

Вильям И. Армс
William Y. Arms

Цифровые библиотеки

Digital Libraries

Издательство Массачусетского технологического института
MIT Press

Кембридж, Массачусетс, Лондон, Англия
Cambridge, Massachusetts, London, England

© 2000 Massachusetts Institute of Technology
© 2000 Перевод Арнауты С.А.
© 2000 Научный редактор Максимов Н.В.

Digital Libraries

William Y. Arms

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Series Foreword	vii
Preface	ix
Acknowledgements	xi
1 Libraries, Technology, and People	1
2 The Internet and the World Wide Web	21
3 Libraries and Publishers	39
4 Innovation and Research	57
5 People, Organizations, and Change	77
6 Economic and Legal Issues	99
7 Access Management and Security	123
8 User Interfaces and Usability	143
9 Text	163
10 Information Retrieval and Descriptive Metadata	187
11 Distributed Information Discovery	207
12 Object Models, Identifiers, and Structural Metadata	225
13 Repositories and Archives	245
14 Digital Libraries and Electronic Publishing Today	263
Glossary	273
Index	285

Оглавление

Оглавление.....	4
Предисловие к серии “Цифровые библиотеки и электронные издательства”	8
Предисловие	9
Благодарности	10
Глава 1. Библиотеки - технологии и люди.....	11
Экономика.....	12
Почему электронные библиотеки?	12
Цифровые библиотеки “доставляют библиотеку” непосредственно к пользователю	13
Компьютер расширяет возможности по поиску и обработке информации.....	13
Информацией можно делиться	13
Информацию легче поддерживать актуальной	13
Информация доступна всегда	13
Становятся возможными новые виды информации	14
Стоимость цифровых библиотек	14
Технические усовершенствования	15
Хранение в электронном виде становится дешевле, чем в бумажном.....	16
Дисплеи персональных компьютеров становятся более комфортабельными в использовании.....	17
Увеличивается количество высокоскоростных сетей.....	17
Доступ к цифровым библиотекам	17
Базовые концепции и терминология	18
Коллекции	19
Данные и метаданные	19
Объекты в цифровой библиотеке	19
Библиотечные объекты (library objects)	19
Представления, распространение и формы хранения цифровых объектов (presentations, disseminations, and the stored form of a digital object)	20
Работы и содержание (Works and content)	20
Люди.....	20
Компьютеры и сети.....	21
Необходимость изменений.....	22
Глава 2. Интернет и веб	24
Интернет.....	24
Интернет-сообщество	26
Научные публикации в Интернет	27
Веб	29
Технология веб	30
Информация в веб	33
Соглашения веб	34
Веб как Цифровая Библиотека.....	35
Глава 3. Библиотеки и Издательства	37
Разделение ресурсов и онлайн-каталоги.....	37
Связывание (linking) онлайн-каталогов и Z39.50.....	39
Рефераты и индексы.....	39
Поиск информации (Information Retrieval)	41
Языки разметки (Markup Languages).....	41
Цифровые библиотеки научных журналов.....	42
Первые попытки	42
Электронные журналы.....	44
Конверсионные проекты и цивилизации (humanities)	46
Библиотеки и инновации	48

Глава 4. Инновации и исследования	49
Модели объектов (Object Models).....	53
Пользовательские интерфейсы и взаимодействие человек-компьютер	54
Обнаружение информации (Information Discovery).....	54
Описательные метаданные: каталогизирование и индексирование.....	55
Обработка текстов на естественном языке (Natural-Language Processing)	55
Нетекстовые материалы	55
Управление и сохранение коллекций.....	55
Организация.....	56
Архивирование и сохранение	56
Конвертирование.....	56
Интероперабельность (Interoperability).....	57
Масштаб.....	59
Надежность и устойчивость	59
Экономические, социальные и юридические аспекты	60
Мировые исследования.....	61
Глава 5. Люди - организация и изменения (Изменения в управлении персоналом)	62
Цифровые библиотеки, созданные пользователями.....	62
Мотивы создателей и пользователей.....	65
Создатели (creators).....	65
Пользователи	66
Профессии, связанные с информацией и происходящие перемены	67
Библиотечные работники	67
Издательства.....	69
Компьютерные профессионалы и веб-мастера	70
Новые формы организаций	71
Консорциумы.....	71
Вторичные информационные сервисы и агрегаторы (aggregators)	72
Университеты и их библиотеки	73
Изменения в университетских библиотеках.....	73
Здания для цифровых библиотек.....	74
Глава 6. Экономические и юридические аспекты.....	76
Экономические аспекты	76
Конкретный случай: научные журналы в электронной форме.....	80
Научные журналы и их авторы.....	82
Законодательные рамки.....	83
Международные аспекты	84
Ответственность (Liability).....	84
Копирайт	85
Частная жизнь, конфиденциальность (privacy).....	88
Патенты на программное обеспечение	89
Глава 7. Управление доступом и безопасность.....	90
Элементы управления доступом.....	90
Пользователи	91
Цифровые материалы	92
Операции.....	94
Последующее использование (Subsequent Use)	94
Возможные политики	95
Проведение политики управления доступом	95
Управление доступом в репозиториях	97
Системы на доверии.....	97
"Безопасные контейнеры"	98
Безопасность цифровых библиотек.....	100
Шифрование (Encryption).....	101
Шифрование закрытыми (private) ключами	102

Шифрование с двумя ключами.....	102
Цифровая подпись.....	103
Распространенность шифрования с открытым ключом	103
Глава 8. Пользовательские интерфейсы и удобство пользования (usability)	105
Удобство пользования и создание пользовательских интерфейсов.....	105
Браузеры.....	107
Базовые функции.....	108
Распространение браузеров за рамки веб	108
Мобильный код (Mobile Code).....	109
Недавние разработки в создании пользовательских интерфейсов.....	111
Концептуальные модели.....	111
Дизайн интерфейса	112
Функциональный дизайн.....	113
Компьютерные системы и сети.....	115
Увеличение скорости отклика (responsiveness) браузеров	115
Создание зеркал (mirroring) и кэширование	115
Пользовательские интерфейсы и надежность	116
Пользовательские интерфейсы для мультимедийной информации	116
Пользовательские интерфейсы и эффективность цифровых библиотек	117
Глава 9. Текстовые материалы.....	119
Разметка, описание страницы и таблицы стилей (Markup, page description and style sheets)	119
Конвертирование текста.....	121
Кодировки символов.....	123
ASCII	123
Unicode	124
Транслитерация (transliteration)	125
SGML.....	125
Упрощенные версии.....	126
Таблицы стилей	129
Языки описания страниц	131
TeX.....	131
PostScript	131
PDF	132
Документы: структура против внешнего вида	133
Глава 10. Обработка информации и описательные метаданные	134
Описательные метаданные.....	134
Каталоги.....	135
Службы реферирования и индексирования	135
Дублинское ядро (Dublin Core).....	137
Автоматическое индексирование	138
Присоединение метаданных к содержанию	140
Методы поиска информации.....	142
Базовые концепции и терминология	142
Обработка текстов на естественном языке (natural-language processing) и компьютерная лингвистика	144
Пользовательские интерфейсы и информационно-поисковые системы (information-retrieval systems).....	145
Оценки эффективности (evaluation)	146
Глава 11. Поиск распределенной информации	148
Поисковые программы для веб	150
Поисковые роботы (web crawler)	150
Поиск в индексе веб-страниц.....	151
Федеративные цифровые библиотеки.....	152
Онлайновые каталоги и Z39.50.....	154

NCSTRL и Dienst.....	155
Исследования альтернативных подходов к распределенному поиску.....	156
После поиска.....	158
Глава 12. Модели объектов, идентификаторы и структурные метаданные	160
Произведения, представления, публикации, экземпляры (Works, Expressions, Manifestations, and Items)	160
Форма представления	161
Мультимедиа	161
Динамические и комплексные объекты.....	164
Идентификация	164
Доменные имена и URL.....	165
Долговременно действующие имена и унифицированные имена ресурсов (Persistent Names and Uniform Resource Names (URN))	166
Компьютерные системы для разрешения (resolving) имен	167
Структурные метаданные и объектные модели	168
Типы данных.....	168
Комплексные объекты (Complex Objects).....	168
Структурные типы	169
Объектные модели, обеспечивающие интероперабельность (Object Models for Interoperability).....	170
Доставка (Dissemination)	170
Глава 13. Репозитории и архивы (Repositories and Archives)	172
Репозитории	172
Веб-серверы	172
Современные репозитории (Advanced Repositories)	173
Метаданные в репозиториях	174
Протоколы, обеспечивающие интероперабельность.....	175
Объектно-ориентированное программирование и распределенные объекты.....	175
Соккрытие данных (Data Hiding).....	177
Унаследованные системы (Legacy Systems).....	178
Архивирование	179
Хранение (Storage)	180
Репликация и обновление (Replication and Refreshing)	181
Сохранение содержания путем преобразования (Preserving Content by Migration).....	182
Цифровая археология.....	183
Создание цифровых библиотек с учетом архивных потребностей (Creating Digital Libraries with Archiving in Mind)	183
Глава 14. Цифровые библиотеки и электронные издательства сегодня	185
Близорукий взгляд на цифровые библиотеки (Myopic view...).....	185
Люди.....	186
Организации	187
Коллекции и доступ	188
Технология.....	189
Исследования и разработки.....	189
Заключительные замечания	190
Глоссарий.....	191
Индекс	200

Предисловие к серии “Цифровые библиотеки и электронные издательства”

Начиная с 1990 года цифровые библиотеки (ЭБ) и электронные издательства постепенно перестали быть экзотической формой деятельности отдельных энтузиастов и начали проникать в традиционные библиотеки и издательства. Катализаторами этого процесса стали как технические новации (персональные компьютеры, Интернет, веб), так и новые требования, которые стали предъявлять к получению высококачественной информации разнообразные пользователи. Большая часть трудностей, возникающих при создании новых библиотек носит скорее социальный, экономический и юридический, чем технический характер.

Над развитием ЭБ работают люди разных специальностей - как традиционные библиотечные работники, так и те, кто не связан с библиотечным или издательским делом. Появление веб вызвало к жизни новую профессию. Исследования в области ЭБ стали самостоятельной дисциплиной, в рамках которой сотрудничают специалисты в области компьютеров, экономисты, социологи, юристы и библиотечные работники. Причем необходимо подчеркнуть именно междисциплинарный характер этого сотрудничества.

Писать обычную книгу о ЭБ - это может показаться противоречием. Большинство исследователей в данной области привыкли полагаться на Интернет как на источник текущей информации, хотя многие по-прежнему склонны полагаться на обычные публикации. Сообщения и статьи об инновациях и исследованиях докладываются на конференциях и широко публикуются в журналах; большинство этих материалов доступно в Интернете без ограничений. Книжный же формат дает возможность для более разностороннего рассмотрения как состояния, так и перспектив данной области. Целью предлагаемой вашему вниманию серии “Цифровые библиотеки и электронные издательства” издательства MIT Press является формирование целостной точки зрения на электронные библиотеки. Мы надеемся создать серию, которая будет ценным источником как для специалистов, так и для дилетантов. Специалисты смогут увидеть как сегодняшние лидеры в данной области оценивают состояние дел, результаты исследований, долговременные тенденции. Дилетанты смогут познакомиться с этой инициативой, которая ведет к новому поколению библиотек и издательств.

Вильям И. Армс

Предисловие

В 1995 году я вместе с несколькими коллегами создал онлайн-ежемесячный журнал D-lib Magazine, который стал “голосом” исследователей в области ЭБ. Мы инициировали этот проект потому, что ЭБ объединяют специалистов многих областей, мало знающих о проблемах коллег. Наша цель - создать журнал, который бы предоставлял экспертную информацию о состоянии в этой области. Компьютерные специалисты часто не осознают всю изощренность методов обработки и анализа информации, которые были развиты библиотечными работниками за многие годы работы. В то же время, библиотекари и издатели могут не знать, что создатели Интернета имеют опыт онлайн-управления информацией уже в течение десятилетий. Как библиотекари, так и издатели понимают, что их области подвергаются воздействию внешних экономических и юридических факторов, но пока, как представляется, глубина такого понимания недостаточна.

Уже первые годы существования D-lib Magazine показали, насколько провидческим был такой подход. Темы публикаций охватывают такие вопросы, как рассказ о последней конференции по метаданным, ход выполнения проекта в Библиотеке Конгресса, принципы создания пользовательского интерфейса для поиска и обработки информации, исторический обзор протокола Z39.50, цифровая библиотека японских народных сказок, подход JSTOR к организации электронных публикаций, методы обработки альтернативных кодировок. Никто не может быть экспертом во всех этих областях одновременно, однако для того, чтобы быть лидером в создании ЭБ, требуется глубокое понимание в каждой из них.

Данная книга является моей попыткой дать полный обзор области, посвященной ЭБ. Безусловно, компьютеры и сетевые технологии принципиально важны, - но это только средства. Реальность ЭБ - это взаимодействие людей, организаций, техники и технологий. Как библиотеки и издатели используют эти новые технологии? Как отдельные люди обходятся без традиционных организационных структур, создавая свои собственные библиотеки? Куда все это ведет? Ответ на последний вопрос прост. Никто не знает. Я пытался избежать спекуляций и сосредоточиться на описании текущего состояния области, тенденций и исследований. “Врезки”, к примеру, посвящены отдельным важным аспектам ЭБ: технологиям, применениям, исследованиям. Их совокупность составляет сердцевину данной книги.

Очевидно, что данная книга отражает мой опыт и мои привязанности. Выбирая примеры, я обычно выбирал те из них, о которых я знаю лично и которые связаны с активностью американских университетов. При этом я отдаю дань профессиональным работам, выполненным по всему миру. Некоторые мои привязанности вполне бессознательные, но в других я отдаю себе полный отчет. Я определенно привязан к идее ЭБ, которые предоставляют свободный доступ к информации. Как читатель, я очарован тем, что Интернет дает мне доступ к огромному массиву информации без всяких ограничений; как автор, я публикую свои статьи он-лайн с тем, чтобы каждый мог получить доступ к моим результатам.

Благодарности

Пользуясь тем, что я знаю очень многих из тех людей, чьи работы описаны в этой книге, я бессовестно обращался к ним за помощью. Amy Friedlander, первый редактор D-lib Magazine, был моим постоянным консультантом. Моя жена Кэролайн познакомила меня с ЭБ еще в 1966 году, когда она была аспиранткой в Массачусетском технологическом институте. Работая сейчас в Библиотеке Конгресса, она помогала в написании ряда глав данной книги. Среди тех, кто прислал свои замечания, Robert Allen, Kate Arms, Steve Cousins, Gregory Crane, Jim Davis, Peter Denning, Jack Dongarra, George Furnas, Henry Gladney, Steven Griffin, Kevin Guthrie, Larry Lannon, Ron Karsen, Michael Lesk, Ralph LeVan, Mary Levering, Wendy Lannon, Clifford Lynch, Harry S. Martin III, Eric Miller, Andreas Paepcke, Larry Page, Normann Paskin, Vicky Reich, Scott Stevens, Terrence Smith, Sam Sun, Nal Varian, Howard Wactlar, Donald Waters, Stuart Weibel, and Robert Wilensky.

Ранние версии главы 4 были доложены на Fundcao Getulio Vargas и на летней школе в Ticer. Набросок главы 8 был представлен на конференции по SGML/XML в 1997 году. Врезка 6.3 позаимствована из американского агентства по авторским правам. Рисунки врезки 7.3 основаны на рисунках Henry Gladney. Список элементов на врезке 10.3 взяты с веб-сайта Dublin Core.

Глава 1. Библиотеки - технологии и люди

Мы переживаем замечательный период в истории библиотек и издательств. Впервые стало возможным построение широкомасштабного сервиса на основе коллекций, которые хранятся в цифровой форме и доступ к которым осуществляется по сети. Информация хранится в компьютере. Сеть соединяет серверы и персональные компьютеры пользователей. В истинно электронных библиотеках не возникает никакой необходимости в бумаге.

Данная книга дает обзор этой новой области. Частично она о технологических аспектах, но в равной мере и о людях, и об организационных проблемах. ЭБ интегрируют в себе проблемы из разных областей, объединяют специалистов с различным опытом и различными подходами. Обсуждаются люди, которые создают информацию и люди, которые ею пользуются - их потребности, мотивы, экономические стимулы. Анализируются глубокие изменения в издательском и библиотечном деле. Описываются новые технологические разработки, большинство из которых базируется на веб и Интернете. Среди тем книги технические аспекты компьютеров и сетей, библиотечное и издательское дело, проблемы экономики и законодательства. В книге постоянно присутствует тема *изменения* с ее социальными, организационными и юридическими аспектами.

Безусловно, в одной книге невозможно глубоко осветить все эти вопросы, так что значительная часть материала останется либо за ее рамками, либо на уровне введения в проблему. Большинство примеров взято из практики США, преимущественно академического и университетского сообщества. При этом показано, что развитие ЭБ происходит повсеместно, в различных формах и из различных источников. Хотя специалисты из американских университетов реально вносят основной вклад, они не являются единственными разработчиками. В каждой дисциплине есть богатство и разнообразие инновационных подходов, которые разрабатываются в различных странах мира.

Информационное определение ЭБ: управляемая коллекция информации в совокупности с соответствующими сервисами, причем информация хранится в цифровых форматах и доступна по сети. Ключевой частью этого определения является слово "управляемая". Поток данных, передаваемый со спутника на Землю еще не библиотека. Те же самые данные, организованные в определенную систему, становятся коллекцией ЭБ. Большинство людей не согласятся рассматривать базу данных финансовых записей отдельной компании в качестве ЭБ, однако совокупность такой информации по множеству компаний уже является частью библиотеки. ЭБ содержат различные коллекции информации для использования большим числом пользователей. Размеры ЭБ могут меняться от крошечного до огромного. Они могут использовать любое пригодное для этого аппаратное и программное обеспечение. Объединяет все это то, что информация находится в компьютерах, а также процедуры поиска, выбора, архивирования, организации информации, - т.е. некоторая общность способов ее предоставления пользователям. В некоторых аспектах ЭБ отличаются от обычных библиотек, хотя в других сходство налицо. Люди не меняются с появлением новой технологии. Они по-прежнему создают информацию, которая должна соответствующим образом организовываться, храниться и распространяться. Им по-прежнему бывает необходимо найти созданную другими информацию и использовать ее для обучения, развлечения и других целей. Однако технология оказывает очень большое влияние на формы, в которых информация хранится и управляется - и это ведет к переменам. Каждый год возрастает разнообразие и число коллекций, доступных в цифровой форме, причем одновременно непрерывно улучшается соответствующая технологическая база. В совокупности эти изменения стимулируют фундаментальные перемены в процессах создания и использования информации людьми.

Для того, чтобы понять эти силы необходимо прежде всего понять людей, которые создают библиотеки. Технология диктует темпы создания ЭБ, но способы использования этих технологий зависят от людей. Источниками большинства новшеств являются два профессиональных сообщества. Одна группа, профессионалы информационного обслуживания,

включает библиотекарей, издателей и широкий спектр провайдеров информации (такие сервисы, как индексирование или рефераты). Другое сообщество включает “компьютерщиков” и их нынешних потомков - разработчиков Интернета. До недавнего времени эти сообщества взаимодействовали разочаровывающе мало, даже сейчас можно встретить специалиста по компьютерам, который ничего не знает об основах библиотечного дела, или библиотекарей, чьи концепции обработки информации безнадежно устарели. Однако с появлением веб глубина сотрудничества и взаимопонимание существенно возросла. Частично это следствие того, что ЭБ становятся признанной областью исследований, но значительно более важным является фактор тесной вовлеченности конечных пользователей. Недорогое оборудование и простое программное обеспечение, которое сделало электронную информацию доступной для каждого и обеспечило читателям прямой доступ к информации без посредников. Много ярких достижений вышло из академических или отраслевых групп, которые разрабатывали ЭБ для собственных нужд. Например, медицина имеет долгую традицию творческих инноваций; пионерская юридическая информационная система была разработана адвокатами для адвокатов; веб изначально был создан физиками для физиков.

Экономика

Технология и социально-экономические аспекты информации обоюдно влияют друг на друга. Технология ЭБ быстро развивается - также как и финансовые, организационные и социальные рамки. Различные группы разработчиков привносят свои, различные социальные привычки и отношения к деньгам. Издатели и библиотекари имеют длительную традицию работы с материальными объектами - главным образом книгами, а также картами, фотографиями, звукозаписями и др. Разработаны экономические и юридические подходы, которые в основном базируются на том, что эти объекты продаются и покупаются. Они инстинктивно пытаются перенести эти концепции и на ЭБ. Компьютерщики и ученые выросли в другой традиции. Их интерес к цифровой информации зародился в то время, когда компьютеры были очень дороги. Только очень обеспеченные исследователи могли создать первые компьютерные сети. Они обменивались с коллегами информацией неформально и открыто, бесплатно. Сети росли - но традиция “открытой” информации сохранялась. Экономические подходы, разработанные для ЭБ являются смесью двух этих походов. Некоторые ЭБ идут по пути традиционных издательств, требуя предоплату доступа к коллекциям и сервисам. Другие используют иную экономическую модель, предполагающую, что доступ к их материалам совершенно свободный. Стоимость создания и распространения информации в этом случае ложится на производителя, а не на пользователя. В этой книге описано много примеров в рамках обеих моделей и попыток найти баланс между ними. Почти наверняка, обе они имеют будущее, хотя окончательное соотношение сил предсказать невозможно.

Почему электронные библиотеки?

Главный довод в пользу создания ЭБ - уверенность в том, что это позволит доставлять информацию лучше, чем это было в прошлом. Традиционные библиотеки являются важной, но не самой совершенной частью общества. Можно ли выполнять их функции лучше? Энтузиасты ЭБ указывают на то, что компьютеры и сети уже изменили способы общения людей между собой. В некоторых областях, убеждают они, значительно удобнее обслужить исследователя или школьника, сидящих за собственным персональным компьютером, чем заставлять их идти в библиотеку. Информация, ранее доступная только профессионалу, теперь открыта для всех. Со своего персонального компьютера пользователь может обратиться к материалам, хранящимся на различных компьютерах в различных частях мира. С другой стороны, все, кроме самых негибких энтузиастов ЭБ, признают тот факт, что печатные документы остаются важной частью цивилизации и их роль в хранении и передаче информации может

меняться только очень постепенно. Несмотря на то, что в ряде случаев, безусловно, удобнее пользоваться информацией в электронной форме, далеко не все рассматривают процесс масштабного перехода к электронной информации как желательный, даже если технически, экономически и юридически это возможно.

Ниже приводится ряд потенциальных преимуществ ЭБ.

Цифровые библиотеки “доставляют библиотеку” непосредственно к пользователю

Для того, чтобы воспользоваться обычной библиотекой, в нее нужно прийти. В университете это может занять несколько минут, но большинство людей живут не в кампусах и не имеют библиотеки поблизости. Уже поэтому многие инженеры и практикующие врачи имеют исключительно плохой доступ к свежей информации.

ЭБ доставляет информацию прямо на рабочий стол пользователя, будь то дома или на работе. Читателю больше не надо ходить в библиотечное здание. Библиотека там, где есть персональный компьютер, подключенный к сети.

Компьютер расширяет возможности по поиску и обработке информации

Бумажные документы удобно читать, но поиск информации может быть весьма сложным. Несмотря на огромное число вспомогательных средств и навыки профессиональных библиотечных работников, использование большой библиотеки может быть затруднительным. Обычное возражение состоит в том, поиск информации в обычной библиотеке стимулирует любознательность и мышление, потому что позволяет найти неожиданные материалы. Истина же в том, что библиотеки полны ценными материалами, однако читатели зачастую обнаруживают их лишь случайно.

Во многих аспектах компьютерные системы уже сейчас лучше ручных способов поиска информации. Они не так хороши, как нам бы хотелось, но они вполне работоспособны и быстро улучшаются. Компьютеры особенно полезны в реферативной работе, которая включает повторные переходы от одного источника информации к другому.

Информацией можно делиться

Библиотеки и архивы содержат много уникальной информации. Размещение цифровых материалов в сети делает их доступными для всех. Большинство ЭБ или электронных издательств размещаются на центральном сервере, возможно, с несколькими дублирующими “зеркалами”, продуманно размещенными в различных точках планеты. Это значительно упрощает дорогостоящую задачу физического дублирования малоиспользуемых материалов, а также позволяет сделать более доступным уникальный документ, для знакомства с которым ранее требовалось приехать в данное хранилище.

Информацию легче поддерживать актуальной

Большинство важных информационных источников необходимо поддерживать в актуальном состоянии постоянно. Печатные источники весьма сложно обновлять, требуется перепечатать весь документ, изъять и заменить все старые экземпляры. Эта трудоемкая задача становится существенно легче когда определенная версия документа хранится в цифровом формате на центральном компьютере.

Многие библиотеки поддерживают онлайн-версии указателей, энциклопедий и других реферативных источников, когда получаемая от издателей новая информация устанавливается на библиотечный компьютер. К примеру, Библиотека Конгресса поддерживает онлайн-коллекцию Thomas, которая содержит проекты всех законопроектов, вносимых в Конгресс.

Информация доступна всегда

Двери цифровой библиотеки никогда не закрываются. Недавнее исследование в Британском университете показало, что примерно половина посещений ЭБ приходится на время, когда библиотечное здание закрыто. Материалы не могут быть выданы не тому пользователю,

поставлены не на ту полку или просто украдены, не могут находиться в удаленном книгохранилище. Возможности коллекции выходят за стены библиотеки. Частные документы в офисе или в библиотеке на другом краю земли так же легко использовать, как и материалы в местной библиотеке.

Безусловно, это не означает, что ЭБ совершенны. Компьютерная система может отказать, связь по сети может быть медленной или ненадежной. Однако в сравнении с традиционной библиотекой, информация в такой форме будет более доступной для пользователя, где бы и когда бы он за ней не обратился.

Становятся возможными новые виды информации

Печать далеко не всегда является лучшим способом записи и распространения информации. К примеру, данные переписи населения удобнее хранить и анализировать в компьютерной базе данных. Спутниковые данные можно обрабатывать различными способами. Математическая библиотека может хранить математические выражения в форме компьютерных символов, которыми могут оперировать такие программы, как Mathematica или Maple.

Даже если форматы совпадают, материалы, изначально созданные для цифрового представления, не идентичны материалам, созданным для бумажных или иных носителей. Произносимые слова оказывают иное воздействие, чем написанные, а онлайн-текстовый материал отличается как от произносимых, так и написанных слов. Хороший автор по-разному использует слова в зависимости от формы публикации, а пользователь находит новые возможности для использования информации. Материалы, созданные для “цифрового мира” могут обладать своеобразным “духом”, который отсутствует в материалах, механически конвертированных в цифровой формат - так, к примеру, художественный фильм, созданный для демонстрации в кинотеатрах никогда не выглядит адекватно, при демонстрации по телевизору.

Каждое из вышеприведенных преимуществ можно наблюдать в уже действующих ЭБ. Другая группа потенциальных достоинств, пока не реализованных, имеет менее очевидные (tantalizing) перспективы. Существует надежда, что ЭБ эволюционируют в будущем от статических репозиторий неизменных объектов, предоставляя широкий спектр сервисов, позволяющих тесно сотрудничать и обмениваться идеями. Технология ЭБ тесно связана с соответствующими технологиями, используемыми для электронной почты и телеконференций, которые исторически не вызвали особого отклика у библиотекарей. В то же время потенциал для конвергенции между этими областями обнадеживает.

Стоимость цифровых библиотек

Последнее потенциальное преимущество ЭБ - экономия финансовых средств. До настоящего времени нет надежных данных о реальной стоимости ЭБ, но некоторые факты вполне очевидны.

Обычные библиотеки дороги. Они занимают дорогостоящие здания на дорогой земле. В больших библиотеках работают сотни сотрудников - высокообразованных, но низкооплачиваемых. Библиотеки никогда не имели достаточно средств для приобретения и обработки всех желаемых материалов. Издательское дело также высокзатратно. Добавляемая к обычной электронная форма публикации также удорожает производство. Для покрытия затрат на развитие новых продуктов, издатели иногда назначают более высокую цену за электронную версию, чем за ее печатный эквивалент.

Сегодняшние ЭБ также дороги - вначале даже более, чем обычные библиотеки. Однако, цена на компоненты ЭБ быстро снижается. По мере того, как базовые технологии развиваются, стоимость непрерывно уменьшается. В особенности быстро падает цена хранения и распределения информации. Снижение затрат не будет равномерным. Некоторые вещи уже сейчас дешевле делать на компьютере, чем традиционными методами. Другие затраты не изменятся или даже возрастут. Суммарно, тем не менее, есть возможность серьезно снизить стоимость издательских расходов и библиотек.

Снижение долговременных расходов - это необязательно хорошая новость для существующих библиотек и издательств. В краткосрочной перспективе серьезную нагрузку на бюджет создает необходимость поддерживать традиционные собрания вместе с новыми цифровыми коллекциями. Поскольку люди и организации быстро оценили преимущества онлайн-доступа и электронных издательств, они предполагают расходовать все большую часть своих средств на компьютерное, сетевое оборудование и цифровые информационные технологии. Однако почти все эти деньги идут не в традиционные библиотеки, а в новые области: компьютеры, сети, веб-сайты, на оплату веб-мастеров.

Издатели сталкиваются с трудностями в связи с тем, что обычная модель ценообразования при продаже индивидуальных экземпляров не совпадает со структурой цены в электронных издательствах. Большая часть расходов в традиционных издательствах приходится на производство и распространение тиража книг, фотографий, видеокассет или других изделий. Цифровая информация - другая. Фиксированная цена создания информации и размещения ее на сервере может быть значительной, а вот стоимость ее использования близка к нулю. Поскольку остаточная стоимость незначительна, большой объем информационных ресурсов в сети открыт для доступа без ограничений. Конечно, не все в сети доступно бесплатно, но то, что открыто для всех реально уменьшает доходы издателей. И это давление неизбежно меняет экономические решения, которые должны принимать авторы, пользователи, издатели и библиотекари.

Технические усовершенствования

Первая серьезная попытка хранить библиотечную информацию в компьютерах, предпринятая в конце 60-х годов, столкнулась с серьезными техническими проблемами, включая высокую стоимость компьютеров, неудобный пользовательский интерфейс и отсутствие сетей. Поскольку стоимость хранения была высокой, первые приложения были рассчитаны на быструю окупаемости даже при малых объемах хранимых данных. Одним из первых успешных примеров было создание в Библиотеке Конгресса формата MARC для создания и ведения машиночитаемых каталогов. Его использование в Центре компьютерных онлайн-библиотек (Online Computer Library Center, OCLC) позволило предоставить доступ к каталожным записям для многих библиотек, что позволило последним сэкономить значительные средства.

Первые примеры информационных сервисов, такие как это каталогизирование в едином центре, законоведческие информационные системы, сервис Medline Национальной медицинской библиотеки, использовали технологии того времени. Малые объемы информации размещали в большом центральном компьютере. Пользователь соединялся с ним с удаленного терминала по медленной линии. Эти системы предъявляли высокие требования к пользователям, которые были вынуждены мириться с головолomным интерфейсом в обмен на ускорение процесса поиска информации или доступ к удаленным источникам.

Эти системы не предполагали получения печатного документа. Все, что могло быть изображено на экранах терминалов, был неформатированный текст (обычно с использованием равноширинного шрифта) с графикой очень низкого качества.

Врезка 1.1 Два первооткрывателя цифровых библиотек

Представление о ЭБ не ново. Прогресс в этой области был достигнут за счет объединения усилий большого числа людей в течение длительного периода времени. Необходимо, однако, отметить несколько авторов, чьи работы оказали существенное влияние на следующие поколения. Среди них Vannevar Bush и J.C.R. Licklider.

В июле 1945 года Bush, тогда директор Американского агентства научных исследований и разработок (U.S. Office of Scientific Research and Development), опубликовал статью под заголовком "Как мы можем думать" в Atlantic Monthly. Эта публикация элегантно демонстрировала те потенциальные возможности, которые технология предоставляла ученым для сбора, хранения, поиска и обработки информации. Многое в этом анализе остается верным и сейчас, 50 лет спустя.

Bush утверждал, что “наши методы передачи и обработки информации исследований отстали на поколение и в настоящее время абсолютно неадекватны”. Он обсуждал недавние технологические новинки и то, как они могут быть использованы в будущем. Он сделал набросок одного технического подхода, названного им Memex. Интересное историческое примечание состоит в том, что Memex базировался на использовании фотографии для хранения информации. Много лет спустя, микрофильмы и микрофиши по своим технологическим характеристикам оказались наиболее подходящими для хранения информации с небольшими затратами.

Bush зачастую называется как первый человек, который озвучил новое видение библиотек в будущем. Это не совсем так, его статья основывается на ранних работах, большинство из которых выполнено в Германии до Второй мировой войны. Значение его статьи в том, что он замечательно четко продемонстрировал связь между информацией и научными исследованиями, а также скрытый потенциал новых технологий.

Atlantic Monthly поместил копию статьи “Как мы можем думать” на своем сайте. Любой, кто интересуется библиотечным делом или научной информацией, должен прочесть ее.

В 60-х годах Licklider был одним из тех сотрудников Массачусетского технологического института, кто изучал влияние цифровых вычислений на будущее библиотек. Как и Bush, Licklider более всего интересовался научной литературой, однако он предсказал многие разработки в современных компьютерных вычислениях.

В своей опубликованной в 1965 году книге “Библиотеки в будущем” Licklider перечислял исследования и разработки, необходимые для того, чтобы создать истинно дружественную к пользователю ЭБ. В то время вычисления с разделением времени (time-shared computing) существовали только в лабораториях, компьютерная память стоила 1 доллар за байт, однако он сделал попытку предсказать, как будут выглядеть ЭБ через 30 лет, в 1994 году. Его предсказание замечательно подтвердилось в основных чертах, хотя, естественно, он не мог предсказать все и каждое изменение, произошедшее за эти годы. В общем, он недооценил возможности, которые приобрели современные мощные компьютеры, использующие сверхбыстрые процессоры и огромные объемы памяти и переоценил возможный прогресс в таких областях, как искусственный интеллект и машинная обработка текстов на естественных языках.

Книгу Licklider трудно найти и она менее известна, чем того заслуживает. К сожалению, это один из весьма важных источников информации о ЭБ, который отсутствует ныне в Интернете.

Когда к этим недостаткам добавлялись внутренне присущие тогдашним дисплеям проблемы - низкий контраст и разрешение, уже не удивляет тот факт, что большинство людей уверенно предсказывали невозможность того, чтобы пользователи охотно читали текстовую информацию с экрана компьютера.

За последние 30 лет эти технические проблемы в значительной степени устранены. В начале 90-х годов ряд технологических разработок устранил последние фундаментальные препятствия для построения ЭБ. Некоторые из необходимых технологий все еще несовершенны, но снижение вычислительной стоимости стимулировало взрывное развитие онлайн-информационных сервисов.

Четыре технических аспекта особенно важны для реализации ЭБ.

Хранение в электронном виде становится дешевле, чем в бумажном

Большие библиотеки болезненно дороги даже для богатых организаций. Расходы на содержание зданий составляют около четверти всех расходов. Многие выдающиеся библиотеки располагаются в огромных старых зданиях с плохим контролем параметров среды. Даже когда есть финансовые ресурсы, бывает трудно найти новое пространство для расширения в центре города или университетского кампуса.

В то время как стоимость создания новых и поддержания старых зданий для хранения печатных изданий будет только возрастать, цена электронного хранения снижается приблизительно на 30% ежегодно. В 1987 году, когда мы начинали работу над электронной библиотекой Mercury в университете Карнеги-Меллона, были заказаны несколько

компьютеров, каждый с 10 Гб дискового пространства. Цена каждого компьютера составляла ок. 120 000 долларов. В 1997 году значительно более мощный компьютер с той же емкостью диска стоил 4000. Таким образом, за десять лет цена упала на 97%. Это серьезный повод ожидать, что к 2007 году цена упадет еще на 97%.

В 1987 году стоимость хранения документов на компакт-дисках была уже ниже стоимости хранения книг в библиотеках. Сегодня хранение большинства видов информации в компьютерной форме существенно дешевле хранения физических объектов в хранилище. В 1987 году стоимость оборудования являлась главным барьером при создании ЭБ. Сегодня цены существенно ниже, хотя хранение больших объектов (таких, как цифровые видеозаписи, обширные коллекции графических образов, высококачественные аудиозаписи) стоит все еще дорого. Через десять лет то оборудование, которое слишком дорогостояще сегодня, будет настолько доступно, что цена вряд ли будет фактором, влияющим на принятие решения.

Дисплеи персональных компьютеров становятся более комфортабельными в использовании

Стоимость хранения - не единственный фактор. (Если бы это было так, библиотеки стандартизовали микрофильмы давным-давно.) До недавнего времени только очень немногие люди находили удобным читать с экрана компьютера. Качество представления документа на экране было слишком низким. Обычной была процедура распечатки бумажной копии. Однако недавно был достигнут существенный прогресс в улучшении компьютерных дисплеев, качестве экранных шрифтов, программного обеспечения для обработки информации. Все больше и больше людей читают прямо на экране - особенно те материалы, которые непосредственно были созданы для представления на дисплее (как веб-страницы). Лучшие дисплеи все еще достаточно дороги, но с каждым годом они становятся все дешевле и все лучше. Еще нескоро станет привычным читать обычные книги на компьютере - но опытные образцы экранов с высоким разрешением уже сейчас обеспечивают удивительное качество изображения.

Увеличивается количество высокоскоростных сетей

Рост Интернета в 90-е годы был просто феноменальным. Телекоммуникационные компании предоставляют как локальный, так и удаленный Интернет-сервис по всей территории США; международные связи достигли практически всех стран мира; каждая сколько-нибудь заметная компания имеет свою локальную сеть; университеты создали свои сети; частные лица могут приобрести недорогой доступ по телефонной линии непосредственно из дома. Даже в США все еще есть много мест, где Интернет недоступен, некоторые страны вообще не имеют доступа. Но уже во многих странах информацию легче получить через Интернет, чем разыскивать обычные книги и журналы традиционными методами.

Компьютеры становятся портативными

Несмотря на то, что ЭБ используют сети, их полезность существенно выросла с появлением ноутбуков. Подключаясь непосредственно с ноутбука к сети, пользователь соединяет ресурсы ЭБ со своей непосредственной работой, хранящейся на этом компьютере. При отключении от сети копии выбранных библиотечных материалов могут быть сохранены для дальнейшего использования.

Мощность и качество дисплеев портативных компьютеров существенно выросли за последние годы - а цена непрерывно снижается.

Доступ к цифровым библиотекам

В обычных библиотеках как правило, требуется, чтобы читатель являлся сотрудником той организации, которая поддерживает ее функционирование. В США многие университеты и ряд других организаций имеют прекрасные библиотеки, однако большинство людей не принадлежит к этим организациям. Большая часть Библиотеки Конгресса доступна для любого человека старше 18 лет, некоторые города имеют отличные публичные библиотеки, но

большинство людей ограничено небольшими коллекциями их местных библиотек. Даже ученые зачастую имеют ограниченные возможности. Доктора в крупных медицинских центрах имеют богатые библиотеки, но врачи в провинции зачастую не имеют ничего. Одним из поводов для IEEE (Institute of Electrical and Electronics Engineers) проявить интерес к электронным издательствам уже на ранней стадии и был тот факт, что большинство инженеров не имело доступа к специализированным библиотекам.

Использование ЭБ требует компьютера, подключенного к Интернету. В США многие организации предоставляют компьютеры практически всем сотрудникам, причем некоторые практикуют это уже много лет. Существуют национальные программы по внедрению компьютеров в школы и публичные библиотеки. Для индивидуального пользователя адекватный доступ к Интернету потребует 2000 долларов за оборудование и около 20 долларов в месяц за оплату подключения, а также некий минимум знаний. За чуть большие деньги можно получить уже мощный компьютер и связь по выделенной линии. Для высокооплачиваемых профессионалов это небольшие деньги, хотя для других это может явиться серьезным барьером. По оценкам, в 1998 году уже 95% населения США проживало в районах, где была техническая возможность доступа в Интернет и этот процент быстро растет.

За пределами США ситуация различная. Только в некоторых странах библиотеки работают так же, как и библиотеки США. А, например, университет в Мексике сообщает, что надежная доставка педагогических изданий невозможно, даже если есть финансовые ресурсы. Некоторые страны весьма хорошо оснащены компьютерами и сетями, однако в большинстве мест стоимость оборудования выше, чем в США, люди не столь богаты, соответствующая инфраструктура не развита, а монополии поддерживают высокие цены на услуги связи. ЭБ доставляют информацию тем людям, для которых недоступны обычные библиотеки, однако Интернет все еще не является общедоступным сервисом повсюду и везде.

Важным фактором, который необходимо учитывать при планировании создания ЭБ, является то, что уровень технологий, используемых пользователями, весьма различен. Некоторые счастливчики имеют на своем рабочем столе последние модели персональных компьютеров с самым современным программным обеспечением, подключенных к Интернету по скоростной линии; у них есть возможность получить помощь от высококвалифицированной службы технической поддержки для настройки оборудования, разрешения проблем, установки новых версий программного обеспечения. Большинство людей, к сожалению, вынуждено довольствоваться меньшим. Их оборудование может быть устаревшим, программное обеспечение уже несовременным, связь неустойчивой и медленной, а их системные администраторы недостаточно профессиональны и слишком перегружены. Именно поэтому одна из важных задач в развитии ЭБ состоит в том, чтобы создавать такие системы, которые позволяли бы пользоваться преимуществами современной технологии в разных ситуациях.

Базовые концепции и терминология

Известно, что терминология часто служит барьером для обсуждения проблем ЭБ. Участники процесса пришли из разных областей и принесли с собой специфическую терминологию этих дисциплин. Некоторые слова имеют такую сильную социальную, профессиональную, юридическую или техническую коннотацию, что их использование в дискуссии между специалистами разных школ и дисциплин практически невозможно. Простые слова имеют различное значение для различных людей. Например, слова “копия” и “публиковать” имеют разное значение для специалистов по компьютерам, адвокатов или издателей. Использование в повседневном английском - и использование в качестве профессионального термина отличаются; в различных диалектах английского в разных странах одно и то же слово имеет разное значение. И при этом нужно помнить, что проблемы ЭБ обсуждаются не только на английском.

Некоторые слова вызывают такое непонимание, что представляется правильным вообще запретить их использование в дискуссиях вокруг ЭБ. Дополнительно к словам “копия” (copy) и “публиковать” (publish), этот список включает термины “документ” (document), “объект”

(object) и “работа” (work). Эти слова необходимо использовать крайне осторожно, сопровождая подробным объяснением их значения в данном контексте. Некоторые другие слова, которые не столь многозначны, имеют совершенно разные значения в разных контекстах. Например, необходимо различать использование терминов “фотография” (образ на бумаге) и “цифровая фотография” (набор битов в компьютере). Конечно, в большинстве случаев, такая точность не требуется. В данной книге я следую практике словоупотребления большинства профессионалов в данной области и использую термины достаточно неформально в тех случаях, когда их значение очевидно из контекста.

Коллекции

ЭБ могут содержать любую информацию, которая может быть представлена последовательностью битов. Иногда это цифровые версии обычных носителей, таких как текст, графика, музыка, звукозаписи, чертеж и спецификации и многое другое. По мере расширения ЭБ их содержание все чаще и чаще уже не просто цифровая копия, но объект, который не имеет “твёрдого” эквивалента, к примеру, данные с научного инструмента, компьютерная программа, видеоигра, базы данных.

Данные и метаданные

Информация, хранящаяся в ЭБ, может быть разделена на данные и метаданные. “Данные” - это общий термин, обозначающий информацию, закодированную в цифровой форме. (Является это слово существительным в единственном или множественном числе - это вопрос соглашения. В данной книге, следуя традиции его использования в вычислительной технике, я рассматриваю это слово как собирательное существительное единственного числа.)

Метаданные - это данные о данных. Многие люди недолюбливают это слово, но оно широко используется. Распространённые категории метаданных включают “описательные (descriptive) метаданные” (например, библиографическая информация), “структурные (structural) метаданные” (информация о форматах и структурах), и “административные (administrative) метаданные” (которые включают права, разрешения и другую информацию, используемую для управления доступом). Ещё одна разновидность метаданных - идентификатор (identifier), который однозначно представляет объект для внешнего мира.

Различие между данными и метаданными часто зависит от контекста. Запись в каталоге и реферат обычно рассматривают как метаданные (т.к. они описывают другие данные), однако будучи в онлайн-каталоге или реферативной базе это - данные.

Объекты в цифровой библиотеке

Пока не предложено устоявшегося термина для обозначения элемента (item) в ЭБ. В этой книге используется несколько терминов. Наиболее общий - это “материал” (material), который обозначает все, что может храниться в библиотеке. Слово “элемент” (item) практически является синонимом. Оба этих термина ничего не сообщают о содержании, структуре или точке зрения пользователя на данную информацию.

Слово “материал” может быть использовано для описания физических объектов или информации в цифровой форме. Термин “цифровой материал” используется, когда качество “цифровой” необходимо выделить. Более точный термин “цифровой объект” (digital object) используется для описания объектов, хранящихся в ЭБ. Обычно цифровой объект состоит из данных, метаданных и идентификатора.

Библиотечные объекты (library objects)

Термин “библиотечный объект” полезен тем, что выражает взгляд пользователя на то, что хранится в ЭБ. Например, статья в онлайн-периодическом издании обычно хранится в компьютере в виде нескольких различных объектов (страницы текста, графика и, возможно, компьютерные программы или связанные материалы на других сайтах). С точки же зрения пользователя, это один библиотечный объект, состоящий из нескольких цифровых объектов.

Представления, распространение и формы хранения цифровых объектов (presentations, disseminations, and the stored form of a digital object)

Форма, в которой информация сохраняется в ЭБ может быть различной в зависимости от формы ее использования. Тренажер для обучения летчиков может храниться в виде нескольких компьютерных программ, структурных данных, цифровых образов и других данных. Это “форма хранения” объекта, - а пользователь воспринимает последовательность образов, управляющих сигналами, сопровождаемых звуком. Некоторые используют термин “представление” для обозначения того, что видит пользователь и в большинстве случаев это вполне допустимо. Более общий термин - это “распространение”, который подчеркивает то обстоятельство, что трансформация из формы хранения в то, что видит пользователь, требует исполнения ряда компьютерных программ.

Когда цифровая информация поступает на пользовательский компьютер, она должна быть преобразована в форму, которая доступна для пользователя - обычно на дисплее, иногда в звуковом сопровождении или с другими эффектами. Такое превращение называется “рендеринг” (rendering).

Работы и содержание (Works and content)

Особенно трудно найти адекватные термины для описания содержания. Английский язык очень гибок и слова имеют различное значение в зависимости от контекста. Рассмотрим пример “the song Simple Gifts”. В зависимости от контекста, фраза может относиться к песне, как произведению (work) со словами и музыкой, к музыкальной записи песни, к исполнению этой песни конкретным исполнителем, к звукозаписи исполняемой песни, к изданию музыкального компакт-диска, к конкретному компакт-диску, к акту воспроизведения музыкальной записи, к воспроизведению цифровой записи этой песни из ЭБ и к многим другим аспектам песни. Эти различия значимы для музыкальной индустрии, поскольку именно они определяют, кто получает деньги, которые выплачиваются за запись или воспроизведение музыкального произведения.

Некоторые исследователи ЭБ пытались ввести некую общую иерархию терминов, которая могла бы быть приложима ко всем работам (объектам, works) и библиотечным объектам. Сама по себе, эта цель весьма важна и полезна, однако связана с большими трудностями. Материалы библиотек столь разнообразны, что любая классификация будет хорошо соответствовать одним коллекциям, но окажется абсолютно непригодной для других.

Тем не менее, термины “работа” и “содержание” полезны. Большинство людей свободно пользуются словом “содержание” и мы в этой книге последуем их примеру. Этот термин полезен во многих контекстах, когда необходимо подчеркнуть то обстоятельство, что библиотечные материалы есть не только биты и байты, обрабатываемые компьютером, - но информация, представляющая интерес для пользователя. Мы не можем точно определить “содержание”, но оно очевидно для нас практически.

В то время как термин “содержание” определен свободно, общий термин “работа” используется более конкретно. Согласно закону США об авторских правах, понятие “литературная работа” сформулировано как абстрактное содержание (например, последовательность слов или нот) независимо от конкретного типа хранения, представления или презентации. В этой книге мы используем это слово именно с таким значением, хотя, возможно, не всегда юридически точно.

Люди

Ряд терминов используются для обозначения людей, связанных с ЭБ.

Создатели (creators) информации для библиотеки включают авторов, композиторов, фотографов, картографов, дизайнеров и всех остальных, кто создает интеллектуальные продукты. Некоторые из них любители, некоторые профессионалы, одни работы созданы группой авторов, другие - отдельными людьми. У них было много различных причин для создания информации.

Другая группа состоит из пользователей (users) ЭБ. В зависимости от контекста, пользователи могут обозначаться различными терминами. В библиотеках их часто называют читателями (readers or patrons); в другой ситуации их могут назвать аудиторией (audience) или потребителями (customers). В традиционных библиотеках создатель и пользователь - иногда одно и то же лицо. Академические, исследовательские и школьные учреждения используют библиотеки для своей работы и публикуют свои результаты в изданиях, которые становятся частью библиотечной коллекции.

Последняя группа людей весьма обширна и включает всех, кто играет какую-либо роль в оказании услуг (помощи) создателям и пользователям: специалисты по компьютерам, библиотекари, издатели, редакторы и многие другие. Их можно назвать информационными посредниками (information managers). С развитием веб появилась новая профессия - вебмастер. Понятно, что один и тот же человек может быть создателем, пользователем и информационным менеджером.

Компьютеры и сети

ЭБ состоит из ряда компьютеров, объединенных в сеть. Основная, базовая сеть - это Интернет, который в процессе развития превратился в гибкий, недорогой, общедоступный инструмент, являющийся главным фактором в развитии ЭБ.

На рисунке 1.1 показаны некоторые типы компьютеров, используемых в ЭБ. Они выполняют три главные функции: помощь пользователям во взаимодействии с библиотекой, хранение коллекций, предоставление сервисов.

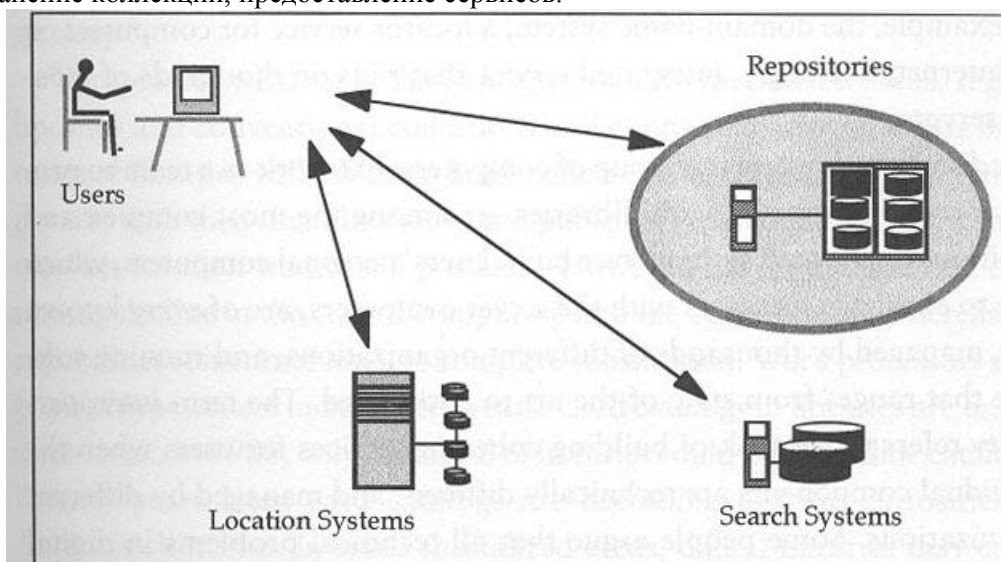


Figure 1.1

В компьютерной терминологии любой, кто взаимодействует с компьютером, является пользователем. Этот широкий термин охватывает создателей (авторов), пользователей библиотеки, профессионалов информатики, и всех остальных, кто имеет доступ к компьютеру. Компьютер, который используется для доступа к ЭБ, называется “клиентом”. Клиент может взаимодействовать с ЭБ и без участия человека, к примеру, роботы, которые автоматически индексируют библиотечные коллекции или сенсоры, которые собирают данные о погоде и передают их в ЭБ.

“Репозитории” (repositories) - это компьютеры, которые хранят информационные коллекции и предоставляют к ним доступ. “Архивы” - это хранилища, которые специально организованы для долговременного хранения материалов.

Два типичных “сервиса” (services), предоставляемых ЭБ - это поиск (search) и определение местонахождения (locate). Поиск включает каталоги, индексы и другие инструменты, призванные помочь пользователю найти информацию. Второй сервис позволяет определить место хранения информации и идентифицировать ее.

В некоторых ситуациях в “цепочке” между клиентом и компьютером, в котором хранится информация, могут участвовать и другие компьютеры. Зеркала (mirrors) и кэш-хранилища (caches store) дублируют информацию для повышения надежности и ускорения доступа. Зеркала дублируют большие массивы информации; в кэше хранится только недавно запрошенная. Прокси (proxies) и шлюзы (gateways) являются мостиками между различными типами компьютерных систем и особенно полезны в системах, которые имеют несовместимые технические спецификации.

Универсальный термин “сервер” (server) используется для обозначения любого компьютера кроме клиента. Данный сервер может выполнять сразу несколько из вышеперечисленных функций, к примеру, хранилища и поиска. И наоборот - одна функция может быть распределена между многими серверами. К примеру, служба доменных имен, определяющая IP адреса компьютеров в Интернете, является единым интегрированным сервисом, который выполняется на тысячах отдельных серверов.

“Распределенная система” (distributed system) - это группа компьютеров, которые работают сообща, предоставляя сервис пользователю. ЭБ является одним из самых сложных распределенных систем, которые когда-либо создавались. Персональные компьютеры пользователей, которые должны обмениваться сообщениями с сервером (локальные сети), - это общеизвестный пример такой системы, которая существует в тысячах различных организаций, работающей на самом различном программном обеспечении. Термин “интероперабельность” (interoperability) относится к задаче построения работоспособного сервиса для пользователей в условиях, когда индивидуальные компоненты технически различны и управляются различными организациями. Некоторые люди вообще утверждают, что все технические проблемы в ЭБ - это лишь аспекты проблемы интероперабельности. Возможно, это преувеличение, но определенно можно утверждать, что достижение интероперабельности является фундаментальной проблемой в построении и функционировании ЭБ.

Необходимость изменений

Если ЭБ так хороши, то что же мешает любой библиотеке стать “цифровой” немедленно? Частично ответ в том, что соответствующие технологии все еще несовершенны, но главная проблема вовсе не в технологии. Не меньшая проблема в способности отдельных людей и организаций найти пути адекватного использования этих технологий, усвоить и принять неизбежные перемены и создать требуемые социальные условия. Информацией занимается много людей, и каждый привносит свой опыт, знания и ресурсы. Для фундаментального изменения системы требуются согласованные изменения в экономических, социальных, юридических взаимоотношениях всех участников.

ЭБ зависят от людей и не могут быть созданы быстрее, чем люди и организации могут адаптироваться. Это относится и к создателям, и к пользователям, и к персоналу ЭБ. Взаимоотношения между этими группами меняются. Пользуясь ЭБ, читатель часть напрямую обращается к нужной информации, не обращаясь в библиотеку лично и не вступая в контакт с библиотечными сотрудниками. Авторы могут выполнить большую часть работы над рукописью. Профессионалам требуются новые навыки и приемы выполнения этих новых обязанностей. Некоторые из этих навыков приобретаются из длительного опыта, другим можно научить. Библиотечные учебные заведения могут изменить свои программы, но пройдут годы и годы, прежде чем изменения затронут всю систему, ведь традиция уходит в глубину веков.

Наиболее правильный подход заключается в том, чтобы - за исключением некоторых очень специализированных областей - традиционные и цифровые библиотеки продолжали сосуществовать. При этом первые будут по-прежнему поддерживать огромные коллекции и параллельно предоставлять цифровые сервисы, а у издателей останется их традиционный рынок сбыта. Однако непрерывное и значительное снижение стоимости компьютерных технологий вкупе с существенным увеличением возможностей приведет к полной замене традиционных библиотек. Текстовые процессоры вытеснили печатные машинки в течение практически 10 лет. Каталоги из библиотечных карточек близки к этому. В 1980 году только половина библиотек могла позволить себе онлайн-каталог. Двадцать лет спустя уже

традиционный каталог становится курьезом в американских библиотеках. В некоторых специальных областях ЭБ могут полностью заменить обычные библиотечные материалы.

Во многих новосозданных организациях были выполнены интересные разработки в области ЭБ. Они могут начать “с чистого листа”, в то время как “старые” организации вынуждены поддерживать традиционные сервисы и одновременно внедрять новые. Один из наиболее вероятных эффектов появления ЭБ - массовый “переток” денег от традиционных поставщиков информации к новым агентствам и в компьютерную индустрию. Естественно, существующие организации будут серьезно противиться любым изменениям, которые уменьшают их значение, но экономические взаимоотношения между участниками уже меняются. Некоторые крупные “игроки” на этом рынке, без сомнения, вынуждены будут сократить свой бизнес или даже уйти из него. Предсказать такие изменения особенно трудно в связи с неопределенностью в таких вопросах, как стоимость ЭБ и электронных издательств, и с необходимостью адаптировать законодательство. Очевидно, что давление рынка приведет к формированию нового порядка. В определенном смысле, для внесения юридических изменений необходимо чтобы рынок устоялся. До этого времени, экономические и юридические неопределенности будут мешать - но они не смогут стать непреодолимым барьером для прогресса.

