составляющую печатного текста, мы создадим масштабную «полнотекстовую» базу данных мировой литературы на тех языках, на которых она была написана. Конечно, это будет значительным прорывом по сравнению с нынешней ситуацией и позволит иметь лучший доступ к нашему совместному печатному наследию, чем это возможно сейчас, но это будет и большим провалом, если мы не сможем достичь гораздо большего.

Веками ученые, зачастую жившие в религиозных сообществах, монахи, раввины и муллы, а также ученые в научных сообществах, занимались созданием дополнительных межтекстовых механизмов получения знаний, таких как конкордансы, указатели и полные академические издания. Сегодня мы можем использовать компьютер для создания множества интеллектуальных механизмов, которые когда-то требовали труда на протяжении всей жизни; конкордансы или списки словоупотреблений могут создаваться мгновенно для каждого оцифрованного текста. Среди других механизмов, которые позволяют нам эффективно работать с текстом, можно назвать аннотирование, мета-базы данных цитат и библиографических ссылок, которые также могут создаваться с помощью компьютера, однако требуют более глубокого понимания печатных текстов. Мы считаем, что функциональный эквивалент академических изданий может создаваться с помощью информации, извлекаемой из отсканированных страниц мирового печатного наследия и побуждающей человеческое знание к новым открытиям. Доступность этих текстов по сети позволит большинству людей, сегодня не имеющих адекватного доступа к ресурсам, которые служат основой творчества, делать новые открытия.

Для того чтобы сделать этот скачок, нам необходимо объединить две ветви технических достижений. Первая – абстрактное знание того, что «важно» понимать о печатных текстах. Этот вид знаний представлен в Принципах Инициативы кодирования текста (Text Encoding Initiative, TEI), разработанных исследователями текстов за последнее десятилетие для «кодирования» цифровых представлений текстов, для того, чтобы формальные и абстрактные качества, важные для исследователей, могли быть «маркированы» в оцифрованном тексте и извлекались бы независимо (см. http://www.tei-c.org/). В результате кодирования таким образом тысяч текстов планируется создать базу данных из единого текста, которая бы отражала все, что касается академических

издателей (Burnard, O'Keeffe, & Unsworth, 2005). Несмотря на то, что эта инициатива разрабатывалась для цифровой эпохи, ей свойственны ограничения, присущие и академическим изданиям: только небольшое количество текстов было или может быть столь интенсивно обработано учеными. Тем не менее, то, что они уже обработали много важных текстов, означает, что у нас есть правила отображения этих жанровых черт, которые считались важными при анализе печатного текста, а также модель для их преобразования в широкомасштабные электронные библиотечные системы (см., например, Crane, 2005).

Вторая ветвь технических разработок – это работа над системами распознавания текстов. В конце 1990-х годов исследователи, которых привлекал, в первую очередь, потенциал извлечения структурированных данных из современных бизнес-документов, правительственных документов и документации в области здравоохранения, а также возможность автоматического получения библиографических метаданных, указателей, аннотаций и баз данных цитат, достигли значительного прогресса в сегментации изображения страницы и распознавания структурированного содержания на основе печатных традиций (Doermann, Rivlin, & Rosenfeld, 1998), (Okun, Doermann, & Pietikainen, 1999), (Doermann, 1998)). Недавно был зарегистрирован прогресс в области декодирования схем, заложенных в печатных традициях многоязычных словарей (Ma, Karagol-Aran, & Doermann, 2003). Продвинутые технологии самообучающихся систем, экспоненциальный рост скорости обработки и снижение затрат на обработку изображений – все это может привести к крупным прорывам в исследованиях.

Нам все еще неизвестен весь спектр того, что необходимо для языка представления значений, заложенных в печатной традиции, как неизвестен и весь набор функций, которые станут возможными. Мы пока не знаем, до какой степени литературе каждой страны понадобятся правила языка представления или оптимальные способы декодирования исторически сложившихся традиций. Мы также не можем сказать, необходимо ли распознавать жанры на уровне книги или раздела книги, или структуры должны строиться на уровне отдельных абзацев или текстовых блоков. В области понимания изображения текста мы еще не уверены в том, что сегментация страницы и статистический анализ отклонений от нормативного рисунка шрифта, разрядки и других характеристик являются достаточными для отображения всех традиций оформления, имеющих семиотическое значение. По мере создания универсальных библиотек электронных книг нам понадобятся итеративные самообучающиеся программы анализа изображения текста.

Выводы

При оцифровке печатного наследия национальные библиотеки и другие институты должны использовать самые лучшие из доступных форматов сканированных изображений, чтобы можно было немедленно начать первую работу по декодированию. Будущие усовершенствования способов извлечения знаний из печатного текста могут итерационно применяться к изображениям. Очень важно,

чтобы смысл, извлеченный из книг, отсканированных на ранних этапах процесса и полученный как методами семантического анализа, так и методами анализа структуры оформления текста, мог использоваться для понимания книг, отсканированных позже. Необходимо поощрять тех, кто занимается обработкой оцифрованных изображений печатных текстов, в том числе частные инициативы, которые могут разрабатывать краткосрочные методы исследования для получения коммерческой выгоды. Что касается базовых ресурсов, электронные изображения сканированной мировой печатной литературы останутся открытыми и доступными, и подобные инновации позволят со временем извлечь более глубокое знание.