В долгосрочной же перспективе никаких барьеров для ЭБ и электронных издательств нет. Технические, экономические, социальные и законодательные трудности есть, но они настойчиво решаются. Мы не можем быть уверенными в окончательной форме, которую примут ЭБ, но очевидно, что они здесь для того, чтобы остаться.

Глава 2. Интернет и веб

Интернет

Как следует из термина, Интернет есть взаимосвязанная группа независимо управляемых сетей. Каждая сеть поддерживает необходимые для взаимодействия технические стандарты - семейство TCP/IP протоколов и общие методы идентификации компьютеров, но со многих точек зрения эти сети существенно различаются. Различные сегменты Интернета используют практически все существующие способы передачи данных, от быстрых и надежных до медленных и ненадежных. Они являются либо частной собственностью, либо функционируют как публичные и оплачиваются различными способами. Интернет еще называют иногда "информационной магистралью". Более точным было бы сравнение с интернациональной транспортной системой, которая включает все, начиная от самолетов до тяжелых грузовиков.

Сети, образующие Интернет, появились по двум путям развития. Первый путь - это развитие локальных сетей, которые создавались для того, чтобы связать между собой компьютеры и терминалы в пределах одного департамента или организации. Большое число оригинальных концепций вышло из исследовательского центра компании Ксерокс в Пало-Альто (Xerox Corporation's Palo Alto Research Center, PARC). В США пионерами в расширении маленьких локальных сетей до размеров целого кампуса стали университеты. Второй путь развития сети заключался в создании национальных сетей. Наиболее широко известный пример - это сеть ARPAnet, которая в середине 80-х годов связала около 150 компьютеров научно-исследовательских организаций. Другие большие сети, такая как Bitnet, сейчас уже почти забыты. Их пользователи пользуются Интернетом.

В связи с тем, что при построении первых сетей были использованы разные технические подходы, связать их между собой оказалось тяжело. В конце 80-х годов университеты и исследовательские организации сошлись на использовании TCP/IP, наборе протоколов, созданном в рамках проекта ARPAnet, что и привело к появлению Интернета в его нынешнем виде. Ключевым явилось решение в 1986 Национального научного фонда США построить высокоскоростную магистраль (network back-bone) для всей страны и поддержать развитие локальных сетей с их подключением к магистрали. При этом в проекте было решено использовать технические разработки, созданные в рамках проекта ARPAnet. Это положило начало созданию стандартов Интернет. Одновременно локальные сети кампусов использовали эти же стандарты. Скоростная сеть университета Карнеги-Меллона, разработка которой закончена в 1986 году, является хорошим примером "ранней" сети. Она базировалась на нескольких методах передачи данных и поддерживала соперничавшие сетевые протоколы фирм Digital, Apple, IBM, но объединяющей основой явился именно протокол TCP/IP.

Очевидно, будущие историки укажут на то обстоятельство, что именно комбинация финансовых, организационных и технических факторов привела к тому, что были приняты именно стандарты ARPAnet. Главный вклад в развитие Интернета сделали несколько компаний, но бремя лидерства взяли на себе 2 государственные организации: Defence Advanced Research Project Agency (DARPA) и Национальный научный фонд.

Технология Интернета описана в книгах. Большинство деталей для пользователей неважны, но базовое понимание будет полезно для разработчиков и пользователей ЭБ. Принципы функционирования TCP/IP и цифровой адресации, известной как IP адреса, описаны на врезке 2.1. Другой способ идентификации компьютера в Интернете - это присвоить ему имя, вроде tulip.mercury.cmu.edu. Такое имя называют доменным (domain name), а система, которая обеспечивает соответствие этих имен с IP-адресами, называется системой доменных имен (domain-name system, DNS). Компьютеры, на которых установлен TCP/IP протокол, обычно

поддерживают стандартный набор базовых приложений, которые называются стеком (suite) протоколов. Наиболее широко используемые из них приведены на врезке 2.2.

Врезка 2.1

TCP/IP

Два базовых протокола, которые и образуют Интернет, - это TCP и IP. Мы так привыкли видеть их вместе, что легко забыть о том, что это два разных протокола.

Интернет-протокол (Internet Protocol, IP) обеспечивает взаимодействие отдельных сегментов сети, из которых состоит Интернет. Каждый компьютер в Интернете имеет уникальный адрес, известный как IP-адрес. Он состоит из четырех чисел, каждое в интервале от 0 до 255, например, 132.151.3.90. В компьютерной форме этот адрес занимает 4 байта. При записи принято разделять эти числа точками. Интернет-протокол позволяет любому компьютеру в сети отправить послание другому компьютеру. Различные части Интернета соединены между собой специализированными компьютерами, называемыми маршрутизаторами (routers). Как следует из их названия, они используют IP-адрес для того, чтобы направить каждое послание в следующий промежуточный пункт до тех пор, пока оно не достигнет места назначения.

Послания в Интернете передаются в виде отдельных коротких пакетов размером в несколько сотен байт. Маршрутизатор получает пакет из одного сегмента сети и направляет его далее. IP-маршрутизатор не имеет представления о том, достиг пакет пункта назначения или нет. Пользователи сети также редко интересуются судьбой данного пакета или устройством сегмента сети. Им требуется надежная доставка всего послания с одного компьютера на другой. Эту функцию выполняет TCP протокол (Transport Control Protocol). Программа-приложение на компьютере-отправителе передает послание программному обеспечению, реализующему функции TCP. Последний разрезает послание на пакеты, указывает в каждом IP-адрес назначения, порядковый номер и посылает его в сеть. В компьютере-получателе получение каждого пакета подтверждается (компьютеру-отправителю), а затем они собираются в исходное послание, которое и передается программе-приложению.

TCP может быть “невидим” для пользователя ЭБ, но именно он отвечает за большинство проблем, которые влияют на скорость и качество связи. Не все пакеты прибывают благополучно. Перегруженный маршрутизатор может просто уничтожить (проигнорировать, drop) часть из них. Если это происходит, компьютер-отправитель не получает подтверждения, и через определенное время (time-out) отправляет эти пакеты вновь. Это и воспринимается пользователем как досадная задержка.

TCP гарантирует безошибочную доставку послания, но не гарантирует срок. Для некоторых приложений последнее обстоятельство важнее отсутствия ошибок. Предположим, один компьютер передает аудиозапись, которая немедленно воспроизводится на компьютере-получателе. Даже если один пакет не придет вовремя, для человеческого уха удобнее “заполнить” провал образующийся из-за этого, чем ожидать повторной передачи этого пакета. Поскольку TCP не пригоден для таких приложений, они используют другой протокол, называемый UDP который также может работать с IP-протоколом. Протокол UDP просто отправляет последовательность пакетов “в надежде”, что они дойдут благополучно, не гарантируя их получение.

Врезка 2.2

Семейство TCP/IP протоколов

Семейство TCP/IP представляет собой группу компьютерных программ (базирующихся на TCP/IP) и имеется на большинстве компьютеров. Он включает в себя следующие программы:

Эмуляцию терминала (Terminal Emulation). Эта программа, известная, как Telnet позволяет эмулировать на компьютере терминалы различных (в т.ч. устаревших) типов для того, чтобы предоставить возможность использования данного компьютера в различных целях. Поскольку Telnet предоставляет интерфейс наиболее низкого и простого уровня, он часто используется для системного администрирования.

Передача файлов (File Transfer). Этот протокол (File Transfer Protocol, FTP) позволяет передавать файлы через Интернет. Поскольку он использует протокол TCP, который гарантирует передачи всех пакетов, становится возможной передача больших объемов информации в файлах.

Электронная почта. Почта в Интернет использует простой транспортный почтовый протокол (Simple Mail Transport Protocol, SMTP). Именно он превратил электронную почту из набора локальных сервисов в единую всемирную службу. Данный протокол представляет собой базовый механизм для доставки почты. За последние годы были предложены важные дополнения, которые позволили включать в сообщение символы других кодировок, отправлять мультимедийные послания и файлы как вложения.

Врезка 2.3

Сеть новостей (NetNews, телеконференции, “нюусы”)

Телеконференции (также известные как Usenet) являются важным и содержательным примером подхода интернет-сообщества к открытому распространению информации. Тысячи новостных конференций, называемые newsgroups, организованы по тематико-иерархическому принципу. Иерархия высшего уровня включает группы comp (компьютеры и все связанное с ними), rec, alt (все, что не вошло в остальные темы, alternative) и ряд других. К примеру, телеконференция rec.arts.theatre.musicals предназначена для обсуждения музыки.

Система телеконференций так децентрализована, что никто не имеет полного списка. Любой человек, который хочет принять участие в конференции, просто отправляет свое послание на специальный сервер новостей, который передает его соседнему серверу, и так далее по всей сети.

NetNews в определенном смысле являет собой противоположность ЭБ, поскольку информацией в ней никто не управляет. Нет абсолютно никаких ограничений на то, кто и что может опубликовать. В результате в своих худших проявлениях сеть телеконференций может способствовать распространению порнографии, ненависти, расизма и просто ложной информации. Тем не менее, телеконференции работают очень хорошо. К примеру, люди, которые пользуются языком программирования Python, создали телеконференцию comp.lang.python, в которой они обмениваются технической информацией, задают вопросы и общаются с разработчиками языка.

Интернет-сообщество

В Интернете впервые появилась практика открытых стандартов. В 1997 году Vinton Cerf и Robert Kahn были награждены национальной медалью за развитие технологии (National Medal for Technology) за их вклад в создание Интернета. В сопроводительном дипломе была отмечена их работа по развитию TCP/IP, но особо была выделена их пионерская роль в экономичном и эффективном внедрении интернет-технологий. Сказано также, что они “стойко поддерживали линию на то, чтобы их межсетевые протоколы были свободно доступны для любого”, а TCP/IP был сознательно создан таким образом, чтобы не зависеть от фирм-производителей и поддерживать все виды передачи между любыми компьютерами.

В традициях Интернета по-прежнему тесное сотрудничество по техническим проблемам, а продолжающееся развитие сети остается в руках инженеров. Для некоторых людей невозможно согласиться с тем, что правительство США способно сделать что-то полезное, но факт в том, что создание Интернета возглавили именно правительственные агентства, причем делали они это зачастую в условиях сильного сопротивления именно тех компаний, которые сейчас получают прибыль от их усилий. Недавно была предпринята попытка переписать историю Интернета, дабы выделить определенные интересы, причем некоторые люди присвоили себе достижения за те разработки, которые изначально выполнялись командой независимых друг от друга разработчиков. Существует разительный контраст между согласованной политикой дальновидных чиновников государственных агентств в развитии Интернета и мешаниной несовместимых стандартов в области коммерческих приложений (таких, к примеру, как мобильная телефония).

Важной характеристикой Интернета является тот факт, что инженеры и компьютерные специалисты, которые развивают соответствующие технологии и методы, являются их основными пользователями. Они общаются по электронной почте, называя обычную почту черепашней (snail mail). Если они пишут статью, они делают это на собственном компьютере. Если это веб-страница, они пишут html-разметку в обычном редакторе, а не пользуются соответствующей программой. Компьютерные специалисты высокого уровня могут проводить больше времени за написанием и созданием презентаций, чем за написанием программ, но программирование является базовым навыком, которым каждый из них должен безусловно обладать.

Научные публикации в Интернет

Традиция публиковать серьезные академические материалы в Интернете насчитывает много лет. Врезки 2.4 и 2.5 содержат информацию о двух важных примерах: серии стандартов Интернета RFC и физических архивах E-Print в Лос-Аламосской национальной лаборатории. К сожалению, оба неудачно названы. RFC означает "Request for Comments" ("запрос комментариев", т.е. некий текст, на который хотят получить отклики заинтересованных специалистов), но в настоящее время это определенно серия технических разработок для Интернета. Она включает различную техническую и информацию и формальные стандарты. Сервис в Лос-Аламосе - это не архив в обычном смысле. Его первичная функция - "сервер препринтов" (preprint server) - место, сайт, где исследователи могли бы публиковать результаты исследований сразу по их выполнении, без задержек, связанных с публикацией в обычных научных журналах.

Врезка 2.4

Инженерные группы по развитию Интернета и RFC.

Инженерные группы по развитию Интернета (The Internet Engineering Task Force, IETF) - это сообщество, которое координирует технические аспекты Интернета. Метод ее работы уникален и подтвердил свою исключительную эффективность как способ организовать общую работу большого числа людей из самых разных (в т.ч. конкурирующих) компаний. Первая необычная особенность заключается в том, что IETF открыто для всех. Любой может прийти на встречу, присоединиться к рабочей группе и имеет право голоса.

Базовый принцип работы IETF - "приблизительный (rough) консенсус и рабочий код". Любой, кто хочет предложить новый протокол или другое техническое новшество, может выпустить рабочую записку (называемую Internet Draft) и необходимые дополнительные данные к ней. Эти материалы должны представлять собой программное обеспечение (с открытым исходным текстом). Эти предложения (Internet Draft) обсуждаются на встречах рабочих групп. Если общее мнение сходится на том, что разработка заслуживает дальнейшего развития, то она может быть предложена как стандарт RFC. Это предложение не может превратиться в официальный стандарт до тех пор, пока полная спецификация и соответствующее программное обеспечение не будут доступны для всеобщего использования.

IETF зародилось в США, но в настоящее время стало интернациональной. Каждый год одна встреча происходит за пределами США. Участники, включая лидеров рабочих групп, собираются со всего мира. Если ранее IETF поддерживалось правительством США, то сейчас это самофинансирующаяся структура. Расходы покрываются взносами за участие во встречах.

Процесс в IETF открыт для любого, кто хочет сделать свой вклад. В отличие от других обществ, занимающихся стандартизацией, чьи рабочие предложения трудно получить, а конечные стандарты сложны и дороги, все рабочие предложения и стандарты (RFC) Интернета доступны он-лайн. Поскольку упор делается на реально работающее программное обеспечение, первый же из двух соперничающих подходов, который продемонстрирует свою работоспособность, имеет высокие шансы быть принятым. В результате, базовые стандарты Интернета исключительно просты.

Недавние встречи IETF привлекли свыше 2000 участников. Однако, за счет того, что работа проходит в специализированных группах по конкретным темам, сохраняется атмосфера, чувство интимности. Практически все участники являются действующими инженерами или

компьютерщиками. Начальники остаются дома. Формальная часть встреч коротка, а неформальная - интенсивная и длинная. Многие важные спецификации родились на этих продолжающихся далеко за-полночь сессиях IETF, когда люди из конкурирующих компаний работали вместе.

Благодаря своему быстрому росту, Интернет всегда потенциально уязвим для технических проблем. И именно существование IETF обеспечивает такую большую гибкость. Если бы Интернет контролировался одной единственной компанией, то уровень технических решений был бы хорош ровно настолько, насколько высок уровень ведущих инженеров этой компании. А в рамках IETF, которая взяла на себя контроль за технологиями Интернета, лучшие инженеры мира работают вместе.

Рабочие предложения (Internet Drafts) представляют собой замечательную серию технических публикаций. В научных и инженерных дисциплинах большая часть информации устаревает очень быстро, но журналы остаются на полках навсегда. В данном случае все наоборот. Каждое рабочее предложение начинается с определенной преамбулы, в которой, в частности, говорится: "Рабочие предложения являются временными документами сроком действия не свыше 6 месяцев и могут быть обновлены или заменены другими документами в любое время. Не принято использовать ссылки на рабочие предложения или цитировать их иначе как на документ "в процессе разработки".

IETF публикует в сети любое исходное рабочее предложение и извещает заинтересованных участников по спискам рассылки. Затем начинается рецензирование и рассмотрение. Отдельные участники представляют свои замечания в соответствующие телеконференции (списки рассылки). Комментарии могут быть от детализированных предложений до уничтожающей критики. К тому времени, когда рабочая группа соберется для обсуждения документа, он станет предметом рецензирования признанными экспертами в данной области.

RFC являются официальными публикациями IETF. Эта серия из нескольких тысяч документов начала формироваться в 1969 году и составляет сердцевину документации по Интернету. Наиболее широкоизвестные RFC образуют базовые стандарты (standards track). Они включают формальные спецификации каждой версии IP протокола, электронной почты, компонентов веб и многое другое. Другие типы RFC включают информационные документы, в которых дается соответствующая информация об отдельных аспектах Интернета.

Дискуссии о научных публикациях редко замечают феномен RFC, хотя очень трудно найти другой такой пример научных публикаций, которые проходили бы такой жесткий процесс рецензирования и так широко изучались экспертами в данной области. RFC никогда не публиковались на бумаге. Исходно доступные по FTP, сейчас они также доступны (по-прежнему в формате "только текст") и в веб.

Что бы не говорилось об их названиях, два этих сервиса являются фундаментально важными для публикации результатов исследований в соответствующих областях. Еще они важны потому, что продемонстрировали тот факт, что ЭБ открывают новые возможности решения традиционных задач. Издавна считалось, что для достижения высокого качества научной публикации необходимо предварительное рецензирование. Процесс разработки рабочих предложений является примером очень интенсивного рецензирования, а его отличительной чертой является то, что рецензии появляются на уже опубликованный исходный вариант предложения. Сервис в Лос-Аламосе вообще не предусматривает рецензирования. В тоже время, оба они доказали свою высокую эффективность как методы научного общения.

Они интересны и с экономической точки зрения. Оба сервиса абсолютно бесплатны для пользователей. Они поддерживаются профессионально, имеют бюджет, но никакой платы ни авторы предложений, ни читатели не вносят.

Оба упомянутых сервиса сформировались еще до слияния с веб. Веб был столь успешен, что многие люди уже забыли, что существуют и другие пути для эффективного распространения информации в Интернете. И сервис Лос-Аламоса и RFC сейчас пользуются технологией веб, но исходно они базировались на электронной почте и протоколе FTP.

Архивы E-print по физике представляют собой яркий пример того, как практические ученые используют возможности Интернет-технологии для создания новых форм научных коммуникаций путем расширения практики обмена препринтами статей. Первый архив был учрежден в 1991 году Paul Ginsparg в Los Alamos National Laboratory для обслуживания потребностей группы, занимавшейся физикой высоких энергий. Позднее были созданы архивы по другим областям физики, математики и смежных дисциплин. На конференции ЮНЕСКО в 1996 году Ginsparg сообщал: "Ныне эти архивы обслуживают свыше 35000 пользователей из 70 стран в процессе более чем 70000 электронных транзакций в день. В некоторых областях физики они уже переросли традиционные исследовательские журналы и превратились в своеобразные конвейеры по публикации и архивированию исследовательской информации."

Первичная функция архивов E-print состоит в публикации результатов исследований, зачастую это предварительная версия статьи, которая позже будет опубликована в обычном журнале. Статьи для публикации готовятся по обычным правилам. Многие физики используют формат TeX, но PostScript и HTML также возможны. Графики и данные иногда размещаются в тексте, а иногда приводятся в виде отдельных файлов.

Статья может быть предложена к публикации в архив по электронной почте, по FTP или через веб. Предполагается, что авторы снабжают статью краткой аннотацией и стандартным набором индексных метаданных. Обработка материалов практически полностью автоматизирована. Архивы предоставляют возможность поиска с помощью электронной почты либо через веб-интерфейс, а также уведомление подписчиков о новых поступлениях по электронной почте. Варианты поиска включают поиск в одном архиве, нескольких архивах, по имени, заглавию или полному тексту реферата.

Технология архивов проста. Они используют стандартные форматы, протоколы и сетевые инструменты, известные и понятные исследователям. Пользовательские интерфейсы создавались прежде всего с целью минимизации усилий, необходимых для поддержания архивов. Предполагается, что авторы и читатели установят соответствующее программное обеспечение на своих компьютерах и будут выполнять определенные процедуры.

Это открытая система, финансируемая за счет ежегодных грантов Национального научного фонда и Департамента энергетики. Авторы сохраняют права на свои работы. (Единственное упоминание об авторских правах содержится на странице с благодарностями и disclaimers.)

Как пишет Ginsparg, "из эксплуатации данной системы необходимо извлечь уроки, которые можно и нужно использовать везде, где автор пишет материалы не с финансовыми целями - но с целью сообщить информацию (для распространения знаний, что приносит этим людям выгоду в плане карьеры и профессиональной репутации)."

Веб

World Wide Web (далее для краткости называемый просто "веб") является заметным успехом в истории развития компьютерных технологий. Его можно сравнить с созданием текстовых процессоров и электронных таблиц. Веб и другие сопутствующие технологии являются критически важными для быстрого роста ЭБ. В этой секции приведен краткий обзор веб и ряда "скрытых" технологий, на которых он базируется. Детали и специфические вопросы выходят за рамки данной книги.

Веб - это связанная коллекция информации, расположенная на многих компьютерах по всему миру в Интернете. Эти компьютеры называют веб-серверами. Некоторые сервера и информация на них поддерживаются отдельными людьми или маленькими группами, включая университетские департаменты и исследовательские центры. Другие являются частью корпоративной информационной службы. Некоторые сайты изначально организованы как ЭБ, другие поддерживаются людьми, которые не считают себя издателями или библиотечными работниками (но делают они это зачастую здорово!) Некоторые веб-серверы имеют значительные коллекции высококачественной информации, другие используются в краткосрочных или частных целях и поддерживаются неформально. Часть серверов используется для целей (таких, как маркетинг, к примеру), которые не имеют отношения к ЭБ.

Технология веб была разработана около 1990 года Tim Berners-Lee с коллегами в Европейском центре ядерных исследований (CERN) в Швейцарии. Популярным его сделала "Мозаика", пользовательский интерфейс, разработанный Marc Andreessen с коллегами в университете Иллинойса в Урбане. "Мозаика" появилась в 1993 году. Через несколько лет вышло множество коммерческих версий этой программы. Кроме нее, наиболее широко используются Netscape Navigator и Internet Explorer фирмы Microsoft. Эти пользовательские интерфейсы называют веб-браузерами или просто браузерами (browsers).

Основу успеха веб можно сформулировать так: он предоставил удобный путь для распределения информации в Интернете. Отдельные люди могут публиковать или получать информацию без предварительной подготовки или помощи извне. Лишь небольшой объем знаний о компьютерах необходим для создания веб-сайта, практически ничего не нужно для того, чтобы пользоваться браузером.

Распространение веб стало удивительным феноменом. И дело не только в его технологической простоте, сам способ, каким он распространялся, устранил практически все барьеры для его использования. То, что произошло в университете Карнеги-Меллона, было типичным: люди копировали программное обеспечение из Интернета на свои компьютеры. Затем они загружали собственную информацию, делая ее доступной для других. Их компьютеры уже входили в состав ЛВС и имели выход в Интернет. Уже через 6 месяцев после появления "Мозаики", три человека создали большую коллекцию информации по статистике, английскому языку и литературе и по экологии. По мере распространения Интернета в мире, огромное число людей получало немедленный доступ к этой информации. Естественно, университеты приняли веб официально, создали профессионально разработанные веб-сайты, на которых можно найти информацию об университетах - но сделали они это только после того, как отдельные люди указали путь.

Одна из причин того, что отдельные люди смогли экспериментировать с веб состояла в том, что программное обеспечение было бесплатным и доступным через Интернет. CERN и университет Иллинойса начали традицию свободного распространения своего программного обеспечения для веб-серверов и пользовательских интерфейсов. Сегодня наиболее распространенным программным обеспечением является Apache, бесплатная версия веб-сервера этого университета. Свободное распространение программ через Интернет создало прекрасный стимул для того, чтобы новая технология была адекватно воспринята. Многие "технари" любят поэкспериментировать, но очень не любят бюрократию. Эксперимент же теряет элементы увлекательной игры, если вы должны подписать чек на оплату или просить начальника подписать лицензию на программное обеспечение.

Другой причиной непрерывного успеха веб стало то, что эта технология предоставила способ доступа и к такой информации, которая изначально создавалась для других целей. Браузеры базируются на протоколе Hypertext Transfer Protocol (HTTP), но они поддерживают и другие протоколы Интернета, такие как FTP, NetNews, электронная почта. Встроенная поддержка протоколов Gopher и WAIS (теперь уже почти забытых) позволила ранним электронным коллекциям сосуществовать с первыми веб-сайтами. Другой механизм, Common Gateway Interface (CGI), позволил браузерам создать мостик между веб и любыми другими программными системами онлайн-хранения информации. Именно поэтому сразу с появлением "Мозаики" стали доступны большие объемы информации.

Начиная с первой версии "Мозаики", наиболее распространенные браузеры были доступны для большинства операционных систем (различных версий Windows, Macintosh и UNIX) и теперь входят в базовый комплект большинства компьютеров. Администраторы веб-сайтов могут быть уверены, что пользователи в любой части мира смогут увидеть информацию с этого сайта именно в нужном формате, на каком бы компьютере они не работали.

Технология веб

Технически веб базируется на четырех простых технологиях: гипертекстовый язык разметки (Hypertext Markup Language, HTML), протокол передачи гипертекста (Hypertext Transfer Protocol, HTTP), MIME-типы данных и универсальный указатель ресурсов (Uniform Resource

Locators, URLs). Каждая из этих концепций важна не только для веб, но и в принципе для интероперабельности ЭБ.

HTML

HTML - это язык для описания структуры и способа представления текстовых документов. Врезка 2.6 описывает простой HTML файл и демонстрирует, как обычный браузер может его обрабатывать и представлять.

Как показано в примере, HTML файл содержит как текст, так и специальные коды, называемые тэгами (tags), которые описывают форму или структуру. Тэги HTML легко узнать по парным идентификаторам < и >. Большинство тэгов HTML парные, причем прямой слэш идентифицирует вторую половину тэга. Например, тэги <title> и </title> заключают в себе текст, который будет интерпретироваться как заголовок. Некоторые тэги указывают на форматирование: <i> и </i> указывают на курсив,
 означает начало новой строки. Другие тэги позволяют специфицировать структуру: <p> и </p> размечают отдельный параграф, а <h1> и </h1> обрамляют заголовок первого уровня. Структурные тэги не уточняют формат, который интерпретируется в этом случае браузером. К примеру, многие браузеры перед абзацем вставляют пустую строку, но эта стилистическая особенность определяется именно программой.

Пример врезки 2.6 также демонстрирует еще две особенности HTML, которые являются жизненно важными для успеха веб. Первая - это легкость, с которой можно включить цветную графику в состав веб-страницы. Тэг является инструкцией, предписывающей вставить из отдельного файла картинку, при этом буквы img означают "картинка" (image), а src - источник (source). Символы после знака равенства обозначают имя файла, в котором хранится картинка. С появлением этой простой команды, "Мозаика" привнесла цветную графику в Интернет. До появления веб Интернет был серым. Распространенные приложения использовали неотформатированный текст без графики. Веб стал первой широко распространенной системой, объединяющей форматированный текст с цветной графикой.

Второй и более важной особенностью является использование гиперссылок (hyperlinks). Веб-страница может быть связана с любой другой страницей в Интернете. Пример врезки 2.6 содержит одну ссылку: тэг . За ним следует кусок текста, оканчивающийся тэгом . В браузере этот текст выглядит выделенным, обычно голубым цветом и подчеркиванием. Соглашение весьма просто. Если что-то подчеркнуто голубым, пользователь может кликнуть (указать) на него и будет выполнен переход на другую страницу. Соглашение весьма легкое как для пользователей, так и для создателей веб-страниц. (В данном примере ссылка указывает на страницу в другом компьютере, фактически это домашняя страница журнала D-Lib Magazine).

Базовая концепция HTML может быть изложена весьма коротко и быстро усвоена, а использование даже простых тэгов позволяет создать привлекательный документ. В начале развития веб каждый поневоле занимался самообразованием. Полезной характеристикой HTML является то, что его синтаксис позволяет мелкие неточности. Другие языки программирования имеют очень жесткий синтаксис. Отсутствие точки с запятой в компьютерной программе приводит к неудаче или неправильному результату. В HTML, если разметка более или менее правильная, большинство браузеров смогут ее воспринять. Простота этого языка также упрощает создание программ для интерпретации и воспроизведения веб-страниц.

Врезка 2.6

Пример HTML

Вот пример того, каким образом простой текст в формате HTML хранится в компьютере с использованием тэгов, определяющих структуру и формат:

```
<html>
<head>
```

```
<title>HTML Example</title>
</head>
```

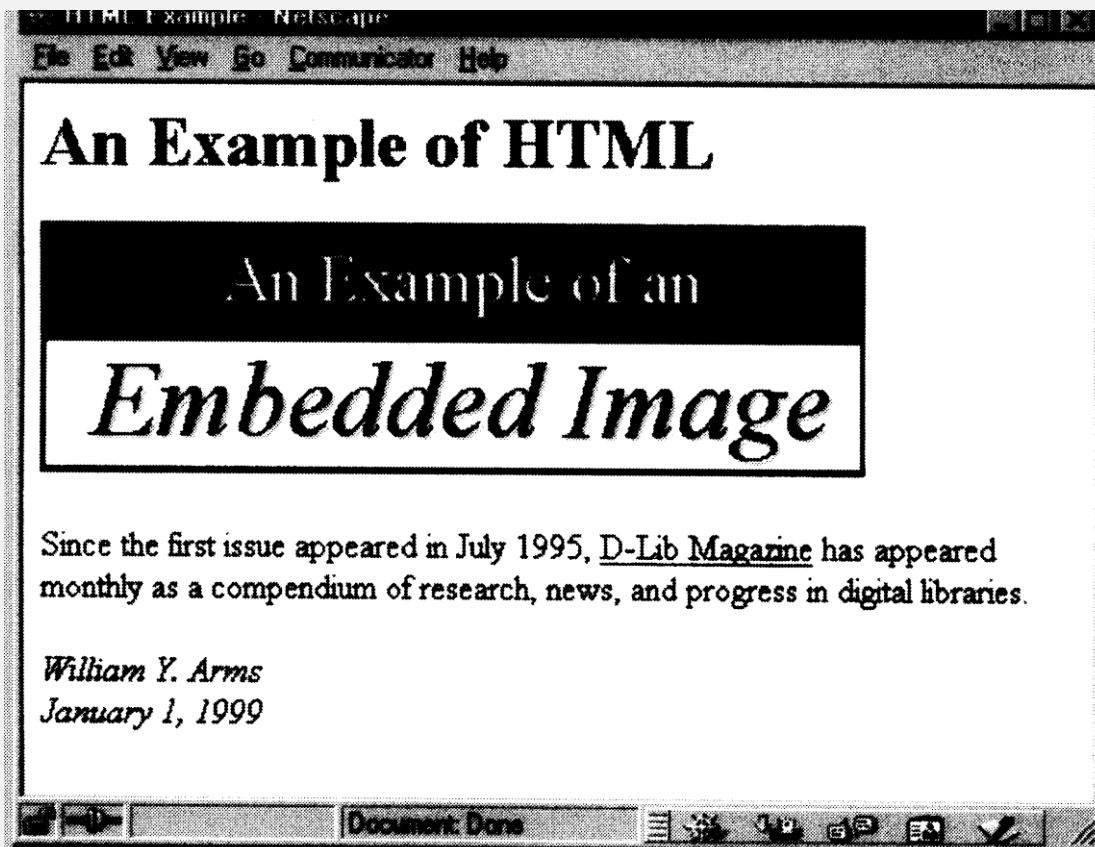
```
<body>
<h1>An Example of HTML</h1>
<img src = "logo.gif">
```

```
<p>Since the first issue appeared in July 1995, <a href = "http://www.dlib.org/dlib.html">D-Lib
Magazine</a> has appeared monthly as a compendium of research, news, and progress in digital
libraries.</p>
```

```
<p><i>William Y.Arms
January 1, 1999</i></p>
```

```
</body>
</html>
```

С помощью браузера этот документ можно увидеть в форме, показанной на рисунке. Реальная форма будет зависеть от браузера, компьютера и выбора опций.



Универсальные указатели ресурсов (URL)

Второй ключевой компонент веб - это универсальные указатели ресурсов (URL). На URL бывает страшно смотреть, но они весьма гибкие. Они предоставляют простой адресный механизм, позволяющий связываться с ресурсами на других компьютерах по всему миру. В примере на врезке 2.6 есть простой URL: `http://www.dlib.org/dlib.html`, в котором `http` означает название протокола, `www.dlib.org` - имя компьютера, а `dlib.html` - имя файла на этом компьютере. Этот URL может неформально интерпретироваться следующим образом: "Используя HTTP протокол, соединиться с компьютером `www.dlib.org` и получить файл `dlib.html`".

HTTP

В информатике протокол - это набор правил, которые используются для обмена сообщениями между компьютерными системами. Типичный протокол включает описание используемых форматов; различные сообщения, последовательности которых должны быть отосланы; соответствующие ответы; условия ошибок и т.д. HTTP - это протокол, который используется для обмена посланиями между веб-браузерами и веб-серверами.

Базовый тип послания в HTTP - это команда `get`. Например, при нажатии на ссылку с URL `http://www.dlib.org/dlib.html` вызывается команда `get`. Неформально она может интерпретироваться как: "Открыть соединение между браузером и веб-сервером, который имеет имя `"www.dlib.org"`. Скопировать файл `"dlib.html"` с этого веб-сервера в браузер. Закрыть соединение."

MIME-типы

Файлы данных в компьютерах представляют собой последовательности битов. Для того, чтобы быть полезными, биты должны интерпретироваться. В примере выше, для того, чтобы корректно отобразить файл `"dlib.html"` браузер должен "знать", что он в формате HTML. Способ интерпретации зависит от того, какой тип данных (data type) содержит файл. Два общих типа данных - это `html` (для текстовых файлов, размеченных в HTML формате) и `jpeg` (для файлов, содержащих закодированные изображения в формате `jpeg`).

В веб, также как и в других Интернет-приложениях, тип данных обозначается согласно схеме, называемой MIME. (Официальное название Типы Среды Интернета, Internet Media Types.) Изначально MIME был создан для описания информации, пересылаемой по электронной почте. Он использует двухчастную кодировку, состоящую из основной и специальной части. Например, `text/ascii` - это тип MIME для текстов в кодировке ASCII, `image/jpeg` - это тип для описания графики в формате `jpeg`, `text/html` - это текст, размеченный HTML тэгами. Существует стандартный набор типов MIME, которые используются многочисленными компьютерными программами, а также дополнительные типы данных, которые могут быть описаны с использованием экспериментальных тэгов.

Важность типов MIME для веб заключается в том, что при передаче по протоколу HTTP командой `get` данные имеют соответствующий тип MIME. Например, файл `"dlib.html"` имеет тип MIME `text/html`. Когда браузер получает файл такого типа, он знает, что соответствующий способ обработки заключается в том, чтобы рассматривать его как HTML текст.

Многие компьютерные системы используют имена файлов как метод указания типов данных. Например, некоторые Windows-приложения используют расширение `.htm` для файлов HTML. Компьютеры под UNIX используют для этих же целей расширение `.html`. Типы MIME являются более гибким и систематическим средством для записи и передачи типов данных.

Информация в веб

Простые компоненты, описанные выше, могут быть использованы для создания множества приложений, причем только часть из них может рассматриваться как ЭБ (т.е. управляемая коллекция разнообразной информации). Вот несколько примеров таких приложений. Многие компании поддерживают веб-сайты, которые предоставляют информацию о компании, ее продукции и сервисах. Можно приобретать авиабилеты через веб-серверы. Ассоциации поддерживают сайты для того, чтобы информировать своих членов. Частные лица имеют свои сайты или персональные странички. Результаты исследований могут быть опубликованы с помощью веб. Многие сайты подходят под определение "мусор". В целом, тем не менее, различные сайты весьма важны для развития ЭБ. Они ответственны за развитие новых

технологий и техники, за проведение экспериментов, за обеспечение единства подходов и стандартизации, за улучшение качества графики и дизайна, за увеличение числа квалифицированных разработчиков, пользователей и веб-мастеров. Размеры веб послужили для многочисленных фирм стимулом для разработки продуктов, значительное число которых сейчас используется для построения ЭБ. Менее приятно то обстоятельство, что успех некоторых веб-сайтов часто приводит к перегрузке некоторые сегменты Интернета, а также порождает обоснованные социальные и юридические сомнения относительно агрессивного поведения в Интернете.

Соглашения веб

Первые веб-сайты были созданы отдельными людьми, которые располагали информацией так, как считали удобным. Однако, уже скоро потребовались соглашения относительно способов организации материалов.

Гиперссылки позволяют создать бесконечное количество вариантов расположения информации на веб-сайте. Однако, пользователям более удобно работать в знакомой структуре. Таким образом, соглашения являются важным элементом при создании ЭБ. Эти конвенции никогда не являются предметом формальной стандартизации, но их широкое распространение позволяет добавить согласованности в веб, который не является согласованным по используемым базовым технологиям.

Врезка 2.7

Консорциум Веб (WWW, W3C)

Хотя центральной организации, контролирующей веб, не существует, должны существовать соглашения относительно базовых протоколов, форматов и их применения, чтобы независимые компьютерные системы могли взаимодействовать между собой. В 1994 году, признавая эту необходимость, Массачусетский технологический институт (MIT) создал Консорциум Веб ("W3C") и пригласил Tim Berners-Lee, разработчика веб, на должность директора. Позднее MIT пригласил иностранных партнеров из французского Национального института по информатике и автоматике и из университета Keio (Shonan Fujisawa campus, Япония). W3C финансируется за счет организаций-членов, которые включают большинство крупнейших компаний, разрабатывающих браузеры, серверы и соответствующие продукты.

W3C представляет собой нейтральный форум, в рамках которого на конференциях и рабочих встречах различные организации работают над разработкой общих спецификаций для веб. Результатом являются коллекции информации для пользователей и разработчиков веб, включая образцы кодов, которые помогают распространению предложенных стандартов. Форум работает в тесной связи с IETF, который разработал такие базовые веб-технологии, как HTTP, HTML и URL.

Поскольку W3C функционирует как нейтральная организация в области, где жестко конкурируют коммерческие компании, его власть зависит от его влияния. Одним из наиболее заметных успехов было быстрое развитие PICS, индустриального стандарта для оценки содержания, который был создан в ответ на обеспокоенность политиков в США возможностью доступа к порнографии и другому нежелательному содержанию для детей и др. групп. Совсем недавно W3C активно участвовал в разработке языка разметки XML.

Хотя Microsoft, Netscape и другие компании могут полагать, что они вырываются вперед, выпуская продукты с нестандартизованными возможностями, такие опции разрушают однородность веб. W3C имеет определенные заслуги в том, что развитие веб-технологий происходит достаточно согласованно.

Веб-сайты

Веб-сайт - это коллекция информации, которую пользователи воспринимают, как отдельную единицу. Иногда веб-сайт соответствует одному веб-серверу. Однако, большой сайт физически может располагаться на нескольких серверах, а один сервер может поддерживать много веб-сайтов.

Домашние страницы (Home pages)

Домашняя страница является вводной страницей коллекции веб-информации. Почти каждый веб-сайт имеет домашнюю страницу. Если в URL не указано имя файла, по общепринятому соглашению сервер открывает страницу с именем index.html. Таким образом, URL <http://www.loc.gov/> интерпретируется как <http://www.loc.gov/index.html>.

Каждый дизайнер имеет свои идеи относительно домашней странички. Однако, как заглавная страница книги обычно следует определенным соглашениям, так и домашняя страница обычно представляет собой обзор веб-сайта (обычно в комбинации с вводной информацией, содержанием и пояснениями по поиску информации).

Маленькие порции информации на данном веб-сайте также могут быть названы домашней страницей. Это касается информации об отдельных департаментах, проектах или сервисах, которые имеют свои собственные страницы в составе веб-сайта компании или организации.

Кнопки

Большинство веб-страниц имеют кнопки с такими надписями, как "next" (следующая), "home" (в начало, домой), "previous" (предыдущая), которые помогают пользователю двигаться (осуществлять навигацию) по сайту. Эти кнопки являются ссылками к отдельным частям сайта.

Иерархическая организация

С точки зрения пользователя, большинство веб-сайтов организованы иерархически. С домашней страницы гиперссылки ведут к нескольким основным разделам; те - в свою очередь, - ведут к более конкретной информации и т.д. Часто кнопки на каждой странице позволяют пользователю вернуться назад на предыдущий более высокий уровень или двигаться на следующую страницу на этом же уровне.

Веб как Цифровая Библиотека

Некоторые люди рассуждают о веб так, словно это было неполноценное подручное средство, которым вынуждены были пользоваться до создания полноценных ЭБ. Один из поводов к такому отношению состоит в том, что для работников других профессий бывает трудно воспринять тот факт, что определенную работу в построении ЭБ выполняли физики из лаборатории в Швейцарии, а не признанные библиотечные работники или компьютерщики. Тем не менее, веб - это не окольная дорога, которой вынуждены пользоваться до появления ЭБ, а центральный элемент в ее построении.

Далее, те, кто не знаком с хорошими онлайн-коллекциями, делают презрительные замечания об информации в веб. Две наиболее распространенные жалобы заключаются в том, что информация плохого качества и что найти ее невозможно. Для этого есть определенные основания, но они весьма далеки от того, чтобы быть истинными. В веб есть огромное количество материалов. Большинство из них действительно малоценны, но многие из веб-серверов поддерживаются регулярно и содержат информацию высокого качества. Поиск

информации в веб может быть сложным, но существуют инструменты и сервисы, которые позволяют пользователю получить большую часть имеющейся информации, проявив минимальную изобретательность.

Тем не менее, сегодня веб скорее в самом начале, чем в конце пути. Несмотря на то, что веб сегодня предоставляет базу для построения ЭБ с лучшими коллекциями, лучшими сервисами и лучшими базовыми технологиями, большинство из современных исследований в области ЭБ могут рассматриваться как простое расширение базовых строительных блоков веб. Мы можем ожидать, что через 25 лет ЭБ будут весьма различными, а начальная стадия развития веб будет вспоминаться с трудом. Возможно, слова "Интернет" и "веб" уйдут в историю или будут обозначать системы, в которых невозможно будет узнать потомков своих прародителей. ЭБ будут вбирать в себя материалы и технологии из разных мест. В течение нескольких лет мы можем, однако, ожидать, что Интернет и веб останутся той базой, на которой будут строиться библиотеки будущего. Как ранние программные продукты на заре эры персональных компьютеров превратились в современные операционные системы, веб может стать основанием для многих поколений ЭБ.

Глава 3. Библиотеки и Издательства

Библиотеки традиционно быстро воспринимают новые технологии, включая микрофильмы, онлайн-информационные сервисы и компакт-диски. (Второй печатной машинкой, импортированной в Бразилию, была машинка для Национальной библиотеки чтобы печатать каталожные карточки.) Последние примеры того, как библиотеки адаптируют новые технологии для своих потребностей - это Интернет и веб.

Однако, базовые коммерческие программные продукты не всегда отвечают нуждам библиотек. Компьютерная индустрия нацеливает свои продукты на большие рынки - бизнес, наука, медицина, военное дело. В США большая часть базовых исследований, на которых базируются такие продукты, финансировалась Министерством обороны или такими компаниями, как IBM. До недавнего времени эти большие рынки проявляли слабый интерес к проблемам управления отдельными фрагментами информации (как в ЭБ) и, как следствие, продукты, реализуемые компьютерной индустрией, не отвечали важным задачам информационного обеспечения. В результате библиотеки были вынуждены заимствовать и адаптировать под свои нужды базовые технологии из других областей.

Разделение ресурсов и онлайн-каталоги

В библиотеках выражение "объединять в сеть" (networking) имеет два значения. До появления современных компьютерных сетей, библиотеки разработали много способов разделения ресурсов, включая межбиблиотечный абонемент, взаимные привилегии по доступу (reciprocal reading privileges) и обмен фотокопиями. Термин, используемый для обозначения таких форм работы, был "объединение в сеть" или "разделение ресурсов"¹. Сервисы, основанные на разделении ресурсов, такие как фотокопирование или доставка документов, предоставляемые Британской библиотекой, улучшали доступ к информации для пользователей всего мира. Однако такое разделение ресурсов ни в коей мере не может соперничать с эффектом использования компьютерных сетей.

Практически каждая библиотека ведет каталог с записями о материалах их коллекций. Каталог помогает читателям найти материал в библиотеке, предоставляет библиографическую информацию и является важным инструментом для управления коллекцией.

Каталогизирование является областью, в которой библиотечные работники используют точную терминологию, часть которой может оказаться малопонятной для посторонних. В тех ситуациях, когда неспециалист может использовать слово "каталог" в своем прямом значении, в библиотечном деле он означает следующее: коллекция библиографических записей, созданная по строго определенным правилам. Чаше, чем повседневное слово "книга", библиотекари используют слово "монография". Многие годы информация, которая включалась в монографическую каталожную запись кодифицировалась согласно определенным правилам ведения каталогов. Англоязычные страны пользуются Англо-американскими правилами каталогизирования (AACR), текущая версия которых известна как AACR2.

Задача внесения в каталог каждой монографии требует значительного времени и опыта. Снижая расходы, библиотеки обмениваются каталожными записями. Главные библиотеки, которые каталогизируют большое количество монографий, такие как Библиотека Конгресса и главные университетские библиотеки, предоставляют свои каталожные записи другим библиотекам бесплатно. Вместо того, чтобы создавать дубликаты записей, большинство библиотек ищет существующие каталожные записи и создают свои собственные только тогда, когда они не смогли найти подходящую для копирования.

¹ "Разделение ресурсов" - технология отдельного (взаимнонезависимого) доступа различных пользователей к одному (общему, "разделяемому") ресурсу. [Прим. пер.]

С конца 1960-х годов обмен каталожными записями был компьютеризирован. Технически, фундаментальный инструмент для представления каталожной записи в компьютере - это формат MARC (Machine-Readable Cataloging). MARC был разработан Henriette Avram с коллегами в Библиотеке Конгресса, изначально для распространения каталожных записей на магнитной ленте. На практике термин MARC cataloging часто используется в общем смысле как для обозначения каталожной записи, так и для компьютерного формата, в которой она хранится. Каталогизирование с использованием протокола MARC распространилось не только на монографии и в настоящее время используется для большинства категорий библиотечных материалов, включая продолжающиеся издания, архивы, неопубликованные статьи (manuscripts). Врезка 3.1 демонстрирует пример записи в формате MARC.

Врезка 3.1

Запись в формате MARC

Рассмотрим монографию, для которой библиографические данные могут быть записаны так: Caroline R. Arms, editor, *Campus strategies for libraries and electronic information*. Bedford, Mass.: Digital Press, 1990.

Поиск в каталоге Библиотеки Конгресса с удаленного терминала, дает следующую информацию в формате MARC:

&001 89-16879 r93

&050 Z675.U5C16 1990

&082 Campus strategies for libraries and electronic information/Caroline Arms, editor.

&260 {Bedford, Mass.} : Digital Press, c1990.

&300 xi, 404 p. : ill. ; 24 cm.

&440 EDUCOM strategies series on information technolog6

&504 Includes bibliographical references (p. {373}-381).

&020 ISBN 1-55558-036-X : \$34.95

&650 Academic Libraries--United States--Automation.

&650 Libraries and electronic publishing--United States

&650 Information technology--United States.

&700 Arms, Caroline R. (Caroline Ruth)

&040 DLC DLC DLC

&043 n-us---

&955 CIP ver. br02 to SL 02-26-90

&985 APIF/MIG

Информация разнесена по полям, каждое из которых имеет трехзначный код. Например, код 440 - это поле названия серии монографий, поля 650 - предметные заголовки (subject headings) Библиотеки Конгресса. Специальные правила говорят библиотечному работнику, какое поле необходимо использовать и как должна интерпретироваться взаимосвязь между элементами.

Реальное кодирование еще сложнее, чем приведенный пример. Полный формат MARC состоит из определенного набора полей, каждое из которых идентифицируется тэгом. Допускаются подполя. Поля обозначаются трехзначным тэгом, подполя - одной буквой. Чтобы получить представление о том, как информация кодируется в формате MARC, рассмотрим поле с заголовком (тэгом) 260 ("выходные данные"), которое начинается как &260. В реальной записи это означает следующее (причем последовательность abc указывает на наличие трех подполей):

&2600#abc#{Bedford, Mass.} :#Digital Press,#c1990.%

Первое подполе, обозначенное тэгом a, указывает на место публикации; следующее, обозначенное как b - издателя; третье, c - дата публикации.

Развитие системы MARC привело к появлению двух важных типов компьютеризированных систем. Первая, предложенная Fred Kilgour (который был основателем OCLC в 1967) - разделенное каталогизирование (shared cataloging). OCLC имеет большую компьютерную систему, которая выросла до более чем 35 миллионов каталожных записей в формате MARC, полученных из Библиотеки Конгресса. Когда член OCLC запрашивает книгу, которую он хотел

бы внести в свой каталог, начинается поиск в базе данных. Если запись найдена, пользователь сгружает запись на свой компьютер и указывает, что эта запись у него имеется. Это называют "каталогизирование копиями" (copy cataloging). Если в базе данных OCLC запись отсутствует, библиотека-пользователь может сама создать ее и передать ее в базу. При такой системе каждый материал описывается для каталога лишь однажды, а интеллектуальные усилия разделяются между всеми библиотеками-участниками. Ведение каталогов в формате MARC и успех OCLC в разделенном каталогизировании нашли последователей по всему миру.

Доступность записей в формате MARC послужила толчком к следующей разработке. Отдельные библиотеки начали создавать свои онлайн-каталоги. В общем случае, основной массив записей был получен методом "каталогизирования копиями". Сегодня практически любая библиотека США имеет свой онлайн-каталог [более точно "интерактивный каталог с публичным доступом" - *прим. пер.*]. На библиотечном жаргоне такие каталоги называются "онлайн-каталог публичного доступа" (Online Public Access Catalog, OPAC). Многие библиотеки приложили большие усилия по переводу старых бумажных каталогов в формат MARC, так что онлайн-каталог является полным. Это позволяет избежать ситуации, когда новые поступления находятся в онлайн-каталоге, а более старые издания - в обычном. Гарвардский университет недавно закончил перевод старого каталога в машинную форму, 5 миллионов записей обошлись примерно в 15 миллионов долларов.

MARC был передовым форматом в то время, когда большинство компьютерных систем представляли текстовые поля только фиксированной длины с заглавными буквами. Он остается актуальным для библиотек, но его возраст хорошо виден. Обсуждение будущего формата MARC осложнено тем огромным объемом средств, которые были инвестированы библиотеками в его внедрение. Что бы не произошло в будущем, MARC был пионерским достижением как в истории компьютерных технологий, так и в истории библиотек. Это ключевой формат, который безусловно должен поддерживаться в ЭБ.

Связывание (linking) онлайн-каталогов и Z39.50

В 80-е годы университетские библиотеки начали объединять свои онлайн-каталоги в сети. Так, в 1984 году в колледже Dartmouth существовала большая ЛВС, а поскольку библиотечный сервер был включен в сеть, то любой житель кампуса мог пользоваться каталогом со своего терминала или компьютера. Так же можно искать информацию и в каталогах других университетов в пределах их ЛВС. Когда же локальные сети были соединены с Интернетом, уже люди со всего мира получили возможность обращаться в эти каталоги. Разделение каталожных ресурсов было первым примером масштабного совместного использования информации с помощью Интернета.

В конце 70-х годов, некоторые библиотеки и исследовательские организации, такие как Библиотека Конгресса, Исследовательская группа по библиотекам (Research Libraries Group), Информационная сеть библиотек Вашингтона (Washington Libraries Information Network) начали проект, известный под названием Linked Systems Project, в рамках которого был разработан протокол, ныне известный как Z39.50. Он позволяет одному компьютеру искать информацию в каталоге, размещенном на другой машине. Первоначально он использовался для поиска записей в формате MARC, но протокол достаточно гибок и не ограничен только этим форматом. Технически, в Z39.50 изложены правила, которые позволяют одному компьютеру искать в базе данных другого и получать найденные записи. Z39.50 широко используется для достижения интероперабельности между различными компьютерными системами.

Рефераты и индексы

Библиотечный каталог является первичным источником информации о монографиях, но он менее полезен для периодики. Каталог дает единственную краткую запись об издающемся

журнале. Поскольку это мало что дает читателю, который хочет найти конкретную статью, появились службы реферирования и индексирования. Типичными примерами являются Медлайн (био- и медицинская тематика), Chemical Abstracts (химия), Inspec (физические науки, включая компьютеры). Эти службы различаются многими деталями, но базовая структура весьма схожа. Профессионалы в конкретных областях читают все статьи из большого списка журналов и либо присваивают индексы, либо составляют рефераты. Некоторые такого рода службы используют тщательно разработанные системы индексов, такие как тезаурус MeSH в Медлайне, разработанный Национальной Медицинской библиотекой. Другие службы менее педантичны. Некоторые генерируют все рефераты сами, другие, как Inspec, используют рефераты от издателей.

Большая часть этих служб начиналась с издания печатных томов для продажи библиотекам, но сегодня практически все виды поиска выполняются компьютером. (Компьютерный поиск по индексам восходит к технологиям пакетной обработки и магнитным лентам.) Некоторые индексные системы поддерживают компьютерные службы, бесплатные для пользователей, другие лицензируют свои услуги третьим сторонам, которые предоставляют различные онлайн-сервисы. Многие крупные библиотеки покупают данные и помещают их на своих серверах. Большинство информации доступно и на компакт-дисках.

После создания онлайн-каталогов, библиотеки начали размещать другие данные, такие как рефераты, индексы и реферативные работы. Эти источники информации могут храниться в центральном компьютере, а найденные записи могут отображаться на удаленных терминалах или персональных компьютерах. Реферативные работы (сборники), состоящие из коротких записей, особенно подходят для подобной формы распространения информации, поскольку пользователи быстро переходят от одной записи к другой и удовлетворительно пользуются простыми дисплеями без графических возможностей. Гибкий и быстрый поиск в данном случае более важен, чем эстетические достоинства монитора. В качестве типичного примера, приведем список информационных источников, доступных в режиме он-лайн в университете Карнеги-Мэллона (УКМ) в 1998 году:

- Библиотечный каталог

- Список журналов

- Библиографические записи архитектурных чертежей и рисунков

- Кто есть кто в УКМ

- Словарь американского наследия (American Heritage Dictionary)

- Рефераты периодики

- ABI/Inform (бизнес-периодика)

- Inspec (физика, электроника, компьютеры)

- Исследовательская директория (исследования и исследователи в УКМ).

Часть этих онлайн-коллекций предоставляет локальную информацию, как "Кто есть кто в УКМ" (университетский справочник по факультетам, студентам и сотрудникам). Руководители библиотек заинтересованы не просто в формальных публикациях или научных материалах. Многие библиотеки предлагают, в частности, множество видов информации от налоговых деклараций до расписания автобусов. Возможности веб-браузеров и полнотекстового индексирования позволяют объединить в единую систему с единым пользовательским интерфейсом как традиционные, так и нетрадиционные библиотечные материалы. Такой подход стал столь обычным, что сейчас уже трудно понять, почему такие примеры объединения информации из разных источников встречались столь редко.

Размещение и поддержание больших объемов информации он-лайн весьма дороги. Несмотря на то, что стоимость компьютерной техники быстро снижается, она все еще существенна. Большие расходы связаны с лицензированием и оплатой персонала, который обеспечивает решение как коммерческих вопросов, так и обработку больших файлов данных. Снижая расходы, библиотеки формируют консорциумы, чтобы одна онлайн-коллекция обслуживала большое число библиотек-пользователей.

Поиск информации (Information Retrieval)

Поиск информации - это центральная задача библиотек. Пользователь - ученый, врач или адвокат - заинтересованы в информации по некоторой теме и хочет найти нужный объект в тематической коллекции. Это требует специального программного обеспечения. В середине 80-х годов в библиотеках появились компьютеры с возможностью полнотекстового поиска в больших коллекциях. Обычно в них прежде всего загружали каталоги библиотеки в формате MARC, а затем - обычные реферативные сборники. Полнотекстовый поиск означает, что пользователь может искать по любому слову в любом месте записи; при этом он не должен знать структуру записи или те правила, по которым она была сгенерирована.

Исследованиям в этой области уже более 30 лет, но базовый подход практически не изменился. Пользователь составляет поисковое выражение, называемое запросом (query). Это может быть как одно слово (как "кактус"), фраза (как "цифровые библиотеки") или длинный запрос (вроде "В каком году Ч. Дарвин путешествовал на "Бигле"?). Задача обработки информации сводится к нахождению объекта в коллекции, который соответствует запросу. Поскольку компьютер не имеет возможности каждый раз анализировать все записи в коллекции по каждому элементу запроса, в компьютере должны быть определенные индексы, которые и позволяют находить нужную информацию.

По мере того, как компьютеры становились все мощнее, а стоимость хранения снижалась, методы обработки информации менялись от тщательной контролируемого поиска коротких записей (вроде каталожных записей или записей в индексах или рефератах) к поиску каждого слова в полных текстах всей коллекции. Раньше высокая стоимость хранения и трудоемкость поиска стимулировали разработку компактных методов хранения и эффективных алгоритмов. В последнее время поисковые программы в веб позволяют искать в больших объемах информации, размещенной на разных компьютерах.

Языки разметки (Markup Languages)

Библиотекарей и издателей объединяет общий интерес в том, чтобы представить в компьютерной форме все богатство печатных материалов. Текстовые документы - больше чем просто последовательность букв, они могут содержать, к примеру, математическую или музыкальную нотацию, знаки других алфавитов, разнообразные шрифты и структурные элементы (заголовки, колонтитулы, сноски и индексы). Желательно было бы закодировать эту информацию и хранить в компьютере вместе с текстом, рисунками, таблицами и другими элементами содержания. Такая кодировка выполняется с помощью языков разметки.

Уже давно ряд организаций, серьезно заинтересованных проблемами обработки текста, разработали схему разметки, известную как стандартный обобщенный язык разметки (Standard Generalized Markup Language, SGML). HTML, формат разметки текста для веб, является просто производным от SGML.

Так как представление документа в SGML не зависит от того, как он будет использоваться, один и тот же текст, размеченный этим способом, может быть отражен во многих формах и форматах: статьях, компакт-дисках, он-лайн, гипертексте и т.д. Эта особенность делает SGML притягательным для издателей, которые желают издать работу в нескольких версиях. Первым примером подобного использования этого языка разметки была подготовка нового издания оксфордского словаря английского языка. SGML широко используется также гуманитариями, которые используют его как метод кодирования структуры текста, который не зависит от компьютерной системы или метода отображения.

Цифровые библиотеки научных журналов

Первые попытки

В конце 80-х годов некоторые издательства и библиотеки попытались создать онлайн-коллекции научных журналов, т.к. технические барьеры, которые делали реализацию таких проектов невозможной, практически исчезли. Стоимость онлайн-хранения снижалась, персональные компьютеры и локальные сети быстро распространялись, становилось доступным хорошее программное обеспечение для построения баз данных. Главной трудностью в построении ЭБ было то, что научная литература была на бумаге, а не в электронном формате, а организации строились вокруг физических носителей, а не вокруг компьютерных сетей.

Одной из первых попыток создать ЭБ в кампусе университета был проект УКМ Mercury Electronic Library в 1987-93 гг. Предпосылкой послужила передовая компьютерная инфраструктура УКМ, которая включала скоростную ЛВС, отличный компьютерный центр. К тому же, в университетской библиотеке была традиция инноваций. Несколько позднее начался проект CORE в Корнельском (Cornell) университете, в котором предполагалось создать по материалам журналов онлайн-коллекцию химических структур. Оба эти проекта выполнялись в сотрудничестве с издательствами, которые выполняли сканирование журналов, создавая тем самым онлайн-коллекции графических образов страниц.

В то время как Mercury задумывалась как система обслуживания, в проекте CORE делался упор на исследование способов создания пользовательских интерфейсов и других аспектов информационных систем для химиков. Хотя в обоих проектах предусматривалась конверсия статей из журналов в битовые изображения, такой подход не выглядит перспективным для ЭБ, он просто отражал тот факт, что издатели не выпускали журналы в других форматах. Широко используется практика "печати в файл", однако этот подход ориентирован именно на печатные материалы, к тому же существует много форматов таких файлов. Зачастую даже корректуру невозможно было внести непосредственно в исходный файл, так что этот метод также не пригоден для ЭБ.

После проектов Mercury и CORE был еще ряд проектов с использованием сканированных изображений. Один из наиболее известных - это проект Tulip издательства Elsevier. В течение 3 лет издательство поставляло группе университетов (включая УКМ и Корнелл) сканированные изображения содержания 43 журналов в области материаловедения. Каждый университет размещал эти материалы на своем сервере и предоставлял локальный доступ.

Все эти три проекта не были долгосрочными. Каждый был жестко ограничен техническими возможностями и представлял собой незначительную по размеру коллекцию. Но они продемонстрировали потенциальные преимущества ЭБ, которые могут быть реализованы на практике.

Следующее поколение разработок в области электронных издательств началось в условиях значительного снижения стоимости хранения электронных данных, что позволяло увеличить объем онлайн-коллекций. Объединение возможностей веб с широким распространением веб-браузеров положило начало длительной эволюции пользовательских интерфейсов. Веб-браузеры не идеальны для ЭБ, хотя они вполне пригодны для начальных экспериментов. В то же время они обладают тем огромным преимуществом, что существуют в версиях для всех стандартных компьютеров и операционных систем и, таким образом, отпадает необходимость в рамках данного конкретного проекта разрабатывать собственный интерфейс для каждого типа компьютера.

Врезка 3.2 Mercury и CORE

Проект Mercury был рассчитан на пять лет и имел целью построение прототипа ЭБ в УКБ. Начавшись в 1988 году, он просуществовал до 1991 года в виде дюжины текстовых баз данных

и небольшого числа изображений страниц журнальных статей по компьютерной тематике. Он представляет собой хороший пример того, как обстояла ситуация до появления веб.

Одной из целей проекта было получения изображений журнальных статей от издателей по лицензиям. Были выделены 4 издательства, которые публиковали 16 из 20 наиболее популярных научных журналов в области компьютеров: Association for Computing Machinery, Institute of Electrical and Electronics Engineers, Elsevier и Pergamon. В ходе выполнения проекта Pergamon стал собственностью Elsevier. Ни одно из издательств не имело машиночитаемых версий своих журналов, но не позволяло сканировать печатные копии для использования в проекте. Таким образом, важной частью проекта было преобразование, хранение и доставка изображений страниц в ЛВС кампуса.

Другой фундаментальной парадигмой проекта Mercury был поиск в текстовой базе данных с целью идентификации информации, которая должна была выводиться на дисплей. Была выбрана ранняя версия Z39.50 в качестве протокола для обмена запросами между клиентскими компьютерами и серверами, в которых хранились индексы. В рамках Mercury появилась концепция "сервера ссылок" (reference server), который хранил информацию об информации, хранящейся на других серверах, списки полей, в которых возможен поиск, индексы и ограничения на доступ. Для того, чтобы предоставлять битовые изображения, в рамках проекта был разработан новый алгоритм хранения графики в сжатом виде, ее передачи по сети, разархивирования и вывода на дисплей - с общим временем обработки 1-2 секунды на страницу.

Поскольку Mercury имел выход в Интернет, а большинство материалов было лицензионными, важное значение приобрели вопросы безопасности. В университете уже существовал набор сетевых сервисов под названием Andrew, так что в рамках проекта Mercury использовались соответствующие возможности по идентификации и распечатке. На другие компьютеры информация доставлялась с помощью электронной почты.

CORE был совместным проектом фирм Bellcore, OCLC, Корнельского университета и Американского химического общества и выполнялся с 1991 по 1995 гг. Было отсканировано около 400 000 страниц, что составляло четырехлетнюю подборку статей из 20 журналов Американского химического общества.

В проекте был использован ряд идей, до сих пор популярных в проектах такого типа. Существовало 2 версии каждой статьи: сканированное изображение и тестовая версия в разметке SGML. Сканированные страницы позволяли убедиться, что выведенная на экран или на принтер страница имеет ту же верстку, что и оригинальная печатная версия. Текст в формате SGML использовался для построения полнотекстовых индексов для обработки информации и быстрого вывода на дисплей. Хранились два отсканированных изображения одной страницы - одна для распечатки, другая для дисплея. Первая была черно-белой с разрешением 300 точек на дюйм. Экранная версия была в "градации серого" с разрешением 100 точек на дюйм.

CORE был одним из первых проектов, в котором были получены результаты, подтвержденные с тех пор многочисленными исследованиями. Один из выводов заключался в том, что технические проблемы по хранению и представлению сложных научных материалов весьма значительны, особенно если оригинал изначально создавался для печати и лишь затем оцифровывался. Проект CORE также высветил проблему масштабируемости. Коллекция CORE, состоящая лишь из 20 журналов, занимала 80 гигабайт. Во всех ранних проектах по оцифровке было обнаружено, что технические проблемы управления большими объемами данных были значительнее, чем это предполагалось.

Несмотря на эти недостатки, в ходе изучения принципов построения пользовательских интерфейсов были достигнуты обнадеживающие результаты. Химикам коллекция нравилась. Хотя они могли предпочесть печатную версию для детального ознакомления со статьей, они обнаружили, что в онлайн-версиях легче выполнять поиск, а изображение на экране достаточно приемлемое. Проект CORE был одним из первых, в котором была подчеркнута важность просмотра (browsing) как возможной альтернативы поиску (searching) - и этот вывод был подтвержден развитием веб.

Электронные журналы

До середины 90-х годов традиционные издательства сомневались относительно онлайн-публикации научных журналов. Несмотря на то, что коммерческое издание на компакт-дисках стало уже частью индустрии, только несколько журналов были доступны он-лайн. Издатели не понимали правил ведения этого бизнеса и опасались, что пострадает реализация печатных версий. Однако к 1995 большие изменения в том, как люди пользуются информацией, стали очевидны. Библиотеки и отдельные пользователи все в больших масштабах пользовались онлайн-сервисами и другими формами электронной информации. Печатные публикации были вынуждены бороться за свою снижающуюся долю практически постоянного по объему бюджета библиотек.

В то время, как оборот издательств остается практически постоянным или чуть снижается, они попадают в ценовую спираль. В прошлом издатели использовали новые технологии для снижения производственных издержек, компенсируя таким образом некоторое снижение оборота. Но в настоящее время потенциал этого процесса практически исчерпан. Издатели понимают, что для того, чтобы расширять или хотя бы сохранять свой бизнес, они должны производить продукты для растущего онлайн-рынка. Электронная информация рассматривается ими как растущий перспективный рынок.

Врезка 3.3

Издательство High Wire Press

Издательство High Wire Press, которое существует при библиотеке Стэнфордского университета, предоставляет онлайн-доступ к ряду лучших научных и медицинских журналов за счет сотрудничества с научными и профессиональными обществами, которые их издают. Успех этого начинания может быть отнесен на счет большого внимания, которое уделялось интересам профессиональных ученых, стремлению сосредоточиться на ключевых журналах и техническому превосходству.

High Wire Press начало в 1995 с онлайн-версии Journal of Biological Chemistry, поскольку несколько членов редколлегии работали в Стэнфорде. Еженедельный номер содержит около 800 страниц и, конечно, никто не читает его от корки до корки. Интерфейс, предложенный High Wire Press, трактует каждую статью как отдельный документ. Есть два способа найти нужный: просматривая содержание номера за номером или поиском. Опции поиска включают поля: авторы, слова в заголовке, полнотекстовый поиск в рефератах и полнотекстовый поиск в оригинальных статьях. Поисковый интерфейс был создан таким образом, чтобы представлять статью так, словно это действительно Journal of Biological Chemistry, а не публикация High Wire Press. Были предприняты большие усилия для того, чтобы адекватно отображать на экране греческие, математические и иные специальные символы.

За три года High Wire Press прошло путь от экспериментов до предоставления онлайн-доступа к сотне журналов, включая Science. Несмотря на свои ресурсы, American Association for the Advancement of Science (AAAS), издатель Science, испытывала определенные опасения, характерные для обществ, которые публикуют всего несколько наименований журналов: что усилия по предоставлению онлайн-доступа приведет к росту штата и ухудшит качество их работы. Партнерство с High Wire Press позволило AAAS разделить соответствующие затраты и иметь возможность сотрудничать со специалистами из Стэнфордского университета. AAAS было восхищено числом посетителей их сайта еженедельно, причем большинство из них не являлось регулярными читателями Science.

Большинство крупных научных издательств быстро перешло к электронным версиям. Врезка 4.3 дает некоторые детали подхода к этой проблеме Association for Computing Machinery (ACM). С небольшими вариациями этой же стратегии следовали Elsevier, Wiley, Academic Press и другие большие коммерческие издательства, Американское химическое общество, университетские издательства, включая MIT Press и другие.

Врезка 3.4

Цифровая библиотека Association for Computing Machinery

Association for Computing Machinery - это профессиональное общество, которое публикует 17 журналов, посвященных компьютерным технологиям. Кроме этого, 38 его профильных отделений проводят большое число конференций, большая часть из которых публикует сборники трудов. Членами общества являются практикующие специалисты в области компьютеров, включая многих из тех, кто создавал Интернет и веб. Они были среди тех первых людей, кто пользовался онлайн-способами связи и они ожидают, что их профессиональное общество будет лидером в процессе перехода к онлайн-журналам.

Традиционно, основной версией журнала была печатная. В 1993 ACM решило, что частью процесса подготовки издания должно быть использование компьютерных систем, с помощью которых создавалась база данных журнальных статей, трудов конференций, популярных журналов (magazines) и бюллетеней (newsletters) - в разметке SGML. Позднее ACM решило конвертировать большое число своих старых журналов для построения ЭБ своих публикаций начиная с 1985 года.

SGML файлы использовались для подготовки печатных версий. При этом ACM готовилось к тому, что ее члены будут иметь доступ к статьям непосредственно в онлайн-базе данных, частично читая непосредственно с экрана, а иногда распечатывая на принтере. Предусматривалась возможность продажи частичной лицензии или полной подписки.

Коллекция, доступная он-лайн с 1997 года, функционирует под веб-интерфейсом, который позволяет просматривать содержание страниц журналов ACM или проводить поиск по автору или ключевому слову. После того, как статья найдена, подписчик может прочитать полный текст. Остальные пользователи должны оплатить получение полного текста, хотя рефераты могут читать бесплатно.

С одной стороны, ACM необходимы доходы, компенсирующие значительные затраты на ЭБ, а с другой, они не хотят предъявлять чрезмерные требования к авторам и читателям. Изначально были избраны два разных коммерческих подхода - к читателям и к авторам.

В 1994 году ACM опубликовало свою политику в области прав интеллектуальной собственности, в которых излагались взаимоотношения общества (как издателя) и авторов. Была предпринята попытка найти баланс между стремлением к получению доходов и интересами авторов. Как признак перемен, эта политика была размещена прежде всего на веб-сервере. Одна из ключевых особенностей этой политики - поощрение такой практики, чтобы большинство журнальных статей прежде всего распространялись через веб.

В качестве источника дохода ACM рассматривает плату за доступ к полному тексту. Члены общества и библиотеки могут подписаться на электронные версии журналов с 20% скидкой по сравнению с печатной версией. Частные лица могут платить за конкретную статью. Структура цен направлена на поощрение подписки на полный набор журналов, а не на отдельные наименования.

Термин "электронные журналы" (electronic journals) обычно используется для обозначения публикаций, которые сохраняют большинство характеристик печатных изданий, но производятся и распространяются он-лайн. Вносит некоторую путаницу то обстоятельство, что этот термин используется как для журналов, которые полностью цифровые и существуют только в онлайн-версии, так и для электронных версий печатных изданий (как это сделано в ЭБ ACM, см. Врезка 3.4).

Многие крупные издательства создали некоторое число чисто онлайн-периодических изданий. Похожие усилия предпринимают другие издательские группы. Часть таких изданий откровенно отказалась от процессов и процедур, принятых в традиционных журналах. Одним из наиболее амбициозных проектов с использованием такого подхода стал On-line Journal of Current Clinical Trials, развиваемый American Association for the Advancement of Science и OCLC. В отличие от других онлайн-изданий, в которых электронная версия была вторичной по отношению к печатной, этот новый журнал изначально задумывался как высококачественное реферируемое, основной версией которого будет именно электронная. (Печатной версии вообще не было.) Поскольку у издательства был полный контроль над изданием, были все возможности для того, чтобы создать и хранить его в такой форме, чтобы

оптимизировать именно электронную доставку и представление, но журнал так и не был воспринят исследователями и врачами. Журнал появился в 1992 году, но так и не смог привлечь то число высококачественных статей, на которое изначально рассчитывался. Такова судьба многих пионерских начинаний.

Более поздние электронные периодические издания объединяют как особенности традиционных журналов, так и преимущества онлайн. Одно из них, D-Lib Magazine, которое мы начали в 1995 году как онлайн-журнал об исследованиях и инновациях в области ЭБ. Если печатные издания должны следовать жестким правилам дизайна и верстки, онлайн-публикации могут дать авторам возможность творческого использования технологических возможностей. Традиционные журналы выходят номерами, каждый содержит несколько статей. Некоторые электронные журналы публикуют статьи по мере поступления, но D-Lib Magazine публикуется пунктуально ежемесячно, причем срок публикации статьи весьма невелик. Времена меняются - и D-Lib Magazine, журнал онлайн- и не привлекающий рецензентов, стал одним из важных журналов в области ЭБ.

Конверсионные проекты и цивилизации (humanities)

Одна из фундаментальных задач исследовательских библиотек - сохранить нынешние материалы как долговременную память завтра. Большие библиотеки имеют замечательные коллекции, которые представляют собой "ткань истории" и цивилизации. В большинстве эти коллекции содержат печатные материалы и другие физические объекты. Развитие ЭБ породило большой энтузиазм относительно возможности "оцифровки" некоторых таких коллекций. Одним из доводов в пользу этого служит плохое физическое состояние старых материалов. Создание цифровой копии сохраняет содержание материала и делает его доступным во всем мире.

Во многих проектах ЭБ предусматривается конверсия существующих бумажных документов в растровые изображения. Печатные документы сканируются постранично. В ранних экспериментах страница записывалась как последовательность черных и белых точек (обычно 300 на дюйм). Затем обычными стали цветные сканеры с высоким разрешением. Хорошие растровые изображения достаточно отчетливы для того, чтобы быть представленными на дисплее высокого разрешения или на бумаге. Поскольку такие сканеры создают огромное количество точек на страницу, используются различные алгоритмы сжатия, чтобы хранить и передавать по сети файлы меньшего размера. Но и в этом случае страничка даже с простой графикой занимает десятки килобайт.

В связи с тем, что движущей силой развития ЭБ являлись научные и технические дисциплины, развитие других существенных потребностей, вроде отображения неанглийских кодировок, игнорировалось. Гуманитарии могли "отстать" от прогресса, но новое поколение ученых-гуманитариев уже воспринимает компьютерные технологии. К счастью, у них есть друзья. Врезка 3.6 описывает проект Фонда Andrew W. Mellon JSTOR, который экономит расходы библиотек и предоставляет доступ к важным журналам широкой аудитории, что вряд ли было бы возможным без ЭБ.

Врезка 3.5

Программы American Memory и Национальная цифровая библиотека (National Digital Library)

Библиотека Конгресса, крупнейшая в мире, имеет восхитительную коллекцию уникальных и неопубликованных материалов. Среди этих сокровищ бумаги 23 президентов США. Редкие книги, памфлеты, документы представляют собой ценный материал для изучения исторических событий, периодов времени и процессов. Миллионы фотографий, карт, нотных сборников, звукозаписей, киноплёнок в различных форматах отражают тенденции, представляют людей и события. До недавнего времени каждый, кто хотел воспользоваться

этими материалами, должен был обращаться непосредственно в здание библиотеки на Капитолийском холме.

American Memory - это пилотная программа с 1989 по 1994, в рамках которой были переведены в компьютерные форматы некоторые части коллекции, которые отбирались с точки зрения ценности для понимания истории и культуры Америки. При этом исследовались особенности работы с различного типа материалами: оттисками (prints), негативами, ранними киноплёнками, звукозаписями и текстовыми документами. Впервые American Memory использовал комбинацию цифровых форматов и компакт-дисков, а также аналоговые видеодиски. В июне 1994 года три коллекции фотографий стали доступны в веб.

Программа национальной ЭБ основывается на успешных результатах проекта American Memory. Ее цель - перевести в электронную форму миллионы объектов и сделать их доступными через Интернет. И снова упор делается на материалы, значимые для американской истории - записные книжки Walt Whitman, документы Континентального Конгресса (Continental Congress) и Конституционной Конвенции (Constitutional Convention), к примеру.

Некоторые из этих коллекций, представляют собой упорядоченные архивы документов или фотографий отдельных лиц или организаций. Некоторые объединены по физическому принципу - коллекции дагерротипов и бумажные оттиски с ранних киноплёнок. Другие представляют собой тематические подборки для учителей или школьников с использованием все источников.

Проект American Memory показал, что школьные учителя с большим энтузиазмом восприняли возможность доступа к подобным материалам, поэтому программа национальной ЭБ прежде всего нацелена на образовательные нужды. Наиболее современный способ работы с коллекциями - это поиск по индексам. Однако многие архивные коллекции были конвертированы без заведения индивидуальной каталожной записи для каждого объекта; вместо этого они внесены в "помощники поиска" (finding aids) - структурные служебные документы, описывающие коллекции как группу материалов без описания индивидуальных объектов. Таким образом, доступ к материалам American Memory организован как комбинация методов, включая поиск библиографических записей для конкретных объектов (если такая запись существует), просмотр оглавлений, полнотекстовой поиск и - в будущем - поиск с использованием "помощников поиска".

Программа имеет важные технические аспекты - в связи с ее масштабностью и наглядностью. Понимая сложность проблем долговременного хранения огромных коллекций, сотрудники Библиотеки Конгресса уделили большое внимание принципам ее организации. Программа Национальной ЭБ серьезно рассматривает проблемы конвертации старых документов в цифровые форматы, отбирая наиболее подходящие форматы для адекватного представления содержания при непрерывном тщательном контроле качества. Текстовые материалы конвертируются дважды: сканируется страница и создается текст в разметке SGML. С каждой фотографии делается несколько изображений, начиная с маленьких "иконок" (thumbnail) до максимального разрешения для архивных целей.

Большинство материалов, отобранных для конвертации, не имеет собственников интеллектуальных прав или иных ограничений на распространение и использование - но для некоторых они существуют. В дополнение к интеллектуальным правам, могут возникать особые условия в связи с требованиями организаций и лиц, предоставляющих оригинальные документы. В случае более старых материалов (в особенности, неопубликованных) зачастую просто невозможно предусмотреть все возможные юридические коллизии, учитывая сложность и дороговизну полного исследования этих аспектов для каждого материала. В результате юристы библиотеки разработали специальную политику и процедуры, в которых пытались сбалансировать ценность того, что материалы будут доступны для использования и риска ненамеренного нарушения чьих-либо прав.

Другие библиотеки признают лидирующую роль Библиотеки Конгресса. Она является активным членом нескольких консорциумов (collaborations), проводит программу грантов и становится спонсором исследований в области ЭБ. Опыт, накопленный в Библиотеке Конгресса, и партнерство с лидирующими технологическими группами помогает ей вносить ощутимый вклад в развитие всего библиотечного сообщества.

Библиотеки и инновации

Библиотеки по природе своей - консервативные организации. Их коллекции и каталоги создаются десятилетиями и веками. Новостные сервисы создаются весьма осмотрительно, поскольку планируются на долгосрочную перспективу. Однако в техническом плане библиотеки бывают даже авантюристичны. Возможно, не все использованные технологии были изобретены в рамках таких проектов, как MARC, OCLC, Linked Systems Project, Mercury, CORE - но использование их библиотеками в широком масштабе на практике определенно было пионерским.

В главе 2 обсуждалось сообщество, которое сложилось, выросло вокруг Интернета и веба. Многие члены этого сообщества лишь недавно открыли для себя онлайн-информацию и совершенно определенно, что библиотеки и издательства предложили многие концептуальные для ЭБ подходы задолго до появления веба.

Врезка 3.6

JSTOR

JSTOR - это проект, который был инициирован фондом Andrew W. Mellon для того, чтобы предоставить академическим библиотекам доступ к ретроспективным коллекциям важнейших журналов. Он преследует как научные, так и экономические цели. Научная цель состоит в создании надежного архива важных журналов и организации широкого доступа к нему. Экономическая - сократить расходы библиотек за счет устранения дублирования коллекций журналов.

Коллекция JSTOR организована по тематическому принципу - экономика, история, философия и т.д. На первом этапе предполагалась охватить около 100 журналов в 15 областях. Для каждого журнала предусматривается хранение всех выпусков от первого до номера пятилетней давности.

Организационно JSTOR была основана в августе 1995 года как независимая и некоммерческая. Ее цель - достичь самоокупаемости за счет оплаты доступа к коллекциям, причем цены устанавливаются на уровне ниже, чем соответствующие расходы на хранение бумажных экземпляров журналов.

Есть три офиса JSTOR. Его административная, юридическая и финансовая активность управляется и координируется из центрального офиса в Нью-Йорке, так же как связи с издателями и библиотеками. Офисы в университетах Мичигана и Принстона поддерживают две синхронно-эквивалентные копии базы данных, совершенствуют техническую инфраструктуру, предоставляют доступ пользователям и непосредственно контролируют конверсию бумажных документов в компьютерные форматы. (Реальное сканирование и составление ключевых слов (keying) осуществляют сторонние организации.) Недавно JSTOR открыл третий зеркальный сервер в Англии, в университете Манчестера.

JSTOR имеет прямые лицензии от издателей или агентств. Подчеркивая ретроспективный характер проекта, JSTOR тщательно избегает конкуренции с издателями, чьи основные доходы - от продаж текущих выпусков. Поэтому доступ к коллекциям предоставляется только академическим библиотекам, подписанным на всю коллекцию полностью. При этом оплата зависит от размеров библиотеки-подписчика. В лучших традициях Интернета условия подписки, оплаты и лицензирования доступны он-лайн.

Все страницы каждого номера каждого журнала сканируются с высоким разрешением (600 точек на дюйм) с обязательным контролем качества. В отличие от ряда других проектов, хранится только одно изображение. Другие версии, как "иконки", создаются динамически при обращении - но не хранятся. Конверсия текста осуществляется с помощью программ оптического распознавания с последующей корректурой. Этот текст используется только для индексирования. Дополнительно для каждой статьи создается файл с оглавлением, включая библиографическое описание, ключевые слова и реферат (если таковой имелся).

Глава 4. Инновации и исследования

Сама область ЭБ и электронных издательств является инновационной. Некоторые новшества приходят как ответ на требования информационного рынка, другие являются результатами систематических исследований в университетах и корпорациях. В этой главе дано как общее представление о процессе исследований, так и обзор текущих разработок. Большая часть затрагиваемых здесь вопросов будет предметом детального рассмотрения в следующих главах.

До недавнего времени инновационный процесс в библиотеках и издательствах был несистематическим. Бюджеты на исследования и разработки зачастую отсутствовали. Существенными исключениями были Национальная Медицинская библиотека США и Исследовательский центр OCLC - они проводили именно исследования новых подходов, не загадывая заранее, каким именно образом эти идеи будут востребованы. Другие библиотеки и издательства подходили к исследованиям с точки зрения практических нужд - как улучшить предоставляемые сервисы и продукты в краткосрочной перспективе. Скорость инновационного процесса лимитируется развитием технологий и доступностью финансовых ресурсов - но также и существующим пониманием запросов пользователей.

Издательства, будучи хорошо организованными для внедрения новых идей, тем не менее выполняли ограниченное число исследований. Большие издательства имеют ресурсы для постоянного инвестирования в новые предприятия, но они рассматривают новые проекты именно как развитие бизнеса, а не исследования. Конечно, процесс перехода к онлайн-доступу к научному журналу не может быть назван исследованием, - но он в такой степени меняет существующий производственный процесс, что одновременная генерация печатной и онлайн-версии превращается в комплексную задачу.

Библиотеки более восприимчивы к инновациям, но они хуже приспособлены для внедрения новых технологий. Как правило, почти весь свой бюджет они расходуют на текущие нужды, а новшества рассматриваются как нечто "дополнительное", а не ключ к будущему. Большие библиотеки имеют большой бюджет, но его роспись столь негибкая, что исследования и инновации остаются сильно недофинансированными.

Программа "Национальная ЭБ Библиотеки Конгресса" может рассматриваться как наиболее важный библиотечный проект в США. Он служит образчиком того, как библиотека расширяет свои функции за традиционные рамки в таких вопросах, как изменение методов управления коллекциями, обработка новых типов материалов, предоставление широкого доступа. Тем не менее Библиотека Конгресса оказывает программе недостаточную финансовую поддержку. Большая часть средств поступает от частных фондов и пожертвований - в результате люди работают по краткосрочным контрактам. Библиотека Конгресса не поддерживает технические разработки вроде MARC или Z39.50; ее наиболее продвинутые проекты не имеют долгосрочного финансирования.

Несмотря на определенный консерватизм, библиотеки, особенно университетские, продвигаются вперед широким фронтом. Структура университетских библиотек препятствует радикальным изменениям, но библиотечные работники отдают себе отчет в том, что за компьютерами - будущее в процессе образования. При этом сотрудники компьютерных департаментов этих университетов ведут себя словно зайцы - они "прыгают" вперед, пытаясь открыть новые направления, потом перепрыгивают куда-то еще и т.д. А большие библиотеки - это в некотором роде черепахи. Они медленно движутся вперед, каждый их шаг подготавливает следующий. Шаги часто делаются в форме конкретных проектов, иногда в сотрудничестве с другими организациями на средства грантов различных фондов. Конечно, отдельные проекты могут выглядеть менее внушительно по сравнению с масштабными правительственными инициативами, но их суммарное воздействие в долгосрочной перспективе представляется ничуть не меньшим.

Многие заметные проекты зародились в университетских библиотеках. Некоторые конвертируют документы в цифровые форматы, другие работают с издательствами по онлайн-доступу. Издательство Стэнфордского университета High Wire Press переносит журналы в Интернет. Университетские библиотеки сотрудничают с Elsevier в рамках проекта Tulip по созданию цифровых версий журналов. Университеты Мичигана и Принстона вносят свой вклад в проект JSTOR по созданию ретроспективных коллекций базовых журналов. Университеты Rutgers и Принстона совместно создают Центр электронных гуманитарных текстов (Center for Electronic Texts in the Humanities).

Большинство проектов стало возможным за счет средств фондов, промышленности, федерального правительства. Например, проект Mercu в университете Карнеги-Мэллона основные гранты получил от благотворительного фонда Pew, корпорации Digital Equipment и от Агентства передовых исследовательских проектов Министерства обороны США (DARPA). Меньшие - но также весьма желанные - взносы сделали другие доноры. Несколько частных фондов щедро поддерживают проекты ЭБ в гуманитарных областях, особенно фонды Andrew W. Mellon и J. Paul Getty. Наконец, хотя бюджет Национального гуманитарного фонда (National Endowment for the Humanities) невелик по сравнению с NSF и DARPA, значительная его часть расходуется на поддержку ЭБ.

Специалисты в области компьютеров к исследованиям подходят серьезно. Они как правило выполняют долгосрочные проекты, конечным результатом которых являются не продукты или сервисы, - но новые концепции или углубление понимания проблемы. В США большая часть денег на такие проекты поступает от федеральных агентств. Самым большим спонсором в области компьютеров является DARPA (чье название несколько раз менялось от DARPA к ARPA и обратно). Второй по значимости - Национальный научный фонд. Несмотря на то, что DARPA является подразделением Министерства обороны и его конечной целью является поддержка военной науки, оно всегда широко подходило к проблемам. Оно поддерживало фундаментальные исследования практически по всем аспектам компьютерных технологий, в особенности необходимо отметить его вклад в создание крупномасштабных экспериментальных систем. NSF отвечает за развитие науки и техники в США, в рамках своих программ он расходует свыше 3,3 миллиарда долларов в год на 20000 научных и образовательных проектов. Он поддерживает проекты в области вычислительной техники и применения компьютерных технологий практически во всех научных дисциплинах.

Многие инженерные и компьютерные компании имеют большой исследовательский бюджет - зачастую не менее 10%. Большинство промышленных разработок имеет целью быстрое создание и внедрение новых продуктов, но из лабораторий таких фирм, как Xerox, Bell, IBM вышло много фундаментальных разработок. Недавно исключительно сильную исследовательскую команду создала фирма Microsoft.

Большая часть базовых технологий ЭБ были разработаны людьми, чьи основные научные интересы лежали совсем в других областях. Интернет, без которого ЭБ выглядели бы совсем иначе, изначально создавался ARPA и NSF. Веб был разработан в CERN, европейской физической лаборатории, которая получила большой грант NSF. Первый веб-браузер, "Мозаика", был создан на средства NSF в суперкомпьютерном центре университета Иллинойса. Некоторые идеи компьютерных технологий - сети, распределенные компьютерные системы, мультимедиа, обработка естественных языков, базы данных, обработка информации, взаимодействие человек-компьютер - важны для управления информацией; большая часть исследований в этих областях имеет значение для текущих разработок в области ЭБ.

Врезка 4.1

The Coalition for Networked Information (Коалиция сетевой информации, CNI)

Фокусом инновационного процесса среди библиотек и издательств США является CNI. Большинство людей, собирающихся на полугодовые конференции этого общества, представляет университетские библиотеки, компьютерные центры. Но там бывают и люди из издательств, компьютерных фирм, национальных библиотек и правительства США. Последние годы было много участников из других стран.

CNI - это совместная форма активности Ассоциации исследовательских библиотек (Association of Research Libraries) и Educause (ассоциации университетских компьютерных

центров). Была основана в марте 1990 года с целью помочь осознать фундаментальную важность скоростных сетей и производительных компьютеров для улучшения качества образования и повышения интеллектуальной производительности. Среди 200 организаций-членов этой коалиции есть учреждения высшего образования, издательства, провайдеры сетевых услуг, производители компьютеров и программного обеспечения, системные компании, объединения библиотек, а также отдельные публичные библиотеки и библиотеки штатов.

В 1991 году, задолго до появления веб-браузеров, CNI была одной из первых организаций, которая создала профессионально поддерживаемый информационный сервис в Интернете. Следующий список был скопирован с их веб-сайта:

- доступ к правительственной информации через Интернет;
- государственная информация о высшем образовании;
- влияние развития компьютерных сетей на академические учреждения;
- идентификация, авторизация и управление доступом;
- ценообразование и анализ "цепочек стоимости" для сетевой информации (measures in the networked information value chain);
- университетские консорциумы для производства и распространения информации;
- создание и хранение коллекций авторефератов и диссертаций;
- электронные доски объявлений на цифровых супермагистральных;
- 51 причина вложить деньги в национальную информационную инфраструктуру;
- акт правительственного центра печатной продукции (Government Printing Office) по созданию широких информационных онлайн-сетей;
- гуманитарные науки и искусство в информационных магистралях;
- институциональная информационная политика (institution-wide information policies);
- рабочая встреча CNI/OCLC по метаданным;
- национальная инициатива по переводу культурного наследия в сеть;
- поиск и обработка информации в сетях;
- создание новых learning communities с помощью компьютерных сетей;
- технология, обучение и гуманитарные науки;
- права на электронный доступ и распространение информации;
- региональные конференции;
- обучение по сети в Калифорнии;
- обучение и изучение по сети;
- защита интеллектуальной собственности в сетевом мультимедийном пространстве;
- трансформация публичных библиотек;
- планируемое отступление (planning retreat) для библиотек и профессионалов в информационных технологиях;
- ресурсы Z39.50.

Этот список иллюстрирует широту интересов CNI и подчеркивает практическое применение ЭБ, коллекций, связей между библиотеками и издательствами, политики в области интеллектуальной собственности. Но список не отражает человеческий аспект CNI, созданный его директором-основателем Paul Evan Peters и продолжаемый его приемником Clifford Lynch. Встречи CNI, на которых общаются люди из разных областей знаний, отличаются особой атмосферой. Люди, чьи профессиональные интересы порой противоречат друг другу, учатся учитывать интересы друг друга и работать вместе. В последние годы прогресс был таким быстрым, что легко забыть тот вакуум, который заполнил CNI, предоставив людям возможность обсуждать взаимные интересы в области сетевой информатики.

В дополнение к целевому финансированию конкретных проектов, федеральные агентства помогают развитию исследовательского сообщества и зачастую координируют внедрение новых технологий. Так, NSF поддерживал становление IETF. Веб-консорциум (W3C) в Массачусетском технологическом институте (MIT) получал финансирование не только от индустрии, но и от DARPA.

До начала 90-х гг. ЭБ не являлись явным объектом исследований, которые могли поддерживать федеральные органы. В 1992 году DARPA начало финансировать проект

Computer Science Technical Reports (Технические отчеты по компьютерным технологиям), который координировался Corporation for National Research Initiative и включал пять университетов: Карнеги-Мэллон, Корнелл, MIT, Стэнфорд и Беркли (Калифорния). Целью было направить усилия сильных компьютерных факультетов этих университетов на исследования в области ЭБ. Однако инициатива, которая реально утвердила ЭБ как самостоятельную область исследований, возникла в 1994 году, когда NSF, DARPA и NASA выступили с Digital Libraries Initiative (Инициатива "Цифровые Библиотеки", DLI).

Эта инициатива ставит своей целью привлечь внимание международного сообщества к достижениям в области ЭБ. Помимо конкретных работ по проектам, эта программа фактически придает форму нарождающейся научной дисциплине. Исследования ЭБ не новы, но до этого они были фрагментарными. Даже сам термин "цифровые библиотеки" был неопределенным. Проект CNI высветил ЭБ как передовое направление исследований. Он поддерживает проведение конференций, публикации и создание академических вакансий для тех людей, кто выражает явный интерес к исследованиям в этой области. Это "утверждение" новой области было весьма важным, т.к. оно укрепило доверие, уверенность в том, что необходимы долгосрочные исследования.

Кроме того, в рамках CNI была внесена ясность в различия между исследованиями и использованием ЭБ. В то время, когда инициатива появилась, некоторые люди полагали, что федеральное правительство предоставляет новый источник денег для создания ЭБ. Но было четко подтверждено, что это именно исследовательская программа. Некоторые разработки уже вошли в стадию внедрения; в некоторых рассматриваются разработки компьютерной техники и программного обеспечения; некоторые по-прежнему чисто экспериментальные.

Становление ЭБ как области исследований сопровождается той опасностью, что исследователи могут сосредоточиться на разрешении теоретических проблем вычислительной техники, экономики, социологии или юриспруденции, забывая, что это все же практическая область, исследования в которой оцениваются прежде всего с точки зрения полезности. К счастью для области, федеральные агентства осознают эту опасность и пытаются держать в фокусе внимания реальные нужды.

DARPA, NASA и NSF формируют свои бюджеты по решению Конгресса в соответствии с их целями и каждый оперирует в совершенно определенной среде. Ни одно из этих агентств не рассматривает библиотеки в качестве первоочередной цели; деньги для CNI поступают из фондов, которые Конгресс ассигновал для исследований в области компьютерных технологий. Как следствие, первая фаза проектов выдвигает на первый план компьютерные аспекты проблемы. Однако, агентства осознают, что ЭБ - это более, чем компьютерная технология. В 1998 году они перешли ко второй фазе и расширили спектр поддерживаемых им направлений деятельности. Новые партнеры - Национальная медицинская библиотека, Библиотека Конгресса, отделение Undergraduate Education NSF и National Endowment for the Humanities. Данная книга была написана до того, как эти новые гранты (направления) были объявлены, но все ожидают, что агентства в рамках своих областей ответственности будут поддерживать более широкий диапазон проектов.

Врезка 4.2

Инициатива "Цифровые библиотеки" (DLI)

В 1994 году отделения компьютерных технологий NSF, DARPA и NASA предоставили средства для шести четырехлетних исследовательских проектов в области ЭБ. Общая сумма составила 24 миллиона долларов, дополнительные средства поступили и из других источников. Предполагалось, что каждый из этих проектов станет вкладом в создание ЭБ и будет сопровождаться исследованиями в смежных областях. Эти шесть проектов приведены ниже вместе с некоторыми пояснениями относительно их содержания:

- Университет в Беркли (Калифорния) создал большую коллекцию документов об окружающей среде штата Калифорния, включая карты, изображения и правительственные отчеты. Следует отметить исследования в области структурированных (мультивалентных, multivalent) документов (концептуальный метод, который рассматривает документ как совокупность слоев информации); Cheshire II (поисковая система, которая объединила

преимущества формата SGML и информацию из записей формата MARC), распознавание образов (в результате такие объекты, как плотины или животные могут быть распознаны на изображении).

- Университет в Санта-Барбаре (Калифорния) сосредоточился на картах и другой географической информации. Эта коллекция называется "Цифровая библиотека Alexandria". Направления исследований включают метаданные географической информации, пользовательский интерфейс для работы с картами (в т.ч. многослойными), методы сжатия и передачи изображений, новые методы анализа того, как люди пользуются библиотеками.

- Университет Карнеги-Мэллона создал библиотеку видеофрагментов под названием Informedia. Исследования сосредоточены на мультимодальном поиске (в котором подбирается информация из различных источников); распознавании речи; распознавании изображений; беглом просмотре (skimming) видео (составление "рефератов" длинных видеофрагментов).

- Университет в Иллинойсе работает с издательствами по построению федеративной (federated) библиотеки научно-технических журналов. Большая часть усилий связана с обработкой документов в формате SGML. В рамках проекта используются суперкомпьютеры для изучения проблем семантической информации в очень больших коллекциях документов.

- Университет Мичигана создает ЭБ на базе коллекции своей библиотеки. Кроме исследования образовательных аспектов, ученые экспериментируют с экономическими моделями и агентным подходом (agent-based) для обеспечения интероперабельности.

- Стэнфордский университет сосредоточился на компьютерной литературе. Сердце проекта - InfoBus, метод комбинирования сервисов из различных источников в единый набор сервисов ЭБ. Другие темы включают моделирование экономических процессов в ЭБ и демонстрационные версии новых пользовательских интерфейсов.

Для уточнения подходов, агентства поддержали ряд рабочих встреч. Некоторые из них были посвящены конкретной исследовательской теме, например, управляемому доступу к информации. Другие преследовали более широкую цель достижения согласованного взгляда на область и выделения ключевых направлений исследований. Небольшая часть федеральных средств использовалась для координации исследований, что исключительно важно для формирования единой области. Ключевые слова - координация и стандартизация. И хотя эти усилия иногда выливались в стандарты, куда важнее их роль в стимулировании исследований и максимизации их воздействия на практику.

ЭБ определенно заявили себя как отдельную область серьезных исследований, некоторые темы приобретают фундаментальный характер. При этом сформировалось сообщество исследователей. Центральная часть данной главы посвящена краткому обзору главных областей исследований ЭБ.

Модели объектов (Object Models)

Одна из важных исследовательских тем - как понимать, интерпретировать объекты в составе ЭБ. ЭБ хранят и распространяют любую информацию, которая может быть представлена в цифровой форме. В результате появляется ряд разнообразных и изощренных проблем в части представления и управления такой информацией.

То, что пользователь видит как одно слово, может быть представлено в компьютере как комбинация файлов и структурных данных в разных форматах. Связи между этими компонентами и пользовательским восприятием объекта иногда называют "моделью объекта".

С точки зрения пользователя, журнальная статья, хранящаяся на веб-сервере, может быть представлена как связный текст с графикой; однако храниться она может как несколько текстовых файлов, несколько графических файлов и, возможно, некоторых программ. Одно изображение может сохраняться в разных видах: как высококачественный архивный образ, как изображение для повседневного использования среднего разрешения и еще раз как маленькая "иконка", которая дает общее представление об изображении, но утрачивает детали. Этот образ может обозначаться единым библиотечным идентификатором - но для компьютера это группа

определенных файлов. Могут существовать множество версий одного и того же объекта. А ЭБ часто имеют "приватные" версии тех материалов, которые готовятся для публичного доступа. После того, как они будут выставлены, могут потребоваться новые версии для исправления ошибок, или материалы могут быть реорганизованы или перенесены на другие компьютеры, или в результате технического прогресса появятся дополнительные форматы.

Способность пользовательского интерфейса и других компьютерных программ предоставить работу (объект, документ) читателю зависит от того, насколько эти программы смогут "понять" связи между различными компонентами и то, как они формируют единый библиотечный объект. Для этого используются структурные метаданные. Языки разметки являются одним из таких методов структурирования текстов. К примеру, на HTML-странице тэг `<img>` является структурным метаданным, которые указывает на расположение изображения.

Большое число ранних работ, посвященных структурным метаданным, было выполнено в библиотеках оцифрованных изображений, музыкальных произведений, видеоклипов и других объектов, конвертированных из других физических носителей. Другой объект текущих исследований - карты. Кроме работы с преобразуемыми данными, есть исследовательский интерес к данным, поступающим в реальном масштабе времени (например, от удаленных датчиков или сенсоров), к проблеме подвижных "агентов", движущихся в сети и к другим категориям цифровых объектов, которые не имеют традиционной физической "основы". Каждый объект представляет собой проблему: как создать, как хранить, как описать, как найти информацию, которую они содержат, как доставить. Эти вопросы сложны по отдельности и еще сложнее в комплексе. Проблема в том, чтобы придумать такую модель объекта, которая представляла бы библиотечные материалы как комбинацию многих форматов и позволила бы различным независимым ЭБ достичь полной интероперабельности.

Пользовательские интерфейсы и взаимодействие человек-компьютер

Может сложиться впечатление, что понимание того, как человек взаимодействует с информацией в компьютере, настолько сложно, что является скорее предметом искусства, чем систематического изучения. К счастью, было показано, что этот пессимизм безоснователен. Развитие веб-браузеров ускорило творческие исследования таких областях, как визуализация сложных наборов данных; послойное представление информации, содержащейся в документах; автоматическое "реферирование" (skimming) для извлечения реферата или для создания гиперсвязей.

Для пользователя за персональным компьютером ЭБ - лишь часть рабочего окружения. Некоторые исследования в области пользовательских интерфейсов пытаются принимать во внимание все окружение целиком, которое обычно включает также электронную почту, текстовый процессор и прикладные программы по специальности пользователя. Дополнительно, окружение, видимо, будет включать широкий круг источников информации, причем не обязательно в цифровой форме: книги, статьи, видеозаписи, карты и фотографии. Показано, что способность пользователя воспринимать (annotate) цифровые объекты, манипулировать с ними и добавлять их к собственным персональным коллекциям, является весьма плодотворной областью исследований.

Обнаружение информации (Information Discovery)

Нахождение и обработка информации - центральные задачи для библиотек; проблема поиска конкретной информации в больших объемах текста - называемая информационным поиском (information retrieval), давно интересует специалистов в компьютерных технологиях. До появления веб проблема просмотра (browsing) не привлекала внимания, несмотря на свою важность. ЭБ выдвинули проблемы обработки и просмотра на передний план, объединив эти

аспекты в единую проблему "обнаружения информации" - как найти информацию. В этой области было выполнено огромное число исследований, здесь мы упомянем только несколько тем.

Описательные метаданные: каталогизирование и индексирование

Большинство лучших систем обнаружения информации использует каталожные или индексные метаданные, создаваемые экспертами с большим профессиональным опытом в соответствующих областях (каталогизаторы в библиотеках, сотрудники специализированных служб по каталогизированию и индексированию). К несчастью, индексирование вручную является дорогим и медленным. Огромные объемы быстро изменяющихся материалов, которые - как ожидается - появятся в ЭБ, потребуют иных подходов. Некоторые метаданные будут генерироваться автоматически, некоторые обученными профессионалами, некоторые - менее опытными людьми, а некоторые будут запрашиваться непосредственно у создателей цифровых объектов. После этого метаданные могут вводиться в соответствующие программы автоматического индексирования.

Исследования в области автоматического индексирования используют программное обеспечение, которое сканирует цифровые объекты, извлекает индексную информацию и строит поисковые индексы. Такие поисковые программы для веб, как AltaVista, Lycos и Infoseek являются продуктами этих исследований, большинство из которых было выполнено еще до того, как ЭБ стали признанной областью исследований.

Обработка текстов на естественном языке (Natural-Language Processing)

Поиск текста значительно облегчается, если поисковая программа понимает некоторые структурные элементы языка. Соответствующие исследования в компьютерной лингвистике включают автоматический разбор предложений (parsing), обработку морфологии (ассоциирование вариантов одного и того же слова) и компиляцию словариков (lexicons) и тезаурусов. Некоторые исследования идут еще дальше, пытаясь оперировать с сущностным содержанием субъектов поиска при обработке информации.

Нетекстовые материалы

Большинство методов обнаружения информации использует текст, но исследователи постепенно прогрессируют по пути поиска конкретного содержания в иных форматах. Распознавание речи только-только становится полезным для индексирования радиопрограмм и аудио треков видеозаписей. Распознавание образов - автоматическое извлечение уникальных признаков из изображений) - также является областью активных исследований, но пока далекой от практического использования.

Управление и сохранение коллекций

Управление коллекциями - тема исследований, которая только еще начинает привлекать то внимание, которого заслуживает. Традиционные библиотеки годами развивали методы, которые позволяли сравнительно маленькому коллективу людей управлять громадными коллекциями материалов. Ранние ЭБ, наоборот, зачастую были весьма трудозатратны. Увлечение созданием цифровых коллекций было столь велико, что необходимость оптимальной организации и сохранения материалов долгое время игнорировалась. Сейчас эти потребности повсеместно признаны.

Организация

Организация большой коллекции онлайн-материалов является комплексной задачей, причем проблемы одни и те же, независимо от того, имеем мы дело с электронными журналами, большим веб-сайтом, библиотекой программного обеспечения, коллекцией карт или большим информационным сервисом. Две основных из них - как загрузить информацию в различных форматах и как организовать ее для хранения и обработки.

Трудности усугубляются тем, что цифровая информация меняется. На заре книгопечатания корректура вносилась непрерывно, так что разные экземпляры одной и той же книги отличались друг от друга. Онлайн-информация также может непрерывно изменяться. Непрерывное внесение небольших изменений никогда не было легким делом, так что в случае больших коллекций может потребоваться полная реорганизация. Многие темы исследований не только весьма важны для достижения интероперабельности между коллекциями, но и в равной степени важны для организации коллекций. В частности, это такие темы, как идентификаторы, метаданные, идентификация применительно как к управлению отдельными коллекциями, так и к взаимодействию между ними.

Архивирование и сохранение

Долговременное сохранение цифровых материалов недавно превратилось в ключевую тему исследований в управлении коллекциями. В то время, когда книги, оставаясь без присмотра десятилетиями, все еще пригодны для пользования, носители цифровой информации имеют значительно более короткий срок службы - зачастую разочаровывающе короткий. Биты информации периодически должны копироваться на новые носители. Дело осложняется тем, что при этом еще происходит замена формата, в котором хранится информация, на новую версию. Форматы текстовых процессоров и графики, которые широко использовались десять лет назад, теперь уже устарели и неудобны в использовании. Для того, чтобы интерпретировать архивную информацию, будущие пользователи должны будут иметь возможность адекватно распознавать форматы и выводить данные на экран.

Конвертирование

Конвертирование физических материалов в цифровую форму хорошо иллюстрирует масштабный эффект. Какой путь наилучший для конвертирования огромной коллекции в цифровой формат? Какое соотношение между ценой и качеством оптимально? Где гарантия, что сегодняшние усилия будут полезны в долгосрочной перспективе? В рамках небольших исследовательских проектов объемом несколько тысяч объектов конверсия воспринимается как неизбежная, хотя и раздражающая необходимость перед тем, как начнется настоящее исследование. Сотрудники сканируют материалы, проверяют результаты на возможные ошибки, создают метаданные, необходимые для конкретного проекта. Но библиотеки и издательства должны конвертировать миллионы объектов, а те, кто будет такую работу делать, вряд ли будут также хорошо мотивированы, как научные работники в рамках небольших проектов. Больше того, при крупномасштабном конвертировании метаданные должны генерироваться автоматически без точного знания того, как в дальнейшем эта информация будет использоваться, а контроль качества играет первостепенную роль.

Некоторые организации разработали эффективные процессы конверсии больших объемов материалов. (Зачастую часть работы выполняется в странах с дешевой рабочей силой.) Однако это конкретный частный метод данной организации. Существует дублирование инструментария и совсем незначительный обмен опытом. Технология оптического распознавания конвертируемых текстов, в которой используется компьютерная идентификация знаков и слов на бумаге, уже достигла высокого и вполне достаточного для практического применения уровня. Некоторые группы исследователей накопили значительный опыт, но он неудовлетворительно систематизирован и слабо используется другими.

Интероперабельность (Interoperability)

С точки зрения компьютерной технологии, большая часть наиболее трудных проблем ЭБ - это аспекты одной большой проблемы: как достичь интероперабельности. Или иными словами, как заставить широкий спектр различных компьютерных систем работать вместе. Интероперабельность включает как синтаксический уровень (где существует поверхностное однообразие подходов к доступу и навигации, но практически полное доверие к человеческому интеллекту в процессе достижения согласованности), так и более глубокий уровень, на котором различные компьютерные системы разделяют понимание информации.

По всем миру создано много независимых ЭБ. Они придерживаются различной политики и базируются на различных компьютерных системах (как современных и совершенных, так и устаревших, давно переживших срок списания). Для обозначения устаревших систем используют термин "унаследованная система" (legacy system) - но ведь любая система устаревает. Как только дается обязательство построить компьютерную систему или создать новый сервис или продукт, это обязательство становится фактором всех будущих решений. Таким образом, любая компьютерная система является "унаследованной", даже если она еще не полностью введена в строй.

Интероперабельность и стандартизация взаимосвязаны. К несчастью, формальный процесс создания международных стандартов зачастую противоречит тому, что требуется для интероперабельности ЭБ. И не только потому, что процесс официальной стандартизации слишком медлителен для быстроменяющегося мира ЭБ; получающиеся стандарты еще и исключительно сложны. Многие международные стандарты никогда не испытывались в реальной жизни. На практике только те стандарты работают, которые часто используются. Иногда стандарт принимается де-факто за счет того, что ведущая исследовательская группа его использует; использование протокола TCP/IP для построения Интернета является примером. Некоторые стандарты принимаются потому, что лидеры сообщества решают следовать некоторому соглашению; формат MARC для каталожных записей являет собой пример такого рода. Иногда общепринятые стандарты создавались по стандартному процессу; MPEG, форма сжатия видеоданных, хороший пример. Другие де-факто стандарты по преимуществу продукты отдельных корпораций; формат PDF (portable document format) фирмы Adobe - совсем свежий пример. (TCP/IP и MARC были созданы тем же сообществом, которое им пользуется, затем стали официальными стандартами, а потом стали улучшаться с использованием официальной процедуры.)

Ниже перечислены лишь некоторые аспекты проблемы интероперабельности:

Пользовательский интерфейс. Предполагается, что пользователь сможет работать с разными коллекциями разных ЭБ. Интероперабельность имеет целью добиться однообразного представления разных коллекций, хотя, безусловно, нет необходимости устранять все различия. Несмотря на то, что коллекция карт отличается от собрания музыкальных записей, пользователь должен иметь возможность легко работать в обоих, в т.ч. выполнять поиск и быть защищенным от "особенностей" компьютерных систем или менеджмента этих коллекций.

Наименование и идентификация. Необходимо несколько способов для идентификации материалов ЭБ. Каждый компьютер в Интернете имеет IP адрес и доменное имя. Универсальный указатель ресурсов (URL) веб расширяет эти имена до индивидуальных файлов. Однако этих возможностей недостаточно. Библиотечные объекты нуждаются в идентификаторах, которые обозначают именно материал, а не его расположение в данный момент времени. Такие независимые от местоположения (location-independent) идентификаторы иногда называют Универсальными Именами Ресурсов (Uniform Resource Names, URN).

Форматы. Во всех известных ЭБ материалы хранятся в цифровой форме. Веб де-факто создал стандарты для новых форматов, к примеру, HTML для простого текста или GIF и JPEG для графики. Однако на этих форматах согласие разработчиков заканчивается. С точки зрения интероперабельности, особую трудность представляет текст. Формат ASCII, в 80-е годы ставший практически стандартом кодировки для компьютеров, имеет ограниченное число символов кроме английских. В настоящее время развивается кодировка Unicode как расширенный набор символов, способный поддерживать очень широкий набор шрифтов. Но он

все еще не поддерживается многими компьютерными системами. Широко поддерживаемый формат SGML используется в некоторых ЭБ как язык разметки и для текста и для метаданных, является столь гибким и сложным, что достичь на его базе интероперабельности весьма нелегко. XML, упрощенная версия SGML, стала популярна в последнее время - возможно ему удастся занять нишу между простым HTML и исчерпывающим SGML.

Метаданные. Метаданные важны во многих аспектах ЭБ, но особенно для достижения интероперабельности. Как уже отмечалось, метаданные часто делят на три категории: описательные (используются для библиографических целей, поиска и обработки), структурные (которые связывают различные объекты или части объектов между собой) и административные (для управления коллекцией и доступом к ней). Для интероперабельности необходим обмен частью этих метаданных между компьютерами. Это требует отдельного соглашения об именах полей метаданных, форматах кодировки и - по крайней мере - соглашение по семантике. (Очевидный пример важности семантики: если в одной коллекции поле "дата" используется для даты создания, а в другой - для даты включения в коллекцию, то это поле представляет относительную ценность.)

Распределенный поиск. Пользователи часто хотят найти информацию из нескольких независимых коллекций одновременно. Они могут быть организованы логично и единообразно, но может существовать различие в описательных метаданных, используемых для поиска. Традиционный подход к поиску информации в различных коллекциях одновременно состоит в том, чтобы во всех них был стандартный набор метаданных и поддерживался один и тот же протокол. Все большее число специалистов осознает нереалистичность таких требований. Должна быть возможность поиска в различных коллекциях, даже если материал в них организован по-разному.

Сетевые протоколы. Перемещение материалов с одного компьютера на другой требует интероперабельности на сетевом уровне. Практически универсальный набор Интернет-протоколов в значительной степени решает эту проблему, но есть "прорехи". К примеру Интернет-протоколы не слишком подходят для передачи непрерывных потоков данных, как видео- или аудиоматериалы, которые должны прибывать в качестве целостного потока в предсказуемые интервалы времени.

Поисковые протоколы (retrieval protocols). В процессе выполнения одной из фундаментальных операций ЭБ, компьютер посылает запрос другому с целью получить некоторые данные (items). Это послание должно быть передано по определенному протоколу, который может быть как простым (вроде HTTP), так и более сложным. Идеальный протокол должен поддерживать идентификацию обоих компьютеров (с целью обеспечения безопасности), высокоуровневые запросы для определения доступных ресурсов, варианты поиска и возможности обработки, методы хранения и модификации промежуточных результатов, а также интерфейс для различных многочисленных форматов и процедур. Наиболее амбициозная попытка достичь этих целей, Z39.50 близок к коллапсу от собственной сложности - и при этом все еще не отвечает всем критериям.