Что мы можем ожидать от подобного направления действий? В отличие от любых действий, предпринимавшихся ранее для сохранения книг, новые виды представления книг будут гораздо более качественными с функциональной точки зрения. Простое сканирование всего лишь «фиксирует бумагу», а «простое» оптическое распознавание текстов дает «только» поиск по всему тексту, но если мы оцифруем печатные книги с использованием систем, учитывающих типографские традиции и включающих полнотекстовый семантический анализ, мы сможем получить добавочные «знания» с помощью систем распознавания изображений. Последующая обработка всей базы знаний может выявить исторические значения слов и идей, содержащихся в литературе предшествующих эпох, поможет понять иностранную литературу иноязычному читателю, а также поможет обычным читателям понять специализированную литературу. Названия и имена зданий, городов, людей, кораблей и других субъектов, имеющих собственные имена, могут быть связаны с данными и изображениями, представляющими их в соответствующее время. Аллюзии авторов на ранние тексты и ссылки на этих авторов в более поздних текстах могут быть связаны, как это делается сейчас только в самых полных академических изданиях.

Если мы при оцифровке позаботимся о том, чтобы обеспечить разметку контента, отображенного на странице, на основе анализа его структуры, то наследие нации будет не только доступным в том же смысле, что и печатное, книга за книгой, но оно будет интегрированным на уровне абзацев, предложений и слов в их соответствующем контексте – заголовках, ссылках, подписях к иллюстрациям, определениях и т.д. Учитывая международные стандарты представления подобных знаний, каждый сможет «генерировать» тот вид текста, который доступен сегодня только для отдельных текстов, участвующих в проектах составления конкордансов, указателей, аннотаций и баз данных цитат. Возможность создавать академические издания любой печатной книги непосредственно в процессе распознавания текста будет заключительным подтверждением того, что мы перешли эту веху. Но потенциал универсального доступа для всех здесь не кончается – социальный эффект этого скорее всего будет огромным, хотя мы пока и не знаем, каким, и интеллектуальный вклад будет заключаться в возможности создания множества дополнительных текстовых продуктов в дополнение к тому, что отдельные ученые в прошлом создавали вручную самостоятельно.

Универсальный доступ к знаниям, содержащимся в огромном количестве печатных материалов во всем мире, является целью, достижимой всего за десятилетие. При благоприятной политике стран и разработке и внедрении стандартов, которые не обсуждались в данном докладе, так как являются предметом отдельных процессов, большая часть этого знания будет бесплатно предоставляться пользователям через пункты доступа, при этом держатели авторских прав будут получать больший доход. При соответствующем внимании к представлению контекстуального знания, содержащегося в изображениях печатных изданий, и к обмену исходными изображениями правительства могут способствовать появлению первых интегрированных всеобъемлющих ресурсов, представляющих всемирную печатную литературу. Творческое начало в человеке и его возможности высвободятся, если устранить барьеры, которые ранее создавал ограниченный доступ к информации, барьеры, из-за которых все, независимо от привилегий, не имели полного доступа ко всемирному печатному наследию, барьеры, которые устанавливали кардинально различные условия доступа к наследию для избранных, имеющих обширный доступ, и подавляющего большинства, практически не имеющего этого доступа.

Рекомендации

1. Печатное наследие должно быть преобразовано в открытые электронные источники ресурсов, предоставляемых правительством; необходимо поощрять творческое использование этих материалов. Таким образом, правительство сможет сделать культурное наследие страны более доступным в рамках стратегии по его сохранению.
2. Необходимо разработать «права публичного заимствования», для предоставления всеобщего доступа ко всем изданиям, находящимся в сфере действия авторского права, но уже распроданным и не переиздающимся. Это может осуществляться посредством сотрудничества правительств различных стран и творческих сообществ. Это будет способствовать устранению барьера, препятствующего массовой демократизации доступа к информации, и поможет сохранить некоторые из исчезающих языков.
3. Необходимо исследовать печатные традиции для извлечения структурированных знаний из печатных текстов и общедоступного их представления. Просто создание изображений печатных текстов, даже с дополнительными библиографическими метаданными и текстом на базе оптического распознавания текстов, является недостаточным. Действуя в этом направлении, правительства должны разрабатывать и внедрять международные стандарты кодирования текстов, способствуя, таким образом, созданию универсальной базы знаний всего мирового печатного наследия.

Ссылки

Burnard, L., O'Keeffe, K. O. B., & Unsworth, J. (2005). Editors' Introduction. In L. Burnard, K. O. B. O'Keeffe & J. Unsworth (Eds.), *Electronic Textual Editing*: Modern Language Association and the TEI Consortium.

Carnegie Mellon University Libraries, & Troll, D. (2005, April 28, 2005 --). *Frequently Asked Questions About the Million Book Project*. Retrieved May 17, 2005, 2005, from http://www.library.cmu.edu/Libraries/MBP_FAQ.html

Crane, G. (2005). Document Management and File Naming. In L. Burnard, K. O. B. O'Keeffe & J. Unsworth (Eds.), *Electronic Textual Editing*: Modern Language Association and the TEI Consortium.

Doermann, D. (1998). The Indexing and Retrieval of Document Images: A Survey. *Computer Vision and Image Understanding, vol.70*, p.287-298.

Doermann, D., Rivlin, E., & Rosenfeld, A. (1998). The Function of Documents. *Image and Vision Computing, 16*, 799-814.

Google. (2005). *Google Print Library Project*. Retrieved May 17, 2005, from http://print.google.com/googleprint/library.html

Internet Archive. (2005). *Canadian Libraries [Project Description]*. Retrieved May 17, 2005, from http://www.archive.org/details/toronto

Ma, H., Karagol-Aran, B., & Doermann, D. (2003). Segmenting and Tagging Structured Content. *Symposium on Document Image Understanding Technology*, 53-64.

Ninch Symposium: April 8, 2003, New York City "The Price of Digitization: Resources", http://www.ninch.org/forum/price.resources.html

Okun, O., Doermann, D., & Pietikainen, M. (1999). *Page Segmentation and Zone Classification: The State of the Art, LAMP-TR-036; CAR-TR-927; CS-TR-4079* (Technical Report): University of Maryland.