Идентификация и безопасность. Ряд наиболее сложных проблем интероперабельности между ЭБ включает идентификацию. Необходимы 3 категории. Первая - это идентификация пользователей. Несмотря на то, что широко используется несколько методов, ЭБ зачастую вынуждены присваивать каждому пользователю "пользовательский ID" и пароль. Вторая категория - идентификация компьютеров. Системы, содержащие ценную информацию особенно финансовые транзакции или конфиденциальную информацию, должны знать, с каким компьютером они соединяются. Один подход - полагаться на IP-адреса, но это крайне ненадежно. Третья категория - идентификация библиотечных материалов. Люди должны быть уверены, что они получают аутентичную версию материала, которая не была изменена (случайно или намеренно). Существует несколько хороших методов идентификации, но они все еще недостаточно распространены для достижения полной интероперабельности.

Семантическая интероперабельность. Это широкий термин, который обозначает необходимость для компьютеров иметь одинаковую семантическую интерпретацию передаваемой информации. Семантическая интероперабельность связана со способностью пользователей находить одинаковые классы цифровых объектов, распределенных по разнородным коллекциям, с учетом различий между сайтами. Полная семантическая

интероперабельность охватывает группу фундаментальных научных проблем, часть из которых исключительно сложна.

Веб предоставляет некий базовый уровень интероперабельности, но именно та простота технологии, которая привела к его широкому распространению, принесла и ряд проблем: возможностей URLs недостаточно для длинных терминов. HTML позволяет представлять ограниченное число типов информации. MIME, который позволяет определять тип каждого объекта, достаточно хорош в настоящее время, однако библиотечная информация намного шире, чем возможности MIME. Пользовательские интерфейсы определяются простотой HTML. Развитие веб-технологии становится большим бизнесом. Некоторые расширения обусловлены объективными вещами, а некоторые - конкуренцией между компаниями. Единственным заметным успехом является язык программирования Java, который позволил сделать огромный вклад в пользовательский интерфейс и устранить многие ограничения HTML.

Парадоксальным образом успех веб также является барьером на пути развития ЭБ. Веб стал унаследованной системой. Если разработчики хотя, чтобы их работа была востребована, они обязаны предусмотреть возможность плавного перехода от сегодняшнего веб к их системам. (А необходимость этого иллюстрируется тем фактом, что нынешние наиболее популярные веб-браузеры не поддерживают URN и это является препятствием для использования этого инструмента для идентификации материалов ЭБ.)

Интероперабельность легко описать, но трудно достичь. Исследователи должны предложить новые концепции, которые открывали бы новые возможности и были легки для внедрения. Новые методы должны быть высокофункциональны (для того, чтобы преодолеть инерцию существующих процессов) и достаточно дешевы. Тщательная проработка возможностей расширения нынешних ЭБ позволит прогрессировать далее, не разрушая существующей базы.

Масштаб

Интероперабельность и управление коллекцией усложняются по мере увеличения масштаба библиотеки. Пользователь может обнаружить, что в огромной библиотеке (вроде библиотеки Гарвардского университета) трудно пользоваться каталогом монографии - по сравнению с маленькой библиотекой колледжа, где каждая рубрика включала всего несколько наименований. Многие согласятся, что по мере расширения веб, индексирующие программы-поисковики вроде Infoseek становятся менее полезными - на простейший запрос они зачастую выдают сотни похожих ссылок и нужный результат трудно обнаружить среди дубликатов и мусора. Трудно исследовать проблемы масштабирования без построения собственно крупных ЭБ. Текущие исследования сосредоточены на технических аспектах, прежде всего надежности и производительности.

Надежность и устойчивость

Вопросы надежности и устойчивости сервиса являются первостепенными для ЭБ. Сложность больших компьютерных систем превосходит нашу способность представлять себе всю картину взаимодействия частей. В достаточно больших системах часть компонентов неизбежно будет неисправной в данный момент времени. Один подход к этим проблемам - дублирование данных. Создание сайтов-зеркал (mirroring), которое предусматривает установку копий коллекции на нескольких сайтах, используется весьма часто. К несчастью, сайты-зеркала очень редко действительно - полные копии оригинала, поскольку объективно возникает задержка при перенесении данных с одного места в другое. Что означает в случае выполнения распределенного поиска информации, если часть коллекции сегодня недоступна или если используется архивная версия, которая содержит уже чуть устаревшую информацию?

Исследования производительности проводятся не только в рамках проблемы ЭБ, но в рамках компьютерных сетей. Интернет в настоящее время распространен практически во всех

странах мира, но еще очень далеко от одинаково высокой производительности повсеместно и постоянно. Кэширование - сохранение временных копий недавно запрошенной (соседним компьютером или пользователем) информации - помогает достичь удовлетворительной производительности во всемирной сети, но является достаточно сложным. Что случится, если - к примеру - временные копии уже устарели? Как решать проблемы обеспечения безопасности и контроля доступа, если есть вероятность того, что информация хранится в общедоступных кэшах в разных точках мира?

Некоторые интересные исследования производительности основывались на концепции "локальности" (locality), когда предполагается, что избранная информация копируется и сохраняется в отдельных точках сети, для которых характерна хорошая связь. Например, Networked Computer Science Technical Reference Library (сетевая компьютерная техническая реферативная библиотека) пользуется набором зон, в каждой из которых хранятся все данные, необходимые для поиска и идентификации информации. За пределы зоны сообщения отправляются лишь для того, чтобы получить доступ к реальным цифровым объектам по местам постоянного хранения.

Экономические, социальные и юридические аспекты

ЭБ существуют в сложных экономических, социальных и юридических рамках. Юридические проблемы могут быть как национальными, так и международными и затрагивать несколько отраслей права, включая такие вопросы, как: права собственности, связь, приватность, непристойности, клевета, национальная безопасность и даже проблемы налогообложения. Социальный контекст включает владение, авторство, акт публикации, аутентичность и целостность. Это совсем нелегкие области исследований.

Среди наиболее сложных для изучения - проблемы экономики. Как ЭБ управляют коллекциями информации, так и коллектив профессионалов нуждается в управлении. Кто платит этим людям? Обычный ответ - платят пользователи или их организации за подписку или за каждое обращение. Поэтому исследования касаются методов оплаты, идентификации и методов контроля за использованием коллекции. В тоже время существование высокопрофессиональных веб-сайтов с открытым доступом указывает на существование иной экономической модели. Появились интересные экономические теории, но реальное понимание экономических факторов приходит от людей, которые реально создают, управляют или используют цифровую информацию. Модели ценообразования лишь кажутся академическими - ошибки реально ведут к тому, что человек теряет работу, а организация теряет свой бизнес.

Библиотеки и издательства зачастую хотят контролировать доступ к своим материалам как с целью проверки оплаты, так и выполнения требований законов об интеллектуальной собственности или условий, на которых были предоставлены материалы. Другие причины - ответ на обеспокоенность распространением клеветы, порнографии или нарушениями частной жизни. Эти методы иногда называют "управление правами" (rights management), однако более точным представляется термин "управление доступом", поскольку проблема явно шире простого контроля за интеллектуальной собственностью или поступлением оплаты. Некоторые методы управления доступом включают шифрование, которое является исключительно сложной областью, в которой намертво сплетены технологические, юридические и политические аспекты.

Влияние ЭБ достаточно сложно определить. Может быть измерена ценность ЭБ или исследований в этой области? К сожалению, несмотря на некоторые попытки, до сих пор неясно, сколько полезной информации может быть получено. Систематических результатов недостаточно и они разрозненные. Проблема эта хорошо известна исследователям рынков. Такие техники, как фокусные группы (focus groups) и обследования (surveys) достаточно эффективны в предсказании того, как небольшие изменения влияют на существующие продукты, но значительно менее эффективны в рассмотрении эффектов от фундаментальных изменений. Это не означает, что подобные измерения не должны проводиться. Невозможно развивать большую систему без надежных управленческих данных. Сколько людей пользуется

каждым сервисом? Насколько они удовлетворены? Какова удельная стоимость сервиса? Сколько стоит пополнение коллекций? В чем задержки? Наличие таких данных о системе весьма важно, но это обычная практика компьютерных исследований, а не исследований ЭБ.

Экономические, социальные и юридические аспекты обсуждаются в конце этой главы не вследствие их неважности, - но именно потому, что они так трудны. Два критерия важны при выборе темы исследования: тема должна быть осмысленной, а проект - исполнимым. Ценность библиотек в образовании, школьном обучении и для общества практически невозможно адекватно измерить количественно. Попытки замерить воздействие ЭБ на раннем этапе их развития серьезно осложнены неполнотой коллекций, быстрым изменением технологии и степенью адаптированности пользователей к новым возможностям. Измерения радикально новых компьютерных систем полезно для истории, интересно как факт - но малополезно с точки зрения планирования будущего. Определение ценности ЭБ в некотором смысле остается делом информированного здравомыслия, а не исследований.

Мировые исследования

ЭБ - это всемирный феномен. Интернет позволяет исследователям из разных стран сотрудничать в повседневном режиме. Исследователи из Австралии и Новой Зеландии уже оценили эти преимущества и вносят существенный вклад. Веб был создан в Швейцарии. Британский проект e-lib придал дополнительный импульс новым инициативам в области электронных публикаций и распространении цифровых материалов. Недавно Европейский Союз и американский Национальный научный фонд (NSF) начали совместное финансирование серии совместных встреч по планированию. В качестве существенного международного вклада можно отметить серию рабочих встреч в Дублине по базисным метаданным (Core metadata) [ныне на русском языке уже принято наименование "Дублинское ядро" - *прим.пер.*], которые описаны в гл.10. Результаты исследований в области ЭБ, которые ежемесячно публикуются в электронном журнале D-lib Magazine, иллюстрируют тот факт, что это в самом деле международное сотрудничество. За последние три года авторами публикации были ученые из более чем десяти стран. Большие, хорошо финансируемые американские проекты, безусловно, важны - но они еще далеко не вся история ЭБ.

Глава 5. Люди - организация и изменения (Изменения в управлении персоналом)

История ЭБ - это история непрерывных изменений. Авторы, читатели, библиотекари, издатели, сотрудники информационных служб усваивают новую технологию со значительной скоростью - и это меняет отношения между людьми.

В каждой организации есть люди, которые хотят пользоваться самым современным оборудованием (даже если оно не совсем подходит). Есть и другие, которые хотят, чтобы все происходило по установленным правилам, традиционно (даже если новые технологии просто замечательные). Иногда это называют разрывом между поколениями ("generation gap"), но это неточно, т.к. воспринимать новые идеи могут люди любых возрастов.

Оставляя в стороне предпочтения отдельных людей, необходимо сказать, что темпы изменений существенно различаются в различных организациях и отраслях. Некоторые корпоративные библиотеки (фармацевтических компаний, к примеру) уже тратят более половины соответствующего бюджета на электронные материалы и сервисы. А вот в гуманитарных библиотеках в центре внимания по-прежнему печатные материалы, манускрипты и другие физические объекты - и такая ситуация сохранится в обозримом будущем.

Наиболее фундаментальное изменение состоит в том, что онлайн-компьютерные технологии изменяют поведение создателей и пользователей информации. Возможности персонального компьютера позволяют человеку с умеренными навыками выполнять процессы, которые раньше требовали тренированных профессионалов. Текстовые процессоры и оборудование для настольных типографий позволили создавать офисную корреспонденцию, которая выглядит так же элегантно, как и настоящие книги. Диаграммы, графики и другие иллюстрации могут создаваться в цвете. Не каждый захочет изучать все эти возможности, поэтому все еще встречаются откровенно невзрачные материалы. Однако многие люди создают элегантные и эффективные материалы без помощи профессионалов - и эта их возможность существенным образом влияет на всех профессионалов информатики, включая издателей, библиотекарей, архивистов, каталогизаторов и индексаторов, вебмастеров.

Некоторые возражают, что новые технологии устраняют необходимость в профессиональном менеджменте информации. Это наивно. Библиотекари и издатели выполняют функции, которые далеко выходят за рамки простого управления физическими объектами. Такие сервисы, как реферирование, редактирование, составление рефератов, индексирование, не являются методами какой-то одной технологии. Несмотря на то, что веб позволяет пользователям размещать свою собственную информацию, большинство людей охотно принимает помощь профессиональных веб-мастеров. Общая потребность в специалистах в информационных технологиях будет сохраняться и, возможно, расти; при этом специфика работы будет меняться по мере развития технологии. Возможное появление новых форм организации такой работы и новых типов профессионалов - тема, открытая для обсуждения.

Цифровые библиотеки, созданные пользователями

Некоторые из наиболее успешных ЭБ были созданы исследователями или группами профессионалов для себя и своих коллег при минимальной поддержке издателей или библиотечных работников. Две из них, физический архив E-print в Национальной лаборатории в Лос-Аламосе и серия стандартов Интернета RFC описаны в главе 2. Врезки данной главы дают еще три примера: библиотека математического обеспечения Netlib; архив Международного консорциума исследований в области политологии (International Consortium for Political Science Research); коллекция классических текстов Perseus. Эти ЭБ хорошо

организованы и активно используются. Они приглашают на работу профессионалов, но лидерство и большинство проектов осуществляет персонал, пришедший из других областей: физики, специалисты по компьютерам, прикладные математики, социологи и филологи.

ЭБ, созданные группами пользователей особенно интересны тем, что их сервисы были созданы так, чтобы удовлетворять конкретные потребности в данной области, дисциплине, - не задумываясь о том, как это принято делать. Когда создатели и пользователи разрабатывают свою собственную систему для работы, они рассматривают обычные в таких случаях вопросы об организации и обработке информации, контроле качества, стандартах, сервисах (все они являются традиционными для библиотек и издательств). Иногда они находят новые и весьма творческие подходы к этим старым вопросам.

Библиотеки и музеи играют особую роль для пользователей-гуманитариев, поскольку они предоставляют ту базу, на которой строится вся работа. ЭБ могут обеспечить более широкий доступ к этим материалам, чем это было в любой физической коллекции. Университеты с прекрасными библиотеками всегда будут иметь преимущество с точки зрения обслуживания процесса подготовки гуманитариев и проведения исследований, но при этом они не должны эксклюзивно пользоваться этими материалами. Британская библиотека начинает оцифровку и публикацию в веб таких своих сокровищ, как Magna Carta или манускрипт Beowulf. В прошлом доступ к таким документам был для студентов и школьников ограничен и они вынуждены были покупать дорогостоящие факсимильные издания. В будущем любой будет иметь возможность видеть прекрасную репродукцию.

Врезка 5.1

Netlib

Netlib - это цифровая библиотека высококачественного математического программного обеспечения. Основанная в 1985 году Jack Dongarra и Eric Grosse (последний остается ее главным редактором), в настоящее время она поддерживается консорциумом в составе AT&T Bell Laboratories, университета Теннесси и Окриджской национальной лаборатории. Существует большое число сайтов-зеркал. В год с них сгружаются сотни тысяч программ, которые используются практически во всех областях научных исследований.

Сначала Netlib специализировался на программном обеспечении, созданном для численного анализа (numerical analysis), особенно на программах для суперкомпьютеров с векторной или параллельной архитектурой. Однако сейчас коллекция включает и другие программы, технические отчеты, статьи, данные тестирования производительности, информацию о конференциях и встречах. Хотя большинство материалов можно взять свободно, в некоторых случаях (при коммерческом использовании) взимается плата.

Техническая история Netlib приходится на период экстенсивного роста Интернета. Он начинался как сервис, базировавшийся на электронной почте. В разное время использовался интерфейс X-Window, анонимный FTP, компакт-диски или сервис Gopher. В настоящее время это веб. Команда Netlib продолжает оставаться среди лидеров в развитии передовых архитектур для организации и хранения материалов в ЭБ.

Способ организации коллекций Netlib весьма прагматичен. Предполагается, что пользователи - это математики и ученые, которые знакомы с областью и будут использовать программное обеспечение в составе собственных разработок. Коллекция организована иерархически, программы группируются по дисциплинам, области применения и источнику. Каждую коллекцию курирует свой редактор, который использует свои знания в данной области для выбора подходящего метода организации. Netlib развивался в форме индексированных записей в специальном формате, отвечающим конкретным задачам данной библиотеки. Коллекция также классифицирована согласно Справочнику доступного математического программного обеспечения (Guide to Available Mathematical Software), который представляет собой кросс-индекс, созданный Национальным институтом стандартов и технологии (NIST).

Врезка 5.2

Межуниверситетский консорциум политических и социальных исследований

Консорциум (Inter-university Consortium for Political Science Research, ICPSR) функционирует с 1962 года, является некоммерческой организацией и базируется в университете Мичигана. Его ЭБ позволяет социологам хранить собранные данные и предоставлять их другим. Сбор данных - это дорогостоящее занятие, но они зачастую теряли актуальность в процессе их организации, документирования, обработки. Коллекция данных охватывает различные области, включая политологию, социологию, демографию, экономику, историю, образование, геронтологию, криминологию, здравоохранение, международную политику и законодательство.

ICPSR является некоммерческим консорциумом. Его членами являются сотни ученых и университетов как из США, так и по всему миру. Их годовая подписка позволяет им обеспечить доступ к коллекциям для своих студентов и сотрудников. Посторонние пользователи могут оплачивать индивидуальные подборки данных.

В год архив пополняется около 200 комплектами данных (data set) (несколько тысяч файлов). Некоторые комплекты весьма объемны. В архиве также хранится документация по этим комплектам и пояснительные записки по проектированию и выполнению исследований, документы ученых, выполнявших эти исследования, информация о том, как именно эти данные были собраны; уточнения в процессе исследования; иная техническая информация, помогающая использовать данные в будущих исследованиях. Коллекция организована иерархически по дисциплинам, так что просмотр весьма прост. Каждый комплект данных описан в короткой записи, содержащей общую каталожную информацию и резюме.

Архив существовал на самых разных компьютерных системах. Текущий интерфейс ICPSR - веб, он может быть использован для просмотра и поиска в каталоге. Комплекты данных доставляются заказчику по FTP, некоторые комплекты доступны на компакт-дисках.

Врезка 5.3

Perseus

Некоторые из больших проектов в области электронной информации возглавлялись признанными членами сообщества ученых-гуманитариев, а некоторые - импровизациями с минимальной официальной поддержкой. Иногда молодым сотрудникам даже приходится преодолевать сопротивление более старших работников факультета. В середине 80-х годов в Гарвардском университете Gregory Crane, молодой сотрудник, специалист в области классической литературы, инициировал проект Perseus, который предусматривает гиперссылки для взаимодействия между собой таких ресурсов как тексты и карты - с инструментарием (например, словарями). Сам Crane хотел предоставить студентам возможность насладиться поэмами греческого поэта Пиндара. И из этой затеи выросла одна из наиболее важных ЭБ по гуманитарным дисциплинам.

В настоящее время коллекции Perseus исчерпывающим образом охватывают период классической Древней Греции и стремительно расширяются на другие периоды греческой, а также римской истории и далее. Материалы включают тексты (на языке оригинала и переводы), графические файлы объектов (к примеру, фотографии греческих ваз) и архитектурных сооружений. Однако наиболее важной особенностью является объем усилий, затраченный на структурирование материалов и база данных связей между объектами.

Особенности содержания коллекций Perseus позволяют легко мигрировать с одной компьютерной системы на другую. Классические тексты весьма стабильны, хотя в новых изданиях могут быть незначительные изменения, а сопроводительные работы (словники, атласы и пр.) живут также долго. Таким образом, усилия по определению аутентичной версии текста, его разметки в формате SGML и связывание с близкими работами представляет собой долговременную "инвестицию", которая будет жизнеспособна на любой компьютерной системе. В штате Perseus никогда не было более одного программиста, он опирается на подходящие общедоступные компьютерные технологии. Вначале использовалась система Hypercard фирмы Apple, с помощью которой создавались высококачественные компакт-диски. По мере развития технологии Perseus легко мигрировал в веб. Единственная оригинальная разработка в рамках проекта - это создание системы, базирующейся на наборе правил, для анализа морфологии греческих и латинских слов.

Долговременное воздействие проекта Perseus предсказать трудно, однако цели весьма амбициозны. В последние годы исследования в гуманитарных областях становятся все более изощренными и специальными. Характерно, что Crane не смог продолжать свою работу в Гарварде, поскольку его проект не рассматривался как серьезная научная работа; он перебрался в университет Tufts. Perseus может не быть подходящим для Гарварда, но легкомысленной затеей его определенно не назовешь. Четыре миллиона греческих слов включают большинство обычно цитируемых текстов. Когда нет подходящей фотографии вазы, Perseus способен найти сотни новых фотографий; пользовательский интерфейс помогает читателю легко получить доступ к переводам и словарям, сохраняя при этом фокус на оригинальном материале. Perseus во все возрастающем объеме используется исследователями как прекрасный источник. Самым большим его достижением, возможно, является демонстрация, с одной стороны, широкой публике очарования классики, а с другой стороны, - ученым-гуманитариям того факта, что популярность и серьезные исследования могут идти рука об руку.

Мотивы создателей и пользователей

Создатели (creators)

Для того, чтобы понять, как библиотеки и издательства меняют требования, необходимо рассмотреть различные мотивы, которые заставляют некоторых людей создавать материалы, а других людей - пользоваться ими.

Одно из распространенных заблуждений состоит в том, что люди создают произведения прежде всего за деньги или из-за доходов, которые они приносят. Хотя многие люди получают средства для существования с помощью создаваемых работ, другие преследуют иные цели. "Происхождение видов" Ч. Дарвина было написано для пропаганды определенной идеи; то же относится к работе Pain "Права человека", "Капиталу" К. Маркса и посланиям Св. Павла к Римлянам. С времен классики некоторые книги, манускрипты, картины, музыкальные произведения и поэмы создавались по заказу правителей. Многие из всемирно-известных зданий, включая Египетские пирамиды и Библиотеку Франции, были созданы потому, что кто-то хотел увековечить свое имя. Фотографии, дневники, поэмы и письма могут создаваться просто для пользы автора, хотя впоследствии они становятся ценными библиотечными объектами. И мало что превзойдет религию, так или иначе стимулировавшую создание такого количества прекрасных материалов - художественных, музыкальных, архитектурных, изобразительных.

Творческий акт может быть лишь моментом в процессе иной деятельности. Судья, формулирующий приговор, создает материал для ЭБ. То же относится к музейному работнику, составляющему каталог выставки или к фармацевту-исследователю, откликающемуся на нужды больных людей. Правительственные агентства создают материалы для публичного использования: навигационные карты, прогнозы погоды, официальную статистику, договора, торговые соглашения и пр. Много материалов сейчас хранится в библиотеках или архивах, созданных для сохранения информации, сопутствовавшей определенным событиям или решениям: криминальная хроника, правительственные документы, церковная хроника, официальная история, фотографии военных действий и пр.

Людей, переводящих оригинальные материалы в цифровую форму, тоже можно назвать создателями. Такая работа включает как конверсию любимого рассказа с размещением на собственной домашней страничке, так и конвертирование миллионов объектов. Реальный акт созидания может выполняться даже машиной, когда она фиксирует спутниковую фотографию.

Другое заблуждение заключается в том, что создатели и организации, в которых они работают, преследуют одни и те же цели. Некоторые работы создаются группой людей, некоторые - индивидуально. Игровой фильм должен быть результатом групповых усилий, но никто не ожидает, что поэма будет создана в рамках комитета. Таким образом, часть создателей (свободные писатели, фотографы, композиторы) работает индивидуально, а другие работают в

организациях и создаваемые ими материалы являются частью ее работы. Когда кого-то принимают на работу, работодатель часто оговаривает творческий характер и права собственности. В этом случае "творчества по обязанности" мотивация индивидуумов может отличаться от целей организации. Киностудия, создающая боевик, может быть мотивирована будущей прибылью, но режиссер может рассматривать это как возможность высказать политические взгляды, а актриса-исполнительница главной роли преследует профессиональные цели.

Создатели, слабо мотивированные финансовой стороной, обычно получают удовлетворение от максимально широкого распространения их работы. Это создает напряженность в отношениях с издателями, чья модель бизнеса обычно предусматривает возможность ознакомления только за деньги. Академические журналы представляют собой другую важную категорию материалов, в которых интересы авторов (либо сообщающих о своих результатах, либо улучшающих свою научную репутацию - в любом случае это требует максимального распространения) могут быть в прямом конфликте с финансовыми интересами издателей.

Пользователи

Пользователи библиотеки также различаются между собой в своих интересах и своем опыте, как и создатели материалов. В особенности разнообразны группы пользователей городских публичных библиотек. Библиотека служит источником чтения для отдыха, информации о вакансиях в поисках работы, расписания пригородных автобусов. Библиотека может предоставлять выход в Интернет, который пользователи используют для доступа к медицинской или юридической информации. В фондах могут быть аудиозаписи детских сказок, а также подходящие материалы для краеведов, случайных посетителей или экспертов.

Оставляя в стороне различия между людьми, индивидуум может нуждаться в разных материалах в разное время. Даже два пользователя с одинаковыми потребностями могут работать в библиотеке по-разному. Один может широко пользоваться каталогами и индексами, другой больше полагается на цитаты и указатели (links). Создатели ЭБ должны удержаться от искушения принять для себя определенный способ использования и, соответственно, создать систему исключительно под него.

Возможно, тем не менее, выделить несколько больших категорий пользователей и эти различия значимы как для цифровых, так и для обычных библиотек. Первая категория - это люди, которые пользуются библиотекой для отдыха; в ЭБ это иногда принимает форму бессистемного просмотра, обычно называемого *surfing* (скольжение). Другая - пользователи, которым нужна вводная, ознакомительная информация о конкретных объектах, например, инженер начинает изучение конкретной области с изучения обзорных работ, а турист ищет информацию о стране, в которую предполагает поехать. Иногда читатель хочет узнать простой факт: текст первой поправки к Конституции США; температура плавления свинца; кто вчера выиграл футбольный матч. Некоторые факты такого рода содержатся в справочных материалах - таких, как карты, энциклопедии, словари. В иных случаях они "спрятаны" в коллекции глубже. Некоторым пользователям требуется исчерпывающая информация по определенной теме - исследователь в области медицины может захотеть найти все ссылки в опубликованных работах на свойства определенного лекарственного средства; адвокат хочет найти все прецеденты, которые могут иметь отношение к данному случаю.

В большинстве таких ситуаций запросы пользователей могут быть удовлетворены различными способами. Например, разные карты и атласы содержат информацию по географии. Однако полная информация по теме требует доступа к специфическим источникам. Эти различия существенны, если рассматривать экономические аспекты информации (поскольку альтернативные источники информации ведут к ценовой конкуренции) или изучать обработку информации (когда исчерпывающий поиск приобретает особую важность).

Профессии, связанные с информацией и происходящие перемены

По мере увеличения объема цифровой информации и постепенной замены традиционных ее форм, изменяется суть информационных профессий. В частности, библиотекари и издатели имеют разные традиции, поэтому неудивительна их различная реакция на изменения.

Библиотечные работники

Чтобы подробно рассмотреть изменения в деятельности библиотекарей, есть смысл отдельно проанализировать разные аспекты и группы: руководители библиотек, опытные сотрудники, воспитание молодых сотрудников, возрастающая роль сотрудников других специальностей (прежде всего, специалистов по компьютерам).

Директора библиотек в новой ситуации находятся под давлением. Руководство большой библиотекой раньше было связано с приятной, интересной работой, престижем и большой зарплатой. Престиж и зарплата остаются - а вот содержание работы резко меняется. ЭБ имеют долговременный потенциал, но создают кратковременную головную боль. Библиотеки обычно "зажаты" растущими ценами. Консервативные пользователи требуют, чтобы все привычные сервисы сохранялись; другие хотят получить цифровые услуги немедленно. Многие директора не получают от вышестоящих руководителей (причем их понимание происходящих в библиотеке изменений весьма слабое) той поддержки, в которой нуждаются. Каждый год часть очень хороших директоров принимает решение оставить свой пост, связанный с таким количеством административных проблем, и поискать работу, в которой они смогут более четко контролировать судьбу.

Опытные библиотечные работники полагают, что ЭБ являет собой и возможности и проблемы одновременно. Серьезная нехватка старших библиотекарей, знакомых с современными технологиями открывает новые возможности для энергичных и творческих личностей. И наоборот, опытные специалисты, которые не успевают за технологией, рискуют отстать. Освоение новых технологий - это больше, чем вопрос простого переобучения. Обучение важно, но неправильно думать, что при этом происходит только замена одного набора статических навыков на другой. Библиотеки нуждаются в людях, которые способны ответить на вызовы, пытливых и открытых для новых идей.

Подготовка молодых сотрудников библиотек происходит в соответствующих учебных заведениях. Библиотекарь - это профессия. В США степень мастера требуется для занятия многих должностей в библиотеке. Многие годы программы библиотечных колледжей были весьма консервативными, основное внимание уделялось базовым навыкам, необходимым библиотекарям среднего уровня и во многих университетах подготовка по этой специальности была наиболее слабой среди других специальностей. В последние годы ряд университетов осознал, что появление ЭБ открывает новые возможности для библиотечных колледжей, с новыми учебными программами и планами научных исследований.

ЭБ нуждаются в профессионалах других специальностей, прежде всего в области компьютерных технологий и юриспруденции. "Встроить" таких специалистов четко структурированный мир библиотек весьма проблематично. К тому же руководители зачастую избегают принимать на работу молодых талантливых работников без соответствующей степени, без необходимости ограничивая свой выбор. Очень немногие из молодых людей, принимающих участие в создании ЭБ видят библиотечный колледж как часть своей карьеры.

Врезка 5.4

Летняя школа Ticer

В 1996 году университет Тилбурга (Нидерланды) организовал летние двухнедельные курсы для повышения квалификации опытных библиотечных работников в области ЭБ. Эта программа, названная Ticer, имела большой успех и вакансии в школу неизменно заполнены полностью. Большинство слушателей из стран северной Европы, но в 1998 году были участники из Малайзии, Японии, Индии, Южной Кореи и других стран. Почти все слушатели - высокопоставленные сотрудники академических или специализированных библиотек.

Университет Тилбурга долгие годы занимает лидирующие позиции в развитии и применении ЭБ, что находит отражение в программе школы, посвященной большей частью сочетанию проблем стратегического планирования и практического внедрения и использования. Большинство лекторов летней школы являются сотрудниками университетских библиотек или их компьютерных подразделений. Ряд специально приглашаемых экспертов придает программе обучения широту и глубину. Главным ресурсом для слушателей является толстая книга с текстами всех лекций, которую получает каждый участник.

Есть и другие особенности, предопределившие успех программы Tiser. Несмотря на то, что условия проживания и обучения весьма хорошие, цена курса вполне приемлемая, а насыщенная социальная программа позволяет в полной мере оценить преимущества совместного проживания и совместной работы 50 участников со всего мира. Программа тесно сотрудничает с датским издательством Elsevier. Сотрудники издательства участвуют в школе и как слушатели и как лекторы.

Программа демонстрирует тот факт, что область ЭБ практически полностью основывается на использовании английского языка. Все материалы школы англоязычные, несмотря на то, что участники могут приезжать и не из англоязычных стран.

Врезка 5.5

Школа информационных систем и менеджмента в Беркли

В начале 90-х годов ряд университетов задумались о проблеме повышения качества подготовки библиотечных работников. В то время, как их бизнес- и юридические школы испытывали наплыв студентов, приносили ощутимый доход и выполняли высококачественные исследования, библиотечные колледжи владели скромное существование. Преподаватели имели маленькие зарплаты и, соответственно, имели низкую продуктивность. Хуже всего, что программы обучения не были рассчитаны на подготовку специалистов, способных адекватно работать по созданию и внедрению ЭБ и электронных издательств.

Колумбийский университет просто закрыл свой библиотечный колледж, т.к. его прибыльность не окупала платы на здания колледжа. Университет Чикаго также закрыл свой колледж. Университет Беркли (Калифорния), как и университет Мичигана пошли другим путем, решив полностью реорганизовать свои колледжи. Очевидно, наиболее драматические изменения произошли в Беркли.

Многие университеты утверждают, что принципиальные исследовательские решения принимаются на факультетах и департаментах. Зачастую это попытка выдать желаемое за действительное, но в Беркли академическое сообщество действительно обладало реальной властью. Первой идеей было последовать за примером Колумбийского университета и просто закрыть колледж, но оппозиция этому плану была достаточно велика, чтобы руководство университета отступило и создало специальную группу по планированию. Отчет этой группы, озаглавленный Проект Школы информационных систем и менеджмента (Proposal for a School of Information Management and Systems) был представлен в декабре 1993 года. В лучших бюрократических традициях, его основные рекомендации звучали как "упразднить" существующий колледж и заново "основать" новую школу на его руинах.

Создавая программы обучения, лучшие университеты ориентируются на два фактора. Первый - это научное содержание: имеет ли эта область достаточное интеллектуальное наполнение, чтобы привлечь на факультет преподавателей высшего уровня, чье имя способно увеличить авторитет и престиж университета? Второй - это образование: привлечет ли программа таких студентов, которые впоследствии смогут взять на себя бремя лидерства? Группа планирования в Беркли утвердительно ответила на оба эти вопроса.

В качестве исходной точки, отчет совершенно ясно отверг традиционную схему подготовки специалистов на степень мастера библиотечного дела. С замечательной прямоотой отчет утверждал: "Степень мастера присуждаемая по результатам обучения по данной программе, ... определенно не будет отвечать требованиям Американской библиотечной ассоциации, скорее, программа послужит моделью для формулирования критериев создания (описания) дисциплины, которая будет преподаваться в Школе".

Отчет был принят и школа начала набирать преподавателей и студентов. С самого начала особое внимание уделялось междисциплинарным исследованиям. Отчеты об исследованиях полны отсылок на соответствующие совместные проекты с другими колледжами и департаментами. В программе подготовке по новой специальности отмечалась связь с "компьютерными технологиями, общей и когнитивной психологией, социологией, экономикой, бизнесом, юриспруденцией, коммуникациями". От студентов изначально требовались существенные компьютерные навыки.

Состав преподавателей также был весьма разнообразным. Первый декан, Hal Varian, был экономистом. Другие преподаватели включали специалистов в области интеллектуального права, бывшего декана факультета информатики. Многие одновременно преподавали и на других факультетах, выполняли исследования, порождая нетрадиционные связи между департаментами.

Пока рано оценивать успешность этого предприятия (как и схожего в университете Мичигана) - но первые результаты обнадеживают. Колледж отвечает высоким требованиям преподавателей и студентов. Исследовательская программа быстро расширяется, получая значительные гранты извне. Через несколько лет первые выпускники составят основу нового поколения профессионалов информатики.

В сравнении с другими профессионалами особенно заметно, как плохо оплачиваются работники среднего звена. Высококласная работа требует высококлассных сотрудников - и в области ЭБ лучшие работники имеют максимальную рыночную стоимость. Неправильно платить программисту больше, чем руководителю отдела, но еще более неправильно видеть, как хорошая библиотека деградирует из-за недостаточной квалификации своих компьютерных специалистов. Успех проекта Мерсигу в университет Карнеги-Меллона в определенной степени предопределялся тем, что технические сотрудники формально являлись работниками университетского вычислительного центра. Работали они в библиотеке над созданием ЭБ, но подчинялись своему руководителю в центре, несли учебную нагрузку, пользовались первоклассным оборудованием и получали такую же зарплату, как и другие профессионалы в области компьютерных технологий. Лишь немногие библиотеки демонстрируют подобную гибкость.

Издательства

Изменения в издательском деле столь же драматичны, как и в библиотечном. После появления книгопечатания в Европе в 15 веке, издательское дело эволюционировало медленно. Традиционной задачей издателя было производство и распространение печатной продукции. Сегодня издательская индустрия по-прежнему получает основной доход от продаж физических объектов - книг, журналов, видео- и аудиодисков. Однако многие издатели видят свое будущее связанным с электронными публикациями.

Издательский бизнес включает как крошечные, так и гигантские предприятия. Большие издательские группы, такие как Time Warner, Reed Elsevier, Thomson входят в число наиболее крупных и прибыльных корпораций в мире, в то время как большинство из 50000 издательств в США в год выпускает не более 10 наименований. Научные издательства имеют ту странную особенность, что часть из них - откровенно коммерческие организации, и их цель - прибыль акционеров. А другие - это некоммерческие общества и университетские издательства, чья основная цель - образовательная.

Успех издательского дома зависит от издателей, которые отбирают материалы для публикации, работают с авторами и отчасти наблюдают за процессом производства. Издательство может рассматриваться как бизнес, но многие работающие в нем люди не стремятся делать карьеру. Они работают потому, что им интересно содержание того, что они публикуют.

Издательства интенсивно используют преимущества специализации. Некоторые, такие как West Publishing (большое юридическое издательство), имеют свою типографию, но в большинстве случаев тираж печатают специализированные предприятия. Кропотливое

редактирование манускриптов зачастую выполняются по одноразовым контрактам. Продаются книги через специализированных распространителей, а журналы - через подписные агентства. Это придает издательствам необходимую гибкость. Когда Elsevier начал готовить издание журналов в формате SGML, оно отбирало те типографии, которые уже готовы работать с SGML.

Недавняя волна слияний в издательском бизнесе породила несколько гигантских компаний, способных развивать крупномасштабные компьютерные проекты. Однако все еще не очевидно, что только крупные издательства способны прибыльно работать в электронном книгопечатании. Развитие веб-технологий означает, что маленькие фирмы имеют реальные шансы успешно работать в этой области, а зачастую и иметь преимущества за счет более тесного сотрудничества между редакторами и авторами. Некоторые наблюдатели полагают, что доходы West Publishing, находящегося в частном владении и весьма осторожно подходящего к электронным технологиям, могут привести к утрате лидирующего положения на рынке.

Некоторые университетские издательства включая Oxford University Press и University of Chicago Press, имеют больше сходства с коммерческими предприятиями. Они издают большое число книг разных направлений, исходя из финансовых соображений и не предоставляя каких-либо преимуществ авторам из этих университетов. Другие университетские издательства имеют альтернативную цель: поддержка образовательного процесса, причем тематика некоторых так узка, что тираж составляет сотни экземпляров. Такие книги никогда бы не привлекли внимания коммерческих издательств. Бюджеты таких издательств просто крошечные, а в последние годы, когда многие университеты вынуждены экономить, поддержка издательств либо существенно снизилась, либо исчезла совсем.

Компьютерные профессионалы и веб-мастера

Компьютерные профессионалы - такая же часть ЭБ, как библиотекари или издатели, но их сотрудничество существенный затрудняет барьер культурных различий. Многие члены Интернет-сообщества выросли "погруженными" в компьютерные технологии и придерживаются неортодоксального стиля работы. Как и в любой другой дисциплине, некоторые настолько глубоко увлечены техническими подробностями, что не видят ничего другого или просто неспособны объяснить технические подробности неспециалистам. Некоторые вообще намеренно сложны для понимания, хотя это встречается достаточно редко. Многие люди хотели бы, чтобы другие знали, что они делают, но работники технических специальностей зачастую с трудом могут объяснить технологические проблемы непрофессионалам.

Интернет-сообщество было основано специалистами, участвовавших в создании ОС UNIX. Значительная часть успеха этой операционной системы обусловлена традицией открытости. UNIX-, Интернет- и веб-сообщества координируют свою работу, поскольку поняли, что такое сотрудничество полезно для всех. Понимание этой традиции открытости является фундаментально важным для уяснения подхода компьютерных специалистов к проблеме ЭБ. Их отношение может представляться утопической мечтой, а некоторые энтузиасты говорят о полностью неконтролируемом информационном обществе, хотя это крайняя точка зрения. Большинство членов Интернет-сообщества имеет здоровое отношение к деньгам; предпринимательство сохраняет свою привлекательность и новые компании нарождаются непрерывно.

Даже для быстро изменяющегося мира компьютерных технологий развитие специальности веб-мастеров является исключительным. В начале 1994 года веб был едва известен, а к лету 1995 года более 1000 участников приехали на первую встречу Federal Webmasters, неформальной группы людей, кто поддерживает веб-сайты правительственных учреждений США. Словечко "веб-мастер" родилось за одну ночь и утвердилось настолько быстро и прочно, что последующие попытки (имея в виду то обстоятельство, что слово "мастер" - мужского рода) придумать что-то иное провалились. Поэтому термин "веб-мастер" относится и к мужчинам, и к женщинам.

Веб-мастер одновременно является издателем, библиотекарем, системным администратором и дизайнером. Эта профессия требует широких знаний в разных областях. В дополнение к традиционным навыкам издателя (отбор, редактирование), требуются навыки в создании пользовательских интерфейсов и работе высокопроизводительных компьютерных систем. Веб-сайт - это лицо организации перед внешним миром. Качество его графики и способ подачи материала так же важны, как и другие PR-материалы, выпускаемые фирмой. (Веб-мастер CNRI полшутя называет свою работу как "Отдел искусств". Будучи профессиональным библиотечным работником, она обладает глубокими техническими навыками и большую часть времени проводит за графическим дизайном.)

В некоторых организациях веб-мастер отбирает или даже создает материалы, которые появляются на сайте. В более общем виде можно сказать, что такие материалы создают отдельные люди или группы сотрудников организации. Задача веб-мастера - отредактировать и отформатировать их, а также структурировать их в рамках единого веб-сайта. Таким образом, "форма" веб-сайта создается именно веб-мастером. К примеру, CNRI поддерживает и сайт Internet Engineering Task Force, содержание которого жестко контролируется секретариатом IETF. Однако дизайн домашней страницы и ее "связывание" с другими материалами - это ответственность веб-мастера.

Компьютерные навыки веб-мастеров могут различаться и зависеть от сложности веб-сайта. Некоторые инструменты создания пользовательских интерфейсов, вроде языка программирования Java, требуют профессиональных программистов. Некоторые сайты очень велики, будучи размещенными на разных мощных компьютерных системах, распределенных в разных местах мира. Действительно популярные сайты, к примеру, Cable News Network посещает в день до 100 миллионов пользователей. Дизайн и поддержка таких систем и их сетевых соединений является задачей для профессионалов в компьютерных технологиях. Если же сайт осуществляет коммерческие транзакции, веб-мастер должен иметь навыки в обеспечении безопасности.

Среди организаций, внесших существенный вклад в развитие ЭБ, много факультетов компьютерных технологий. Веб-мастера могут рассчитывать на них в случае возникновения неожиданных задач, таких как создание веб-сервера или индексированный поиск на сайте. В других организациях веб-мастер вынужден быть мастером на все руки. Многие сайты обслуживают организации, в которых вообще нет компьютерных специалистов. Существуют специализированные компании, которые предоставляют веб-сервис для таких фирм, включая создание веб-страниц, техническая поддержка компьютерной техники и выполнение административных функций (вроде регистрации доменного имени).

Новые формы организаций

Консорциумы

Управление крупными коллекциями весьма дорого и трудозатратно. Библиотеки могут сэкономить, объединившись в консорциумы для создания и поддержания общих коллекций. Это также позволяет экономить издателям, поскольку они должны вести переговоры и оказывать поддержку меньшему числу клиентов. В США библиотечные консорциумы организовывались по территориальному принципу (в отдельных штатах, к примеру, в Огайо) и по отраслям знания. В Европе, где большинство университетов - государственные, созданы национальные консорциумы, которые предоставляют услуги ЭБ академическому сообществу. Хороший пример сотрудничества являет собой консорциум Melvyl, созданный университетом Калифорнии еще до появления веб (изначально для совместного пользования каталожными и индексными записями).

Врезка 5.6

Melvyl и цифровая библиотека California

Девять кампусов Университета Калифорнии зачастую ведут себя как девять отдельных университетов. Кампусы в Беркли и Лос-Анжелесе рассматриваются как самостоятельные университеты, но организационно это части одного гигантского университета. Уже многие годы они совместно пользуются цифровой библиотечной системой Melvyl. Руководит Melvyl почти все это время один человек - Clifford Lynch.

Ядро Melvyl - компьютерный каталог всех библиотек во всех девяти кампусах, библиотеки штата Калифорния в Сакраменто и библиотеки Калифорнийской академии наук. В каталоге свыше 8,5 миллионов названий манускриптов, представляющие 13 миллионов единиц хранения. Кроме книг, он включает также карты, фильмы, музыкальные произведения, файлы данных и звукозаписи. База данных периодики включает около 800 000 наименований газет, журналов и пр., включая фонды других крупных библиотек штата Калифорния. Melvyl предоставляет доступ к многочисленным файлам с рефератами и индексами, включая базы данных Медлайн и Inspec. В 1995 в фонды Melvyl добавлены битовые изображения публикаций IEEE, связанные гиперссылками с базой Inspec. Для пользователей, имеющих доступ к этой базе, надпись "Доступно изображение" означает, что эта запись имеет ссылку на растровый образ, который может быть выведен на экран. Использование файлов книг и периодики свободно для всех, а доступ к базам данных статей ограничен только для студентов и сотрудников университета.

Melvyl всегда был среди пионеров использования новых библиотечных технологий. Разработка Z39.50 в значительной степени связана с ним. Команда Melvyl занимается поддержкой и созданием локальной сети, связывающий все кампусы университета.

Успех Melvyl повлиял на решение администрации университета Калифорнии организовать собственные службы ЭБ. Каждый из кампусов имеет свою библиотеку. В 1997 году, после двухлетнего периода планирования, университет создал ЭБ California (Калифорнийская цифровая библиотека). По своему статусу она равна остальным библиотекам университета, а ее директор, Richard Lucier по положению - равен остальным девяти директорам. Начальный бюджет составил 10 млн. долларов в год, но ожидается его существенное увеличение.

Экономические преимущества в такой централизации электронных служб для университета очевидны: устранение дублирования, переговоры по лицензированию, увеличение фонда для приобретения информационных источников. Melvyl стал частью новой ЭБ, которая значительно более амбициозна. Она предполагает принимать активное участие в процессе трансформации традиционного процесса обучения. Предполагается, что ЭБ California будет вести обширную исследовательскую программу и сотрудничать с различными организациями для повышения роли библиотек в процессе учебного процесса и научных исследований. По меньшей мере, университет Калифорнии получил прекрасный набор электронных сервисов; в лучшем же случае ЭБ California изменит жизнь ученых и студентов.

Вторичные информационные сервисы и агрегаторы (aggregators)

Термин "вторичные информационные" охватывает широкий набор сервисов, включая каталоги, индексы, рефераты, а также помощь в поиске информации. Некоторые из таких сервисов (вроде Chemical Abstracts) созданы в рамках профессионального общества; другие (вроде ISI, выпускающего Current Contents и Science Citation Index или Bowker, публикующего Books in Print) всегда были коммерческими. OCLC занимает особую нишу, организации-члены объединяются вокруг задачи совместного каталогизирования.

Эти организации одновременно и доступны для изменений и хорошо адаптированы для расширения своего бизнеса в область ЭБ. Их преимущества: годы опыта в компьютерных технологиях, хороший маркетинг, большие коллекции ценных данных. Многие имеют значительные финансовые резервы или являются дочерними фирмами крупных конгломератов, обладающих достаточными средствами. Почти каждая из этих организаций связывает свое будущее с интеграцией вторичной и первичной информации. В результате существует большое число проектов между издательствами и вторичными информационными службами.

Агрегаторы - это службы, которые подбирают пакеты изданий от разных издательств и предоставляют пользователям как единый пакет (обычно они продают эти пакеты библиотекам). Корни некоторых из них растут из ранних онлайн-информационных систем, таких как Dialog или BRS. Эти службы приобретают (лицензируют) индексы и базы данных, размещают их на центральном компьютере с оригинальным интерфейсом и продают права доступа. В настоящее время крупные агрегаторы могут владеть лицензиями от 5000 издательств, а пользователю предлагать единую лицензию.

Университеты и их библиотеки

Изменения в университетских библиотеках

Как и большинство других организаций, университеты испытывают трудности при переменах. Это большие, консервативные организации с кастовыми предрассудками, затрудняющими командную работу. Культурные различия между гуманитариями и "технарями" общеизвестны, но подобные различия существуют и между студентами, библиотекарями и специалистами-компьютерщиками. Сотрудники факультетов считают ниже своего достоинства общаться с коллегами, не являющимися сотрудниками, библиотекари общаются на собственном жаргоне, а компьютерные специалисты рассматривают объем технических знаний как единственное мерило ценности человека.

Факультеты, которые обычно возглавляются заслуженными работниками, являются мощной силой, пытающейся сохранить статус-кво. Закрытие департамента под любым предлогом или причине, рассматривается как акт "академического вандализма". Когда корпорация закрывает убыточную фабрику, ее акции обычно растут. Когда Колумбийский университет закрыл свой библиотечный колледж, ничьи акции не выросли. Существует очень немного причин для изменений и очень громкая оппозиция против них.

Тем не менее, университеты являются источниками непрерывных нововведений. Главы 2 и 3 дают ряд примеров разработанных в университетах новых технологий и их использования задолго до коммерческого сектора. Поток высокотехнологичных компаний, которые двигают американскую экономику, питается за счет небольшого числа университетов, включая Стэнфорд, Беркли, Гарвард, MIT, Корнелл и Карнеги-Мэллон.

Инновации в больших учреждениях требуют стратегического перераспределения ресурсов. Новые инициативы требуют нового финансирования, которые могут быть изысканы в результате трудного выбора. Процесс перераспределения связан с проблемами. Директор музея Ashmolean в Оксфорде как-то пошутил, что лучший путь для музея - продать часть коллекции для того, чтобы сохранять оставшееся; правда, он сомневается, что сохранит свой пост, если примет такую стратегию. Некоторые деканы соглашались с сокращением некоторых специальностей для того, чтобы оставшиеся были более продуктивными. Наконец, ресурсы перераспределены. Год за годом часть бюджета, расходуемая на компьютеры, сети и их обслуживание (включая персонал) возрастает с темпом ок. 1% в год.

Пока еще неясно, способно ли такое перераспределение в долгосрочной перспективе привлечь больше ресурсов для существующих библиотек или университеты будут развивать новые информационные службы вне рамок библиотек. Сигналы поступают неоднозначные. Парадоксальным образом сейчас, когда ценность хорошей информации высока как никогда, снижается роль университетских библиотек среди источников информации. Университетская библиотека с ее коллекцией книг и журналов является лишь одним компонентом в процессе информационного обмена. Пятьдесят лет назад студенты и преподаватели имели ограниченное число источников. Сегодня в их распоряжении дюжины методов коммуникации. Персональные компьютеры, Интернет, видеоконференции позволяют отдельным студентам обмениваться большими объемами информации. Технология становится такой простой, что студенты способны создавать и распространять информацию без помощи профессионалов - и делать это вне формальной структуры библиотек.

Если библиотеки становятся центрами новых информационных служб, они должны произвести внутреннее перераспределение ресурсов. Обсуждение библиотечного бюджета обычно сфокусировано на растущей стоимости материалов, перегруженности помещений, стоимости компьютеров - но главная статья расходов, это оплата персонала. Некоторые университеты предпринимают честные попытки оценить стоимость их библиотек, но правильный анализ типичной библиотеки показывает, что 25% идет на покупку литературы, 50% на оплату работников и 25% на здания и оборудование. Если библиотеки, пытаясь воспользоваться возможностями электронной информации, будут увеличивать зарплату привлеченным профессионалам, они должны реорганизовать штатную структуру. Это не вопрос того, чтобы угрозами заставить оставшихся работать больше или простого совершенствования библиотечных технологий; требуется именно фундаментальная реструктуризация.

Здания для цифровых библиотек

Проблемой, трудной для всех библиотек, но особенно - для университетских, является задача планировки библиотечных площадей (зданий). В то время, как ЭБ находятся в фокусе исследований и активно создаются во всем мире, для многих библиотек главной проблемой является недостаток места. Например, в декабре 1993 года фонд высшего образования (funding councils for higher education) Великобритании выпустил доклад по университетским библиотекам, известный как доклад Follett. С финансовой точки зрения основной рекомендацией было разработать большую программу строительства зданий. Это особенно актуально для Британии, поскольку число студентов в университетах резко выросло и большая часть места используется для учебного процесса и проживания студентов.

Расходы на новые здания для хранения редко используемых документов трудно установить, но старые привычки меняются медленно. Новые здания были построены во многих местах, включая национальные библиотеки в Лондоне и Париже. Многие библиотеки сохраняют бумажные каталоги в элегантных дубовых кабинетах, пытаясь удовлетворить запросы нескольких высокопоставленных пользователей, несмотря на то, что существует онлайн-каталог, а новые карточки не добавлялись уже несколько лет.

Для того, чтобы воспользоваться традиционной библиотекой, читатель должен туда прийти. Некоторые библиотеки доставляют книги или фотокопии привилегированным пользователям, но даже они должны жить поблизости и быть лично известными работникам библиотеки. Наоборот, пользователи ЭБ не должны посещать какое-либо определенное место. Библиотекари, веб-мастера и другие профессионалы, занятые поддержкой коллекций ЭБ, имеют свои рабочие места, но они совершенно необязательно должны встречаться с пользователем. Публичная библиотека Нью-Йорка должна быть в Нью-Йорке, но Нью-Йоркская ЭБ может хранить свои коллекции на Бермудских островах.

Сейчас трудно сказать, каким должно быть хорошее здание для библиотеки в будущем. Практика эксплуатации и анализ жалоб нового здания Британской библиотеки в Лондоне показывает, что может происходить при плохой планировке. Библиотечные здания обычно эксплуатируются не менее 50 лет, но сейчас никто не может предсказать, как будут выглядеть научные библиотеки всего через несколько лет. Таким образом, при проектировании новых зданий самым главным является гибкость подхода. Архитекторы должны учесть коммуникационные потребности, которые сейчас лишь едва намечаются, все рабочие места должны быть подключены к ЛВС общего пользования и соответствующему энергоснабжению. Одновременно это же здание должно быть пригодным и для традиционных библиотечных полок.

Врезка 5.7

Реконструкция юридической библиотеки Гарварда

Langdell Hall, главная библиотека Гарвардской юридической школы, является не просто действующей и обеспечивающей учебный процесс, но это еще и огромные коллекции по

истории юриспруденции. В течение 1996-1997 гг. библиотека была полностью реконструирована (расходы составили 35 миллионов долларов). Этот проект иллюстрирует проблемы строительства новых библиотечных зданий в период быстрых изменений.

Во время составления плана были сделаны два важных допущения: библиотека не предоставляет компьютеров в общий доступ, а на каждом рабочем месте есть возможность подключить ноутбук (портативный компьютер). В предположении, что каждый пользователь придет со своим компьютером, все читательские места были снабжены розетками и возможностью выхода в сеть. Всего в здании было установлено 540 портов Ethernet, 1158 выходов в сеть и почти все они были немедленно включены (часть оставлена в резерве). Несколько смягчив свою позицию "никаких публичных компьютеров", Гарвард установил около 100 ПК, включая один компьютерный класс и две небольших учебных лаборатории.

Даже для Юридической школы сумма в 35 млн. долларов является весьма значительной. Допущение, что пользователи будут пользоваться личными ноутбуками и что создание большой ЛВС позволит гибко реагировать на возможные изменения, оказалось неточным. Однако главное заблуждение даже не было ясно осознано. Реально было увеличено место для размещения пользователей, - исходя из того, что слушатели по-прежнему будут посещать библиотеку для занятий. При этом было известно, что преподаватели предпочитали работать в своих кабинетах, а не ходить в библиотеку. Много лет назад библиотекарь Harry S. Martin III констатировал: "Наша цель - держать преподавателей вне библиотеки." Предполагалась обеспечить высокий уровень услуг за счет онлайн-информации и курьерской службы. При этом при увеличении числа рабочих мест для пользователей в Langdell неявно предполагалось, что студенты и слушатели юридической школы по-прежнему много лет будут приходить в библиотеку для серьезной работы.

Не все деньги были истрачены на технологические новации. Библиотека представляет собой элегантное старинное здание с коллекциями исторических книг и ценных произведений искусства. Большинство бюджета реконструкции ушло на лифты, подъемники, обогреватели, кондиционирование, водопровод и сантехнику. Не забыли и о внешнем облике. Были отреставрированы старинные столы, новые люстры заменили неэффективные прозрачные потолки, надписи на латыни и барельефы, символизирующие закон, также были отреставрированы. В результате получился внушительный образ замечательной юридической библиотеки, в котором объединились традиционные представления о библиотеке и возможности современных технологий.

Глава 6. Экономические и юридические аспекты

Эта глава более, чем все другие, представляет собой беглый обзор исключительно сложной области, к тому же, отражающий мою личную точку зрения, которая, очевидно, со временем изменится. Однако, я надеюсь, что базовые положения сохранятся.

ЭБ развиваются в обстановке, где интересы пользователей, библиотечных технологий тесно вплетены в неоднозначный экономический, социальный и юридический контекст. Эти аспекты должны изучаться в комплексе и не могут быть поняты в отдельности. Издательства и библиотеки следуют своим внутренним правилам и процедурам, которые разрабатывались много лет. Но по мере развития электронных издательств и ЭБ практика бизнеса и юридические рамки быстро меняются.

Поскольку ЭБ основываются на технологиях, некоторые люди надеются, что техника способна решить все проблемы. Другие люди предполагают, что всего можно достичь, приняв новые законы. Оба подхода недостаточны. Технология вносит свой вклад, но она не может решить социально-экономические вопросы. Изменения законодательства могут быть полезными, но плохие законы еще хуже, чем их отсутствие. Законы эффективны тогда, когда кодифицируют те рамки, которые люди понимают и которые они готовы принять. По тем же причинам, бизнес-модели для ЭБ потерпят неудачу, даже если они учитывают интересы всех сторон. Скрытая проблема в том, чтобы создать такие социальные рамки для использования информации, которые были бы понятными и общепринятыми. Если эти рамки позволят разумным людям выполнять свои задачи, разумные люди будут их соблюдать. Изменения экономических и юридических рамок должны следовать за этим.

Экономические аспекты

Библиотеки и издательства - это крупные деловые предприятия. Огромная индустрия занята созданием информации с целью извлечения прибыли: кинофильмы, газеты, коммерческие фото, романы, учебники, программное обеспечение, звукозапись и пр. Некоторые оценки говорят о том, что эта индустрия составляет до 5% экономики США. В 1997 году библиотека Гарвардского университета имела бюджет свыше 50 млн. долларов и 1000 сотрудников. В библиотеке Конгресса работает 4500 человек. Time Warner и другие издательские группы являются главными игроками на мировом рынке акций. В 1996 году Thomson Corporation заплатила свыше 3 млрд. долларов за West Publishing. Биржевая стоимость акций Интернет-компаний Yahoo еще выше, хотя прибыль ее крошечная.

Соображения издателей о деловых и юридических рамках онлайн-информации определяются страхом и жадностью. Жадность произрастает из веры в то, что издатели смогут извлечь огромные суммы денег из электронной информации, нужно только понять, как это сделать. Страх обусловлен тем, что изменение экономических условий может разрушить традиционные источники прибыли, организации разорятся, а люди потеряют работу. С точки зрения компьютерных профессионалов, такой страх вполне обоснован. Множество компаний, которые достигли успеха, выпуская один вид продукции, затем потерпели неудачу. Такие крупные компании, как Univac, CDC, Burroughs, Honeywell были уничтожены миникомпьютерами. Миникомпьютерные компании, вроде Prime, Wang, Data General не выжили при переходе к персональным компьютерам. Ранние компании по выпуску ПК также умерли, как и большинство первопроходцев в области программного обеспечения. Даже IBM утратила свое превосходство. Увидим ли мы такой же сценарий развития в электронной информации? Издатели и информационные службы, которые доминировали на традиционном рынке, совсем необязательно останутся лидерами и на новых.

Какой бы ни была организация, большой или маленькой, коммерческой или нет, она обязана покрывать свои издержки. Каждая стадия в создании, распространении и использовании ЭБ весьма дорого стоит. Авторы, фотографы, композиторы, дизайнеры и редакторы должны получать вознаграждение за свою работу по созданию информации. По многим причинам вознаграждение - в основном финансовое. Издатели, библиотекари, архивисты, книготорговцы, подписные агентства и специалисты-компьютерщики тоже требуют оплаты. До сих пор не достигнут консенсус относительно того, как платить за информацию в Интернете. Были испробованы почти все мыслимые методы. Для простоты, различные модели могут быть разделены на открытый доступ (средства поступают от создателей или производителей информации) или ограниченный (пользователи платят за доступ к коллекциям).

В главе 5 отмечалось, что люди, участвующие в ЭБ и электронных издательствах, руководствуются различными мотивами. В частности, говорилось о том, что авторы и издатели имеют различные финансовые цели. Когда автор мотивирован денежными соображениями, его интересы совпадают с интересами издателя. Цель - получить доход, прибыль, вопрос только в том, сколько полагается автору, а сколько издательству. В случае если автор руководствуется иными мотивами, нежели денежный, а издатель преследует цель покрыть издержки и заплатить дивиденды держателям акций, их цели объективно противоречивы.

Замечательная особенность веб в том, что есть огромные массивы полностью бесплатной информации, когда пользователь не должен платить за доступ и использование. Многие предсказывали, что открытыми останутся только "внутренние" материалы веб, однако сейчас в сети имеется много открытой информации, издержки по подготовке и доступу к которой несут сами создатели. Многие авторы и поставщики информации заинтересованы прежде всего в том, чтобы их материалы можно было увидеть, поэтому они готовы к тому, чтобы расходовать на эти цели часть бюджета. Среди создателей, которые вкладывают свои финансовые ресурсы для обеспечения открытого доступа к материалам можно выделить исследователей (заинтересованных в профессиональном признании), правительственные агентства, маркетологов всех сортов, людей, чье хобби подталкивает их в этом направлении и пр.

В том случае, если создатели мотивированы максимально широким распространением их материалов, они предпочтут свободный доступ. Но деньги для поддержания их коллекций все равно должны откуда-то поступать. Гранты - это один из важных источников финансовых средств. Проект Perseus в значительной степени существует благодаря грантам различных фондов. В Библиотеке Конгресса проекты American Memory и National Digital Library финансируются как за счет собственных средств, так и за счет грантов. Гранты обычно даются на короткий срок, но их можно получить еще раз. Архив E-print в Лос-Аламосе ежегодно получает гранты NSF, а Netlib - от DARPA. Фактически, финансирование через гранты стало для этих ЭБ институционализированным.

Фирмы, специализирующиеся на поиске информации в веб, вроде Yahoo, Infoseek, Lycos предоставляют свои материалы свободно, получая доходы от рекламы. Они использовали финансовую модель, которая использовалась для телевидения в США. Телевизионные сети платят создателям, а транслируют программы бесплатно. Телезритель формально не платит ничего, доходы получают от рекламы.

Переломной точкой в развитии ЭБ было появление поисковых программ. Некоторые службы такого рода пытались взимать деньги, однако создатель Lycos Michael Mauldin предложил свободный доступ - в результате у конкурентов не осталось выбора, он уже не могли брать деньги за услуги такого же вида. Сегодня поисковые службы бесплатны. После нескольких тяжелых лет, в настоящее время эти фирмы прибыльны за счет поступлений от рекламы и лицензирования права предоставления доступа к их услугам.

Исследовательские коллективы особенно заинтересованы в том, чтобы их работы были широко доступны. Многие научные организации поддерживают собственные веб-сайты, на которых в настоящее время большинство новых научных результатов появляется даже раньше, чем в печати. Серия Internet Draft, поддерживаемая CNRI, оплачивается из различных источников. За счет оргвзносов за участие в конференциях IETF выплачивается заработная плата сотрудникам, которые выполняют технические функции при публикации материалов. А сам доступ ко всем материалам бесплатный.

Правительственные департаменты представляют собой другой важный источник открытых коллекций. Они предоставляют много краткосрочной информации (такой, к примеру, как возможная траектория движения ураганов, поступающей из Национальной службы погоды), но многие их коллекции (к примеру, коллекция международных договоров, поддерживаемая Trade Compliance Center департамента коммерции) имеют долгосрочную ценность.

Частные лица также поддерживают открытые библиотечные коллекции. Эти отличные коллекции посвящены спорту, хобби, клубам по увлечениям. Есть частнопубликуемая поэзия, романы. Оплата за счет самого производителя материалов не является чем-то новым. То, что обычные издательства могут рассматривать как мелочную суету, часто служит источником прекрасных Интернет-коллекций.

Когда издатель или ЭБ хочет получать плату с пользователей, доступ к коллекциям почти всегда ограничивается. Доступ возможен только после получения денег. Технические аспекты описаны в следующей главе. В данной главе мы приводим обзор деловой практики в этой области.

Книжные и журнальные издательства традиционно предполагают оплату со стороны пользователей. Когда экземпляр книги продается частному лицу или библиотеке, деньги делятся между книготорговцем, издателем, автором и другими участниками. Производство кинофильмов придерживается этой же модели. Затраты на производство и распространение копий фильма окупаются за счет входных билетов в кинотеатры, доходов от продажи и проката видеокассет. Распространение этой модели на онлайн-информацию - введение оплаты в зависимости от времени пользования. Большинство пользователей информационной систем Lexis и Westlaw платят повременно. Альтернативный подход к оплате онлайн-информации, который иногда используют, состоит в том, чтобы устанавливать оплату в зависимости от количества одновременно работающих компьютеров (пользователей). В Интернете и веб эти методы выглядят достаточно искусственными.

Альтернативой может быть оплата за материал, переданный пользователю. Некоторые издатели предоставляют полный текст статей за плату, скажем 10 долларов за статью, причем оплата может производиться кредитной карточкой (хотя такой метод оплаты потенциально опасен и дорог технически). Таким образом, существует исследовательская проблема создания автоматической системы оплаты за информацию, полученную по сети. Цель таких систем - создать систему расчетов в Интернете, которая была бы предоставляла возможности осуществления безопасных и низкозатратных транзакций. Предполагается, что это позволило бы даже маленьким организациям открывать свои сетевые сервисы без создания собственной громоздкой системы получения оплаты. Если бы такие системы были созданы, это привело бы к огромному объему небольших по объему транзакций финансовых средств. Как физические объекты существуют в качестве отдельных единиц, так и можно попытаться создать "формальную" единицу электронной информации.

В настоящее время концепция автоматической платежной системы весьма неопределенна. Основной формой оплаты материалов из ЭБ является подписка, включающая определенную оплату за доступ к определенному набору материалов. Неограниченный доступ также возможен при соблюдении определенных условий. Журнал Wall Street Journal разработал хорошую систему коммерческой подписки на свою онлайн-версию. Многие научные журналы сейчас предлагают подписку на электронные журналы для библиотек. Такие общества, как Association for Computing Machinery продает подписку как библиотекам, так и частным лицам. Некоторые из ранее описанных ЭБ начинались с грантов и постепенно достигали самоокупаемости за счет подписки. По этому пути шли JSTOR и Inter-university Consortium for Political and Social Research.

Напрашивается еще одна любопытная параллель с ТВ. В телеиндустрии США были опробованы две альтернативы доходам от рекламы. Первый, оплата за просмотр, предполагает оплату конкретных программ. Эта практика получила определенное распространение, но не стала общепринятой. Второй подход, используемый кабельными компаниями, заключается в оплате ежемесячной подписки на определенное число ТВ-каналов. Последняя модель ведения бизнеса представляется наиболее успешной.

Представляется, что для пользователей ЭБ, как и для телезрителей, предпочтительной является регулярная выплата заранее известных сумм денег. Подписка представляет

определенные преимущества как издателям, так и пользователям. Библиотеки и другие подписчики знают цену заранее и могут адекватно планировать свой бюджет. Издатели могут оценить предполагаемые доходы. Кроме того, подписка позволяет обойти обычную для издательств проблему, когда популярные издания приносят прибыль, а узкоспециальные - убыточны.

Для библиотек подписка является привлекательной формой оплаты еще и потому, что предполагает активное использование их фондов. Они не заинтересованы в том, чтобы издания пылились на полках. ЭБ устраняют многие из существовавших в обычных библиотеках барьеры - и не хотелось бы, чтобы взамен воздвигались новые (как это будет при введении повременной оплаты).

Экономический фактор, которым электронная информация отличается от обычных публикаций, заключается в том, что цена не зависит от того, насколько широко используется данный материал. Многие этапы подготовки публикации - подготовка, рецензирование, редактирование, верстка - одни и те же как для обычной, так и для электронной публикации. Однако для цифровой информации после того, как материал помещен в ЭБ, цена распространения одного экземпляра крошечная. Экономически это значит, что цена практически полностью состоит из фиксированных расходов, остаточная стоимость почти равна нулю. В результате, если объем продаж цифровых продуктов позволяет покрыть производственные расходы, то все остальные продажи приносят чистую прибыль. До этого момента продукт приносит убытки.

Продажа физических материалов (книг, компакт-дисков, фотографий) предполагает цену отдельного экземпляра. Производство каждой копии стоит денег и пользователь чувствует, что это определенным образом отражается в цене. В случае цифровой информации пользователь "в руки" ничего не получает - и это один из доводов против повременной (или в зависимости от объема полученной информации) оплаты. Подписка предполагает фиксированную цену за пользование материалами в соответствии с фиксированной ценой из производства.

Другое экономическое различие электронного книгоиздания от обычных заключается в том, что цена на электронные услуги должна включать в себя расходы на изменения. В долговременной перспективе ЭБ и электронные издательства будут дешевле обычных - но сейчас они весьма дороги. Организации, пытающиеся закрепиться в этом бизнесе, должны сохранять свои традиционные процессы, инвестируя при этом в новые технологии в то время, когда эти новые системы весьма дороги, а уровень спроса относительно низок.

Многие электронные публикации являются версиями печатных материалов. Когда издатели устанавливают цены на электронные издания, они хотели бы знать, каким образом их продажа повлияет на распространение печатного тиража. Если словарь доступен он-лайн, как это повлияет на объем продаж через книжные магазины? Если журнал доступен он-лайн, как это повлияет на число подписчиков? Имеющийся десятилетний опыт пока не позволяет сделать однозначных выводов по конкретным видам изданий. На макроэкономическом уровне очевидно, тем не менее, что электронные материалы становятся существенной статьей расходов в бюджетах библиотек. Обычные и электронные издания зачастую конкурируют за одни и те же деньги. Если один издатель получает больше прибыли от электронных изданий, это, скорее всего, за счет уменьшения продаж другой печатной продукции. Этот общий переток денег от печатных к электронным изданиям является одной из сил, побуждающих издателей переходить к электронной информации.

Некоторые довольно драматические примеры представили вторичные информационные службы. Такие продукты, как Chemical Abstracts и Current Contents весьма незначительно изменили свое основное содержание с конца 80-х годов, но тогда они выходили преимущественно на бумаге - а сейчас доминируют электронные версии. При этом издатели предоставили пользователям возможность выбора из компакт-дисков, магнитных лент, онлайн-доступа и печатных экземпляров. Любое издательство, которое остается вне электронного рынка, должно учитывать вероятное снижение доходов по мере того, как электронные издания изменяют структуру расходов.

К счастью, библиотеки и издательства не должны платить за один из наиболее дорогостоящих компонентов ЭБ - за компьютерные сети общего назначения, связывающие

персональные компьютеры в различных организациях. Инвестиции в телекоммуникации и оборудование идут за счет средств других бюджетов.

Повременная оплата и подписка не являются исключительными способами получения доходов. Рассмотрим практику кабельных каналов, которые распространяют программы, созданные телесетями. Сначала телекомпании жестко сопротивлялись. Политическое лоббирование (прежде всего, Ted Turner) привело к нынешнему состоянию, когда кабельные каналы могут демонстрировать все программы, но обязаны отчислять часть доходов в виде оплаты. Когда радиостанция транслирует аудиозапись, реальные доходы композиторов, исполнителей, студий звукозаписи и других участников процесса распределяются по сложной согласованной системе. Британская футбольная лига получает доходы за счет продажи исключительных прав на трансляцию будущих игр.

Все упомянутые подходы имеют право на существование, т.к. представляют собой прагматичные попытки разрешения сложной проблемы - как люди могут создавать и распространять различные виды информации, получая при этом компенсацию от пользователей, извлекающих определенную выгоду.

Конкретный случай: научные журналы в электронной форме

Научные журналы в электронной форме представляют собой интересный объект изучения, поскольку это одна из первых областей, в которых электронные издательства получали доходы от библиотек и индивидуальных пользователей. Он высветил определенную напряженность между коммерческими издательствами (чья цель - максимизация прибыли) и авторами (кто хотел бы максимально широкого распространения своих работ).

Академические издательства и университетские библиотеки являются естественными партнерами, но переход к электронному книгоизданию породил долговременные трения. Как это показано на врезке 6.1, библиотеки зачастую считают, что издатели запрашивают слишком дорого за свои журналы, хотя во многом университеты сами создают себе проблемы. Многие исследователи и библиотечные работники надеются, что ЭБ создают новые возможности для научных публикаций, более доступных и не столь дорогих для библиотек. В тоже время, коммерческие издательства испытывают давление со стороны своих акционеров, настаивающих на повышении прибыли из года в год.

Врезка 6.1

Экономика научных журналов

Прибыль, получаемая коммерческими издательствами от научных журналов, явилась предметом пристального внимания в последние годы.

В США федеральное правительство использует бюджетные деньги для финансирования исследований в университетах, государственных лабораториях, медицинских центрах и других организациях. Обычный способ публикации результатов таких исследований состоит в том, что ученый пишет статью и посылает в научный журнал.

Первая стадия процесса публикации состоит в том, что редактор (на общественных началах или сотрудник редакции) отсылает статью на рецензирование другим ученым. Рецензенты, работающие без оплаты, критически читают статью на предмет уровня, ошибок и дают свои рекомендации относительно возможности публикации. Этот процесс известен как рецензирование (peer review). Если редактор принимает решение опубликовать статью, он дает автору возможность внести в нее изменения согласно замечаниям рецензентов.

Перед публикацией, большинство издателей выдвигает авторам ряд требований. Обычно требуется, чтобы права на статью были переданы издательству. Дополнительно многие издательства настаивают на том, чтобы эти результаты нигде не публиковались до появления статьи. В результате без каких-либо финансовых затрат издатель получает контроль над работой.

Хотя некоторые журналы (обычно издаваемые научными обществами) имеют индивидуальных подписчиков, главный журнальный рынок - это академические библиотеки. Подписка на многие журналы весьма дорогая - в среднем выше 1000 долларов в год и ежегодно

возрастала на 10-15% в течение 90-х годов. Тем не менее, библиотеки чувствовали себя обязанными выписывать эти журналы, т.к. студенты и преподаватели нуждались в опубликованных в них результатах.

Экономическая система в издании научных журналов весьма странная. Исследователи, большинство рецензентов, а также частично библиотеки получают средства за счет налогоплательщиков, но копирайт принадлежит издателю. Университеты должны взять на себя большую часть вины за создание такой ситуации. Поскольку это их преподаватели выполняли исследования и их библиотеки покупали журналы, простое изменение политики части университетов могло бы сэкономить миллионы долларов, которые сейчас рассматриваются как справедливое вознаграждение издателей. Недавно университеты начали совместную работу по изменению такой ситуации.

Скрытый довод, который вынудил университеты к действию, заключается в том, что научные статьи публикуются не только для того, чтобы сообщить результаты исследований, но и для того, чтобы повысить профессиональную репутацию авторов. Академическая репутация основывается на опубликованных работах, статьях и монографиях. Профессиональное признание, построенное на публикациях, трансформируется в получение грантов, престижных вакансий, известность. Основная трудность в научной карьере заключается в том, чтобы достичь академической позиции, что достигается прежде всего публикациями в рецензируемых журналах. Некоторые ученые также пишут учебники, которые весьма прибыльны с финансовой точки зрения, - но все успешные ученые пишут научные статьи или монографии (в гуманитарных дисциплинах). Дежурная шутка: "Наш декан не умеет читать, но считать он умеет точно". Считает же декан именно статьи в рецензируемых журналах. Издатели весьма рады получать прибыль предоставляя декану такую возможность. Предполагается, что каждый преподаватель обязан опубликовать несколько статей, вне зависимости от того, насколько важны полученные им результаты. Поскольку престиж является следствием числа статей, возникает соблазн опубликовать несколько слегка различающихся статей на основании одних и тех же экспериментальных результатов. Исследования показывают, что большая часть статей никогда не цитируется другими учеными. Многие статьи абсолютно излишни.

Есть признаки того, что ситуация меняется. ЭБ предоставляют ученым иные способы сообщать о результатах их исследований. Рецензирование играет большую роль в отбраковке плохих статей и исправлении ошибок, но на это требуется много времени. Традиционный процесс рецензирования, редактирования, печати и распространения занимает до года и более. Статья в журнале, хранящемся на книжной полке в библиотеке доступна только посетителям, готовым предпринять соответствующие усилия по ее поиску - отчет в Интернете доступен каждому. В некоторых дисциплинах развивается новая традиция (pattern): результаты публикуются в виде открытых онлайн-отчетов или препринтов, при этом традиционные журналы используются с архивными и карьерными целями.

Подписка на онлайн-журналы

Издательства, специализирующиеся на научных журналах в различных областях науки, технологии и медицины включают как коммерческие компании (Elsevier, John Wiley, Springer-Verlag), так и образовательные общества (например, American Association for the Advancement of Science). Эти издательства интенсивно мигрируют в область электронной информации и - соответственно - первыми сталкиваются с экономическими проблемами. До недавнего времени общим был подход, по которому электронные версии появлялись параллельно с печатными. С 1996 года многие издательства публикуют электронные версии своих печатных журналов в Интернете. Онлайн-версии похожи, но зачастую не идентичны печатным. В некоторых отсутствуют отдельные части (вроде "Писем к редактору"); в некоторых добавляются дополнительные данные, которые слишком объемисты для публикации в печатной версии.

Многие издатели выбрали знакомую экономическую модель продажи годовых подписок отдельным библиотекам или их консорциумам. Среди издательств, предлагающих параллельные подписки на электронные и печатные версии такие как Academic Press, American

Chemical Society, Association for Computing Machinery, Elsevier, American Physical Society, Johns Hopkins University Press, Springer-Verlag, Wiley. High Wire Press предоставляет онлайн-услуги для маленьких издательств, которые публикуют высококачественные журналы, но не имеют возможности инвестировать средства в создание компьютерных систем. Общая политика заключается в предоставлении свободного доступа к поисковым индексам и рефератам и оплачиваемом доступе к полным текстам статей. Оплата может быть как в виде подписки, так и за отдельную статью.

Подписки, которые предоставляют доступ к онлайн-коллекциям, требуют соглашения между издателем и подписчиком. Если подписчиком является библиотека, такое соглашение заключается в форме контракта. Хотя детали могут различаться, большая часть пунктов в таких документах одинакова. Некоторые приведены ниже.

Материал. Когда библиотека подписывается на годовой комплект журнала, она получает все выпуски за этот год и может хранить их бессрочно. Если библиотека подписывается на электронную версию годового комплекта, обычно она получает лицензию для доступа к онлайн-коллекции издательства, содержащую текущие выпуски журнала и те предыдущие, которые были конвертированы в цифровой формат. Библиотека снимает проблемы с доступом к ретроспективным выпускам, но она теряет доступ к любым номерам, если не возобновляет подписку на следующий год или издательство уходит из бизнеса.

Пользователи. Если подписка продана библиотеке, пользователи этой библиотеки должны быть идентифицируемыми. Некоторые библиотеки имеют четко определенное сообщество читателей - например, весьма легко идентифицировать тех, кому разрешено использование материалов в корпоративной библиотеке или постоянных читателей библиотек. Но большое число университетов и колледжей имеет много студентов, которые проходят неполные курсы или сотрудников, которые работают в больницах или иных местных организациях и приписаны к библиотеке. Другой пример - публичные библиотеки, по определению открытые для всех. К счастью, в большинстве библиотек традиционно существует процедура выдачи читательского билета. Простой подход при заключении контракта состоит в том, что доступ разрешен всем, у кого есть читательский билет и кто физически находится в здании библиотеки.

Цены для организаций различного размера. Должен ли университет с 5000 студентов платить ту же цену за подписку, что и колледж с 500 или университет штата с 100000? Какой должна быть цена для маленькой исследовательской лаборатории в составе гигантской корпорации? Простых ответов на такие вопросы не существует.

Ценообразование относительно печатной подписки. Если доступны и электронный и печатный варианты, как должны соотноситься между собой цены на них? В долгосрочной перспективе электронная информация дешевле в производстве за счет экономии на типографских расходах, бумаге и распространении. В краткосрочной перспективе электронные издательства требуют существенных новых инвестиций. Вначале некоторые издательства попытались продавать электронные версии дороже печатных; в настоящее время достигнут консенсус, что цена должна быть чуть ниже. Издатели продолжают экспериментировать с ценами для того чтобы заинтересовать библиотеки в подписке на большее число наименований журналов.

Распечатка. Наиболее сложным вопросом является вопрос о том, как читатель может пользоваться онлайн-журналом. Распечатка - это лишь пример. Издатель онлайн-журнала может спокойно относиться к тому, что некоторые читатели будут распечатывать копии для собственного пользования; однако он будет резко возражать против коммерческого распространения таких копий. Согласие о процедуре, которая бы разрешала первое, но препятствовало второму, до сих пор не достигнуто.

Модель институциональной подписки была перенесена с печатных журналов на онлайн-версии без рассмотрения некоторых скрытых противоречий между издателями и пользователями.

Научные журналы и их авторы

Разница в целях авторов и издателей отражается в тех ограничениях, которые некоторые издательства накладывают на авторов. Многие из них требуют, чтобы автор ограничил распространение этих результатов. Естественно, издательства не публикуют работ, ранее

напечатанных в других журналах; однако многие отказываются печатать даже такие результаты, которые просто были кратко доложены где-то еще - на конференции или на веб-сайте. Другие разрешают разместить текст статьи на своем сервере, но требуют удалить текст из свободного доступа после того, как статья опубликована (пример - E-print Archives). Такие ограничения раздражают авторов и нет подтверждений тому, что они способствуют росту доходов. Раздражение увеличивается за счет требований издательств передать копирайт, что может оставить ученых лишь с ограниченными возможностями пользоваться результатами собственных исследований; а налогоплательщиков (на чьи деньги происходит весь процесс) оставить вообще ни с чем.

В Association for Computing Machinery (ACM) мы попытались найти баланс между заинтересованностью авторов в максимально широком распространении его работы - и необходимостью прибыльности. Достигнутый компромисс выглядит временным, но мы надеемся на то, что он сохранится несколько лет. Текущая версия политики предусматривает традиционную передачу копирайта от автора к издательству и подтверждает, что ACM может использовать эти материалы по своему усмотрению. В то же время, она дает авторам больше возможностей. В частности, они могут размещать материалы на частных серверах как до, так и после публикации. В быстрорастущей области, которой посвящены публикуемые ACM журналы и препринты (которые всегда были весьма важны) многие материалы зачастую размещаются в открытых коллекциях (поддерживаемых частными лицами или исследовательскими департаментами). В этой стратегии есть свой риск. Поскольку препринты весьма важны для членов ассоциации, ACM поощряет практику их открытого размещения, хотя в долгосрочной перспективе это может разрушить журнальный рынок.

Законодательные рамки

Эта книга - не учебник юриспруденции (хотя этот раздел проверяли два адвоката), но некоторые юридические аспекты настолько важны, что она была бы неполной без их обсуждения. Поскольку система законодательства создает общие рамки для упорядоченного развития онлайн-сервисов, не удивительно, что многие юридические вопросы имеют прямое отношение к ЭБ. Это относится к таким областям прав, как заключение сделок, интеллектуальная собственность, законы в области коммуникаций, право на частную жизнь, налоги, международное законодательство, клевета и порнография.

Юридическая ситуация в США усложняется множественной юрисдикцией - конституция, международные договоры, федеральные законы, законы штатов - а также прецедентами, создаваемыми судами разных уровней. Многие законы, например, контролирующие непристойность и порнографию (obscenity), - федерального уровня. Люди, размещающие эротические материалы на сервере в Калифорнии, где это разрешено законом, будут преследоваться в Луизиане, где подобное запрещено. Некоторые области законодательства, такие, как копирайт, регулируются на федеральном уровне, хотя отдельные важные аспекты этой темы никогда не были предметом рассмотрения Верховного Суда США. Если же правила рассмотрения конкретных случаев были установлены судами нижних инстанций, их решения не являются обязательными для других судов, которые могут рассматривать ситуацию самостоятельно.

Каждое из двух сообществ, которые занимаются в настоящее время созданием ЭБ в Интернете - специалисты в компьютерных технологиях и профессионалы информатики - имеют свои традиции ответственного использования обобществленных ресурсов. Традиции эти разные, что приводит к появлению проблем. Но даже наихудшие проблемы такого рода бледнеют перед наплывом пользователей Интернета, не отягощенных никакими традициями. Некоторые из них откровенно неконструктивны и злобны, другие намеренно пользуются неформализованностью Интернета для своих целей, третьи беспечны и не задумываются о соответствующих нормах поведения.

До конца 80-х годов Интернет был "заповедником" ученых и исследователей. Была определенная политика по поводу того, кто может им пользоваться и как именно. Более важно

было то, что пользователи сами думали о своем поведении. Любой нарушитель норм немедленно получал гору гневных посланий и попадал под иные формы давления других участников. Социальная конвенция была абсолютно неформальной, но она работала до тех пор, пока большинство пользователей были членами небольшого сообщества и узнавала соответствующие нормы от своих коллег.

Конвенция перестала работать по мере расширения Интернета. В университете Карнеги-Меллона мы заметили изменения у студентов, которые узнавали Интернет не в университетах.

Многие положения законодательства являются общими для Интернета, а не только для ЭБ. В настоящее время Интернет-сообщество работает над техническими методами контроля за недобросовестным использованием электронной почты (что приводит к перегрузке сетей и раздражает добросовестных пользователей). Порнография и азартные игры - это две области, где сталкиваются коммерческие интересы и нормы морали, адвокаты гражданских свобод и более консервативные группы.

Международные аспекты

Поскольку Интернет - сеть международная, в ЭБ можно войти из любой точки мира. Поведение, которое считается безупречным в одной стране, является незаконным в другой. Например, в США разрешено владение оружием, но ограничено использование средств криптографии. В большинстве европейских стран ситуация прямо противоположная.

Отношение к свободе слова также сильно отличается в разных странах мира. В каждой стране есть собственные законы, ограничивающие свободу выражения и доступ к информации, например, призывы к свержению существующего строя, клевета, расовая ненависть, расизм, нарушение частной жизни, разглашение секретной информации. В США право на информацию зафиксировано в Первой поправке к Конституции и является основным демократическим принципом, при этом суды последовательно расширительно трактуют право на свободу слова. Тем не менее, ограничения есть даже в США. Предполагается, что каждая форма активности контролируется соответствующей юрисдикцией. Германия имеет суровые законы против нацизма, арабские страны - против богохульства. Но Интернет контролировать сложно. К примеру, в Финляндии долгие годы существовал анонимный сервер электронной почты. Наконец, под огромным международным давлением, финский суд заставил тех, кто поддерживал сервер называть имена людей, публиковавших особенно возмутительные материалы анонимно.

Интернациональная суть Интернета создает трудности и для электронной коммерции, включая электронную информацию. В США потребители и поставщики уже защищены законами, регулирующими коммерцию между отдельными штатами, такую как финансовые платежи или торговля по каталогам. Более сложная ситуация во всемирных сетях, когда цифровая информация легко может быть изменена, а участники транзакции даже не должны называть страну, из которой они действуют.

Ответственность (Liability)

Ответственность за содержание библиотечных материалов является важнейшим юридическим и социальным аспектом деятельности ЭБ. Общество ожидает от создателей материалов, что они несут ответственность за их содержание, также как и от людей, принимающих решения о содержании. Однако ЭБ не смогут развиваться, если юридическая ответственность будет возлагаться на тех, в чьи обязанности входит только хранение и передача информации.

Благодаря большому уважению, которым пользуются библиотеки, в большинстве демократических стран они обладают привилегированным положением с точки зрения законодательства. Практически невозможно требовать ответственности библиотеки за клеветнические или подрывные утверждения, содержащих в книгах, которые они хранят. Соответственно, законы в области телекоммуникаций защищают операторов связи, которые не обязаны подслушивать телефонные разговоры; более того, это прямо запрещено.

Традиционно ответственность за содержание ложится на авторов и издателей, знакомых с ним, а не на библиотеки. Это позволяет последним собирать материалы всех культур и всех периодов, не беспокоясь о возможных проблемах с нарушением копирайта, частной жизни и пр. Многие люди воспринимают такой подход как правильный и вполне приемлемый так же и для ЭБ. До тех пор, пока ответственность за содержание лежит на организациях, которые их создают и распространяют, нет оснований ожидать, что библиотеки или иные сервисы в Интернет будут отслеживать содержание той информации, которую они передают.

Ответственность провайдеров услуг является одной из центральных тем акта Digital Millennium Copyright, принятого в 1998 году (врезка 6.2). Этот акт снимает ответственность за нарушение авторских прав с онлайн-услуг, включая библиотеки и образовательные инициативы. Как водится, создание на базе этого акта соответствующего закона - это нелегкий процесс, который не закончится и с его принятием. ЭБ приветствуют появление этого акта, но интересно будет посмотреть, как суды будут его интерпретировать.

Копирайт

Наиболее эмоциональные аргументы в связи с онлайн-информацией касаются взаимодействия экономики и закона об авторских правах. Такая аргументация появляется при развитии любой новой технологии. На заре книгопечатания соответствующих законов не было вообще. Пьесы Шекспира были беззащитны перед пиратским распространением. В XIX веке в США существовала защита авторских прав только для американских авторов; книги европейских авторов бессовестно копировались несмотря на возмущение Диккенса и других авторов.

В США закон об авторских правах касается любых видов работ, включая тексты, фотографии, компьютерные программы, музыкальные произведения, видео- и аудиозаписи. Главным исключением являются произведения, созданные правительственными служащими. Исходно правами владеет либо автор, либо его работодатель. В других странах используют другие подходы; в некоторых странах (Франция, Канада) автор имеет персональные "моральные права" (moral rights), которые не могут быть переданы. Исторически авторские права имеют ограниченный срок, включая определенный срок после смерти автора. Конгресс США несколько раз продлевал этот срок, последний раз в связи с опасением Голливуда в связи с истечением авторских прав на ранние кинофильмы - и это грустный пример того, как интересы общества рассматриваются как вторичные по сравнению с финансовыми интересами отдельных корпораций.

Врезка 6.2

Акт Digital Millennium Copyright

В 1998 году Конгресс США одобрил акт Digital Millennium Copyright, который внес существенные изменения в законодательство об авторском праве. Почти весь он посвящен цифровым произведениям (works) в сетях. При беглом знакомстве акт выглядит как разумный баланс между коммерческими интересами тех, кто хочет продавать цифровые произведения и свободным распространением информации в интересах прежде всего библиотек и образования.

Согласно акту, провайдеры онлайн-услуг в США, включая библиотеки и образовательные учреждения, в значительной степени защищены от обвинений в нарушении авторских прав, которые произошли из-за их незнания. Чтобы установить этот факт, должны быть выполнены определенные правила: организации должны предоставлять пользователям информацию об авторских правах; иметь определенные правила по пресечению подобных нарушений; соглашаться с требованиями убрать подобные материалы; соблюдать технические меры в соответствии с промышленными стандартами; сообщать в Офис по авторским правам о событиях, подпадающих под действие данного акта. Акт явным образом разрешает многие виды деятельности, которые фундаментально важны для ЭБ. Он позволяет провайдерам пользоваться веб-сайтами, гиперссылками, поисковыми программами. Он признает, что провайдеры делают копии материалов с техническими целями (например, для кэширования или для передачи материалов на другие сайты). Эти действия актом разрешены и провайдеры не

отвечают за действия пользователей, если указанные условия и правила соблюдались, а отмеченные проблемы своевременно устранялись.

Для университетов и других высших учебных заведений акт делает важное исключение в том смысле, что в общем порядке организация несет ответственность за действия своих сотрудников. Признается, что администраторы не могут отследить все действия сотрудников и студентов - а, значит, и не могут нести ответственности.

Акт запрещает попытки обойти технические методы, используемые держателями копирайта для ограничения доступа к материалам. Он также запрещает создание или распространение средств, вызывающих нарушения нормальной работы этих технологий. При этом акт признает несколько исключений, причем все они весьма сложны и требуют тщательной интерпретации: разработчики программного обеспечения могут "обходить" (reverse) инженерные системы защиты для достижения интероперабельности, исследователи могут исследовать способы шифрования и методы обеспечения безопасности систем, государственным организациям, отвечающим за безопасность, может быть предоставлено специальное разрешение обходить меры безопасности; библиотеки могут знакомиться с материалами с целью принять решение о приобретении. Наконец, пользователям разрешено идентифицировать и отключать технические устройства (или методы), которые собирают частную информацию о пользователях и процессе работы.

Акт устанавливает правила менеджмента информацией о копирайте данной работы (автор, исполнитель, владелец копирайта). Эта информация не может быть случайно изменена или удалена.

Врезка 6.3

Эпизоды истории авторского права

В витринах Агентства по защите авторских прав правительства США выставлена серия стендов, которая иллюстрирует процесс изменения законодательства в решениях судов различных уровней, включая Верховный Суд. Видно как год за годом юридические прецеденты уточняли и проясняли закон и позволили распространить его на области (фотография, радио, компьютеры), которых не было на момент принятия Конституции и основных законов. Некоторые из этих судебных решений приведены ниже.

Wheaton против Peters (1834). Это дело создало прецедент, благодаря которому был установлен принцип, что авторские права не являются своего рода естественными правами, а возникают в соответствии с законодательным актом об авторских правах и подчиняются условиям, налагаемым в соответствии с ним. (This landmark case, established the principle that copyright is not a kind of natural rights but rather is the creation of the copyright statute and subject to the conditions it imposes.)

Baker против Selden (1880). Этот прецедент устанавливает, что закон об авторских правах защищает то, что написал автор и способ выражения идей. Но он не защищает сами идеи.

Литографическая компания Burrow-Giles против Sarony (1884). Решение суда распространило действие закона на другие объекты, кроме текстовых. В данном случае это была фотография О.Уайльда.

Bleistein против литографической компании Donaldson (1903). Это случай о цирковых афишах. Суд решил, что они являются объектами авторского права независимо от того, имеют ли они эстетическую ценность и приносят ли удовольствие своим видом.

Компания Fred Fisher, Inc. против Dillingham (1924). Это спор о сходстве двух музыкальных фрагментов. Суд заключил, что бессознательное (непреднамеренное) копирование может привести к нарушению авторских прав.

Nichols против корпорации Universal Pictures (1931). Суд постановил, что не является нарушением закона об авторских правах заимствование для фильма идеи и персонажей из успешной бродвейской пьесы.

Sheldon против корпорации Metro-Goldwyn Pictures (1936). Суд постановил, что "плагиатора не извиняет указание на те фрагменты работы, которые он создал сам, а не украл".

C.Ricordi & Co. против Paramount Pictures (1951). Случай о том, как должны интерпретироваться возобновленные авторские права и права на работы "по мотивам" - на примере романа "Мадам Баттерфляй" John Luther Long, пьесы Belasco по этому роману и оперы

Пуччини по мотивам этой пьесы. Суд постановил, что закон об авторских правах защищает только вновь добавленные части работы.

Warner Bros. Pictures, Inc. против Columbia Broadcasting Systems, Inc. (1955). Суд решил, что образ Sam Spade в работе "Мальтийский сокол" относится к ключевым для развития сюжета, но не является элементом, к которому можно применить закон об авторских правах.

Mazer против Stein (1954). Суд решил, что закон не распространяется на утилитарные и полезные предметы, в данном случае лампы-скульптуры. Есть возможность защиты зарегистрированных художественных, графических, скульптурных фрагментов утилитарного объекта.

King против Mr. Maestro, Inc. (1963). Это случай по поводу речи Мартина Л.Кинга (мл.) "У меня есть мечта". Несмотря на то, что он выступал с этой речью перед огромной толпой и одновременно она транслировалась по радио и ТВ, суд решил, что публичное "исполнение" не означает факта публикации и эти авторские права на речь могут быть оформлены только как на "неопубликованную работу".

Letter Edged in Black Press, Inc. против Public Building Commission of Chicago (Комиссия по общественным зданиям г. Чикаго) (1970). Этот случай, касавшийся публичной демонстрации скульптуры Пикассо, был учтен в последующем законодательстве.

William Electronics, Inc. против Artic International, Inc. (1982). Это случай о правах на видеоигру. Суд постановил, что компоненты игры подпадают под действие закона об авторских правах, а компьютерная память "только чтение (read-only)" может рассматриваться как копия.

Norris Industries, Inc. против International Telephone and Telegraph Corp. (1983). Суд постановил, что даже если Агентство по защите авторских прав отвергает работу, как не подлежащую защите, пользователь может возбудить дело и добиваться решения суда.

Даже эти основные решения не могут рассматриваться как окончательные. Многие из этих случаев вообще не рассматривались в Верховном суде и могут быть оспорены. Так, недавно федеральный суд принял решение, которое явным образом противоречит предыдущему решению суда по делу *King против Mr. Maestro*.

Собственник авторских прав на работу имеет исключительное право делать копии, создавать связанные (derivative) с ней произведения и распространять копии любыми способами за деньги или бесплатно. Эти обстоятельства позволяют авторам быть уверенными в том, что их работы не будут каким-либо образом (случайно или намеренно) искажены. Издателям это позволяет выпускать продукцию без опасений за то, что рынок будет разрушен за счет появления копий из других источников.

Хотя держатель авторских прав обладает значительным контролем над тем, каким образом может быть использован материал, закон не является абсолютным. В законе США существуют две важные концепции: "первая продажа" (first sale) и "честное использование" (fair use).

Понятие "первая продажа" применимо к физическим объектам, например, книгам. Держатель прав может контролировать процесс продажи и устанавливать цену; однако если покупатель приобрел экземпляр книги, он является ее собственником и может продать или распоряжаться любым иным способом не получая специального разрешения.

"Честное использование" является законным правом, которое предусматривает некоторые возможности использования защищенной законом информации без разрешения держателя прав. По этому основанию авторы обзоров или студенты имеют право цитировать короткие фрагменты работы, а статья или часть книги могут быть скопированы для личной работы. Границы применимости этого права четко не определены, но есть четыре основных критерия: цель и характер использования (включая коммерческий или некоммерческий образовательный характер), тему (предмет) защищаемой работы, объем и важность использованного фрагмента по отношению ко всей работе, последствия такого использования на цену и рыночный потенциал работы. Поскольку эти критерии достаточно неопределенны, судьи имеют возможность трактовать их различно. В общем порядке, "честное использование" позволяет воспроизводить части работы, но не всю целиком; снимать одну копию, но не несколько; частное, но не коммерческое использование. Точное определение может быть только на основе юридического прецедента. Судебное рассмотрение (тяжба) настолько дорогостоящая процедура, что многие важные аспекты проблемы авторских прав (даже относительно

традиционных печатных материалов) никогда не рассматривались в суде или слушались только в суде низшей инстанции.

Концепции "первой продажи" и "честного использования" не могут быть легко перенесены на ЭБ. Если "первая продажа" еще может быть применена к физическим носителям электронных материалов (компакт-диски, например), то аналогии для информации, доставляемой по компьютерным сетям, нет. Параллели для концепции "честного использования" между физическими носителями и онлайн также неопределенны. Это лишь отдельные примеры тех случаев, которые привели к серии попыток изменить закон об авторских правах, как американских, так и международных. Хотя большинство людей согласны с тем, что такой закон должен представлять собой баланс между правом на доступ к информации и экономическими интересами создателей и издателей информации, нет согласия по поводу того, каким именно должен быть такой баланс. Более того, с каждой стороны существуют экстремисты. Корыстные предприниматели, мотивированные жадностью и страхом, попытались в одностороннем порядке "протолкнуть" соответствующий законопроект. Несколько лет назад в Конгресс был внесен законопроект, имевший целью изменить или прояснить закон об авторских правах в отношении онлайн-информации в интересах нескольких заинтересованных групп. Некоторые коммерческие организации лоббировали драконовские правила контроля за информацией и уголовные санкции за все виды действий, не прошедших явной авторизации. В ответ группы, защищающие общественные интересы, возражали, что нет доказательств того, что "честное использование" снижает прибыли, а необоснованные ограничения информационных потоков будет иметь пагубные последствия. До 1998 года проблема так и не была решена, что - возможно - и хорошо. Существующее законодательство было достаточно адекватным для первой фазы развития ЭБ и электронных издательств. Главной трудностью было - осознать внутренние проблемы. Естественно, уточнение законодательства было желательным, но полезнее было отслеживать складывающуюся практику, чем немедленно вводить жесткие законы. Законодательство 1998 года, описанное на врезке 6.2, вероятно, достаточно хорошо для выживания ЭБ.

Частная жизнь, конфиденциальность (privacy)

Библиотеки - по крайней мере, в США - четко осознают право пользователей на конфиденциальность. Никто не должен знать, что конкретный читатель брал книги "чувствительного" содержания (например, о "неприятных" заболеваниях). Библиотеки охотнее шли в суд, чем сообщали в полицию, о том, читал ли посетитель книги о коммунизме. Многие штаты имеют законы, которые прямо запрещают библиотекам собирать данные, которые нарушают частную жизнь читателей. Интернет-сообщество имеет схожие традиции. Хотя корпорации имеют законное право интересоваться тем, что делают их служащие, большинство людей на технических должностях ожидают, что в обычных обстоятельствах их электронная почта и файлы на персональных компьютерах рассматриваются как частные.

Большая часть технологий, лежащих в основе ЭБ, используется и для электронной коммерции. Рекламодатели и торговцы заинтересованы в сборе информации о пользователях (зачастую без их ведома), более того, они продают такую информацию друг другу. Используемые для этого технологии (вроде регистрации входа пользователя на сервер) также могут быть использованы для наблюдения за поведением людей без их ведома.

Как об этом говорилось во врезке 6.4, ЭБ должны собирать определенные данные. Надежная информация необходима для настройки компьютерных систем, разрешения проблем или планирования. С определенными предосторожностями полезная статистика может быть собрана без установления персональных данных пользователей. Но не все предпринимают такие предосторожности. Разбираясь с проблемой в ЛВС, системный администратор иногда может захотеть просмотреть любой файл на сервере или проверить любое послание, прошедшее по сети. Допустим, он случайно наткнется на глубоко личную или криминальную информацию. Каким должно быть правильное поведение в подобной ситуации? Что говорит об этом закон?

Врезка 6.4

Цифровые библиотеки и частная жизнь (конфиденциальность)

Любой, кто предоставляет услуги, нуждается в информации об этом процессе. Управление компьютерной системой, такой как ЭБ, требует данных о производительности, надежности, использовании ресурсов, о тенденциях и пиковых значениях показателей. Библиотеки и издательства нуждаются в статистике использования коллекций и сервисных служб. Разработчики пользовательских интерфейсов зависят от знаний о том, каким образом люди взаимодействуют с библиотекой. Обеспечение безопасности требует внимательного отслеживания действий пользователей и анализа отклонений.

Эту информацию нелегко собрать. Веб-сайты регистрируют данные о том, как часто обращаются к тому или иному файлу. Широко распространенной практикой на сайтах является размещение счетчиков. Поскольку каждая картинка хранится в отдельном файле, то даже единственный читатель, обратившийся к одной страничке, может существенно увеличить счетчик числа обращений. Эта статистика важна для правильного конфигурирования компьютерной системы, - но больше ни для чего. Менеджер онлайн-журнала хотел бы знать, сколько людей читали журнал с разбивкой по конкретным материалам. Число обращений к страничке оглавления и к первым страницам статей является полезной информацией. Но это не прямые измерения, которые не могут различить, обращается к одной статье несколько раз один и тот же человек - или ее прочитали несколько пользователей. Измеряется число обращений к компьютерной системе, но невозможно оценить, сколько читателей воспользовались копиями материалов, находящихся в кэшах или на сайтах-зеркала. Если пользователи идентифицируются, это существенно облегчает сбор точной статистики. Но в этом случае проблемой становится конфиденциальность.

В университете Карнеги-Меллона мы решили эту проблему таким способом: каждый раз, когда пользователь обращается к материалу, генерируется временная запись, включающая IP-адрес компьютера читателя. Эти временные записи хранятся в специальной части файловой системы, которая никогда не подвергается резервному копированию (в т.ч. на магнитную ленту). Раз в неделю записи анализируются и создается отчет, из которого уже невозможно идентифицировать конкретного пользователя. Он содержит такую информацию, как число различных пользователей каждой коллекции и общее число поисков. Затем временные записи и любые следы IP-адресов уничтожаются.

Патенты на программное обеспечение

Несмотря на то, что в законодательстве есть свои ошибки, ему приходится иметь дело с быстрорастущей областью вычислительной техники и Интернет. Самым плохим примером, исключением являются патенты на программное обеспечение. Лишь некоторые области права столь далеки от реальности, которую они призваны регулировать. В слишком большом числе случаев патентный офис США выдает патенты, которые любой профессионал признает глупым или бессмысленным. До недавних пор в патентном ведомстве даже не привлекали опытных специалистов для проверки заявок. По-прежнему выдаются патенты на излишне широкие патентные формулы, на концепции, которые уже давно известны, или новые способы использования стандартных процедур.

Одна из причин такого беспорядка заключается в том, что концепция изобретения, на которой базируется соответствующий закон, идет еще от Архимеда, который выскочил из ванны с криком "Эврика!". Но эта концепция не подходит для вычислительной техники. Новые идеи в программном обеспечении вызревают постепенно. Сообщество компьютерных профессионалов достаточно гомогенное. Люди учились в одних и тех же университетах, работают на одной технике и программном обеспечении. Существует свободный обмен идеями по многим каналам. В результате разные группы параллельно работают над одними проблемами и используют одни и те же стандартные процедуры одним и тем же способом.

Успех Интернета и быстрое распространение ЭБ базировалось именно на открытом обмене идеями. Патентный закон, который оперирует такими понятиями, как секретность, тяжба, противостояние - способен лишь затруднить этот процесс.

Глава 7. Управление доступом и безопасность

(Эта книга использует термин "управление доступом" (access management) для описания контроля доступа к ЭБ. Но используются и другие термины, в частности, некоторые предпочитают говорить об "правилах и условиях" (terms and conditions). В издательском деле, где в центре внимание обычно прибыль, общим является несколько странное выражение "управление правами" (rights management). Каждый из этих терминов имеет свои особенности, но в принципе это синонимы.)

В этой главе рассматриваются два взаимосвязанных вопроса: методы контроля за теми, кто обращается к материалам в ЭБ и обеспечение безопасности в компьютерных сетях.

Очевидный повод для контроля - экономический. Поскольку издатель предполагает получить доход от своих продуктов, он предоставляет доступ только тем пользователям, кто заплатил. Может показаться, что в иных случаях контроль за доступом излишний, но есть и другие доводы в пользу такой практики. Некоторые материалы могут быть предоставлены в ЭБ с определенными условиями (которые могут быть связаны с внешними событиями, к примеру временем жизни конкретных лиц). Организации в своих коллекциях могут хранить и такую информацию, которую они хотели бы сохранить в тайне (коммерческие секреты, полицейские записи, секретная правительственная информация). Даже если коллекция открытая, необходим контроль за процедурами дополнения, изменения или удаления материалов и метаданных. В хорошей ЭБ должны существовать записи, фиксирующие все изменения, так, чтобы коллекцию можно было восстановить в случае каких-либо ошибок или сбоев компьютерной техники.

Неопределенность является постоянным фактором в процессе управления доступом. Люди с компьютерным образованием иногда полагают, что каждый объект может быть описан метаданными, в которых перечислены все права, разрешения и другие факторы, относящиеся к управлению доступом. Люди с библиотечным образованием, особенно те, кто поддерживает исторические коллекции или архивы, знают, что эта процедура требует огромного количества времени и зачастую просто невозможна. Такие проекты, как American Memory Библиотеки Конгресса предусматривают конвертацию миллионов документов из исторических коллекций. Может показаться естественным, что в случае этих достаточно старых материалов срок авторских прав закончился - однако это далеко от истины. Срок окончания действия авторских прав на опубликованные материалы связан с датой смерти автора. Ее зачастую трудно установить, кроме того, библиотекари не знают, был тот или иной документ коллекции опубликован или нет.

Как отмечалось в гл.6, многие законы, которые относятся к ЭБ, такие как закон об авторских правах, имеет существенные и явные нечеткости, неопределенности. Соответственно, и политика управления доступом, которая основывается на этих законах, страдает от этой неопределенности. По мере того, как нечеткости будут устраняться в новых законах, договорах или судебных прецедентах, эта политика должна будет адекватно уточняться.

Элементы управления доступом

На рисунке 7.1 показаны общие элементы, полезные для понимания процесса управления доступом. Как это изображено слева, информационные менеджеры разрабатывают политику доступа. Эта политика связывает пользователей (вверху) с материалами ЭБ (внизу). Процесс авторизации (в центре) уточняет доступ (справа). Каждая из этих частей требует специального рассмотрения. Политика доступа должна базироваться на соответствующих законах и соглашениях с третьими сторонами (таких, как лицензии от держателей авторских прав). Пользователи должны быть идентифицированы, а их роль в процессе доступа четко определена. Цифровые материалы коллекций должны быть идентифицированы, а их

аутентичность четко установлена. Сам доступ выражается в терминах разрешенных операций (действий).

Когда пользователь запрашивает доступ к цифровому материалу, запрос проходит определенный процесс управления доступом. После идентификации пользователя в результате процесса аутентификации (authentication), процедура авторизации разрешает или отказывает пользователю в доступе или выполнении определенной операции.

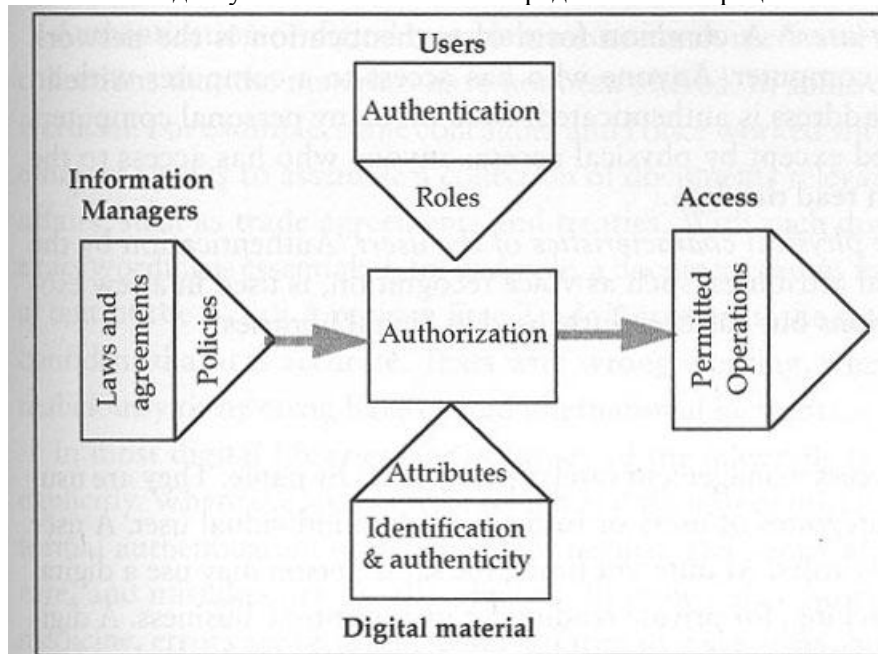


Figure 7.1

Ответственность за доступ лежит на менеджере ЭБ. Это может быть библиотекарь, издатель, веб-мастер или создатель информации. Часть ответственности может быть делегирована. Если библиотека контролирует материалы и предоставляет их пользователям, она устанавливает правила, которые обычно определяются внешними ограничителями, такими как юридические, лицензии от издателей, соглашения с донорами материалов. Если издатель размещает материалы и лицензирует доступ к ним - в таком случае он является менеджером; однако в данном случае он может передать часть ключевых операций (например, авторизация пользователей) другим.

Пользователи

Идентификация

Для идентификации пользователей используется много различных способов. Некоторые из них весьма просты, но легко обходятся. Другие, более безопасные, достаточно сложны. Методы идентификации могут быть разделены на четыре главные группы:

Что знает пользователь? Один стандартный и широко используемый метод идентификации состоит в том, что каждому пользователю присваивается имя (login) и пароль. Это широко распространенная практика, но весьма уязвимая. Пароли легко украсть, а люди часто выбирают такие простые пароли (чтобы легче запоминать), что их очень легко раскрыть.

Чем владеет пользователь? Для идентификации используют два вида карт: магнитные, которые используют специальные считывающие устройства и электронные "умные" карты, которые выполняют определенную программу идентификации. Умные карты - одна из лучших систем идентификации: они весьма безопасны и легки в использовании.

Где пользователь? Общепринятая форма идентификации - сетевой адрес компьютера. Любой, кто имеет доступ к компьютеру, чей адрес прописан в соответствующем файле - автоматически

получает право доступа. Данные на многих персональных компьютерах не защищены (за исключением простого физического доступа), поэтому любой, кто имеет доступ к компьютеру, может прочесть эти данные.

Каковы физические характеристики пользователя? Это идентификация с использованием физических особенностей пользователя - распознавание голоса и пр. Обычно используется в специальных областях, в ЭБ практически не встречается.

Роли

В документах по управлению доступом пользователей редко называют по именам. Обычно их делят на категории пользователей или на роли, которые они выполняют. Читатель может выполнять много ролей. В разное время человек может пользоваться ЭБ для обучения, для частного чтения, дополнительного бизнеса. ЭБ могут иметь различные политики в отношении одного и того же человека в зависимости от его роли. Различают следующие роли:

Член группы. Пользователь может быть, например, сотрудником европейского Института Физики или студентом военно-морской академии.

Местоположение. Пользователь может пользоваться компьютером в библиотеке Карнеги в Питтсбурге или находиться в Новой Зеландии.

Подписка. Читатель может являться подписчиком Journal of the Association for Computing Machinery или работать в университете, который имеет лицензию на использование всех JSTOR коллекций.

Программа автоматического индексирования. Пользователь может использовать веб-поисковик или другие компьютерные программы.

Оплата. Пользователь может иметь кредитный счет в системе Lexis или просто заплатить 10 долларов за доступ к материалу.

Большинство пользователей ЭБ - это люди, использующие персональные компьютеры. Однако пользователем может быть и компьютер, работающий без присутствия человека, программа, которая индексирует веб-страницы или программа-дубликатор, которая обновляет зеркальный сайт. Некоторые сайты могут явным образом запрещать доступ автоматических программ, а некоторые дают им большие привилегии.

Цифровые материалы

Идентификация и аутентификация материалов

Для управления доступом цифровые материалы должны быть четко идентифицированы. Идентификация предусматривает связь некоторого конкретного имени (идентификатора) с каждым объектом материала. Это общий подход как для ЭБ, так и для электронных издательств.

Аутентификация цифровых материалов гарантирует как пользователям, так и менеджерам коллекций, что материалы защищены от изменений. В некоторых случаях это принципиально важно. Например, однажды я с группой коллег работал в правительственном агентстве США над созданием коллекции документов, относящихся к иностранным отношениям (торговые соглашения и договоры). В таких документах важно каждое слово; если, например, провозглашено, что данный документ является официальным текстом Соглашения по Североамериканской зоне свободной торговли, то пользователь должен быть уверен, что это так. Тексты с ошибками, совершенными сознательно или случайно, вызывали международные инциденты.

В большинстве ЭБ точность материалов явным образом не верифицируется. Где уровень доверия достаточно высок, а цена ошибки невелика, формальная аутентификация вообще не требуется. Намеренное изменение документов бывает редко, а ошибки обычно очевидны. Однако, в некоторых областях, таких как медицина, ошибки влекут серьезные последствия. ЭБ

в этих дисциплинах должны серьезно рассматривать формальные средства аутентификации материалов.

Для подтверждения точности объекта, он может снабжаться цифровой подписью (digital signature), более подробно об этом ниже. Цифровая подпись подтверждает, что файл или иной набор битов не изменялся с того момента, как была вычислена подпись. Врезка 7.1 описывает практику использования цифровых подписей в Агентстве по авторским правам США.

Другая группа аутентификации связана с термином "водяные знаки" (watermarking). Эти технологии используются издательствами для предупреждения и выявления незаконного копирования. Идея заключается во внедрении в материал определенного кода, который не мешает пользоваться им, практически незаметен, но позволяет установить собственника. Примерно так телеведущие корпорации добавляют к телевизионной картинке логотип компании. Цифровые "водяные знаки" могут быть полностью невидимы для пользователей и их практически невозможно бесследно удалить.

Атрибуты цифровых материалов

Политика управления доступом зачастую по-разному рассматривает различные материалы в зависимости от их атрибутов. Они могут храниться в виде административных метаданных вместе с объектом или быть полученными из других источников. Некоторые атрибуты могут вычисляться (к примеру, может быть подсчитан размер объекта). Ниже даны типичные примеры:

Разделение на субколлекции. Материалы в коллекциях часто разделяются на открытые для доступа полностью или ограниченно. Издатели могут отделять полный текст статей от индексов, рефератов и рекламных материалов. Веб-сайты, кроме области, открытой для всех, могут иметь разделы только для сотрудников данной организации.

Лицензирование и другие внешние обязательства. В ЭБ могут содержаться материалы, лицензированные у издателей или являющиеся объектами особых условий доступа. Например, такие материалы Библиотека Конгресса получает посредством процесса copyright deposit.

Физические и временные свойства. Политика ЭБ может основываться на времени с даты публикации или физических свойствах объекта (вроде его размера). Некоторые газеты предоставляют открытый доступ к избранным статьям вскоре после опубликования, но требуют лицензии для доступа более поздним материалам.

Тип контента. Политика ЭБ может варьироваться в зависимости от формата или типа материала. Например, оцифрованный звук, тексты, компьютерные программы и изображения могут рассматриваться отдельно.

Врезка 7.1

Электронная регистрация и хранение авторских прав

Агентство по авторским правам (Copyright Office) США является самостоятельным департаментом Библиотеки конгресса. Согласно федеральному закону библиотека должна получать для своей коллекции по два экземпляра любой работы, изданной в стране. Если получение и хранение печатных экземпляров является обязательным, то закон об авторских правах от 1978 не требовал обязательной регистрации авторских прав работ, хотя такая регистрация поощрялась и давала значительные преимущества.

Метод регистрации авторского прав прост. Владелец прав посылает два экземпляра работы в Агентство вместе с регистрационной формой и оплачивает определенную сумму. Таким образом заявляются права и работа рассматривается в соответствии с процедурой. После одобрения выдается регистрационный сертификат. Опубликованные работы затем передаются в Библиотеку Конгресса, которая решает использовать их для своих коллекций или для программ обмена с другими библиотеками.

В 1993 году Агентство по авторским правам и Corporation for National Research Initiatives начали работу над системой, известной как CORDS (Copyright Office Electronic Registration, Recordation and Deposit System, Система Агентства по авторским правам по электронной регистрации, ведению регистра и хранению) для регистрации и хранения электронных работ.

CORDS зеркально воспроизводит традиционную процедуру. Процедура предложения работы для регистрации состоит в заполнении веб-формы и пересылки самой работы через Интернет (с процедурами контроля безопасности). Оплата производится отдельно. Для идентификации заявок на регистрацию используются цифровые подписи. Используя закрытый (личный) ключ заявитель передает заявку, работу, цифровую подпись, открытый ключ и соответствующие сертификаты. Цифровая подпись проверяется Агентством, в дальнейшем именно она будет использоваться при разрешении любых споров относительно копирайта на данную работу.

Атрибуты могут присваиваться на различном уровне. Если все материалы коллекции имеют одинаковые свойства, логично присвоить атрибут всей коллекции целиком. С другой стороны, бывают моменты, когда части объектов могут иметь различные свойства. Зачастую права на изображения отличаются от прав на текстовый материал, в который оно включено - и это должно четко разграничиваться. Некто может передать в библиотеку частную переписку, но потребовать, чтобы доступ к некоторым письмам был ограничен. ЭБ должны иметь возможность связывать соответствующие атрибуты с коллекциями, субколлекциями, отдельными библиотечными объектами или с элементами индивидуальных объектов.

Операции

Политика управления доступом зачастую определяет или ограничивает операции и действия, которые может выполнять авторизованный пользователь над библиотечными материалами. Можно выделить два следующих аспекта:

Вычислительные (computing) операции. Эти операции определяются в терминах компьютерных технологий, таких как запись данных в компьютер, выполнение программы, передача данных по сети, отображение на дисплее, распечатка, копирование с одного компьютера на другой.

Пределы пользования. Пользователю может разрешаться извлекать отдельные материалы из базы данных, но он не имеет права копировать ее целиком.

Вышеназванные операции могут контролироваться техническими средствами. Однако в отношении большей части содержания это практически невозможно. Отметим следующие политики управления доступом:

Цель операции. Пользователи могут авторизовываться в соответствии с целью выполнения определенных операций. Цель использования может быть, например, коммерческой, образовательной или правительственной.

Интеллектуальные операции. Операции могут различаться по интеллектуальному использованию объектов коллекции. Наиболее важными в данном случае являются правила создания новой работы на основе содержания другой. Критерии такого использования необходимо рассматривать как с внутренней, так и внешней точек зрения.

Последующее использование (Subsequent Use)

Системы управления доступом должны обеспечивать как прямые операции, так и последующее использование материалов. Прямые операции - это действия, инициированные (выполняемые) репозитарием (хранилищем) или иной компьютерной системой, которая действует как агент по управлению коллекцией. Последующее использование охватывает все операции, которые могут быть осуществлены с материалом, вышедшим из-под контроля ЭБ, включая все способы копирования (от копирования компьютерных файлов до фотокопирования бумажных документов). С интеллектуальной точки зрения, последующее использование может включать извлечение коротких фрагментов, создание производных материалов и плагиат.

В том случае, когда материал передан на персональный компьютер, технически трудно предотвратить такие действия пользователя, как копирование, хранение или распространение полученных данных. Это можно сравнить с фотокопией документа. Если речь идет о коммерческой информации, потенциальный вред таких действий на уровень доходов очевиден.

Естественно, издатели вынуждены думать о том, что пользователи могут распространять незаконные копии материалов. В пределе, если с единственного проданного экземпляра снять копию и поместить в Интернете - то весь объем продаж может свестись к этому единственному экземпляру. Зачастую, отвечая на эти законные опасения, ЭБ разрешают пользователям доступ к отдельным записям, но не позволяют копировать коллекции целиком. Несмотря на то, что это все равно приводит к некоторому снижению доходов, это действенный барьер, препятствующий подрыву экономических интересов издателей путем глобального копирования.

Возможные политики

Неформальное определение политики как набора правил о том, что и с какими материалами может делать авторизованный пользователь. Для ЭБ типичными являются следующие политики:

- Открытая публикация. Любой желающий может прочесть материал, но только сотрудники могут его изменить.
- Издатели онлайн-журналов предоставляют доступ ко всем материалам только подписчикам. Остальные могут прочесть оглавление и рефераты, к полному тексту доступ только за плату.
- Секретные материалы правительственных организаций (например, "особо секретные") имеют жесткую политику относительно того, кто может получить доступ, при каких условиях и что может с этими материалами делать.

Содержание политики редко бывает таким простым, как в приведенных примерах. Например, журнал D-Lib Magazine придерживается политики открытого доступа, но авторы материалов являются владельцами авторских прав. Политика предусматривает свободное ознакомление и копирование для частного использования. Однако, некоторые виды последующего использования (в т.ч. создание производных работ или продажа копий с целью извлечения прибыли) требуют разрешения от владельца прав.

Поскольку политика управления доступом может быть сложной, требуется формальный метод для обмена этими данными между различными компьютерами. Пожалуй, наиболее полная работа по этой проблеме была выполнена сотрудником компании Xerox Corp. Mark Stefik. Он разработал Цифровой Язык Прав Собственности (Digital Property Rights Language), который позволяет описывать права, условия и плату за использование электронных работ. Цель языка - сформулировать, обозначить атрибуты материалов, политику доступа, включая последующее использование. С его помощью менеджер коллекции может указывать порядок и условия копирования, переноса, обработки, распечатки и других операций. Язык позволяет назначать цену за каждую операцию и предусматривает связь с электронной программой контроля оплаты. Используемая в языке нотация основана на Lisp, языке программирования, используемом для обработки естественного языка. Некоторые полагают, что более простые нотации для ЭБ будут использовать XML. Реальной проблемой является, однако, не нотация, а то, насколько эффективным будет использование этого языка в широкомасштабной практике.

Проведение политики управления доступом

Управление доступом - это не просто вопрос разработки определенной политики. Информационные менеджеры хотели бы добиться реального исполнения этой политики, а, значит, нуждаются в соответствующих методах и формах обеспечения этого.

Некоторые (но не все) политики могут обеспечиваться техническими средствами. Есть определенные технические меры, которые позволяют реализовать политику в отношении тех, кому разрешено изменение материалов или поиск в хранилище. Не существует технических мер против плагиата или нарушения конфиденциальности или мер, гарантирующих исключительно образовательную цель использования материалов. Эти политики, вполне

резонные и целесообразные, особенно трудно подкрепить адекватными техническими мерами; для этого требуются социальные и юридические инструменты.

Существует зависимость между жесткостью мер обеспечения политики и удобствами для пользователей. Технические методы могут раздражать. Лишь немногие отказываются ввести пароль в начале сессии, но никому не нравятся повторные запросы пароля или иные формы идентификации. Информационные менеджеры иногда идут на смягчение ограничительных мер в интересах пользователей. Это ведет к росту рынка даже несмотря на то, что часть доходов теряется из-за несанкционированного доступа. Издатели, которые проводят агрессивную политику ограничений, могут удовлетворять своих пользователей и даже получать суммарно больше прибыли чем другие издательства. Эта стратегия используется в настоящее время в отношении большей части программного обеспечения для компьютеров (см. Врезка 7.2). Данные издательств, например, High Wire Press начинают подтверждать, что сходные результаты получаются и в случае издания электронных журналов.

В случае, если технические меры достаточно мягкие, эффективными могут быть социальные и юридические методы. Среди социальных мер можно отметить повышение уровня образованности пользователей относительно целей такой политики, а также формирование мотивации и установок на ее соблюдение. Эти цели требуют проведения такой политики, которая была бы понятной и легко реализуемой. Пользователи должны быть информированы о ее содержании и понимать, какое поведение является желательным. Полезным инструментом является размещение "объявления о доступе" (access statement) в случае материалов с ограниченным доступом. Это текст, в котором разъясняется политика, например, "Копирование материалов с коммерческими целями запрещается." Другие нетехнические методы обеспечения политики доступа могут быть более жесткими. Если, например, сотрудник организации повторно нарушает лицензионное соглашение или нарушает правила, которые обязался соблюдать, издатель может отозвать лицензию. В крайнем случае одна юридическая акция, сопровождаемая адекватным освещением в средствах массовой информации может вынудить многих вести себя более ответственно.

Врезка 7.2

Политика управления доступом в отношении программного обеспечения

Опыт в области программного обеспечения для персональных компьютеров является примером того, что случается, когда пытаются жестко проводить политику, которая неприятна для пользователей.

Программное обеспечение обычно лицензируется на определенный компьютер. Плата за лицензию предусматривает использование программы только на одном компьютере, но программу легко скопировать. Производители теряют прибыли от нелегального копирования, в особенности, если это делается в широких масштабах.

На заре распространения персональных компьютеров производители программного обеспечения пытались контролировать такое копирование техническими средствами. Один подход заключался в том, чтобы распространять программы на носителях, которые нельзя было легко скопировать. Это называлось "защита от копирования". Каждый раз при запуске программы пользователь должен был вставить оригинальную дискету. Это была эффективная мера, но эффект был не таким, как ожидали производители. Это было неудобным для законных пользователей. Установка на жесткий диск была неудобной, а архивное копирование затруднено. Пользователи стали отказываться от такого программного обеспечения - в результате производители, которые проводили наиболее жесткую политику "защиты от копирования" уступили в конкуренции с поставщиками, которые этими методами не пользовались.

Одной из первых компаний, которая осознала, что можно использовать не только технические меры, была Microsoft. Она стремительно разбогатела на продаже программных продуктов, которые не были защищены от копирования техническими методами. Была проведена большая работа по обеспечению политики иными, нежели техническими, методами. Умелый маркетинг, включая поддержку пользователей, практику предоставления дешевых обновлений (upgrade), поощрил пользователей к приобретению лицензий. Использовались

социальные меры по образованию людей, а юридические - для преследования наиболее злостных нарушителей.

Нелицензионное копирование все еще наносит вред производителям, однако компании, которые концентрируют свои усилия на удовлетворении запросов своих клиентов, имеют возможность выживать и развиваться.

Управление доступом в репозиториях

В большинстве ЭБ введены отдельные политики на уровне репозитория и на уровне коллекций. Хотя они различаются в деталях, общая схема показана на рис. 7.1. ЭБ представляют собой распределенные компьютерные системы, в которых информация циркулирует от одного компьютера к другому. Доступ на уровне репозитория эффективно контролируется "на месте", когда же материал покидает его, мало какие методы контроля будут действенны.

Концепция "последующего использования" уже обсуждалась; как только информация поступает на компьютер пользователя, контроль со стороны менеджера ЭБ становится весьма затруднительным. В сетях существуют и дополнительные проблемы - это многочисленные копии на узлах сети (кэш, зеркала и другие серверы, недоступные для контроля).

В настоящее время большинство ЭБ находят удовлетворительным сочетание технических мер управления доступом на уровне репозитория и социальных и юридических мер для контроля "последующего использования". Обычно комплекса этих мер достаточно, однако некоторые издатели опасаются, что отсутствие контроля приведет к потере доходов. Таким образом, есть интерес к созданию технических мер контроля копирования и последующего использования, в частности после того, как материал покинул репозиторий. Эти методы можно разделить на две категории: "системы на доверии" (trusted systems) и "безопасные контейнеры" (secure containers).

Системы на доверии

Репозитории представляют собой системы на доверии. Менеджеры ЭБ обладают определенной уверенностью, что программно-аппаратный комплекс и административные процедуры обеспечивают адекватный уровень безопасности для хранения и предоставления доступа к ценной информации. Могут существовать и другие системы, связанные с репозиторием, с равным уровнем доверия. В пределах сети таких доверительных систем ЭБ могут использовать те же методы контроля, что и для отдельных репозиториях. Процедуры и атрибуты могут распространяться на все системы с уверенностью в их эффективности.

Использовать системы на доверии в сетях нелегко. Каждый элемент системы должен поддерживать высокий уровень безопасности - но таким же должен быть уровень безопасности при передаче данных между различными компьютерами. По этой причине системы на доверии обычно используются в весьма ограниченных пределах или в специализированных компьютерных системах. Если все компьютеры функционируют в составе одной "команды" или "команды" работают по одним и тем же жестким правилам, то административных проблем обычно не возникает. Примерами больших систем на доверии могут служить компьютерные сети, которые поддерживают банкоматы.

Но ничего такого нельзя сказать о персональных компьютерах пользователей и о том, как они управляются. Правильнее всего не доверять им. С учетом этого ранние попытки внедрения систем на доверии в ЭБ будут сведены к случаям использования специализированного оборудования ("умные карты" (smart cards) или принтеры безопасности (secure printers) или выделенным (dedicated) серверам, использующим жестко контролируемое программное обеспечение.

"Безопасные контейнеры"

Поскольку сеть безопасной не является и сети на доверии использовать затруднительно, некоторые группы разработали так называемые безопасные контейнеры для передачи информации в Интернет. Цифровые материалы доставляются пользователям в виде пакетов, которые содержат данные и метаданные о правах доступа. Часть или вся информация пакета зашифрована. Доступ к ней требует цифрового ключа, который может быть получен от электронной платежной системы или другой системы аутентификации. Преимущество такого подхода в том, что он предоставляет определенные возможности контроля за последующим использованием. Пакет данных может копироваться и передаваться третьим лицам, но содержание недоступно без ключа. Такая система описана во врезке 7.3.

Врезка 7.3 Cryptolopes

Система Cryptolopes фирмы IBM используется безопасные контейнеры для того, чтобы позволить пользователям покупать и продавать их содержание через Интернет. Рисунок дает представление о том, как структурируется информация в Cryptolopes.

Информация передается зашифрованным блоком, называемом "контейнер". Поставщик информации помещает ее в контейнер, который может быть открыт получателем только после выполнения определенных требований по управлению доступом, например, оплаты за использование информацией. Информационное содержание в контейнере никогда не отделяется от информации об управлении доступом и об оплате. Таким образом, в дальнейшем контейнер может быть передан третьей стороне, которая также должна внести оплату, если хочет его открыть. (Каждый пользователь должен иметь код для того, чтобы открыть контейнер.)

Кроме шифрованного содержания, контейнер Cryptolopes может содержать открытую текстовую библиографическую информацию, который дает пользователю описание содержимого. Она может включать реферат, источник, автора, время последнего обновления, размер, цену и условия продажи. Если пользователь решает открыть контейнер Cryptolopes, он получает цифровой ключ. Для того, чтобы увидеть свободную информацию, пользователь "кликает" на соответствующее поле - и содержание появляется на экране. Для доступа к платной информации пользователь должен принять условия соглашения, содержащиеся в открытой части контейнера.

Перечень материалов (bill of materials)	
Открытый текст	
Зашифрованные инструкции по проверке отпечатков пальцев и водяных знаков (Enrypted fingerprinting and watermarking instructions)	
Зашифрованный фрагмент документа	Запись-ключ
Зашифрованный фрагмент документа	Запись-ключ
Зашифрованный фрагмент документа	Запись-ключ
Правила и условия (Terms and Conditions)	
Защита целостности и подписи	
Содержимое контейнера Cryptolopes может быть динамическим, поскольку потенциально система может использовать JavaScripts, фрагменты на языке Java и другие "живые" элементы в составе контейнера. В целях стандартизации фирма IBM лицензировала упоминавшийся выше Digital Property Rights Language фирмы Xerox для изложения условий использования и оплаты содержимого.	

Безопасные контейнеры не имеют для пользователя никакой ценности до тех пор, пока у него нет криптографического ключа. Это требует широкого распространения систем безопасности и методов электронных платежей. До недавних пор распространение таких сервисов было весьма медленным, так что издатели имели очень маленький рынок доставки информации в безопасных контейнерах.

Безопасность цифровых библиотек

В оставшейся части этой главы рассмотрим некоторые основные методы обеспечения безопасности, которые используются в компьютерных системах, объединенных в сети. Это методы общего назначения, используемые далеко за пределами ЭБ. Однако последние нуждаются и в специальных методах в связи высокой степенью децентрализации сетей поставщиков и пользователей информации.

Безопасность начинается с системных администраторов, людей, которые устанавливают и поддерживают компьютерные системы и сети, которые их соединяют. Их честность не должна вызывать подозрений, поскольку у них есть права доступа ко всем частям системы. Хороший системный администратор будет организовывать сети и файловые системы таким образом, чтобы пользователи имели доступ только к конкретной информации. Он управляет выдачей паролей, устанавливает брандмауэры (firewalls) для изоляции отдельных сегментов сетей, пользуется диагностическими программами для выявления неисправностей. Он архивирует информацию таким образом, чтобы система могла быть восстановлена после таких инцидентов, как выход оборудования из строя, пожар или незаконное вторжение.

Интернет не является безопасным. Содержание передаваемых пакетов может быть легко прочитано. Это часто делается с законными целями, такими как устранение неисправностей, но причины могут быть и не столь честными. Главный вопрос в обеспечении безопасности состоит в том, как построить безопасное приложение в небезопасных сетях.

Поскольку Интернет небезопасен, безопасность в ЭБ начинается с отдельных компьютеров, которые составляют библиотеку и данных в них, причем исключительное внимание должно уделяться интерфейсам между компьютерами и локальными сетями. Для многих персональных компьютеров единственным методом обеспечения безопасности является физическое ограничение числа пользователей. На других компьютерах установлены различные виды программного обеспечения для обеспечения безопасности, обычно это просто пароль и идентификатор пользователя (логин). Когда компьютером пользуется много людей, необходим специальный контроль для того, чтобы определить кто может читать или записывать в каждый конкретный файл.

Следующий шаг в построении защиты заключается в контроле интерфейса между локальной сетью и Интернетом и создании определенных барьеров от нежелательного вторжения. Наилучший барьер - это отключение от сети вообще. Более продуктивный путь состоит в соединении локальной сети с Интернетом через специальный компьютер, называемый "брандмауэр" (firewall). Его задача - отслеживать каждый проходящий пакет и отказывать в доступе тем, который могут нарушить безопасность. Брандмауэры могут отказывать запросам извне на соединение с компьютерами в защищаемой сети, могут отвергать пакеты, отформатированные в соответствии с протоколами, не входящими в список допустимых протоколов. Хорошо управляемый брандмауэр может быть достаточно эффективным способом блокирования вторжений.

Университеты находятся в самом эпицентре компьютерных и сетевых технологий уже многие годы. Несмотря на разнообразие их пользователей, они добились успеха в обеспечении адекватной безопасности в сетях из тысяч компьютеров (в университетских кампусах, например). Конечно, инциденты с проявлением агрессивного, антисоциального, злобного поведения случаются в каждом кампусе, но крупные нарушения весьма редки.

При условии тщательного администрирования, подключенные к сети компьютеры могут быть достаточно безопасными. Однако существует множество путей, используя которые асоциальные личности могут попытаться нарушить безопасность. В университетах большинство проблем связано с теми, кто в нем работает или учится. Обычно это кража идентификатора и пароля. Более сложные методы вторжения основаны на сложном программном обеспечении. Каждая операционная система имеет встроенные методы обеспечения безопасности, но ошибки разработчиков и программистов могут порождать "прорехи". Некоторые из наиболее полезных для ЭБ программ, такие как веб-сервера и электронная почта, относятся к числу наиболее уязвимых. С учетом этого каждый, кто строит ЭБ, должен отдавать себе отчет в том, что полная безопасность не может быть гарантирована. При достаточном прилежании угроза невелика, хотя вероятность появления "прорехи" всегда

сохраняется. Менеджеры ЭБ должны иметь сбалансированный подход к проблемам обеспечения безопасности. Хотя абсолютная безопасность недостижима, "умеренная" безопасность вполне может быть достигнута при небольших расходах (при этом требуется ум и внимание).

Шифрование (Encryption)

Шифрование - это общее наименование группы различных методов, которые используются для хранения и передачи конфиденциальной информации, кодируя ее таким образом, что внешне она выглядит абсолютно бессмысленной и случайно - до тех пор, пока не выполнена процедура расшифровки. Если к шифрованной информации получают доступ посторонние, то никакого ущерба они не причинят. В ЭБ шифрование используется для передачи конфиденциальной информации через Интернет, при этом часть информации настолько закрытая, что и хранится в зашифрованном виде. Пароли - это типичный пример информации, которая должна всегда находиться в зашифрованном виде. В большинстве ЭБ пароли - это единственная информация, нуждающаяся в такой защите.

Базовая концепция шифрования показана на рис. 7.2. Данные, которые необходимо сохранить в секрете, X, поступают в процесс шифрования, который выполняет математическое преобразование и создает зашифрованный набор данных, Y. При этом зашифрованный набор имеет то же число битов, что и исходный. Выглядит он как случайный набор битов, но после обратного шифрования процесса расшифровки получается исходный набор X. Эти два процесса шифровки-расшифровки выполняются одной компьютерной программой или специализированным устройствами.

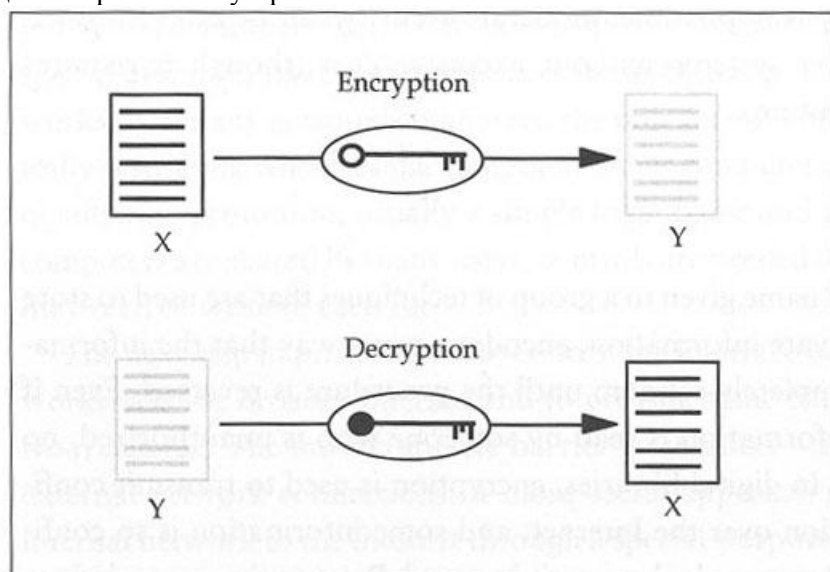


Figure 7.2

В обычных методах шифрования используют пару цифр, известных как "ключи". Один ключ используют для шифрования, другой - для расшифровки. Методы шифрования зависят от используемых процессов и способа выбора ключа. Математические описания процесса не являются секретом; основным секретом является ключ. Ключ - это набор битов, обычно от 40 до 120, иногда более. Длинные ключи существенно более надежны, нежели короткие, поскольку попытки вскрыть ключ занимают соответственно больше времени - добавление каждого следующего бита к ключу увеличивает время необходимое для расшифровки ключа примерно в два раза.

Исторически использование шифрования ограничивалось только мощностью компьютеров. Эти методы требуют значительных вычислений при шифровании и расшифровке. Ранние версии DES, метода, описанного на врезке 7.4, требовали установки специальных устройств к

каждому компьютеру. На сегодняшних быстрых компьютерах все не так серьезно, хотя время на шифровку-расшифровку больших объемов информации остается весьма значительным. Эти методы очень хороши для защиты коротких сообщений (вроде паролей) или исключительно секретной информации. Для защиты больших объемов информации они подходят меньше, особенно если имеет значение время на обработку.

Шифрование закрытыми (private) ключами

Это группа методов, в которых ключ, используемый для шифрования и расшифровывания сообщения - один и тот же. Другое название этих методов - шифрование одним ключом или шифрование секретным ключом. Метод DES относится к этой группе.

Понятно, что безопасность достигается лишь в том случае, если ключ хранится в тайне. При передаче данных с одного компьютера на другой необходимо найти абсолютно надежный способ передачи ключа. Поэтому этот метод используется в приложениях, внутри которых надежные сервисы обмениваются информацией.

Врезка 7.4

Стандарт шифрования данных (Data Encryption Standard, DES)

Метод DES, основанный на использовании конфиденциального ключа, изначально был разработан фирмой IBM, а в 1977 году стал стандартом США. Вычисления, необходимые в этом методе, весьма медленные, особенно при программной реализации, но этот метод вполне пригоден для многих целей. Современные компьютеры могут шифровать около миллиона байт в секунду.

Этот стандарт использует 56-битный ключ. Он делит набор данных на 64-битные блоки и шифрует каждый отдельно. Из 56-битного ключа генерируются 16 отдельных ключей меньшей длины. Суть алгоритма заключается в 16 последовательных успешных трансформациях 64-битного блока информации с использованием этих коротких ключей. При дешифровке используются те же 16 ключей для выполнения обратных операций. Выглядит этот алгоритм достаточно просто, но в нем есть свои тонкости. Наиболее важная в том, что последовательность битов после процесса шифрования выглядит как случайная последовательность, из которой невозможно сделать никаких предположений о данных или ключе шифрования.

Фанатики утверждают, что 56-битный ключ может быть вскрыт простым перебором, но это огромная задача. Для нужд ЭБ DES вполне адекватен.

Шифрование с двумя ключами

При использовании шифрования с закрытым ключом компьютер-отправитель и получатель оба должны знать ключ. Поэтому возникают проблемы безопасной передачи ключа. При шифровании с двумя ключами вся информация может передаваться по сети. По этой причине этот метод еще называют шифрованием с открытым ключом (public-key). Даже если перехвачено все сообщение целиком, зашифрованная информация остается закрытой.

Наиболее широко известным из этой группы методов является т.н. метод Rivest-Shamir-Adleman (RSA). Первый ключ является открытым, а второй держится в секрете. Пользователь А, желающий отправить секретное сообщение пользователю В, шифрует данные с использованием открытого ключа пользователя В. При получении данных пользователь В дешифрует их с использованием закрытого ключа, известного только ему.

Система с двумя ключами имеет множество преимуществ и один главный недостаток. Этот недостаток заключается в том, что необходимо быть уверенным, что используется действительно открытый ключ данного конкретного индивида. Обычный способ решения - передача задачи генерирования открытых ключей специальной организации, которую называют уполномоченным по сертификации (certification authority). Последний генерирует сертификаты, которые представляют собой специальные подписанные сообщения, в которых

содержатся имя индивида и его открытый ключ. Эта система работает удовлетворительно до тех пор, пока не нарушена секретность уполномоченного по сертификации.

Цифровая подпись

Как уже отмечалось, цифровая подпись может использоваться для удостоверения того, что компьютерный файл не изменялся. Она базируется на концепции математической хэш-функции, в результате приложения которой к последовательности байтов в файле генерируются число фиксированной длины.

Наиболее широко используется хэш-функция MD5, выполняющая специальное преобразование битов в файле и генерирующая в конце псевдослучайное 128-битное число. MD5 может использоваться для файлов любой длины. Если два файла отличаются хотя бы на один бит, значение числа MD5 будет абсолютно иным. И наоборот, если два файла имеют одно и то же характеристическое число, то вероятность того, что это разные файлы бесконечно мала. Таким образом, простейший способ установить, подвергался ли файл изменениям - это вычислить с помощью MD5 характеристическое число во время создания файла. Затем при необходимости это число вычисляется вновь и сравнивается с исходным. Если они совпадают, файл почти наверняка тот же самый.

MD5 имеет много преимуществ, в частности, высокая скорость вычислений для больших файлов; однако, как для любого секретного метода, всегда существует вероятность того, что некий вундеркинд откроет метод генерации файлов с одинаковыми характеристическими числами. Во время написания данной книги были некоторые признаки, что MD5 может быть обойден подобным способом. В любом случае, можно использовать и другие математические функции такого рода.

Характеристическое число не дает никакой информации о том, кто его вычислял. Цифровая подпись идет на шаг дальше в том, чтобы гарантировать аутентичность библиотечного объекта. После вычисления характеристического числа, оно шифруется с использованием открытого ключа собственника этого объекта. Таким образом, цифровая подпись создается с помощью уполномоченного по сертификации и открытого ключа. Перед тем, как проверить характеристическое число, пользователь дешифрует цифровую подпись с использованием открытого ключа. Если число совпадает с исходным, то объект не изменялся, а цифровая подпись была сгенерирована с использованием соответствующего открытого ключа.

У цифровой подписи есть один недостаток - несмотря на то, что пользователей интересует целостность объекта, они говорят не о битах, а о содержательной информации. Например, патентный офис США уделяет большое внимание интеллектуальному содержанию, вроде слов в тексте, но не заботится о том, присоединяет компьютерная система некоторую контрольную последовательность в конец файла или о том, изменялся ли шрифт, которым набирался текст. А цифровая подпись покажет на изменения файла и в том случае, если изменился всего лишь шрифт текста. Тем не менее, до сих пор никто не предложил более эффективного способа проверки аутентичности содержания, нежели битовый.

Распространенность шифрования с открытым ключом

Поскольку математическое обоснование метода было разработано свыше двадцати лет назад, можно было ожидать широкого использования этого метода уже в течение многих лет. Печально, но это не так.

Существуют большие технические проблемы, трудности управления ключами, их генерацией, как хранить закрытые ключи; какие предосторожности должны быть предприняты на случай, если секретность агентства, генерирующее ключи, будет нарушена. Однако главная причина задержки - политическая.

Во-первых, существует проблема патентования программного обеспечения (обсуждалась в гл.6). Шифрование с открытым ключом является одним из примеров компьютерной технологии, которую практически все специалисты признают изобретением. Методы шифрования - это не пустяки и их создатели хотели бы видеть свои разработки

запатентованными. К сожалению, держатели патентов и их агенты следуют по пути ужесточения лицензионной политики, что замедляет внедрение этих методов.

Более серьезное обстоятельство заключается в противодействии правительственных агентств США. ЦРУ утверждает, что методы шифрования представляют собой военную тайну и их распространение (экспорт) может подорвать безопасность США, Полиция (прежде всего, ФБР) заявляет, что общественная безопасность зависит от их способности перехватывать и читать сообщения в Интернете (после получения соответствующего разрешения на такие действия в конкретных случаях). Аргумент об экспорте трудно защищать, принимая во внимание то обстоятельство, что соответствующие работы были широко опубликованы, а серьезные японские и европейские фирмы предлагают свои продукты для шифрования. Аргумент общественной безопасности более сложен, но и он легко отражается тем обстоятельством, что американское общество не доверяет ни техническим способностям, ни административным процедурам ЦРУ и ФБР. Люди хотят иметь возможность передавать конфиденциальную информацию без того, чтобы за ними наблюдали.

Весьма характерно, что эта глава заканчивается разговором о том, что технические решения сдерживаются политическими аргументами. Это перекликается с вопросом, который проходит сквозь всю тему ЭБ и особенно актуален в проблеме управления доступом. Люди, технологии и административные процедуры связаны между собой теснейшим образом. Успех ЭБ должен базироваться на единстве всех этих аспектов, а не только на технологии.

Глава 8. Пользовательские интерфейсы и удобство пользования (usability)

Человек, пользующийся услугами библиотеки, обычно называется "клиент" или "читатель". В вычислительной технике обычными являются термины "пользователь" или "конечный пользователь". Какое слово не используй, пользы от ЭБ будет не много, если ими неудобно пользоваться.

Взаимоотношения человек-компьютер являются объектом интенсивных исследований таких дисциплин, как когнитивная психология, графический дизайн, риторика, математическое моделирование компьютерных систем. Некоторые работы имеют целью углубить теоретическое понимание того, как люди взаимодействуют с компьютерами с тем, чтобы использовать человеческие методы обработки информации для создания более совершенных компьютерных систем. Другие исследователи помогают пользователям осознать принципы построения компьютерных систем, стимулируя таким образом более продуктивное использование сервисов и информации, которые можно из этих систем получить. Эта глава не столь амбициозна. Она сосредоточена на широко используемых в настоящее время методах и предлагает некоторые примеры перспективных экспериментальных систем.

Изменения - одна из центральных тем в ЭБ, и изменения являются одной из проблем при создании пользовательских интерфейсов. Традиционные библиотеки совсем не просто использовать эффективно; однако, они изменяются медленно и за годы работы пользователи накапливают опыт. ЭБ развиваются столь быстро, что новые сервисы, новые коллекции, новые интерфейсы появляются каждый месяц. Как и новая головная боль... Пользователи не в восторге от перспективы непрерывно переучиваться - но изменения представляются совершенно необходимыми.

Частично вследствие высокой скорости развития ЭБ, заложенные в их рабочие модели оценки навыков пользователей сильно различаются. Большинство разработок ЭБ вышло из университетов, где много специалистов. Сотрудники и библиотекари всегда под рукой и могут помочь пользователям. Системные администраторы конфигурируют компьютеры, устанавливают программное обеспечение и отслеживают изменения на рынке. Но с распространением Интернета ЭБ начинают пользоваться люди, у которых нет такого опыта и которые не хотят тратить время на изучение вещей, которые им - возможно - больше никогда не понадобятся. Это вызывает напряжение: для специалистов требуются продвинутые сервисы, которые позволяют опытным пользователям работать быстрее и эффективнее. Но ЭБ должны быть доступны и людям с минимальной подготовкой.

Удобство пользования и создание пользовательских интерфейсов

Обсуждая проблемы удобства пользования компьютерными системами, легко сосредоточиться лишь на проблемах создания интерфейса между пользователем и компьютером. Но это свойство самой системы. Все компоненты ЭБ должны работать согласованно, делая библиотеку эффективной и удобной для клиентов, библиотекарей и системных администраторов.

На рисунке 8.1 показан один из подходов к обсуждению удобства пользования и созданию пользовательских интерфейсов. Любой пользовательский интерфейс компьютерной системы базируется на концептуальной модели, которая описывает способ использования данной системы. Ниже приведены некоторые концептуальные модели, которые используются в ЭБ. (На практике, большая часть ЭБ комбинирует концепции из нескольких подобных моделей.)

В классической модели библиотеки пользователь проводит поиск в каталогах или индексах, выбирает объекты и затем ищет их в книгохранилище.

В базовой веб-модели пользователь следует по ссылкам между файлами.

Протокол Z39.50 поддерживает концептуальную модель поиска в коллекции и сохранения результатов с возможностью их последующей обработки и анализа.

концептуальная модель	дизайн интерфейса
	функциональный дизайн
	данные и метаданные
	компьютерные системы и сети

Рисунок 8.1

На правой стороне рисунка приводятся уровни дизайна, которые являются составными частями любой модели. Дизайн интерфейса охватывает такие проблемы, как: что появится на экране и как пользователь будет этим манипулировать - шрифты, цвета, логотипы, меню, кнопки и работа с клавиатурой. Типичные функции включают выделение частей цифровых объектов, поиск в списках или сортировка результатов, обращение за помощью, манипулирование объектами, представленными на экране. Эти функции становятся возможными за счет данных и метаданных, которые поступают из ЭБ на основе компьютерных систем и сетей. Врезка 8.1 детализирует то, как эти аспекты удобства пользования воплощаются в приложениях.

Почти все современные компьютеры имеют интерфейс, базирующийся на стиле, разработанном фирмой Apple для своих компьютеров Макинтош, созданных на базе ранних исследований PARC фирмы Xerox. Они используют метафоры (иконки) файлов и папок на рабочем столе. Характеристики интерфейса включают перекрывающиеся окна, меню, указывающий прибор (типа мышь). Несмотря на многочисленные попытки улучшения, этот стиль интерфейса остается доминирующим на компьютерном рынке. Пользователи первых Макинтошей спустя 15 лет легко работают на последних версиях операционных систем Microsoft. Были внедрены новые соглашения, возможности аппаратной части сильно выросли - но базовый стиль интерфейса не изменился.

Концептуальная модель - рабочий стол. Файлы рассматриваются как документы, которые можно передвигать по рабочему столу, помещать в папки или хранить на диске. Каждый пользовательский интерфейс Windows использует эту концептуальную модель, хотя дизайн может быть разным (например, Apple использует мышь с 1 кнопкой, Microsoft с двумя, а Unix обычно с тремя). Функции, которые поддерживаются в рамках этой модели, включают открытие и закрытие файлов и папок, их выбор, передвижение с одного места на другое и т.д. Эти функции практически идентичны у разных производителей, хотя могут различаться в способах реализации этих функций. Концепция рабочего стола требуется, чтобы приложения ассоциировались с файлами данных. Системы Microsoft и Unix используют соглашения об именах файлов; например, имена файлов, которые предназначены для использования с программой просмотра PDF файлов должны иметь расширение pdf. Apple хранит такие метаданные в отдельных структурах, скрытых от пользователей.

Врезка 8.1

Перелистывание (Page Turning)

Пример введения опции "перелистывание" (переход на соседнюю страницу) иллюстрирует аспекты разработки пользовательского интерфейса, показанные на рис. 8.1. Такие проекты, как JSTOR, American Memory, предусматривают создание коллекций цифровых объектов, содержащих отсканированные изображения страниц из книг и других печатных материалов. Концептуальная модель в данных проектах предусматривает, что пользователь взаимодействует с объектами теми же способами, что и с книгой. Обычно страницы будут прочитываться подряд, но пользователь может вернуться назад или перейти на другую страницу по своему выбору. Некоторые страницы, например, содержание или индексы, могут идентифицироваться как специальные. Поскольку большинство персональных компьютеров снабжается дисплеями, которые меньше, чем размер печатной страницы и имеют меньшее

разрешение, концептуальная модель предусматривает возможности изменения размера изображения (zoom) и просмотра страницы по частям (panning).

Дизайн интерфейса определяет реальное "содержание" экрана, в т.ч. рамки, иконки, цвета и визуальные ключи (clues). Он предполагает также решения о том, какими индивидуальными функциями сможет воспользоваться пользователь. Когда мы создавали опцию "поворот страницы" в университете Карнеги-Меллона, была предусмотрена возможность максимизации площади экрана для того, чтобы иметь возможность увидеть образ страницы целиком. Большинство манипуляций контролировались с клавиатуры; клавиши-стрелки использовались для просмотра частей изображения, большего, чем размер экрана, а клавиша Tab - для перехода на следующую страницу. Альтернативный вариант дизайна может предусматривать использование виртуальных кнопок на экране - но при этом остается меньше места непосредственно для изображения страницы.

Функциональный дизайн должен предоставить возможности, эквивалентные реальному повороту (переворачиванию, перелистыванию) страниц в книге, включая возвращение на первую страницу, следующую, предыдущую и последнюю. Понятно, что это будут функции, связанные с реальным содержанием конкретной страницы, например, переход к странице с определенным номером (в печатном оригинале) или к оглавлению. Для обеспечения функций увеличения/уменьшения и просмотра по частям, необходимы дополнительные функции интерфейса, которые позволяют двигать площадь изображения по экрану и изменять увеличение.

Функции пользовательского интерфейса зависят от типа цифровых объектов в коллекции и, в особенности, от структурных метаданных. В типичном случае образы страниц хранятся в архивированных файлах, которые могут обрабатываться в любой последовательности. Для того, чтобы обеспечить изменение последовательности просмотра страниц, структурные метаданные должны идентифицировать первую страницу и последовательность последующих изображений страниц. Переход к странице с определенным номером (в печатном оригинале) требует структурных метаданных, связывающих нумерацию страниц с последовательностью образов страниц, поскольку весьма редко бывает, чтобы первый образ страницы в наборе файлов соответствовал странице книги с номером 1. Просмотр по частям и изменение масштаба требуют метаданных, которые отражают размеры каждой страницы.

Пользовательский интерфейс хорош только тогда, когда он удовлетворительно функционирует на данной компьютерной системе. Задержки при передаче образов страниц по сети могут раздражать пользователя. Один из возможных путей сократить это ожидание - выполнять (потенциальные) запросы пользователя по пересылке страниц из репозитория на пользовательский компьютер ранее чем сам читатель запросил их, так, чтобы по крайней мере, следующая страница уже находилась в памяти. Эта процедура известна как pre-fetching (предварительная доставка). В случае опции "поворот страницы" в проекте Карнеги-Меллона, приоритет был отдан достижению быстрого ответа - около секунды на чтение и отображение образа страницы, даже в случае передачи по загруженной сети кампуса. Это привело к разработке своеобразного "конвейера", когда первая часть страницы уже выводится на экран пользователя, а последняя часть еще не считана с жесткого диска в репозитории.

Браузеры

Создание программ-браузеров (прежде всего, "Mosaic", разработанной в 1993 г.) стимулировало разработки более совершенных интерфейсов для сетевых применений. Несмотря на то, что браузеры разрабатывались для веб, они так гибки, что их используют как интерфейсы практически для всех приложений в Интернете, включая ЭБ. Перед тем, как были созданы универсальные браузеры, разработчики были вынуждены разрабатывать отдельные пользовательские интерфейсы для каждой модели компьютера или операционной системы. Эти интерфейсы нуждались в коррекции при каждом изменении операционной системы, что является трудоемкой задачей, которая практически никогда не выполнялась с должным качеством. Принимая веб-браузеры в качестве универсального инструмента для

взаимодействия с пользователями, разработчики ЭБ могут сосредоточиться на задачах организации потоков информации, оставляя проблемы компьютерного "железа" и операционной системы браузеру.

Базовые функции

Основной функцией браузера является поиск и получение файла с удаленного веб-сервера и его представление на пользовательском компьютере.

- Для нахождения файла на веб-сервере браузер нуждается в URL. Последний может вводиться пользователем или содержаться в гиперссылках на HTML странице или храниться в "закладках".
- Из URL-строки браузер извлекает тип протокола. Если это HTTP, то браузер извлекает из URL доменное имя компьютера, на котором хранится файл. Далее браузер посылает HTTP-запрос, ожидает ответа и затем закрывает соединение.
- Если все происходит нормально, ответ содержит файл и MIME-тип. Для представления файла на пользовательском компьютере браузер определяет MIME-тип и вызывает соответствующие процедуры. Эти процедуры могут быть встроенными или реализовываться внешними программами, вызываемыми браузером.

Простой веб-браузер может выполнять лишь поддержку HTTP-протокола и процедуры по представлению страниц в HTML-формате. Поскольку HTTP и HTML весьма просты, веб-браузер может быть весьма несложной программой.

Распространение браузеров за рамки веб

Браузеры разрабатывались для веб и каждый из них поддерживает базовые протоколы и некоторый набор стандартных форматов. Однако браузеры могут использоваться для предоставления других сервисов без изменения интерфейса. Большой частью своего успеха веб обязан браузерам, можно даже сказать, что стремительное расширение Интернета имеет ту же причину.

"Мозаика" и его потомки поддерживали три типа соответствий: для данных, для протоколов и для исполнения программ.

Данные

Для каждого типа данных браузер ассоциирует способ обработки файлов этого типа. В современные браузеры встроена поддержка нескольких типов, включая текст, HTML, GIF, при этом пользователи могут расширять этот список, используя внешние приложения и плагины (plug-ins). Внешние приложения представляют собой отдельные программы, которые вызываются для обработки этого типа данных. Файл передается в это приложение как данные. Например, браузеры не имеют встроенной поддержки для файлов PostScript, но на большинстве компьютеров имеются программы просмотра файлов в этом формате, которые и используются как внешние приложения. Когда браузер получает файл в формате PostScript, он запускает соответствующую программу и передает ему файл, который необходимо отобразить на экране. Концепция плагинов сходна с внешними приложениями за исключением того, что они не являются самостоятельными отдельными программами. Они используются для представления файлов в нестандартных форматах в составе HTML-страниц.

Протоколы

Основной протокол веб - HTTP, но браузеры поддерживают и другие протоколы. Некоторые из них, такие как Gopher и WAIS, были важны исторически, поскольку предоставляли браузерам доступ к более старым информационным службам. Другие, такие как NetNews,

электронная почта и FTP остаются актуальными и сейчас. Слабым местом большинства браузеров является то, что список поддерживаемых протоколов ограничен и не может быть расширен. В результате, не существует простого способа добавить протокол Z39.50 (или иные) в браузер.

Исполнение программ

Сообщение в формате HTTP может делать больше, чем просто поиск статического файла с информацией на сервере. Оно может запускать на сервере программы и возвращать их результаты на браузер. Наиболее ранний метод выполнения этих задач заключался в использовании интерфейса CGI, который предоставляет браузеру простой путь для выполнения программ на удаленном компьютере. Программы CGI обычно называют CGI-скриптами (scripts). CGI представляет собой механизм, используя который большинство поисковых программ веб пересылают запросы от браузера к поисковой системе. Издатели хранят свои коллекции в базах данных и используют CGI-скрипты для предоставления доступа пользователям. Неформальная интерпретация следующего URL:

`http://www.dlib.org/cgi-bin/seek?author='Arms'`

следующая: "На компьютере с доменным именем `www.dlib.org` выполнить программу `seek` из директории `cgi-bin`, передав в качестве параметра имя автора `author='Arms'` и вернуть результат". Программа будет выполнять поиск в базе данных записей со словом `Arms` в поле "автор".

Одним из ранних применений CGI было связывание браузеров со старыми базами данных и иной информацией. По странному стечению обстоятельств, сейчас, когда веб стал доминирующей системой, эта роль сохранилась. Разработчики новых ЭБ зачастую используют CGI как метод связывания старых (веб) систем со своими новыми разработками.

Мобильный код (Mobile Code)

Разработка и продажа браузеров быстро превращается в большой бизнес и внесение новых возможностей в эти программы происходит непрерывно. Некоторые из них определенно удобны для пользователей; другие служат скорее для целей маркетинга. Улучшения включают повышение производительности, развитие HTML, встроенную поддержку новых форматов (относительно недавно был добавлен формат JPEG), и упрощение способов добавления новых форматов. Два изменения можно рассматривать, как более существенные - это улучшение функций безопасности и введение мобильного кода (который позволяет серверам посылать на клиентские машины компьютерные программы, которые исполняются браузером).

Мобильный код позволяет дизайнерам веб-сайтов создавать страницы, которые включают программы. Врезка 8.2 описывает один подобный подход, который использует апплеты (applets) - маленькие программы, написанные на языке программирования Java. Апплеты могут быть скопированы с веб-сайта и исполняться на машине-клиенте. Поскольку Java является полномасштабным языком программирования, он может быть использован для выполнения сложных операций, вроде пересылки идентификационных форм с веб-сайта на браузер с тем, чтобы пользователь мог напечатать логин и пароль, который затем апплетом шифруется и безопасно передается обратно на сервер.

Врезка 8.2

Java

Java - это универсальный язык программирования, который был специально разработан для создания распределенных систем, в частности, пользовательских интерфейсов, в сетевом окружении.

Традиционный подход к созданию пользовательского программного обеспечения, которое может исполняться на разных типах компьютеров (например, веб-браузеры и почтовые

клиенты), заключался в создании различных версий. Даже если версии писались на стандартном языке, вроде С, различия между операционными системами Windows, Macintosh и Unix заставляли разработчика создавать несколько версий пользовательского интерфейса для конкретного типа компьютеров. В лучшем случае, такой подход дорогостоящ и трудозатратен; в худшем новые программы делают невозможным исполнение других программ или привносят вирусы.

Компьютерные программы, написанные на языке высокого уровня, известны как "исходный код" (source code). Обычно написанная программа затем компилируется в машинный язык конкретного компьютера. Компилятор Java иной. Вместо того, чтобы генерировать машинный код для определенной компьютерной системы, он переводит исходный текст программы в промежуточный код, называемый Java-байт кодом (bytecode Java), который предназначен для программного окружения, называемого Виртуальной Машиной Java (Java Virtual Machine). Для исполнения кода на конкретном компьютере требуется следующий шаг, который каждую команду в байтовом коде транслирует в машинный код данной машины. Современные браузеры поддерживают Виртуальные Машины Java и имеют встроенные Java-интерпретаторы.

Апплеты Java являются короткими программами, которые скомпилированы в файл байт-кодов Java и могут передаваться браузерам по сети, обычно с помощью протокола HTTP. Браузер распознает файл как апплет и передает данные интерпретатору Java для выполнения.

Поскольку Java представляет собой универсальный язык программирования, почти все компьютерные процедуры теперь могут включаться в веб-приложения. Java также предоставляет программистам набор инструментов, которые могут включаться в программы, в т.ч.:

- а) базовые элементы программ, такие как строки, числа, методы ввода-вывода, структуры данных;
- б) соглашения о построении апплетов;
- в) сетевые сервисы, такие как URL, TCP-сокеты и IP-адреса;
- г) справочные системы для создания программ, которые могут использовать иные языки, нежели английский;
- д) средства обеспечения безопасности, включая цифровые подписи, управление открытыми и закрытыми ключами, управление доступом и сертификаты;
- е) компоненты программного обеспечения, называемые JavaBeans, которые могут встраиваться в другую архитектуру программного обеспечения;
- ж) средства подключения (connections) к базам данных.

Поскольку базовые функции встроены в браузер, они не должны передаваться через Интернет с каждым апплетом; более того, они могут быть написаны на машинном коде данного компьютера и исполняться быстрее, чем при интерпретации байт-кода.

Java потенциально исключительно полезна для ЭБ, но и она несовершенна. Некоторые ее дефекты носят обычный технический характер: это большой, сложный и трудный для изучения язык программирования. Интерпретируемые языки всегда исполняются медленно, и Java не исключение. Отсутствие решений, которые исключали бы передачу в составе апплетов вирусов и заражение компьютеров пользователей, также затрудняет внедрение этого языка. Однако крупнейший недостаток Java - не технический. Его разработчик, фирма Sun, пыталась разработать стандартный язык. К несчастью, другие фирмы, например, Microsoft, создали свои, несовместимые варианты.

Значение Java не только в том, что он является источником мобильного кода. Для HTML страниц существует альтернатива включения скриптов с инструкциями, написанными на языке JavaScript. Последний проще, чем Java, а скрипты исполняются быстрее. Типичное применение JavaScript - проверка данных, вводимых пользователем в реальном масштабе времени, без пересылки на сервер. (Java и JavaScript представляют собой совершенно разные языки программирования. Сходство имен - это чисто маркетинговый ход. Java более широко известна, однако JavaScript имеет ряд преимуществ; при этом оба они широко используются.)

Недавние разработки в создании пользовательских интерфейсов

Создание пользовательских интерфейсов для ЭБ - частично искусство, частично наука. Успех интерфейса совершенно определенно зависит от инстинктов, интуиции дизайнеров, от их понимания пользователей. Новые концепции в репертуаре дизайнеров появляются непрерывно. В этой секции дан краткий обзор некоторых новых идей и тем исследований. Глава 12 описывает исследования в области структурных метаданных. Эта тема исключительно важная для пользовательских интерфейсов, поскольку способ моделирования цифровых объектов и ассоциированные с ними структурные метаданные представляют собой "сырье" с которым оперирует (взаимодействует, акт) пользовательский интерфейс.

Концептуальные модели

Несколько исследовательских групп разработало концептуальные модели, призванные помогать пользователям работать в огромных коллекциях, которые ныне доступны он-лайн. В Интернете существуют определенные ориентиры, карты, признаки. Гиперссылки, которые являют собой сердце веб, могут вести в неожиданные места и пользователь может легко заблудиться. Пользователи ЭБ зачастую имеют слабую подготовку и работают поодиночке. Это указывает на необходимость создания интерфейсов, основанных на концептуальных моделях, которые будут вести пользователей по четко очерченным путям - хотя часть пользователей замечательно легко адаптируется к задаче поиска информации. Наблюдения показывают, что опытные пользователи способны удовлетворить большую часть своих потребностей в информации из сети, однако менее опытные пользователи зачастую теряются и испытывают трудности при работе с найденной информацией. Врезка 8.3 описывает два исследовательских проекта, в которых изучались новые концептуальные модели для ЭБ.

Врезка 8.3

Новые концептуальные модели: DLITE и Pad++

Были выполнены интересные эксперименты по изучению новых концепций пользовательских интерфейсов, отличных от традиционного "рабочего стола". Поскольку это были исследовательские проекты, интерфейсы эти вряд ли будут использованы в реальных продуктах; однако, они весьма важны для развития концепций разработки следующего поколения интерфейсов и для иллюстрации процесса дизайна, который базируется на систематическом анализе пользовательских потребностей и ожиданий.

DLITE - это экспериментальный пользовательский интерфейс, разработанный Steve Cousins в Стэнфордском университете в рамках инициативы "Цифровые библиотеки". Он создавался в качестве пользовательского интерфейса системы Stanford InfoBus и использует концепции объектно-ориентированного программирования. Каждый объект рассматривается как CORBA-объект. (InfoBus и CORBA описаны в главе 13).

Концептуальная модель DLITE, основанная на результатах анализа задач, которые выполняет пользователь ЭБ, позволила сформулировать следующие требования:

ЭБ состоят из гетерогенных (неоднородных) коллекций, которые должны быть доступны из Интернета.

Результаты одного сервиса могут быть входными данными для другого.

Интерфейс должен быть "открытым", чтобы новые ресурсы могли легко интегрироваться с существующими.

Ресурсы должны сохраняться длительное время.

Пользователи должны иметь возможность взаимодействовать.

Модель DLITE описывает ЭБ в терминах компонентов четырех основных типов - документов, запросов, коллекций и сервисов. Эти компоненты представлены иконками, которыми можно манипулировать непосредственно из пользовательского окна на экране. Например, перетаскивание запроса на поисковый сервис инициирует процедуру поиска и приводит к созданию коллекции результатов. DLITE позволяет создавать пользователям

интерфейсы, ориентированные на конкретные задачи, за счет агрегирования экранных элементов интерфейса.

Функциональный дизайн DLITE определили два обстоятельства: быстрый отклик и возможность добавлять новые сервисы с минимальными усилиями. С учетом этих требований, функциональный дизайн пользовательских интерфейсов подразделяется на клиентов и серверы.

Клиенты выполняют манипуляции с компонентами, причем может одновременно выполняться несколько таких процессов. Сервер предоставляет интерфейс для внешних сервисов и может функционировать даже в отсутствие клиентских процессов.

Возможность гибкого расширения создается за счет существования сервера. Когда добавляется новый сервис, на сервере должен быть запрограммирован новый интерфейс; при этом клиенты не нуждаются в модификации. Поддержка различных компьютерных и операционных систем обеспечивается за счет создания различных клиентов.

Pad++ - это концепция пользовательского интерфейса, предложенная Ken Perlin из университета Нью-Йорка, была развита исследователями из нескольких университетов и исследовательских центров. Фундаментальная концепция заключается в том, чтобы просматривать большие коллекции информации в разных масштабах. Pad++ рассматривает метафору "рабочего стола" значительно шире, чем просто компьютерный дисплей, скорее это коллекция документов, распределенная на огромной стене. Действия пользователя основываются на знакомых концепциях изменения масштаба и просмотра по частям. Пользователь может уменьшить масштаб и увидеть всю коллекцию целиком (с минимумом деталей), увеличить некоторую часть для знакомства с разделами или еще "добавить увеличения", чтобы рассмотреть детали. Такой пространственный подход требует значительных исследований закономерностей человеческого восприятия. Поскольку люди имеют обычно хорошую пространственную память, Pad++ рассматривает форму и положение как ключевые элементы для того, чтобы помочь пользователям анализировать и легко позднее вызывать (находить) найденную ранее информацию.

Когда в Pad++ используется минимальный масштаб, детали отражаются с использованием некоего порогового значения. Иначе программа вынуждена была бы оперировать с огромным количеством крошечных деталей. При достижении некоего масштаба, маленькие детали заменяются (сливаются) в одну новую или не показываются вовсе. В ходе процесса, называемого "семантическим масштабированием" (semantic zooming) изменяется вид объектов в соответствии с их текущим размером. Этот подход знаком из картографии, когда на картах маленького масштаба изображается город "вообще", а не отдельные здания.

Pad++ не рассматривается как универсальный интерфейс для всех приложений. В некоторых случаях - например, для знакомства и работы с большими иерархически организованными коллекциями ЭБ, - он может служить в качестве основного; для других целей он может использоваться лишь для визуализации и совместно с другими концепциями пользовательских интерфейсов в составе обычных "оконных" систем.

Pad++ представляет собой интересный пример исследований в области пользовательских интерфейсов, в частности, изучения взаимосвязи между производительностью и удобством использования. Было создано несколько версий концепции Pad++ - как отдельное приложения, так и компоненты веб-браузеров. Каждая содержит внутреннее средство контроля производительности. Операции вывода (rendering, построение изображения на экране дисплея) нормируются таким образом, чтобы частота обновления кадра оставалась постоянной и при просмотре по частям и при увеличении/уменьшении. Когда интерфейс начинает функционировать медленно, средние детали воспроизводятся с меньшим разрешением; если система простаивает, глубина прорисовки деталей возрастает.

Дизайн интерфейса

Частично дизайн интерфейса представляет собой искусство, но в ходе недавних исследований были сформулированы общие принципы. Согласованность "картинки" (consistency in appearance), методов контроля, методов функционирования являются важными для пользователей. Им нужна обратная связь; им необходимо понимать, что делает

компьютерная система и почему они видят определенные результаты. Они должны иметь возможность прервать операцию или отменить ее выполнение (с возвращением к исходной ситуации, undo). Обработка ошибок должна быть простой и доступной для понимания. Опытным пользователям должны предоставляться shortcuts, тогда как для начинающих должны присутствовать простые, определенные опции. Помимо всего сказанного, пользователи должны обладать чувством контроля за происходящим.

Контроль - особенно над графическим дизайном и представлением материалов - создает постоянные трения между разработчиками ЭБ и пользователями. Многие дизайнеры хотели бы, чтобы пользователи видели материалы именно так, как они это запрограммировали. Они хотят контролировать качество графики, шрифтов, размеры окон, расположение информации в окне и все остальное, что важно для хорошего дизайна. К сожалению для дизайнеров, браузеры от них не зависят, они не могут знать, какой именно браузер у данного клиента, какой компьютер, насколько хорошая связь, какой дисплей - большой или маленький. Пользователь может реконфигурировать свой компьютер, может предпочесть более крупный шрифт или окно меньшего размера, может перейти в текстовый режим (если связь слишком медленная). Таким образом, хороший дизайн должен быть эффективен в различном компьютерном окружении. Лучшие дизайнеры обладают особой сноровкой в построении интерфейсов, удобных для пользователей и привлекательно выглядящих на различных компьютерах. Однако другие разработчики испытывают трудности при переходе от традиционных носителей (где они могут контролировать ситуацию полностью) к ЭБ и веб. Общая ошибка - перегруженность, усложненность, в результате чего интерфейс практически неработоспособен без быстрых сетей и мощного компьютера.

Примером необоснованной усложненности являются браузеры Netscape и Microsoft, двух лидирующих на этом рынке фирм, поставляющих продукты, которые сознательным образом делаются не полностью совместимыми. В результате, если дизайнер хочет использовать специальную опцию или особенность, пользователь должен специально предупреждаться о том, что эта опция может некорректно работать с некоторыми браузерами. Пользовательский интерфейс, который прекрасно работает под одним браузером, может стать сущим мучением под другим. Инстинкт конкуренции у производителей взял верх над соображениями удобства пользователей.

Функциональный дизайн

ЭБ - это распределенные системы, в составе которых работает много компьютеров. Исследования в области функционального дизайна направлены на выбор между тем, какие функции будут выполняться на определенных компьютерах и как последние будут связаны между собой. Между репозиторием (где находятся коллекции) и пользователем (использующим, как правило, веб-браузер) находится целый слой компьютерных программ-интерфейсов, которые иногда называют программным обеспечением промежуточного слоя (middleware). Оно функционирует как посредник между пользователем и репозиторием, интерпретирует инструкции первого и сообщает их последнему; получает информацию из репозитория, организует ее и доставляет на пользовательский компьютер. Зачастую это программное обеспечение поддерживает дополнительные сервисы, такие как аутентификация и контроль ошибок, но наиболее важной ее задачей является согласование сервисов, затребованных пользователем с возможностями репозитория.

Рисунки 8.2 и 8.3 демонстрируют две типовые конфигурации. В каждом случае пользователь за персональным компьютером использует стандартный веб-браузер в качестве оконечного интерфейса для доступа к ЭБ, а именно, к репозиторию и поисковой системе.

На рисунке 8.2 программы-интерфейсы представлены в виде отдельных программ (обычно это CGI-скрипты), которые исполняются на сервере. Браузер посылает сообщение интерфейсу используя стандартный протокол (допустим, HTTP). Скрипты могут исполняться как удаленный сервис или могут взаимодействовать с данным сервисом с помощью обычных протоколов. В некоторых случаях для связи между скриптами интерфейса и поисковой системой необходимо использование протокола Z39.50. Подробности технической реализации важны для достижения

гибкости и производительности системы, однако с функциональной точки зрения они эквивалентны. Эта конфигурация имеет тот недостаток, что каждое действие пользователя должно передаваться по сети для обработки на удаленном компьютере; пользователь при этом вынужден ждать ответа.

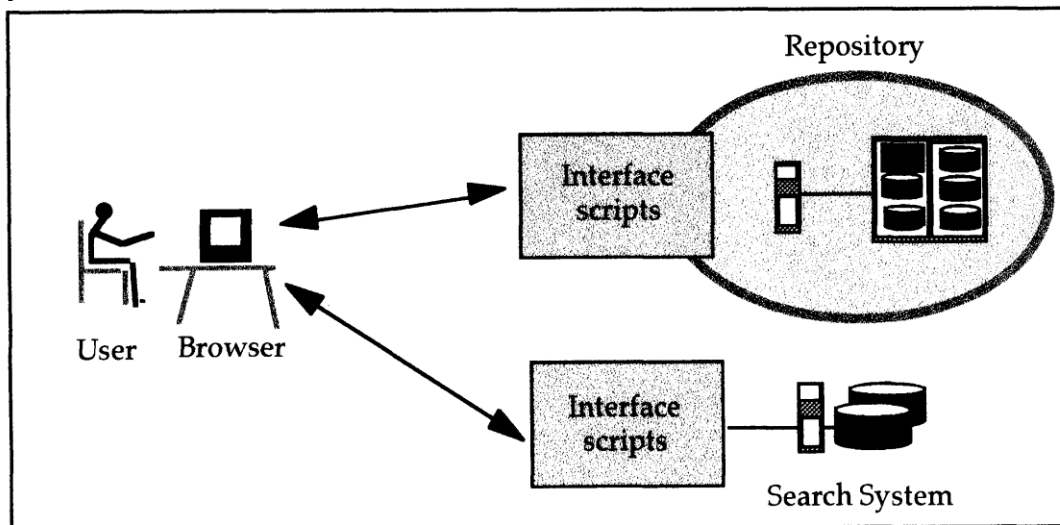


Figure 8.2
User interface with interface scripts.

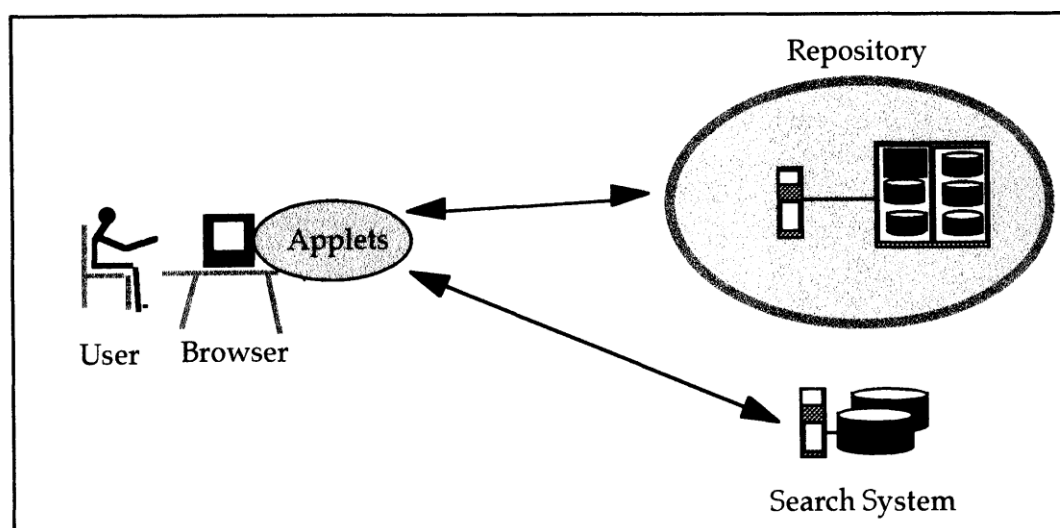


Figure 8.3
User interface with mobile code (applets).

Рисунок 8.3 иллюстрирует более современную конфигурацию, в которой браузер на пользовательском компьютере исполняет мобильный код. Этим кодом могут быть апплеты Java. По умолчанию апплеты хранятся на удаленном сервере; они загружаются на пользовательский компьютер по запросу автоматически. Эта конфигурация имеет много преимуществ. Поскольку код интерфейса выполняется на пользовательском компьютере, он может быть очень гибким. Протоколом между пользовательским интерфейсом и сервисом в этом случае может быть любой протокол, как стандартный, так и разработанный специально для этого приложения.

Недавно была выдвинута интересная концепция под названием *"профиль презентации"* (*presentation profile*). Менеджеры ЭБ устанавливают определенные правила (*guidelines*), задающие порядок представления объекта для пользователя. Например, профиль может

рекомендовать два возможных способа воспроизведения объекта, отдельно для маленького размера и отдельно для большого. Интерфейс пользователя может (как правило) следовать этому профилю, но может использовать и другие подходы.

Компьютерные системы и сети

Производительность компьютерных систем и сетей имеет существенное значение для удобства пользования. Разработчик ЭБ может сделать определенные допущения о том, каким оборудованием располагает пользователь (естественно, по умолчанию предполагается, что в принципе имеется компьютер, подключенный к Интернету). Однако даже это простое допущение предполагает множество различных ситуаций. Некоторые ПК мощнее других, качество дисплеев также сильно различается. Некоторые люди имеют личные компьютеры, которые они могут конфигурировать по своему желанию; другие вынуждены работать на общем компьютере или на разных. Любая ЭБ должна предполагать, что пользователи работают на разном оборудовании с использованием разных операционных систем и окружения (environments), включая разные версии Windows, Macintosh и Unix. Различия между операционными системами могут быть минимизированы за счет использования в качестве пользовательского интерфейса веб-браузера, но различия в производительности устранить не так просто.

Удобство пользования компьютерной системой зависит от того, как быстро она откликается на команды. Разработчик ЭБ имеет очень мало (или практически никакой) информации о качестве сетевого соединения пользователя с ЭБ, которое может изменяться от исключительно быстрого до ужасно медленного, причем даже в лучшем варианте иногда все равно возникают задержки. Интерфейс, который замечательно работает в сети на 10 Мбит/с, может быть бесполезен при работе через модем, когда связь редко лучше 28 Кбит/с. Таким образом, ЭБ должны сбалансировать эффективность использования расширенных возможностей (которые требуют быстрого оборудования и современного программного обеспечения) с предоставлением услуг для менее оснащенных пользователей.

Увеличение скорости отклика (responsiveness) браузеров

В веб-браузерах предусмотрен ряд мер для улучшения скорости отклика (как это видно с точки зрения пользователя). Один из них - это внутреннее кэширование. Информация, которую использовали один раз, возможно, будет необходима еще (к примеру, когда вы кликаете на кнопку "back" (назад) веб-браузера); часть графики (логотипы, кнопки и пр.) часто повторяется. Веб-браузеры сохраняют на диске компьютера пользователя недавно полученные файлы. Это могут быть файлы HTML, графики или мобильного кода (например, JavaScript). Когда пользователь запрашивает файл, браузер сначала проверяет кэш, прежде чем обращаться к удаленному серверу.

Другие методы повышения скорости отклика предусматривают параллельное выполнение многих операций. Браузер показывает первую часть длинного файла еще до того, как тот получен полностью; изображения сначала демонстрируются с низким разрешением, а детали прорисовываются по мере поступления данных. Несколько отдельных потоков данных могут быть запрошены параллельно. В итоге это во многом позволяет улучшить производительность Интернета.

Создание зеркал (mirroring) и кэширование

Дублирование данных широко используется для повышения производительности систем в Интернете.

Если предполагается, что коллекциями ЭБ будут пользоваться люди по всему миру, то можно предусмотреть создание копий коллекций на отдельных сайтах - к примеру, два в

Европе, два в США, один в Австралии и один в Японии. Это называют созданием зеркал. В общем виде, пользователь обращается к ближайшему сайту, который, вероятнее всего, обладает наибольшей производительностью для данного пользователя. Этот подход так же позволяет пользователю обратиться к дублеру, если один сайт вышел из строя или недоступен определенный сегмент сети.

Дублирование определенной информации (в отличие от традиционных библиотек) относится и к процедуре кэширования (caching). Кэш - это место, где сохраняется недавно запрошенная информация для того, чтобы избежать задержки при повторном обращении. Изначально кэш создавали аппаратно за счет использования специальных микросхем с высокой скоростью доступа в контроллерах жестких дисков в компьютерах. Теперь такие решения используются, например, в системе доменных имен для увеличения скорости и надежности преобразования доменных имен в IP-адреса.

ЭБ также используют кэш в различных случаях. Например, можно организовать несколько областей для хранения недавно использованных документов. ЭБ, которые хранят большие коллекции на дешевых, но медленных дисковых массивах могут создавать кэши на более быстрых дисках.

Все методы дублирования данных подвержены той опасности, что версии документов могут отличаться и некоторые пользователи в результате могут получить некорректные материалы. Это требует разработки процедур, предусматривающих удаление дублированной информации по истечении установленного времени или проверку источника на предмет изменений.

Кроме того, кэш уязвим с точки зрения обеспечения безопасности. Компьютерная система безопасна настолько, насколько безопасно ее самое уязвимое звено. Кэш легко может стать таким звеном.

Пользовательские интерфейсы и надежность

Лишь редкие компьютерные системы полностью, абсолютно надежны. ЭБ зависят от большого числа подсистем, расположенных в Интернете. С точки зрения числа независимых компонентов вообще удивительно, что любая всемирная система вообще работает.

Хорошо сконструированная компьютерная система предоставляет пользователю информацию о ходе выполнения задач. Одной из простейших форм такой обратной связи является анимация, которая продолжается во время ожидания ответа; это - по крайней мере - говорит пользователю, что что-то происходит. Более продвинутые методы предусматривают отражение части процесса, которая уже выполнена и оценку необходимого времени для завершения. В любом случае, пользователь должен иметь возможность прервать затянувшуюся операцию.

Термин "изящная деградация" (graceful degradation) описывает методы, которые в случае задач, занимающих много времени, пытаются предоставлять пользователю хотя бы частичное удовлетворение. Простейший вариант - это перейти в текстовый режим браузера. Некоторые методы доставки графики позволяют выводить на дисплей "черновик", все детали прорисовываются по мере поступления данных позднее.

Пользовательские интерфейсы для мультимедийной информации.

Состав материалов в ЭБ становится все более разнообразным. Пользовательские интерфейсы, разработанные для текстовых коллекций, могут быть неадекватными для коллекций музыки, карт, программного обеспечения, графики, статистических данных или видеоигр. Однако, в любом случае применяются следующие общие принципы:

Резюмирование (Summarization). Когда пользователь должен выбрать из нескольких объектов, правильным является предоставление о каждом из них (его содержании) информации общего характера. Для книги такая информация включает имя автора и название. Для графики обычно достаточно иконок (размером с ноготь большого пальца, отсюда термин thumbnail). Для более

редких видов информации, вроде компьютерных программ или фильмов, такую резюмирующую информацию в рамках ЭБ придумать трудно.

Сортировка и категоризация. Библиотеки организуют информацию классифицируя или упорядочивая, однако для многих видов материалов (например, музыкальные нотации) метод оптимальной организации неочевиден.

Презентация. Многие виды информации требуют специального оборудования для наилучшего представления пользователю. Высококачественное цифровое видео, к примеру, требует специальных дисплеев, мощных компьютеров, высокоскоростной связи и соответствующего программного обеспечения. ЭБ должны иметь все эти возможности, если они доступны. Для пользователей с ограниченными возможностями должны предусматриваться иные способы презентации.

Не существует единого метода создания ЭБ для различных типов информации. Пользовательские интерфейсы также должны различаться в зависимости от классов материалов и - возможно - от категории пользователей. Врезка 8.4 представляет характерный пример пользовательского интерфейса: ЭБ оцифрованных видеофрагментов Informedia.

Врезка 8.4 **Informedia**

Informedia - это программа университета Карнеги-Меллона под руководством Howard Wactlar, которая посвящена исследованию ЭБ видеоматериалов - конкретно, новостных и документальных программ телекомпаний. Видеоматериалы автоматически разбиваются на короткие фрагменты (вроде отдельных сюжетов в программе новостей). Суть исследования - в разработке автоматических методов извлечения информации из видеообъектов для минимизации участия человека. ЭБ содержит более 1000 часов цифровых видеоматериалов (полученных от CNN, открытого британского университета (British Open University), публичной телестанции WQED (г. Питтсбург) и других источников).

Один из основных вопросов в связи с аудио и видеоматериалами - это удобный и быстрый обзор содержимого коллекций. Читатель книги может обратиться к содержанию или просто быстро пролистать ее - но "пролистать" видео невозможно. В рамках проекта Informedia был разработан метод под названием video skimming для автоматического извлечения важных слов и изображений, которые совокупно и представляют собой "видеореферат", дающий представление о содержании всего видеоматериала. Исследователи добились больших успехов на этом пути.

Пользовательский интерфейс в рамках Informedia базируется на концептуальной модели поиска по индексам, получения набора результатов и последующий их просмотр. Запросы могут вводиться как с клавиатуры, так и голосом (с использованием программ распознавания речи). После поиска интерфейс представляет ранжированные результаты. Каждый видеоклип представляется неким изображением. При наведении на него курсора появляется текст пояснения. Изображение автоматически выбирается из содержимого видеоматериала, а пояснение создается программой обработки естественного языка (natural-language processing). Кликнув на изображение, пользователь может увидеть весь видеофрагмент. Интерфейс включает в себя программу просмотра видео с соответствующими инструментами управления (вроде клавиш "play" и "stop"), имитирующую домашний видеоплеер. Он достаточно интуитивен для того, чтобы использоваться без предварительной подготовки. Предусмотрена максимальная обратная связь для пользователей. Большая часть испытаний выполнялась с участием студентов г. Питтсбурга.

Пользовательские интерфейсы и эффективность цифровых библиотек

За последние несколько лет качество интерфейсов значительно улучшилось. В настоящее время предполагается, что пользователь может начать продуктивную работу без специальной подготовки. Важнее другое - в Интернете имеются многочисленные примеры отличных

интерфейсов, которые могут использоваться в качестве образцов. Стандарты графического дизайна также становятся все выше с каждым годом. (Оформление онлайн-журнала D-Lib Magazine за многие годы подвергалось редизайну трижды. Каждая реализация была элегантной и современной - однако уже через год требовалась "подтяжка лица".)

Эффективная поддержка пользователей - это больше, чем простые косметические ухищрения. Элегантный дизайн, оптимальная функциональность, гибкость - все это создает ощутимую разницу в эффективности использования ЭБ. Если системой трудно пользоваться, пользователи могут просто не найти важные данные, неправильно интерпретировать результаты поиска или просто опустить руки. ЭБ хороша настолько, насколько хорош ее интерфейс.

Глава 9. Текстовые материалы

Текстовые материалы занимают особое место во всех библиотеках, включая ЭБ. Хотя иногда действительно лучше один раз увидеть, чем прочитать тысячу слов, зачастую наилучший способ передачи важных и сложных идей - это слова. Богатство концепций, детали, точность передачи идей с помощью текста просто замечательные.

Текстовые документы в ЭБ могут поступать из разных источников. Некоторые могли быть созданы для онлайн-использования, некоторые получены конверсией печатных материалов, некоторые могут быть оцифрованными звуковыми дорожками к фильмам и телепрограммам. Текстовые записи выполняют также специальную функцию метаданных, которые описывают другие материалы.

Эта глава рассматривает каким образом текстовые документы представляются для хранения, распечатки и вывода на дисплей. Записи метаданных и методы поиска текстовых документов описаны в следующей главе.

Разметка, описание страницы и таблицы стилей (Markup, page description and style sheets)

Методы хранения текстовых материалов должны представлять два разных аспекта документа: его структуру и его реальный вид.

Структура текста описывается такими элементами, как знаки, слова, параграфы, заголовки. Она идентифицирует части документов, которые выделены, материалы в таблицах, сносках и все, что связывает разные части между собой. Структура текста в компьютерах обычно представлена в виде спецификации разметки (markup specification). За последние годы наиболее популярным языком структурной разметки стал SGML (Standard Generalized Markup Language).

Реальный вид документа - это то, как он выглядит на экране или будучи напечатанным на бумаге. Он тесно связан с выбором формата: шрифт и его кегль, поля, межстрочные интервалы, выделение заголовков, расположение рисунков и представление математической и иных специализированных нотаций. В печатной книге имеет значение также выбор бумаги и тип переплета. Для того, чтобы хранить и представлять документы именно в том виде, как они должны выглядеть, используются языки описания страниц. В этой главе описаны три различных подхода к описанию страниц: TeX, PostScript и PDF.

Структура и внешний вид связаны с дизайном документа. В обычном книгопечатании дизайнер создает набор спецификаций, которые описывают, как должны выглядеть все элементы отдельно с исчерпывающими правилами для разрешения всех возможных ситуаций. Эти спецификации позволяют верстальщику создать профессионально отформатированный документ на основе рукописи с правкой редактора и(или) дизайнера.

Рисунок 9.1 иллюстрирует процедуру создания электронных журналов в некоторых издательствах. Полученные от авторов статьи размечаются тэгами SGML, которые описывают структуру и материал статьи. Так называемые таблицы стилей (style sheet) специфицируют то, как должен выглядеть каждый элемент. Разметка SGML и таблица стилей являются входными данными для соответствующего программного обеспечения, которое создает форматированный документ.

На заре ЭБ часто обсуждался вопрос - заменят ли они печатные книги. Вначале дискуссия концентрировалась на проблеме удобочитаемости. При каких обстоятельствах люди предпочтут читать с экрана, а не с бумаги? По мере накопления опыта было достигнуто осознание того, что компьютеры и книги не эквивалентны. Каждый имеет свои уникальные

достоинства. Вычислительная техника позволяет легко выполнять поиск без затрат ручного труда. Зато в печатных изданиях вне конкуренции "человеческий фактор" - книги имеют небольшие размеры, их можно читать где угодно, раскладывать на столе, носить в руках, не требуется специального оборудования для чтения (кроме очков).

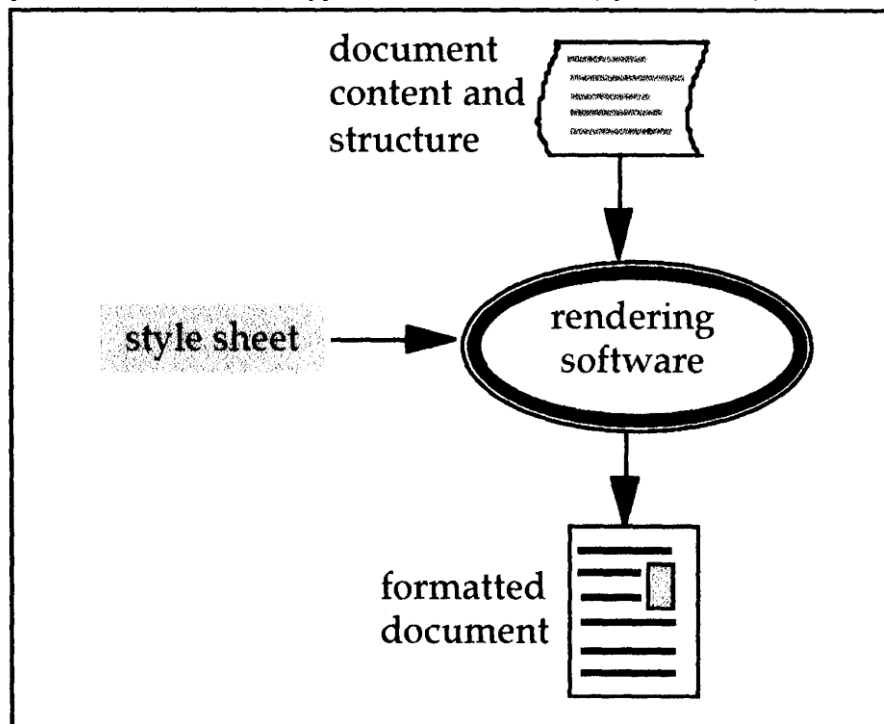


Figure 9.1

The relationship between structure and appearance.

Поскольку цифровые тексты и печатные материалы служат для различных целей, некоторые издатели создают как печатную, так и онлайн-версию одного и того же материала. Рисунок 9.2 показывает, как язык разметки может использоваться для "управления" текстом, который будет и напечатан и выведен на экран компьютера. С помощью отдельных таблиц стилей единый документ со структурной разметкой может быть представлен различными способами для различных целей. Он может быть показан на экране или напечатан. Разметка и графический дизайн могут различаться, но они базируются на едином источнике и представляют одно и то же содержание.

Жестко контролировать внешний вид документа с помощью разметки SGML и таблиц стилей достаточно сложно. Способы представления, организации и интерпретации текстуальных материалов имеют много нюансов. Языки разметки могут выразить почти все структуры; однако разнообразие структурных элементов в документах может быть огромным, при этом различаться могут детали внешнего вида, выбранные как автором, так и дизайнером. Например, многие журнальные статьи публикуемые в изданиях Association for Computing Machinery содержат математическую нотацию. Авторы обычно предоставляют статьи в формате издательской системы TeX, которая обеспечивает адекватную верстку математических формул. В ходе издательского процесса этот формат конвертируется в разметку SGML. При создании изображения с помощью таблиц стилей выдача уже может по форме отличаться от исходного авторского оригинала.

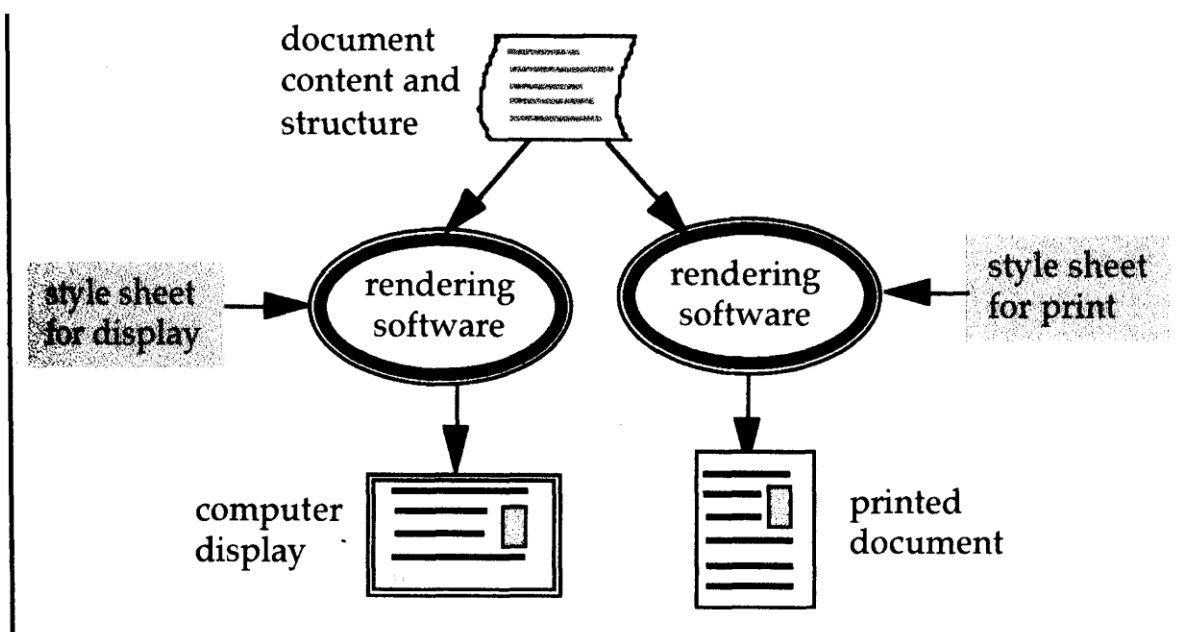


Figure 9.2
Alternative renderings of a single document.

Конвертирование текста

Сегодня большинство документов создается в цифровой форме с помощью компьютеров, однако библиотеки полны документов, которые ныне существуют только на бумаге. Соответственно, возникает проблема конвертирования печатных документов в компьютерные форматы. Проекты конвертации важных документов учитывают как собственно их выявление, так и идентификацию структуры представления материала и создание образов.

Базовая технология для конвертирования - это сканирование. Документ переводится в набор точек, каждая из которых имеет свой код яркости. В простейшем виде различаются только черные и белые точки. При разрешении 300 точек на дюйм (по горизонтали и вертикали) хорошее изображение может быть получено с большинства печатных страниц. Если повысить разрешение до 600 точек на дюйм, или если кодировать 8 уровней серого, четкость изображения становится превосходной и можно адекватно воспроизводить полутоновые иллюстрации. Высококачественная художественная работа требует по меньшей мере 24 бита на точку для адекватного воспроизведения цветовой палитры. При этом создаются гигантские по размеру файлы. Они обычно архивируются для удобства хранения и обработки, но даже простой черно-белый текстовый файл требует не менее 50 000 байт для одной страницы.

Сканированные страницы воспроизводят внешний вид печатных страниц, но представляют текст как картинку. Для многих приложений это не может заменить размеченный или даже простой ASCII текст. В частности, невозможно проводить поиск конкретных слов. В серьезных исследованиях иногда необходимо работать с оригиналом и точно знать, как он выглядел. Для иных целей предпочтительнее электронная версия, которая отражает структуру. Например размеченный текст более удобен, чем изображение оригинала для задач текстологического анализа, поскольку текст может быть размечен таким образом, чтобы выделять его лингвистическую структуру или исторических предшественников. Таким образом, следующим этапом конверсии является создание электронного текста на базе сканированного изображения.

Оптическое распознавание символов (Optical character recognition, OCR) - это технология превращения отсканированных изображений знаков в электронные знаки. Она основана на отделении отдельных знаков и последующем сравнении с математической моделью этого знака. Несмотря на десятилетия исследований, распознавание остается вероятностным

процессом. Количество ошибок зависит от четкости оригинала. Если исходный печатный текст чистый и разборчивый, число ошибок не превышает 1%. Для низкокачественных материалов эта величина может быть существенно выше. Для многих целей величина в 1% слишком велика, поскольку это означает наличие достаточно большого числа неправильных знаков на каждой странице. Для устранения ошибок было предложено много подходов. Один из них - параллельное использование нескольких программ распознавания с последующим сравнением результатов. Другой подход состоит в использовании словаря для проверки полученных текстов. Тем не менее, все случаи, когда требуется высококачественная конверсия, требуют вмешательства профессионального корректора-человека. В ряде систем компьютерная программа выделяет подчеркиванием непонятные слова, предоставляя человеку право выбора точного варианта. Примером организации, в которой хорошо отлажен процесс конвертации, является UMI, в которой каждый год обрабатывается огромное количество диссертаций. Поскольку они обычно представлены чистыми разборчивыми оригиналами, уровень ошибок в архивах UMI очень низок, учитывая тот факт, что после распознавания материалы просматривают корректоры.

Врезка 9.1

Оксфордский словарь английского языка

Второе издание Oxford English Dictionary, опубликованное в 1989 году, представляет собой прекрасный пример использования языка разметки для описания структуры и содержания таким образом, что один и тот же материал может быть использован в качестве исходного для целого ряда продуктов.

Первое издание словаря было предпринято James Morris с коллегами свыше сорока лет назад. Фотография XIX века показывает его за работой в Оксфорде в Scriptorium в окружении горы бумаг. Словарь явился вехой в лексикографии, но существовал только в статичной печатной форме. Было доказано на практике, что и вручную его можно поддерживать в актуальном состоянии.

Первая стадия создания нового издания состояла в переводе текста оригинального издания в базу данных с указанием всех типографских различий, нагруженных семантической информацией. Заглавные буквы, шрифты жирного и курсивного начертания, скобки, более мелкие типы (smaller type) и другие приемы форматирования отражают тонкости, которые более нигде не зафиксированы. Была написана сложная компьютерная программа для извлечения этой скрытой семантической информации и разметки текстовых элементов тэгами SGML. В работе принимали участие специалисты по лексикографии (lexicographers) из Оксфорда, компьютерные лингвисты из университета Waterloo (Онтарио, Канада) и IBM.

Новый словарь существует в виде компьютерной базы данных, в которой разметка SGML так же важна, как и слова. Лексика обновляется непрерывно. С минимальными усилиями были выпущены различные издания словаря, включая компакт-диск, другие цифровые версии и печатные книги.

После распознавания отдельных слов следующая стадия конверсии заключается в уточнении структуры документа и разметке заголовков и остальных структурных элементов. Несмотря на активные исследования, эта стадия также требует участия корректоров и редакторов.

Альтернативным методом конверсии, который широко применяется на практике, является перепечатка документов с последующей ручной разметкой. Он зачастую дешевле. Поскольку эта технология весьма трудозатратна, она обычно используется в странах с дешевой рабочей силой. Один из крупнейших проектов конверсии - это American Memory Библиотеки Конгресса. Документы для конвертации отбираются из исторических коллекций, при этом они зачастую менее разборчивы, чем новые печатные материалы. Конверсия выполняется подрядчиками, которые гарантируют заданный уровень точности, но свободны в выборе любого метода конверсии. Все подрядчики пришли к выводу, что дешевле перепечатывать материалы.

Кодировки символов

ASCII

Базовыми элементами текста являются знаки, например А или 5. Важно проводить различие между концепцией знака, как структурного элемента и различными представлениями одного знака при хранении в компьютерах или изображении для чтения. Знак - это абстрактный концепт, независимый от метода кодировки при хранении или выводе на экран компьютера.

Компьютеры хранят знаки как последовательность битов, каждый отдельный знак кодируется различной последовательностью. Первые компьютеры имели коды для 26 английских букв (иногда только для прописных), 10 цифр и нескольких знаков пунктуации и специальных знаков. Внутреннее представление знаков при хранении текстов в компьютерах до сих пор базируется на этом ограниченном наборе символов. Большая часть современных компьютеров основывается на кодировке ASCII (American Standard Code for Information Interchange).

Первоначально каждый знак в ASCII кодировался семью битами. Эта семибитная кодировка известна как стандартный ASCII. Например, буква А кодируется семибитной последовательностью 1000001. В двоичной записи это число 65, поэтому было принято соглашение, что в стандартном коде ASCII число 65 представляет букву А. Из 7 бит можно составить 128 различных комбинаций - и стандартный ASCII ассоциирует отдельные знаки с числами от 0 до 127. При этом числа от 0 до 31 зарезервированы для служебных знаков (например, "перевод каретки"). В таблице 9.1 приведены ASCII-коды от 32 до 127, называемые печатным набором знаков ASCII. (Пробел рассматривается как печатный знак.)

Набор печатных знаков ASCII превратился в фактический стандарт. Этот код используется в огромном количестве компьютеров и приложений. Если приоритетом является интероперабельность, - 96 печатных знаков ASCII используются всегда. Они же являются допустимыми знаками в HTML и многих системах электронной почты. Почти все клавиатуры, дисплеи и программное обеспечение интерпретируют эти коды одинаково. Существует расширенная версия ASCII с использованием 8 бит, она предоставляет дополнительные коды 128-255 для отображения других знаков.

Таблица 9.1
Печатный знаки семибитного ASCII

32	[space]	52	4	72	H	92	\	112	p
33	!	53	5	73	I	93	]	113	q
34	"	54	6	74	J	94	^	114	r
35	#	55	7	75	K	95	-	115	s
36	\$	56	8	76	L	95	`	116	t
37	%	57	9	77	M	97	a	117	u
38	&	58	:	78	N	98	b	118	v
39	'	59	;	79	O	99	c	119	w
40	(	60	<	80	P	100	d	120	x
41	)	61	=	81	Q	101	e	121	y
42	*	62	>	82	R	102	f	122	z
43	+	63	?	83	S	103	g	123	{
44	,	64	@	84	T	104	h	124	
45	-	65	A	85	U	105	i	125	}
46	.	66	B	86	V	106	j	126	~
47	/	67	C	87	W	107	k	127	□
48	0	68	D	88	X	108	l		
49	1	69	E	89	Y	109	m		
50	2	70	F	90	Z	110	n		
51	3	71	G	91	[	111	o		

Unicode

В текстовых материалах используется значительно больше знаков, чем в стандартном печатном ASCII, который базируется на английском языке. Некоторые европейские языки имеют дополнительные буквы или используют диакритические знаки. Даже старинный английский требует дополнительных знаков. Греческий и русский языки имеют различные алфавиты. Японский, китайский и корейский используют не буквы, а иероглифы Han, которые обозначают целые слова. На самом деле, знаков ASCII не всегда достаточно даже для современного английского. Некоторые дисциплины (математика, музыка, химия) используют жестко формализованные нотации, которые требуют большого количества знаков. Полнота и понятность текстов зачастую зависит от точного использования принятой нотации.

Продукция компьютерной индустрии продается по всему миру, и производители вынуждены признавать необходимость поддержки национальных алфавитов и иных знаков. В этой области лидирует Microsoft. Поскольку все алфавиты невозможно представить с помощью 256 знаков 8-битного кода, были предприняты попытки представлять больший набор знаков с помощью большего числа бит. Недавно один такой подход был принят за стандарт, который обязали соблюдать большинство производителей компьютеров и программного обеспечения. Он называется Unicode.

Первоначально в Unicode каждый знак представляется 16 битами, что достаточно для 65,536 знаков. Тем не менее, несмотря на огромные усилия целого ряда признанных специалистов, комбинации знаков, используемые в некоторых языках, не могут быть представлены с помощью Unicode.

Врезка 9.2

Наборы знаков (scripts), представленные в Unicode

Версия 2.0 стандартного Unicode содержит 16-битные коды 38,885 отдельных знаков. Кодировка организована не по языкам, а по скриптам (scripts). Если несколько языков используют близкие наборы знаков, набор символов, достаточный для этой группы идентифицируется как один набор. К примеру, латинский набор содержит все знаки, используемые в английском, французском, испанском, немецком и близких языках. Unicode 2.0 также поддерживает арабский, армянский, бенгали, кириллицу, грузинский, греческий, Вормофо, Devanagari, Gujarati, Gurmukhi, Han, Hangul, Hebrew, Hiragana, Kannada, Katakana, Lao, Malayalam, Oriya, Phonetic, Tamil, Telugu, Thai и Tibetan наборы. Дополнительно к этим исходным наборам есть группа коллекций символов, включая цифры, диакритические знаки, знаки пунктуации, общие символы, математические, технические символы, стрелки, элементы псевдографики и др. В скором времени появится поддержка некоторых современных языков (например, Ethiopic, Sinhala), а в будущем возможна поддержка древних алфавитов (например, Aramaic, Etruscan, runes).

Особенно важным аспектом Unicode является поддержка иероглифов Han, которые используются в китайском, японском и корейском языках. Unicode совместим с Unihan, который был создан в результате выполнения более раннего проекта и использовался для представления знаков этих языков.

Успех Unicode определяется не просто усилиями ученых-лингвистов. Разработчики тщательно изучили совместимость Unicode с существующим программным обеспечением. Если бы пришлось вносить изменения во все программы, Unicode никогда не был бы принят. Существует специальное представление знаков Unicode, называемое UTF-8, которое делает возможной постепенную трансформацию приложений, базирующихся на ASCII для использования всего набора скриптов Unicode.

UTF-8 - это кодировка, которая использует от одного до шести байт для представления каждого знака Unicode. Наиболее широко используемые знаки представлены одним байтом,

менее широко - двумя и так далее до шести байт. Важнейшая особенность заключается в том, что каждый печатный знак ASCII представляется одним байтом, который совпадает с кодом соответствующего знака ASCII. Таким образом, одна и та же последовательность байт может быть интерпретирована и как знаки Unicode (в представлении UTF-8) и как набор печатных знаков ASCII. Например, страница HTML-текста, созданная с использованием знаков ASCII не потребует исправления для представления с использованием программного обеспечения, ожидающего данных в кодировке UTF-8.

Транслитерация (transliteration)

Unicode - это не единственный метод, которые используются для представления расширенного набора знаков в компьютерах. Другой подход - это транслитерация - конвертация знаков одного алфавита в знаки другого. Например, немецкое "о с двумя точками сверху" иногда транслитерируются как "oe". Фонетическая система pinyin часто используется для представления китайского языка (в особенности, разновидности Mandarin) в знаках английского алфавита. Транслитерация была полезной тогда, когда использовались печатные машинки с физически закрепленными наборами знаков. С появлением компьютеров необходимость в транслитерации отпала и скоро станет седой историей.

Библиотеки использовали широкий набор алфавитов задолго до специальных исследований по скриптам. В результате существуют огромные каталоги в формате MARC (и других форматах), содержащих кодировки в pinyin и других системах, помимо Unicode. Сейчас стоит устрашающая задача конвертации или обеспечения совместимости.

SGML

Языки разметки использовались уже на заре вычислительной техники для описания структуры текста и формата его представления. Сегодня наиболее широко распространены языки разметки семейства SGML.

SGML - это не один язык разметки, это система, базирующаяся на единой спецификации разметки. Отдельная спецификация, заданная в рамках SGML называется "определением типа документа" (document type definition, DTD). Многие издательства и агентства разрабатывают свои собственные DTD. Два примера наиболее важных для учебных библиотек DTD - это Text Encoding Initiative и Encoded Archival Description, описанные на врезке 9.3.

Врезка 9.3

DTD для обучения

Text Encoding Initiative представляет собой раннюю и очень тщательную попытку представления существующих текстов в цифровой форме, с упором на нужды гуманитарных дисциплин. Тексты прошлого создавались на бумаге, пергаменте, папирусе, камне вручную, печатно и пр. Документы содержат аннотации, изъятия и поврежденные участки, которые не могут быть прочитаны. Встречаются знаки, которые практически не используются.

SGML оказался эффективным для описания таких документов, но разработка подходящих DTD является трудной задачей. Технически несложно было бы создать одну, но сложную DTD, которой, однако, никто бы не пользовался. Решение было найдено в создании целого семейства DTD с использованием одинаковых компонентов. Базовый набор тэгов определяет элементы, которые вероятнее всего понадобятся во всех документах и, следовательно, должен входить во все DTD. Каждый DTD содержит основной набор тэгов, которые выбираются из наборов, разработанных для прозы, драмы, поэзии, словарей и файлов данных. Обычно для данного документа достаточно одного основного набора. Кроме того, доступны дополнительные наборы тэгов для специальных целей. (Автор называет это моделью чикагской пиццы: каждая пицца содержит сыр и томатный соус, основу из теста и дополнительные ингредиенты.)

Термин "вспомогательные поисковые данные" (finding aid) относится к большому количеству списков, описаний изобретений, индексов и других текстовых документов, созданных архивами, библиотеками и музеями для описания своих коллекций. Некоторые такие данные содержат более подробную информацию, чем обычные каталожные записи; другие менее подробны и могут не включать детальную информацию о каждом предмете коллекции. Некоторые весьма коротки, другие занимают сотни страниц.

Encoded Archival Description (кодированное описание архивов, EAD) представляет собой DTD, используемое для кодировки электронных версий документов в архивах. Первая версия EAD была разработана командой из университета в Беркли (Калифорния) и базировалась на опыте Text Encoding Initiative. Результаты отражали соглашения и практику, сложившиеся в архивном деле, а также жестко структурированную структуру вспомогательных данных. Большая часть информации была представлена иерархическими взаимосвязями; однако существует много других особенностей, которые необходимо четко представлять, разрабатывая кодировку для вспомогательных поисковых данных для использования в ЭБ.

EAD был одобрен многими архивистами, которые совместно приложили большие усилия для его совершенствования, сравнения с существующими вспомогательными поисковыми данными и подробного его документирования. Он стал высокоспециализированным средством, созданным в ответ на потребности этого профессионального сообщества в обмене и обеспечении доступа к информации.

DTD создается на основе общих концепций сущностей (entities) и элементов. Он определяет, какие сущности и элементы допускаются для определенных классов документов и устанавливает базовую кодировку.

Сущности специфицируются идентификаторами, начинающимися со знака амперсанда и оканчивающимися точкой с запятой, например:

&alph;

&logo;

В типичном DTD допустимые сущности включают большую часть знаков ASCII и некоторые другие (называемые знаковыми сущностями); однако любой символ или группа символов может быть определена как "сущность". Имя сущности - это просто имя. В стандартном наборе знаков, α представляет собой сущность, используемую для кодировки первой буквы греческого алфавита; однако DTD может использовать этот код для различных целей. Разработанные различными издательствами DTD определяют до 4000 различных сущностей, представляющих все специальные символы и варианты, используемые в научной литературе.

Сущности представляют собой цепочку символов, которые могут группироваться в элементы. DTD могут определять любой набор знаков как имя элемента. Элементы размечаются двумя тэгами в угловых скобках, причем прямой слэш используется для индикации окончания тэга. Например, в проекте Text Encoding Initiative тэги и обрамляют текст, удаленный из рукописи. Например, чтобы разметить слова "men and women" как удаленные из текста, нужно записать:

men and women

Примеры элементов включают различные типы заголовков, сносок, выражений и ссылок. Элементы могут выстраиваться в иерархии. Каждый DTD имеет свою грамматику, которая определяет допустимые связи, как набор правил, выполняемых компьютерной программой.

Упрощенные версии

SGML утвердился как гибкий подход для записи и сохранения высококачественных текстов. Его гибкость позволяет создателям текстовых материалов разрабатывать DTD для их собственных целей. Врезки 9.1 и 9.3 описывают некоторые примеры. Издатели научных журналов разрабатывают свои DTD, которые они используют для разметки журнальных статей в процессе создания и для хранения в компьютерных системах. Проекты ЭБ, такие как JSTOR и American Memory, используют DTD, созданные в процессе работы над Text Encoding Initiative.

Недостатком гибкости SGML является сложность программного обеспечения, необходимого для его исполнения. Даже программу реализации простого DTD трудно написать; а задача создания универсальной программы, способной генерировать любой DTD, объединять информацию из различных таблиц стилей и представлять информацию в том виде, как она будет выведена на экран или на печать - это задача практически невыполнимая. К тому же, рынок для подобного программного обеспечения весьма мал. Лишь одна компания проявила настойчивость и создала универсальный продукт для разметки SGML. Но и этот продукт несовершенен; он не полностью реализует возможности SGML, исполняется лишь на некоторых типах компьютеров и использует свои собственные таблицы стилей. Таким образом, полный SGML непригоден для использования в ЭБ, которые ориентируются на достижение интероперабельности.

HTML

Веб стимулировал разработку упрощенных версий SGML. HTML, язык разметки, используемый для веб, может рассматриваться как неортодоксальный DTD. С многих точек зрения, однако, он развивается на основе философии SGML, поскольку смешивает структурную информацию и форматы. Сначала HTML предназначался для структурной разметки, оставляя браузерам самим определять форматирование текста при выводе на дисплей. Однако последующие версии содержат дополнительные опции, которые предоставляют разработчикам веб-страниц контроль над способами предъявления материалов на дисплее или принтере. Врезка 9.4 иллюстрирует как далеко ушел HTML от чистого языка структурной разметки.

HTML продолжает развиваться. В зависимости от точки зрения, добавления могут рассматриваться как попытки сделать HTML более мощным или как попытки сделать добавление сложности и неудобства. Никто не станет оспаривать ценности некоторых новых функций, таких как возможность включения графических изображений тэгом `<img>` (который впервые появился в "Mosaic"). Ценность других дополнений - дело вкуса. Простые команды форматирования, вроде `<center>` (появившаяся в Netscape), не слишком утомительны, однако другие нововведения существенно усложнили ситуацию. Особенно это относится к таблицам и фреймам. В результате автор уже не может освоить HTML за утро, а программист не напишет HTML-редактор за неделю.

Трения между структурной разметкой и контролем форматирования в HTML становятся все серьезней. Многие разработчики веб-страниц, добиваясь их совершенства, используют все возможности HTML. В работе дизайнеров наметилась одна неприятная тенденция - использовать структурные элементы для управления размещением; многие веб-страницы созданы как одна гигантская таблица. Опытные дизайнеры создают привлекательные страницы, используя такой прием. Менее опытные навязывают пользователям страницы, которыми неудобно пользоваться. Наихудшее случается тогда, когда дизайнер закладывает макет страницы, которая не входит в размер дисплея при данном разрешении или использует установки, противоречащие пользовательским (назначая, к примеру, слишком крупный кегль шрифта).

Почти все новые возможности в HTML были предложены разработчиками новых браузеров с целью улучшить свой продукт и сохранить конкурентоспособность. Некоторые из них действительно полезны, другие излишни. Консорциум W3C и IETF попытались как-то структурировать эту работу и сформулировать некие рабочие стандарты, но реальный контроль над HTML принадлежит двум главным разработчикам браузеров - фирмам Netscape и Microsoft.

Врезка 9.4

Особенности HTML

Большинство тэгов HTML - парные. Например, `<h1>Text</h1>` означает, что текст между тэгами является заголовком первого уровня. Некоторые тэги единичны, например `<hr>`,

обозначающий горизонтальную линию. Элементы могут "вкладываться" друг в друга, например, список может содержать много параграфов, а таблица содержать другие таблицы.

Примеры, приведенные ниже, иллюстрируют возможности HTML и то, как HTML комбинирует структурную разметку с форматированием и поддержкой онлайн-приложений. (Таких возможностей намного больше, а некоторые из приведенных имеют много опций.)

Тэги, используемые для описания структурных элементов документа, включают:

<body> (тело документа)

<p> параграф

<h1>, <h2>, ... <h6> заголовки - шесть стандартных уровней

 выделение (обычно курсивом)

, , <dl> неупорядоченный, упорядоченный список и список определений (definition)

<table> таблица

Тэги, определяющие внешний вид документа на экране или принтере:

 (с новой строки)

<i> курсив

 описание используемого шрифта

<center> выравнивание по центру

<pre> уже отформатированный текст.

Тэги для онлайн-приложений:

<a> (якорь - используется с гиперссылками)

 (расположение файла с графикой)

<form> (онлайн-форма).

XML

XML - это вариант SGML, который предназначен для заполнения промежутка между простотой HTML и мощностью SGML. Простота - это ключ к успеху HTML, но она же является его слабостью. С каждым новым добавлением HTML становится менее элегантным, более трудным в использовании, добавляются проблемы с интерпретацией различными браузерами. SGML, наоборот, столь гибок, что может использоваться для описания любых текстов - но эта гибкость достигается за счет сложности. Даже спустя много лет лишь несколько специалистов действительно владеют SGML, а программное обеспечение общего назначения встречается весьма редко.

XML является подмножеством SGML, разработанным специально для веб. В основу дизайна заложены два критерия: он должен быть достаточно простым с точки зрения создания программ, использующих XML; пользователи и системы должны иметь возможность безболезненного перехода от HTML к XML.

Базовый набор символов для XML - это 16-битный Unicode, в частности, кодировка UTF-8. Она позволяет использовать документы, созданные в ASCII, но также поддерживает многие языки и наборы символов. Для простоты некоторые символьные сущности заранее определены, например, < и > для "меньше, чем" и "больше, чем". XML не определяет конкретный метод представления математических символов, но существуют отдельные попытки в этом направлении, например, MathML.

Стандартный HTML может восприниматься как XML с минимальными модификациями. Одна из них заключается в том, что тэги всегда парные. Например, в HTML тэги <p> и </p> размечают начало и конец параграфа, но тэг </p> является необязательным, если затем следует новый абзац. В XML тэг-окончание нужен всегда. Другие существенные отличия касаются тэгов HTML, которые не размечают никакого содержания. Например,
 она вызывает переход на новую строку. Стандартный HTML не использует тэг </br>, а в XML эта же функция выполняется парой тэгов
 и </br> или специальным коротким тэгом
.

Поскольку XML базируется на SGML, каждый документ основывается на DTD, хотя DTD не обязательно должны быть определены явным образом. Если файл содержит заранее

неопределенную пару тэгов, которая размечает некую секцию документа, программа автоматически добавляет их к DTD.

Разработчики XML старались добиться широкого признания их разработки. Их стратегия следовала основной философии Интернета. Разработка проводилась как открытый процесс на сайте W3C. С самого начала члены команды разработчиков писали демонстрационные версии программного обеспечения и свободно рассылали заинтересованным лицам и организациям. Лидирующие корпорации, Microsoft и Netscape, предоставили свою поддержку. Эта сознательное стремление к консенсусу и общей поддержке привело к успеху; представляется, что XML получит широкое распространение.

Таблицы стилей

Языки разметки описывают структуру документа. SGML и XML используют тэги для описания семантики документа и его составных частей, но не описывают внешний вид документа. Таким образом, тэги SGML могут быть использованы для идентификации секций текста (в виде заголовков глав, например), но они не определяют, что глава начинается с новой страницы и печатается определенным шрифтом с определенным выравниванием текста.

Существует необходимость в инструменте, который позволял бы преобразовывать размеченные SGML или XML документы, в соответствии с дизайнерской спецификацией. Например, при подготовке журнальной статьи в соответствии с издательским DTD может возникнуть желание представить ее различными способами: для обычной печатной версии и в экранном формате для передачи через Интернет и вывода на дисплей. В первом случае статья будет преобразована в формат для фотонаборного автомата. Во втором случае будет выполнено преобразование в форматы, поддерживаемые веб-браузерами, обычно HTML или PDF.

Этот процесс требует, чтобы структурные тэги разметки были переведены в такие, которые могут использоваться для печати или вывода на экран. Для этого используют таблицы стилей. Например, DTD может определить элемент текста, как заголовок второго уровня, размеченный парой тэгов `<h2>` и `</h2>`. таблица стилей может содержать указания на то, что этот текст должен отображаться шрифтом Times Roman 13 кегля жирного начертания с выравниванием "влево". Также будут указаны другие важные характеристики, например, межстрочный интервал и порядок размещения заголовков, если она находится внизу страницы. Таблица стилей содержит детализированные инструкции по представлению любого документа, размеченного с помощью конкретного DTD.

При создании таблицы стилей большие возможности SGML являются скорее недостатком. Читатели привыкли пользоваться красиво оформленными и сверстанными книгами. Большая часть этого благоприятного впечатления обусловлена профессиональными навыками верстальщика, который выполняет форматирование. Большая часть его работы - это искусство. Человеческий глаз весьма чувствителен; детали форматирования статистических таблиц или расположения объектов в книге по искусству никогда не будут сведены к механическим правилам. Легко создать очень сложные таблицы стилей, которые, тем не менее, не дадут удовлетворительного результата при представлении сложных документов.

Поскольку SGML лишь задает рамки, был разработан общий метод определения конкретных таблиц стилей для любых DTD, названный Document Style Semantics and Specification Language (DSSSL). Это исключительно сложная задача. В настоящее время программы для реализации DSSSL написаны лишь для нескольких простых DTD, а общая задача выглядит невыполнимой. Значительно легче создать удовлетворительную таблицу стилей для одного DTD, который используется в четко определенном контексте. Многие книги и журналы, напечатанные с использованием разметки SGML и специализированных таблиц стилей, выглядят превосходно.

Применительно к HTML формальная концепция таблиц стилей отсутствует. Разметка комбинирует структурные элементы (такие, как различные типы списков) с тэгами форматирования (например, курсив или жирный шрифт). Тэги дают общее представление о внешнем виде, которое каждый браузер интерпретирует и адаптирует для используемого дисплея. Тот внешний вид, который воспринимает пользователь, формируется как комбинация разметки дизайнера веб-страницы, соглашений по форматированию, заложенных в браузер и

опций, выбранных автором. Авторы HTML-документов, которые хотели бы больших возможностей управления, могут вставлять в документы различные скрипты, апплеты или плагины.

Врезка 9.5 описывает CSS и XSL, которые представляют собой методы создания таблиц стилей для HTML и XML (соответственно). Разработчики XML, осознавая, что самым большим препятствием на пути широкого признания и внедрения XML является набор средств управления внешним видом документов, предусмотрели поддержку и CSS и XSL. Еще рано говорить об успехе, но такое сочетание выглядит многообещающим. Эти методы основываются на важной концепции, направленной на создание четких и точных правил интерпретации и обработки стилей, указанных дизайнером, в случае несогласия с ними пользователя.

Врезка 9.5

Каскадные таблицы стилей (Cascading Style Sheets, CSS) и расширяемый язык стилей (Extensible Style Language, XSL)

Языки разметки описывают структурные элементы документа. Таблица стилей определяет то, как элементы будут выглядеть на экране или на бумаге. CSS был разработан для использования с разметкой HTML. XSL является расширением CSS для XML. Рассматривая XML как упрощенную версию SGML, которая дает возможность простого перехода от HTML, XSL основывается на Document Style Semantics and Specification Language (DSSSL). Любые стили CSS могут быть конвертированы в XSL чисто механически. Сначала предполагалось, что XSL будет частью DSSSL. Сейчас XSL является только спецификацией, но есть надежда, что по мере широкого внедрения XML, XSL обретет такую же важность.

В CSS правило определяет стиль, который должен быть применен к указанным элементам документа. Например, простое правило

```
h1 {color: blue}
```

для элемента, размеченного как h1 (HTML заголовок верхнего уровня) выбран голубой цвет. В общем случае, каждое правило содержит "селектор" (selector), который выбирает некий элемент текста, и "декларацию" (declaration) (в скобках и состоящую из двух частей, разделенных двоеточием: свойства (property) и значения(value)), которые специфицируют стиль данного элемента.

Таблица стилей CSS представляет собой список правил. Существуют соглашения по написанию этих правил. Например,

```
h1, h2 {font-family: sans-serif; color: blue}
```

устанавливает, что заголовки первого и второго уровней будут отображаться голубым шрифтом sans-serif.

Разметка HTML-документов определяет структуру, которая может быть представлена в виде иерархии. Таким образом, заголовки, параграфы и списки являются элементом тела HTML; списки могут быть вложенными в другие списки. Правила таблиц стилей CSS предусматривают наследование стилей из других правил. Если явное правило для некоего элемента отсутствует, используется правило для элементов, расположенных выше в иерархии. Например, рассмотрим следующую пару правил.

```
body {font-family: serif}
```

```
h1, h2 {font-family: sans-serif}
```

Заголовки h1 и h2 являются элементами тела HTML и имеют явное правило - они будут показаны шрифтом sans-serif. Поскольку в этом примере отсутствуют явные правила для параграфов или списков, они наследуют стили тела, которые выше по иерархии. Таким образом, они будут отображаться шрифтом serif.

Таблица стилей должны быть связаны с HTML страницей. У дизайнера есть несколько возможностей, включая встраивание в заголовок страницы или ссылку на отдельный файл, содержащий таблицу стилей. Каждый браузер имеет свою внутреннюю таблицу стилей, которая может изменяться пользователем. Пользователь может создавать свои таблицы стилей.

Поскольку к одной странице могут применяться несколько таблиц стилей, может возникнуть конфликт между правилами. Для разрешения таких ситуаций был разработан простой механизм. Основной принцип - если правила конфликтуют, выбирается какое-то одно,

остальные игнорируются. Правила, в которых элементы определены явно, имеют приоритет над наследуемыми правилами. Наиболее противоречивым является соглашение, что приоритет имеет правило, установленное дизайнером - по отношению к правилам пользователя. Пользователь, который желает обойти это правило дизайнера, должен использовать правило-флаг "important". Несмотря на то, что это спорный принцип, он позволяет создавать специальные таблицы стилей - например, с особо большим кеглем шрифта для пользователей с плохим зрением.

Языки описания страниц

Поскольку и создатели, и читатели придают большое значение внешнему виду документа, желательно иметь методы, которые конкретизируют внешний вид непосредственно, без применения структурной разметки. Такие методы сильно различаются между собой в деталях, однако преследуют одну и ту же цель - представление текстуальных материалов с таким графическим качеством и точностью, как в лучших документах, отпечатанных традиционными методами. Это непросто. Немногими вещами в жизни так же приятно пользоваться, как хорошо отпечатанной книгой. За годы развития типографские методы, способы верстки, качество бумаги и переплета были доведены до совершенства. Вначале методы форматирования разрабатывались для вывода на принтер, но со временем вывод на экран стал столь же важен. В этой части книги рассмотрены три языка описания страниц: TeX, PostScript и PDF. Они предназначены для разных целей и по разному их решают, но все они представляют примеры реальных языков, которые широко применяются и дают хорошие результаты.

TeX

TeX был разработан Donald Knuth в 1980 году. Его цель - высококачественная печать текстов математической тематики. Кроме того, что математика использует многочисленные специальные знаки, математическая нотация - это сложные выражения, которые обычно не могут быть представлены простой (линейной) последовательностью знаков. TeX содержит правила для кодировки математики в виде последовательности ASCII знаков для выдачи, хранения и обработки в компьютере и тэги форматирования для представления на экране.

Большинство пользователей работают с одной из двух разновидностей TeX: гладкий (plain) TeX и LaTeX. Каждый из них определяет набор тэгов форматирования, которые предусматривают обработку большинства стандартных ситуаций.

Близким союзником TeX является система для разработки шрифтов под названием Metafont. Knuth имел много хлопот, разрабатывая для различных компьютерных систем отдельные версии своего стандартного шрифта.

TeX остается незаменимым для подготовки математических статей. Он широко используется авторами в математике, физике и смежных областях.

PostScript

PostScript был первым продуктом фирмы Adobe Systems, которая отделилась от фирмы Xerox в 1984 году. Это язык программирования, используемый для создания графических образов выдачи на распечатку. Лишь немногие люди когда-либо программировали на этом языке, однако многие компьютеры имеют подпрограммы вывода на печать, которые обрабатывают текст или графику, создавая некий эквивалент программы на PostScript. Эти программы затем могут пересылаться на контроллер принтера, который исполняет PostScript программу и создает контрольную последовательность для непосредственного управления принтером.

Встроенная поддержка шрифтов является сильной стороной PostScript. Большая часть первоначального успеха компьютеров Macintosh фирмы Apple обусловлена сочетанием растровых технологии (bit-mapped design) для дисплеев и принтеров PostScript, которые

позволили получать распечатки с качеством, ранее доступным лишь на очень дорогих компьютерах. Как при лазерной печати, так и на экране дисплея, знаки строятся из маленьких точек. Простой лазерный принтер использует 300 точек на дюйм, типографская машина имеет разрешение 1200 точек на дюйм и выше, а большая часть компьютерных дисплеев имеет разрешение 75 точек на дюйм. Шрифты, которые привлекательно выглядят на экране - отнюдь не те же самые, что используются при печати, изображение текста на дисплее должно согласовываться с его разрешением. Обычно вывод на дисплей обеспечивается функциями операционной системы, которые отличаются от команд PostScript для принтера. Несмотря на все старания разработчиков, это может привести к различным разрывам страниц и другим неприемлемым искажениям.

Хотя PostScript является прежде всего языком описания графической выдачи и первоначально предназначался для управления лазерными принтерами, программы PostScript используются и в других приложениях. Одно из применений PostScript состоит в хранении и обмене представлениями любого текста или графики. PostScript не идеален для этой цели, поскольку имеет много версий. Кроме того, приложения должны обрабатывать файл исходя из возможностей компьютерных систем, на которых они будут исполняться.

PDF

Фирма Adobe использовала опыт разработки языка PostScript для создания Portable Document Format (формат переносимых документов, PDF), который не зависит от компьютерной системы. PDF является наиболее популярным языком описания страниц и объединяет богатые возможности форматирования и набор инструментов для создания, хранения и изображения документов в нем.

PDF широко используется в коммерческих системах управления документами, но некоторые ЭБ неохотно пользуются им. Причина одна - техническая. PDF наилучшим образом подходит для представления документов, созданных из компьютерных оригиналов. Также в нем можно сохранять битовые изображения, фирма Adobe выпускает системы оптического распознавания знаков для создания PDF файлов. Однако большинство достоинств PDF пропадает, когда он используется для хранения файлов изображений, - их размер чрезмерно велик, а гибкость (которая особенно желательна для ЭБ) утрачивается.

Некоторые ЭБ и архивы отвергают PDF потому, что это коммерческий формат. Существует определенный страх в связи с принятием PDF для использования в ЭБ, связанный с тем, что возможные проблемы в будущем будут столь большими, что проще использовать один из стандартизованных форматов. Эти доводы представляются ошибочными. Первоначальный успех PDF связан с корпоративной Америкой, которая отдает предпочтение коммерческим продуктам, обеспеченным надежной технической поддержкой. Академическое и образовательное сообщества должны признать, что формат, поддерживаемый корпорацией, может быть более стабилен в долгосрочной перспективе, чем официальные стандарты, которые не используются в популярных высококачественных программах. Полное описание PDF опубликовано, а широкое использование этого формата в коммерческих приложениях гарантирует, что программное обеспечение будет доступно, даже если фирма Adobe выйдет из бизнеса или откажется его поддерживать.

Врезка 9.6 **Формат PDF**

Формат PDF фирмы Adobe является весьма важным. Это также интересный пример того, как корпорация может создавать стандарт, распространять его по всему миру и делать на этом хороший бизнес. PDF - это формат представления документов, независимый от приложения и компьютерной системы. Документ PDF состоит из страниц, образованных текстом, графикой, изображениями и поддерживающими данными. Однако, страницы PDF могут использоваться не только для статического представления и распечатки страниц, поскольку эта форма

поддерживает гиперссылки, поиск и другие возможности, которые доступны в электронной форме.

PDF является развитием языка программирования PostScript, также созданного фирмой Adobe. Способ создания PDF-файла заключается в преобразовании потока данных, которые обычно направляются на принтер. Другой вариант - это конвертирование из других форматов (PostScript и др.). Файлы могут храниться, передаваться по сети, выводиться на экран или на принтер.

PDF имеет множество технических достоинств. При просмотре на экране компьютера большинство документов выглядит исключительно разборчиво, словно напечатанные. За исключением случаев, когда они содержат графику, это файлы умеренного размера. Если на компьютере отсутствует шрифт, использованный при создании документа, встроенный описатель шрифтов (font descriptor) позволяет программе просмотра PDF-файлов генерировать некий искусственный шрифт, который обычно весьма близок к оригиналу. Хотя он с трудом различает некоторые типы файлов и необычные шрифты, PDF стал важным форматом для онлайн-документов.

PDF предоставляет пользователю много возможностей, включая создание гиперссылок (как внутри документа, так и на внешние URL). Программа просмотра PDF снабжена возможностями поиска слов в тексте, хотя поиск между документами проблематичен. Поддерживаются аннотации. Поддержка принтеров великолепна. Существует даже возможность воспрепятствовать распечатке или иному использованию документа пользователем способами, которые не одобрены создателем документа.

После некоторых колебаний в начале, фирма Adobe проницательно повела маркетинговую политику, частью которой стало создание прекрасных программ для просмотра PDF файлов на любых типах компьютеров, которые свободно доступны в Интернете. Фирма зарабатывает деньги на продаже средств создания PDF файлов. Adobe является собственником спецификации на PDF, однако - будучи заинтересованной в максимальном распространении этого формата в качестве универсального средства обмена - она свободно выдает разрешения на создание PDF файлов, разработку приложений, предусматривающих выдачу результатов в PDF формате и на написание программного обеспечения для чтения и обработки файлов. Конкуренты могут создавать продукты, которые будут соперничать с Adobe, но фирма сознательно идет на такой риск.

Документы: структура против внешнего вида

Эта глава начиналась с обсуждения двух требований к хранению документов в ЭБ: представление структуры и представление внешнего вида. Они не должны выглядеть как альтернативы или конкуренты. Как равноправные потребности, они обе заслуживают внимания. В некоторых приложениях для обеих целей используется одно представление; но многие ЭБ хранят две версии каждого документа. Текстовые материалы являются сердцевинной ЭБ и электронных издательств. Авторы и читатели весьма требовательны, но уже существуют и методы удовлетворения этих потребностей.

Глава 10. Обработка информации и описательные метаданные

Помощь пользователям в поиске информации относится к базовым службам ЭБ. Эта глава, первая из двух, посвященных этой общей задаче, начинается с обсуждения описательных метаданных - каталогов, индексов и другой общей информации, которая используется для описания объектов в ЭБ. Затем следует обсуждение методов полнотекстового поиска. Глава 11 дополняет обсуждаемые здесь концепции проблемами распределенного поиска - например, поиск информации, которая распределена в разных коллекциях или на нескольких компьютерных системах.

Эти две главы сосредоточены на методах поиска конкретной информации. Прямой поиск - это только одна стратегия, используемая людьми при поиске. Просмотр (неструктурированное исследование массива информации) - также популярный и эффективный метод неожиданных открытий. Большинство традиционных библиотек упорядочивает свои коллекции по тематике для облегчения просмотра. Девеевская десятичная классификация (Dewey Decimal Classification) и классификация Библиотеки конгресса предоставляют как предметную информацию, так и иерархическую структуру, которая может быть использована для организации коллекций. Наиболее широко используемый сервис веб, Yahoo, с фундаментальной точки зрения, является классификацией веб-ресурсов. ЭБ с их гиперссылками пользуются смешанными стратегиями, объединяющими поиск и просмотр.

Диапазон потребностей пользователей ЭБ иллюстрирует, почему поиск информации является такой сложной проблемой и почему не существует единого подхода, который подошел бы всем пользователям или годился для всех материалов. Целью настоящего поиска является обнаружение всего, относящегося к заданной теме. И профессионалы обычно начинают работу с исчерпывающего поиска информации в данной области. В "поиске по известному предмету" (known-item search) пользователь ищет конкретную вещь, например, первую главу книги "Моби Дик". Зачастую полученные результаты недостаточны и поиск должен быть продолжен, - однако особенностью этого вида поиска является то, что существует определенная, известная вещь, которую пользователь узнает, когда найдет. Другой вид поиска - "поиск фактов" (searches for facts). Факты (например, название столицы Барбадоса, атомный вес ртути) - это специфические единицы информации, которые могут быть найдены во многих источниках. Поиск факта закончен в тот момент, когда искомый факт найден. Источник факта должен быть надежным, хотя может существовать много других возможных источников. Следующий вид поиска - "обзорный по данной теме". Запросы типа "Как обычно организуются свадьбы?" или "Что такое шифрование с открытым ключом?" могут приводить ко многим релевантным источникам информации, но обычно пользователю нужна лишь малая часть. Как только нужная информация найдена, пользователь может захотеть получить информацию о связанных предметах. В процессе исследования такие виды поиска как, например, общий вопрос - "Существуют ли более поздние исследования, которые базируются на результатах, приведенных в этой работе?"

Описательные метаданные

Большинство методов поиска информации не ищет реальных объектов в коллекциях, обычно они работают с описательными метаданными об этих объектах. Метаданные об объекте могут представлять собой каталожную запись, запись в индексе или реферат. Описательные метаданные обычно хранятся отдельно от описываемых объектов, но иногда они являются их частью.

Описательные метаданные обычно текстовые; однако можно использовать метаданные в иных форматах, например, изображения, звукозаписи, карты и компьютерные программы.

Единый каталог может объединять записи обо всех видах материалов, носителей и форматов. Это позволяет пользователям ЭБ находить материалы на любых носителях в ходе поиска по текстовым записям.

Описательные метаданные обычно создаются профессионалами. Библиотечные каталоги и научные индексы аккумулируют усилия подготовленных сотрудников, предпринимаемые в течение десятилетий и даже веков. Этот экономический факт критически важен для понимания сегодняшних тенденций. С одной стороны, важно продолжать эти усилия и накапливать опыт. С другой - есть побудительный мотив найти более дешевый и быстрый способ генерации метаданных либо с использованием автоматического индексирования, либо с использованием компьютерных возможностей для совершенствования человеческих навыков.

Каталоги

Каталожная запись содержит краткую общую информацию о библиотечном объекте. Слово "каталог" применяется к записям, которые имеют predetermined структуру и организованы в соответствии с определенными систематическими правилами. Реферат (abstract) представляет собой краткое резюме (summary) содержания более длинного документа. Другие типы индексных справочников менее формализованы, чем каталожные, но более структурированы, чем рефераты.

Библиотечные каталоги выполняют много функций помимо поиска информации. Некоторые каталоги предоставляют полную библиографическую информацию, которая не может быть явным образом получена непосредственно из объекта. Это в т.ч. информация об авторах или происхождении музейных объектов. Для управления коллекциями каталоги содержат служебную информацию - например, о том, где хранится данный материал. Каталоги обычно много меньше, чем коллекции, которые они представляют; в обычных библиотеках материалы хранящиеся на километрах книжных полок, описываются записями, которые располагаются в каталожных шкафах в одной комнате. Индексы ЭБ могут "отзеркаливаться" для увеличения производительности и надежности.

Информация в каталожных записях подразделяется на поля и подполя, которые обозначаются тэгами. Например, может быть поле "фамилия автора" и подполе "имя автора". Англо-американские правила каталогизирования и формат MARC используются для различных типов материалов, включая монографии, продолжающиеся издания и архивы. Поскольку создание каталожной записи - процесс трудоемкий, он выполняется однажды (зачастую в национальной библиотеке, вроде Библиотеки Конгресса), а запись распространяется другим библиотекам с помощью такого механизма, как OCLC.

Роль формата MARC и связанных с ним правил каталогизирования в ЭБ является объектом обсуждения. Насколько адекватны традиционные правила каталогизирования для поддержки новых форматов, типов носителей и методов издания? Ведение каталогов с помощью MARC сохраняет свою значимость для обычных материалов. Библиотекари распространили его на некоторые другие типы объектов, которые можно найти в ЭБ, но этот формат пока не принят ни одной организацией, за исключением традиционных библиотек.

Службы реферирования и индексирования

В науке и технике принято больше полагаться на рефераты и индексы, чем на каталоги. В каждой научной дисциплине есть своя служба, которая помогает пользователям искать информацию в специальных журналах - например, Medline для медиков и биологов, Chemical Abstracts для химиков и Inspec для физики, вычислительной техники и близких отраслей. Каждая такая служба индексирует статьи из большого списка журналов. Описание каждой статьи включает базовую библиографическую информацию (автор, название, дата и пр.) и предметную информацию, специально организованную для поиска. Поскольку они развивались во времена, когда компьютеры были медленными и дорогостоящими, информация

структурировались в расчете на простой текстовый поиск; однако эти описания доказали свою полезность и в более мощных и гибких системах.

Врезка 10.1

Медицинские предметные заголовки (Medical Subject Headings, MeSH)

Национальная медицинская библиотека США предоставляет услуги по поиску информации по медицине и смежным областям с 60-х годов. Поскольку в медицине (сложной области с гигантской терминологией) весьма часто бывают случаи, когда одна и та же концепция описывается в различных терминах, библиотекой был создан контролируемый словарь, известный как MeSH, который представляет собой тезаурус предметных заголовков для каждой из 400000 индексных записей о статьях и книгах, добавляемых ежегодно. Эти поисковые заголовки могут использоваться для поиска информации. В дополнение к MeSH (который содержит около 18000 первичных заголовков), есть отдельный тезаурус химических терминов (ок. 80000). Термины в MeSH организованы иерархически. Вверху общий термин, такой, как "анатомия", "организм", "болезнь". Вниз по иерархии анатомия подразделяется на 16 тем, начиная с "участки тела" и "мышечно-скелетной системы". "Участки тела" дальше подразделяются на секции, например, "ахилл", "спина" и пр. Некоторые секции имеют дополнительное членение. Термины связаны между собой тысячами перекрестных ссылок.

Успех MeSH зависит от профессионализма тех, кто поддерживает тезаурус и прививает предметные термины документам, а также от пользователей и информационных работников, которые понимают медицину и способны формулировать запросы с использованием терминов и структуры MeSH.

Научные пользователи зачастую хотят видеть информацию по определенному предмету. Вследствие языковых различий простой предметный поиск ненадежен, если только нет индексной информации, которая четко описывает каждый предмет. Предметной информацией может быть реферат, ключевые слова, предметные термины (subject terms) или иное. Некоторые службы предлагают авторам самим добавлять ключевые слова или рефераты, но это ведет к существенной несогласованности. Значительно лучше пользоваться услугами профессионалов.

Эффективным, но дорогим подходом является использование "контролируемого словаря" (controlled vocabulary). Он заключается в том, что если для описания понятия существует несколько терминов, используется только один из них. Сотруднику выдается утвержденный список терминов и правила его применения. Остальные термины запрещены. Этот подход используется в Библиотеке Конгресса (служба предметных заголовков) и в службе MeSH заголовков Национальной медицинской библиотеки.

Использование контролируемого словаря требует наличия подготовленных сотрудников, а также опытных пользователей, поскольку термины в запросах должны согласовываться с терминами, используемыми при индексировании. Медицине в США особенно повезло в смысле наличия кадров опытных библиотечных работников. В ЭБ существует тенденция предоставлять пользователям соответствующие инструменты, позволяющие находить информацию непосредственно, без помощи профессионалов. Тезаурусы (например, MeSH или Art and Architecture) могут использоваться для взаимного уточнения терминов индексирования и поиска.

Врезка 10.2

Тезаурус по искусству и архитектуре (Art and Architecture Thesaurus)

Тезаурус был создан фондом J.Paul Getty как контролируемый словарь для описания и поиска информации по искусству, архитектуре, декоративному искусству и материальной культуре. Он содержит свыше 120000 терминов для объектов, текстовых материалов, изображений, архитектурных и культурных форм всех культур и всех периодов,

преимущественно западной цивилизации. Тезаурус может использоваться в архивах, музеях и библиотеках для описания коллекций и при поиске материалов.

Серьезная работа над этим проектом началась в начале 80-х годов, когда Интернет был еще в мечтах и планах; однако данные создавались с использованием гибкого формата, который допускал различные версии, включая варианты для открытого доступа через веб, печатные книги и разные компьютерные форматы. Фонд специально организовал тезаурус таким образом, чтобы его могли использовать компьютерные программы для поиска информации и обработки естественного языка.

Тезаурус подразделяется на семь категорий, каждая содержит иерархию терминов. Категории следующие: ассоциированные концепции, физические атрибуты, стили и периоды, агенты (agent), деятельности (activity), материалы, объекты. Отдельная концепция представлена кластером терминов, один из которых определяется как дескриптор (т.е. предпочтительный термин, preferred term). Тезаурус предоставляет не только термины для обозначения объектов, но и словарь, используемый для их описания (например, стиль, период, форма, цвет, конструкция, использование) и образовательные концепции, такие как теории и критика.

Цена разработки и поддержания большого специализированного тезауруса велика. Даже в устоявшейся области, такой как искусство и архитектура, терминология постоянно меняется, а технический персонал должен внедрять новые технологии. Только потому, что Фонд весьма богат, разработка тезауруса продолжается уже много лет.

Дублинское ядро (Dublin Core)

С 1995 года международная группа под руководством Stuart Weibel из OCLC работает над разработкой простого набора элементов метаданных, которые могли бы применяться для широкого набора библиотечных материалов. Набор таких элементов, созданный группой, известен под названием Дублинское ядро по названию города Dublin в штате Огайо, где расположена штаб-квартира OCLC и где состоялась первая рабочая встреча. Несколько сотен людей участвовали в рабочих встречах Dublin Core или в обсуждениях посредством электронной почты. Присущий им дух сотрудничества является примером того, как индивидуумы с различными интересами могут работать вместе. Пятнадцать элементов, которые они отобрали, приведены на врезке 10.3.

Простота одновременно является и сильной, и слабой стороной Дублинского ядра. В то время, когда традиционные правила каталогизирования сложные и длинные, требуют профессиональной подготовки для эффективного использования, Dublin Core может быть описан весьма просто. Однако простота конфликтует с точностью. Команда много работала над этим парадоксом. Первоначальная цель заключалась в создании простого набора элементов метаданных для неподготовленных людей, которые публикуют электронные материалы с описанием своих результатов. Некоторые продолжают придерживаться этого минималистского подхода, они хотели бы видеть простой набор правил, которыми мог бы воспользоваться любой. Другие предпочитают ориентироваться на преимущества более тщательно разработанных правил каталогизирования и согласны на увеличение трудоемкости и стоимости. Они указывают на то, что дополнительные структурные элементы позволяют добиться большей точности в метаданных. Например, если поле "предмет" (subject) заполнялось с использованием Девеевского десятичного классификатора, было бы полезно отразить этот факт в метаданных. Для дальнейшего повышения эффективности метаданных при обработке информации, предлагается присваивать некоторым элементам некоторые "значения" (values). Это может быть определенный набор, список типов, которым могут руководствоваться составители индексов.

Врезка 10.3 **Дублинское ядро**

Следующие пятнадцать элементов составляют набор метаданных Дублинского ядра. Все элементы являются необязательными и могут повторяться. (Это описание было создано на основе официальных определений группы разработчиков Dublin Core с их разрешения.)

Заголовок (Title) Название, присвоенное ресурсу создателем или издателем.

Автор (создатель, Creator) Человек или организация, изначально ответственная за интеллектуальное содержание ресурса (в случае рукописного документа это авторы; исполнители, фотографии или иллюстраторы в случае визуальных ресурсов).

Предмет (Subject) Тема ресурса. Обычно предмет выражается в ключевых словах или фразе, описывающей предмет или содержание ресурса. Приветствуется использование контролируемых словарей и формальных схем классификации.

Описание (Description) Текстовое описание содержания ресурса, включая реферат в случае документов или описание содержания в случае визуального ресурса.

Издатель (Publisher) Организация, ответственная за создание ресурса в его нынешней форме - например, издательский дом, факультет университета или корпорация.

Участник создания материала (Contributor) - человек или организация, которые не являются авторами (не обозначены в элементе "автор"), но внесли значительный интеллектуальный вклад в ресурс, но чей вклад вторичен по отношению к любому человеку или организации, указанной в числе авторов - например, редактор, переводчик, иллюстратор.

Дата (Date) Дата, указывающая на создание или появление (в доступном виде) ресурса.

Тип (Type) Категория ресурса - например, домашняя страничка, роман, поэма, статья, препринт, технический отчет, эссе, словарь.

Формат (Format) Формат представления данных ресурса (обычно указывается тип программного обеспечения и - возможно - тип компьютера, которые могут быть необходимы для отображения и работы с ресурсом).

Идентификатор (Identifier) Набор букв или цифр, который обычно используется для уникальной идентификации ресурса. В случае сетевых ресурсов примерами являются URL и URN.

Источник (Source) Информация о вторичном источнике, из которого был получен настоящий ресурс.

Язык (Language) Язык, на котором изложено интеллектуальное содержание ресурса.

Связь (Relation) Идентификатор вторичного ресурса и его связь с настоящим ресурсом. Этот элемент позволяет связывать между собой близкие ресурсы, а также описания ресурса, которые необходимо показать. Примеры - издание книги и глава книги.

Охват (Coverage) Характеристики местонахождения и временной продолжительности ресурса.

Права (Rights) Утверждение об авторских правах и управление ими; идентификатор, связанный с таким утверждением; идентификатор, связанный с сервисом, представляющим информацию об управлении правами на данный ресурс.

Таким образом, можно сказать, что стратегию Дублинского ядра развивают "минималисты" с одной стороны, и "структуралисты" с другой. Первые исходят из первоначального намерения создать систему, полезную для людей без специальной подготовки. Вторые хотят создать систему более сложную, требующую исчерпывающего описания и подготовленных сотрудников.

Автоматическое индексирование

Каталогизирование и индексирование стоят дорого, если выполняются профессионалами. По опыту известно, что создание и распространение одной записи стоит около 50 долларов. В некоторых областях, например, медицине или химии, требования к информации достаточно велики - но эти дисциплины составляют исключение. Обычно каталогизирование монографий ограничено общей записью, без детального описания отдельных тем. Многие объекты в музеях, архивах и специальных коллекциях библиотек индивидуально не каталогизируются и не индексируются.

В ЭБ стоимость каталогизирования отдельных объектов оценить сложно. Коллекции могут быть очень большими и непрерывные изменения объектов могут препятствовать инвестициям в каталогизирование. Каждый объект может существовать в нескольких версиях, которые быстро сменяют одна другую. Единичный объект может состоять из множества других объектов, каждый из которых изменяется независимо. Новые категории объектов возникают непрерывно, в то время как другие исчезают. Зачастую пользовательское восприятие объекта является результатом исполнения компьютерной программы и отличается с каждой итерацией. Эти факторы увеличивают сложность и стоимость каталогизирования материалов ЭБ.

С учетом сказанного, можно полагать, что профессиональное каталогизирование и индексирование будут занимать в ЭБ не столь важное место, как в обычных библиотеках. Альтернативой является использование программ для автоматического индексирования. Записи, генерируемые этими программами, обычно низкого качества, зато весьма дешевы. Мощным поисковым системам еще предстоит долгий путь для того, чтобы компенсировать низкое качество индивидуальных записей. Поисковые программы для веб, которые строят свои индексы автоматически, доказывают это. Записи не слишком хороши, но успех поисковых сервисов указывает на их полезность. По крайней мере, они лучше, чем ничего.

Большинство разработок, которые привели к созданию автоматического индексирования, было выполнено в рамках исследований в области распознавания смысла (text skimming). Типичной проблемой в этой области является организация электронной почты. Пользователь имеет много посланий и хочет организовать их по предметному принципу. Предполагается, что компьютерная программа их прочитывает и относит к определенным предметным областям. Это достаточно сложная задача для людей - и исключительно сложная для компьютеров. Однако уже сейчас достигнут значительный прогресс. Программы в документе отыскивают "ключи" - структурные элементы (например, поле "Тема" в послании электронной почты), лингвистические ключи или просто ключевые слова.

Автоматическое индексирование зависит также от ключей, которые будут найдены в документе. Первый пример врезки 10.4 - об успешном проекте, поскольку скрытая информация веб-документов содержит полезные ключи. Программа индексирования AltaVista могла идентифицировать заголовок и автора. Например, страница включает следующий размеченный элемент:

<title>Digital library concepts</title>

Автор вставил эти тэги для того, чтобы указать веб-браузеру заголовок, представляющий статью. В равной степени они полезны и для программ автоматического индексирования.

Врезка 10.4

Автоматическое индексирование

Две записи, воспроизведенные ниже (слегка отредактированная версия записей из AltaVista 1997 года), являются типичными примерами индексных записей, созданных автоматических поисковыми программами в веб.

Digital Library Concepts

Key Concept in the Architecture of the Digital Library. William Y. Arms Corporation for National Research Initiative Reston, Virginia...

<http://www.dlib.org/dlib/July95/07arms.html> - size 16K - 7-Oct-96 - English

Repository References

Notice: HyperNews at union.ncsa.uiuc.edu will be moving to a new machine and domain very soon. Expect interruptions. Repository References. This is a page.

<http://union.ncsa.uiuc.edu/HyperNews/get/www/repo/references.html> - size 5K - 12-May-95 - English

Первый пример демонстрирует лучший результат автоматического индексирования. Он содержит имя автора, название, дату и расположение статьи в электронном журнале. Для многих целей такая запись вполне адекватно заменяет запись, составленную профессионалом.

Второй пример иллюстрирует некоторые недостатки. Трудно было бы хуже индексировать эту веб-страницу. Информация о местоположении и дате, вероятно, точная. Однако заголовок весьма странный, а содержание записи - это лишь несколько первых слов страницы.

Одно из потенциальных применений языков разметки, вроде SGML или XML, в том, что структурные тэги могут использоваться программами автоматического индексирования для построения записей для поиска информации. В тексте документа фраза "Marie Celeste" может быть именем человека, названием книги, песни, корабля, издательства, пьесы или вообще не являться именем. При наличии структурной разметки фраза может быть идентифицирована и помечена. Таким образом, информация, содержащаяся в разметке, может быть использована для дифференцирования различных категорий информации, таких как автор, название или дата.

Автоматическое индексирование - быстрое и дешевое. Точная стоимость является коммерческим секретом, но приблизительно это доли цента на запись. За ту же цену, за которую профессионал создает одну запись, компьютерная программа генерирует 100000 или более записей. Это делает экономически целесообразным индексирование огромных объемов информации в Интернете и даже проведение повторного индексирования через определенные промежутки времени.

Создатели каталогов и индексов могут приводить в соответствие расходы и доходы. Наиболее дорогими методами создания описательных метаданных являются традиционные подходы, используемые для библиотечных каталогов и в реферативных и индексных службах. Структуралистский подход Dublin Core будет занимать промежуточную позицию по стоимостному критерию, сохраняя большинство преимуществ при умеренной стоимости. Минимальный подход Dublin Core будет еще дешевле - но не бесплатным. Автоматическое индексирование обеспечивает наихудшее качество при ничтожной стоимости.

Присоединение метаданных к содержанию

Описательные метаданные должны быть присоединены к материалу, который они описывают. В прошлом метаданные обычно хранились отдельно, как внешний каталог или индекс. В этом способе много достоинств, но требуются связи между метаданными и описываемыми объектами. Некоторые ЭБ двигаются в ином направлении, сохраняя метаданные и данные вместе (либо встраивая метаданные с объектом или в виде двух отдельных связанных объектов). Этот подход удобен для распределенных систем и для долговременного архивирования, поскольку гарантируется, что компьютерные программы будут иметь доступ к данным и метаданным в одно и то же время.

Механизм ассоциирования метаданных с веб-страницами был предметом серьезного обсуждения. Для HTML-страниц простейший подход состоит в использовании специального HTML-тэга <meta>, который появился из описания элементов Дублинского ядра для HTML. Таблица 10.1 иллюстрирует использование этих тэгов. (Выбор тэгов <meta> - это системообразующее решение. В рамках проекта Dublin Core специально не определяется, каким именно образом метаданные связываются с материалом).

Таблица 10.1

Метаданные, представляемые тэгами HTML <meta>

```
<meta name>="DC.subject" content="dublin core metadata element set">
<meta name>="DC.subject" content="networked object description">
<meta name>="DC.publisher"> content="OCLC Online Computer Library Center, Inc.">
<meta name>="DC.creator"> content="Weibel, Stuart L., weibel@oclc.org.">
<meta name>="DC.creator"> content="Miller, Eric J., emiller@oclc.org.">
<meta name>="DC.title"> content="Dublin Core Element Set Reference Page">
<meta name>="DC.date"> content="1996-05-28">
<meta name>="DC.form"> scheme="IMT" content="text/html">
<meta name>="DC.language" scheme="ISO639" content="en"<
<meta name>="DC.identifier" scheme="URL"
content="http://purl.oclc.org/metadata/dublin_core">
```

Поскольку эти тэги не могут использоваться в иных, нежели HTML, типах файлов и быстро становятся громоздкими, ряд организаций в составе консорциума W3C разрабатывают более общий подход, называемый "среда для описания ресурсов" (Resource Description Framework, RDF). RDF описан на врезке 10.5.

Врезка 10.5 RDF

RDF представляет собой метод обмена метаданными, разработанный консорциумом W3C с привлечением некоторых других концепций (включая формат PICS для генерирования меток для материалов особого содержания, например насилие и порнография на веб-страницах). Команда Dublin Core работает в тесном контакте с разработчиками RDF.

Схема метаданных, например, Дублинское ядро, может рассматриваться в трех аспектах: семантическом, синтаксическом и структурном. Семантика описывает способы интерпретации концепций, таких как "данные" или "создатель". Синтаксис уточняет способы формулирования метаданных. Структура определяет связи между элементами метаданных (например, день, месяц и год как компоненты даты). RDF предлагает простую и универсальную структурную модель для выражения синтаксиса. Он не обуславливает использованную семантику схемами метаданных. Для описания схемы метаданных и для обмена информацией между компьютерными системами и различными схемами используется XML.

Структурная модель состоит из ресурсов (resources), свойств типов (types property) и значений (values). Рассмотрим простое утверждение, что Шекспир является автором пьесы "Гамлет". В схеме метаданных Дублинского ядра это может быть представлено следующим образом:

| Ресурс | | Тип свойства | | Значение |
|--------|---|---------------------|---|--------------|
| Hamlet | → | creator (создатель) | → | Shakespeare |
| | → | type (тип) | → | play (пьеса) |

Другие схемы метаданных могут использовать термин "автор" вместо "создателя" или использовать термин "тип" в другом значении. Таким образом, разметка RDF может явным образом продемонстрировать то, что приведенные метаданные составлены на основе базовой схемы Дублинского ядра:

```
<DC:creator>Shakespeare</DC:creator>
<DC:type>play</DC:type>
```

Для того, чтобы завершить этот пример, необходимо более точно идентифицировать "Гамлета". Если предположить, что он находится по воображаемому URL <http://hamlet.org/>, полная RDF-запись в разметке XML выглядела бы следующим образом:

```
<RDF:RDF>
  <RDF:description RDF:about = "http://hamlet.org/">
    <DC:creator>Shakespeare</DC:creator>
    <DC:type>play</DC:type>
  </RDF:description>
</RDF:RDF>
```

Разметка этой записи явно показывает, что "описание" и "о" (about) определены в схеме RDF, а "создатель" и "тип" - в разметке Дублинского ядра (DC). Для завершения записи необходим еще один шаг - схемы RDF и DC должны быть определены именами XML.

Структурная модель RDF позволяет ресурсам иметь свойства типов, относящихся к другим ресурсам. Например, база данных может включать запись о Шекспире, содержащую метаданные о нем, например, когда и где он жил и различные способы произношения его имени. В таком случае свойства типов схемы DC:Creator должен быть описан так:

```
<DC:creator RDF:about = "http://people.net/WS/">
```

По этому способу могут быть созданы метаданные любой сложности из простых компонентов. Используя подходы RDF для синтаксиса и структуры, а также представление

XML, компьютерные системы могут связывать метаданные с цифровыми объектами и обмениваться метаданными между различными схемами.

Методы поиска информации

Поиск информации представляет собой область, в которой компьютерные и информационные профессионалы работают вместе уже много лет. Это все еще актуальная область исследований, которая - с точки зрения ЭБ - является одним из немногих примеров дисциплины с систематической методологией для измерения и сравнения производительности различных методик.

Базовые концепции и терминология

Различные методы поиска информации используют несколько простых концепций для поиска в больших объемах данных. "Запрос" представляет собой последовательность знаков, описывающих искомую информацию. Каждое слово запроса называется "поисковым термином". Запрос может состоять из одного поискового термина, фразы на естественном языке или из сложного выражения с использованием специальных символов.

Некоторые методы поиска сравнивают запрос с каждым словом в тексте, не делая различия между функциями различных слов; это так называемый "полнотекстовый поиск". Другие методы идентифицируют поля библиографической или структурной информации (например, автор или заголовок) и проводят поиск только в этих полях (например, "author" = Gibbon"); это называют "поиском по полям". Оба эти метода достаточно эффективны, поэтому современные системы обработки информации часто используют их в комбинации.

Если поиск по полям требует неких методов идентификации полей, полнотекстовый поиск такой поддержки не требует. Учитывая мощность современных компьютеров, полнотекстовый поиск может быть эффективен даже в необработанном тексте. Однако, неоднородные тексты переменной длины, стиля и содержания искать таким методом трудно, а результаты могут не воспроизводиться или быть неадекватными. Хотя юридические информационные системы Westlaw и Lexis базируются на полнотекстовом поиске, при наличии описательных метаданных большинство служб предпочитает либо поиск по полям либо полнотекстовый поиск в рефератах или иных метаданных.

Некоторые слова, как общеупотребительные существительные, союзы и вспомогательные глаголы, встречаются столь часто, что не представляют ценности при поиске информации. Большинство систем имеет списки часто используемых слов, которые в запросах игнорируются. Это так называемые списки "стоп-слов". Отбор таких слов достаточно труден. Он явным образом зависит от языка текста; также он может быть связан с предметом материала. По этим причинам некоторые системы вместо заранее составленных списков используют статистические методы для выделения наиболее часто встречающихся слов, которые игнорируются. Но даже такие системы несовершенны. Всегда существует опасность того, что некоторые запросы - например, "To be or not to be?" - могут быть отвергнуты потому, что все слова запроса находятся в списке стоп-слов.

Основой для поисковых процедур используемых для сравнения поисковых терминов с содержанием текстовых документов, являются "инвертированные файлы", описанные на врезке 10.6.

Врезка 10.6

Инвертированные файлы (inverted files)

Инвертированный файл представляет собой список слов в наборе документов и их расположение в этих документах. Ниже приведена небольшая часть такого файла:

| Слово | Документ | Расположение |
|--------|----------|--------------|
| abacus | 3 | 94 |
| | 19 | 7 |
| | 19 | 212 |
| actor | 2 | 66 |
| | 19 | 200 |
| | 29 | 45 |
| aspen | 5 | 43 |
| atoll | 11 | 3 |
| | 34 | 40 |

Видно, что слово "abacus" является 94 словом в третьем документе, 7-мым и 212 в 19-ом документе; слово "actor" стоит 66-м по счету во 2-ом документе; 200-м в 19-ом и 45-м в 29-ом документе и т.д.

Инвертированный файл может использоваться для поиска в группе документов для каждого поискового термина. В вышеприведенном примере поиск на слово "actor" найдет в файле, что слово появляется в документах 2, 19 и 29. Простая ссылка на инвертированный файл обычно быстро выполняется и не занимает ресурсы компьютера.

Большинство таких списков дают расположения слов в документе. Это важно для отображения результатов поиска, в особенности для длинных документов. Может быть показана подходящая часть документа с выделенными поисковыми терминами.

Поскольку инвертированные файлы включают каждое слово из набора документов (за исключением слов из "стоп-списка"), их размеры весьма велики. Такие файлы для типичных материалов ЭБ даже после архивирования (сжатия) могут составлять до половины размера всех документов. Таким образом инвертированные файлы предоставляют наиболее быстрый путь для обнаружения каждого вхождения данного слова в коллекции документов. Большинство методов обработки информации используют эти файлы.

Поиск с использованием Булевых операторов

Поисковый запрос с использованием Булевых операторов состоит из двух или более поисковых терминов, соединенных логическими операторами, такими как, and, or, not. Рассмотрим запрос "abacus and actor" применительно к инвертированному файлу из врезки 10.6. Запрос включает два поисковых термина, соединенных Булевым оператором. На первой стадии выполнения такого запроса считывается инвертированный список для термина "abacus" (документы 3 и 19) и для "actor" (документы 2, 19 и 29). Затем эти списки сравниваются между собой. Оба слова встречаются только в документе 19, который и является единственным удовлетворительным ответом на запрос. В случаях, когда инвертированные списки короткие, такой поиск с несколькими поисковыми терминами почти также быстр, как и простой запрос. Однако требования к вычислительной мощности существенно возрастают для больших коллекций и сложных запросов.

Инвертированные файлы могут использоваться для расширения базовой концепции поиска с логическими операторами. Поскольку расположение слов в документах зафиксировано в инвертированных списках, их можно использовать для поисков, которые задают относительное положение двух слов, например сначала "West" и затем "Virginia". Можно их использовать и для поисков с использованием усечений слов (truncation) - т.е. всех слов, начинающихся на определенную букву. Во многих поисковых системах запрос "comp?" позволит найти все слова, которые начинаются с последовательности букв "comp", в частности, compute, computing, computer, computers, computation. К сожалению, такой подход не позволяет различать "посторонние" слова, начинающиеся с этой же последовательности, например, "company".

Поиск с логическими операторами является мощным инструментом, но позволяет находить только точные соответствия. Поиск слова "library" не найдет "libraries". Запросы "John Smith" и "J.Smith" будут рассматриваться как различные, хотя каждый понимает, что это одно и то же. Существует ряд методов для таких ситуаций.

Современный подход отказывается от попытки находить документы, которые абсолютно точно отвечают запросу, и пытается определить некую меру сходства между запросом и каждым документом. Предположим, что общее число различных слов в наборе документов равно n . Данный документ может быть представлен вектором в n -мерном пространстве. Если документ содержит данное слово, вектор принимает значение 1 в соответствующем направлении; в противном случае его значение равно 0. Запрос может быть также представлен вектором в этом же пространстве. Степень полноты, с которой документ отвечает запросу, измеряется степенью близости друг другу этих векторов. Это может быть угол между этими двумя векторами в n -мерном пространстве. После того, как эта величина рассчитывается для каждого документа, их можно ранжировать от максимально близкого к запросу до минимального. На этой общей концепции базируется несколько методов ранжирования. Некоторые вероятностные методы используют статистическое распределение слов в коллекции. Они основаны на предположении, что точные слова, выбранные автором для описания искомой темы или пользователем для формулировки запроса были выбраны с определенной уверенностью, хотя могут использоваться и другие слова.

Обработка текстов на естественном языке (natural-language processing) и компьютерная лингвистика

Слова в документе - это не просто последовательность знаков; это единицы языка (например, английского), и они упорядочены во фразы, предложения и абзацы. Обработка натурального языка является разделом кибернетики, который используют компьютеры для интерпретации и манипуляции словами, как частью языка. Инструменты "проверка правописания", которые используются в текстовых редакторах дают пример широко известного приложения; они используют методы обработки естественного языка для предложения альтернативных вариантов правописания в тех случаях, когда не могут распознать данное слово.

Одно из достижений компьютерной лингвистики (раздел лингвистики, который имеет дело с грамматикой и лингвистикой) - разработка компьютерных программ, которые способны выполнить грамматический разбор (parse) практически любого предложения с хорошей точностью. Анализаторы разбирают структуру предложения, относят слова к частям речи (существительное, глагол, прилагательное и пр.), группируют слова во фразы и статьи; идентифицируют структурные элементы (существительное, глагол, объект и пр.). Для этих целей лингвистам потребовалось уточнить их понимание грамматики, научиться распознавать значительно больше нюансов правописания, чем содержится в обычных грамматиках. Значительные исследования в области обработки информации были посвящены использованию сочетаний существительных (noun phrases). Во многих контекстах, содержание предложения может быть определено извлечением существительных, их фраз и проведения поиска по ним. Работы такого рода не ограничивались одним английским языком, а проводились для многих языков.

Анализ грамматической структуры требует понимания морфологии слов с тем, чтобы варианты, образованные от одного корня, например единственное и множественное число существительного, формы глагола, могли точно распознаваться. Для обработки информации зачастую более эффективно сводить морфологические варианты к общему корню и использовать его в качестве поискового термина. Эта процедура называется нормализацией (stemming). Она эффективнее усечения (truncation), поскольку способна различать слова с различным значением (такие, как "computer" и "company"), и понимая при этом, что слова

"computer" и "computing" - однокоренные. Поскольку в английском языке слова почти всегда начинаются с корня, эта процедура может быть выполнена для слов, отобранных процедурой "trimming" (т.е. для слов, которые начинаются с одинаковой последовательности знаков). При этом уточнение производится по последним нескольким знакам. В некоторых других языках, например, немецком, также можно сначала отбирать слова, начинающиеся с одинаковой последовательности.

Компьютерная лингвистика создала ряд словарей и других инструментов, включая тезаурусы и словники, для обработки естественного языка. Словник содержит информацию о словах, их морфологических вариантах и грамматическом использовании. Тезаурус связывает слова по значению. Некоторые из этих инструментов универсальны; другие (как например, Art and Architecture Thesaurus или MeSH в Медлайне) созданы для конкретных областей.

Словники и тезаурусы могут значительно облегчить обработку информации, распознавая слова не как некие последовательности символов. Например, они могут распознать синонимы (автомобиль и машина), связать общий термин и конкретный пример (наука - химия), связать технический термин и общеупотребительный вариант ("cranium" и "brain" для обозначения мозга).

Словники или тезаурусы всегда в работе и не могут быть закончены уже просто "по определению". Языки непрерывно меняются, а терминология в быстроразвивающихся областях меняется особенно стремительно.

Пользовательские интерфейсы и информационно-поисковые системы (information-retrieval systems).

Эффективность информационно-поисковых систем зависит от способности пользователей правильно работать с этими инструментами. Если пользователем является подготовленный сотрудник медицинской библиотеки или адвокат, чье образование включало знакомство с поисковыми системами, - эти критерии выполняются. Неподготовленные пользователи обычно значительно хуже формулируют запросы и хуже понимают результаты.

Концепция векторного пространства и вероятностные методы обработки информации наиболее эффективны в случае длинных запросов. Интересно было бы использовать в качестве запроса длинный текст, например, реферат, что было бы эквивалентным запросить систему найти документ, который бы точно подходил к нему. Многие современные поисковые системы очень эффективны, когда используются таким образом, но методы, основывающиеся на векторном пространстве или лингвистических методах требуют для своего успешного применения осмысленных, содержательных запросов.

На практике статистика показывает, что большинство запросов состоит из одного слова. Одной из причин этого является то, что большинство пользователей свои первые запросы выполняли в системах с Булевыми операторами, которые выдают ответы, точно подходящие под запрос - понятно, что в случае длинных запросов точное совпадение практически невероятно. Другой характеристикой ранних поисковых систем, которая приводила к коротким запросам, являлась зависимость от вычислительной мощности компьютера. Многие люди сохранили эти привычки, приобретенные на таких устаревших системах.

Однако тенденция использования коротких запросов не может быть объяснена только историческими причинами или теми крошечными окошками для ввода текста запроса, которые иногда предоставлялись в некоторых системах. Эта закономерность повторяется практически везде. Впечатление такое, словно существует запрет на длинные запросы. Другая неблагоприятная характеристика пользователей, также широко встречающаяся - лишь немногие люди читают даже простейшие инструкции. ЭБ не смогли наладить подготовку своих пользователей для выполнения эффективных поисков и они не используют весь потенциал, заложенный в существующих ныне системах.

Оценки эффективности (evaluation)

В области поиска информации существует давняя традиция оценки эффективности. Два давно известных критерия - точность (precision) (процент релевантных ответов, т.е. в какой степени набор полученных данных отвечает требованиям запроса) и полнота (recall) (процент правильных записей, т.е. насколько полно по запросу были найдены все удовлетворяющие записи). Каждый из этих критериев относится к результатам единичного поиска в заданном объеме информации. Предположим теперь, что в коллекции из 10000 документов есть 50 документов по определенной теме. Идеальный поиск должен выявить эти 50 и отвергнуть остальные. Реально поиск находит 25 документов, из которых 20 по теме (релевантные) и 5 посторонних. Тогда точность составит $20/25=0,8$, а полнота $20/50=0,4$.

Точность измерить значительно проще, чем полноту. Для вычисления точности понимающий человек просматривает каждый документ и решает, релевантен он или нет. В приведенном примере необходимо было проверить лишь 25 документов. В случае полноты нет иного выхода, как просматривать всю коллекцию на предмет выполнения поискового критерия - а это нетривиальная задача.

Критерии полноты и точности были очень важны для разработки методов поиска информации, но приняты они были тогда, когда компьютеры были медленнее и дороже, чем сейчас. Тогда поиск информации состоял из одного поискового запроса к большому массиву данных, а успех или неудача поиска представляла собой однократное событие. В настоящее время поиск обычно интерактивен. Пользователь формулирует запрос, проводит первоначальный поиск, просматривает результаты, уточняет запрос и повторяет цикл. На жаргоне информационных профессионалов это называется "поиск с человеком в петле обратной связи" (searching with a human in the loop). Полнота и точность такого поиска должна определяться по суммарным показателям, а не по одному запросу.

Критерии эффективности, такие, как точность и полнота, измеряют технические аспекты компьютерных систем. Они не могут измерить то, как пользователь взаимодействует с машиной или что считается адекватными результатами поиска. Многие более новые поисковые программы используют стратегию ранжирования всех возможных ответов. Это обеспечивает высокий уровень полноты (за счет большого количества нерелевантных ответов). Кроме того, нужно выработать критерии для измерения эффективности ранжирования при присвоении высшего ранга наиболее релевантному ответу.

Пользователи ищут информацию по различным поводам и используют различные стратегии поиска. Иногда они ищут конкретные факты, иногда исследуют тему. Весьма редко они сталкиваются со стандартной задачей: найти все ответы по четко определенной теме с минимальным количеством посторонних записей. При интерактивном подходе пользователи проходят серию шагов, комбинируя поиск, просмотр, интерпретацию и фильтрацию результатов. Эффективность поиска информации зависит от целей пользователя и от того, насколько полно ЭБ отвечает этим целям.

Врезка 10.7 **Tipster и TREC**

Tipster был продолжительным проектом, финансируемым DARPA, направленным на улучшение качества обработки текста. В центре внимания было несколько проблем, причем все они весьма важны для ЭБ: определение документов (что объединяет поиск информации в хранящихся документах и идентификацию релевантных документов из потока нового текста), извлечение информации (расположение конкретной информации в тексте), укрупнение (сокращение размеров документа или коллекции).

За годы своего существования этот проект эволюционировал от стандартной обработки информации к развитию компонентов, которые могли бы отвечать конкретным задачам и разработке соответствующей архитектуры, которые бы объединяли эти компоненты. Эта попытка была весьма амбициозной и базировалась на концепции, заимствованной из объектно-ориентированного программирования - определять стандартные классы для базовых

компонентов текстовых материалов (главным образом, документов, коллекций, атрибутов и аннотаций).

На ежегодных конференциях по обработке текстов (Text Retrieval Conferences, "TREC conferences") исследователи демонстрировали свои методы на стандартных тестовых текстах. Эти конференции представляют собой выдающийся пример количественного исследования ЭБ и проводятся под патронажем Национального института стандартов и технологий (NIST) и с помощью многих других организаций. Организаторы создали корпус из нескольких миллионов текстовых документов - более 5 гигабайт данных. Исследователи рассматривают свою работу как попытку создания стандартного набора задач (standart set of tasks). Одна задача - поиск корпуса документов по темам, сформированным группой из 20 воображаемых пользователей. Другой тест оценивает системы, которые выполняют стандартные запросы к потоку входящих документов. Среди участников этих конференций есть представители крупных компаний, мелких поставщиков информационных систем и университетских лабораторий.

Конференции предоставляют возможность для сравнения производительности различных методов, включая такие, как автоматическое составление тезаурусов, сложные методы определения "весов" терминов (term weighting), обработку естественного языка, обратная связь по релевантности, передовые способы машинного обучения. В последние годы оценивались такие аспекты обработки информации, как работа с китайскими текстами, речевые документы, кросс-языковой поиск. Другое направление составляют эксперименты с методами оценки интерактивного поиска.

Глава 11. Поиск распределенной информации

Одним из аспектов большой проблемы интероперабельности является поиск информации, распределенной на различных компьютерных системах. ЭБ в мире управляются различными организациями с различными стилями менеджмента и отношением к коллекциям и технологиям. Некоторые библиотеке обладают лишь ничтожной частью нужных пользователю материалов. Таким образом, пользователю нужна информация из разных коллекций и сервисов, предоставляемых различными источниками. Как может пользователь найти и получить доступ к информации в условиях, когда существует множество ее источников?

Распределенные вычисления (distributed computing) - это общий термин для технических аспектов координации отдельных компьютеров таким образом, чтобы они представляли согласованные услуги. Они требуют, чтобы различные компьютеры поддерживали некие общие технические стандарты. В случае, например, распределенного поиска, пользователь может искать во многих независимых коллекциях с помощью одного запроса, затем сравнивать результаты, выбирать наиболее интересные и отбирать нужные материалы из коллекций. Помимо общесетевых стандартов это требует методов идентификации коллекций, соглашений по формулированию запросов, техник пересылки запросов, значения возвращаемых результатов, а также методов обнаружения объектов, которые были открыты в результате поиска. Эти стандарты могут быть вполне формальными, благословленными официальными организациями, а могут быть локальными, разработанными маленькими группами исследователей или просто соглашением по использованию определенного коммерческого продукта.

Идеальный подход мог бы состоять в разработке исчерпывающего набора стандартов, который мог бы быть принят для ЭБ. К сожалению, это нереально, особенно учитывая стоимость и скорость происходящих изменений. ЭБ сейчас в состоянии "кипения". Каждая библиотека пытается улучшить свои коллекции, сервисы, системы - и при этом нет двух одинаковых библиотек. Изменение частей системы для поддержки новых стандартов требует времени. К тому времени, когда изменения будут внесены, может появиться новая версия стандарта или сообщество может вообще двинуться в ином направлении. Абсолютная стандартизация - это мираж.

С точки зрения интероперабельности, ЭБ сталкиваются с проблемой развития распределенных компьютерных систем в мире независимо управляемых и технически различных компьютеров. Это требует форматов, протоколов, систем безопасности просто для обмена сообщениями. Но это также требует семантических соглашений для интерпретации этих сообщений. Главная проблема в нахождении побудительных мотивов к добровольному объединению независимых ЭБ.

Принятие общих методов дает ЭБ дополнительные возможности, но это требует затрат. Некоторые затраты - финансовые: покупка оборудования и программного обеспечения, подбор и обучение персонала. Однако зачастую основные затраты - организационные. Очень редко можно изменить один изолированный аспект ЭБ. Введение новых стандартов требует соответствующих изменений в существующих системах, изменений в организации работ и взаимосвязей между поставщиками, а также других изменений.

Рисунок 11.1 иллюстрирует концептуальную модель, которая может быть полезной для размышлений об интероперабельности. В данном случае модель используется для сравнения трех методов распределенного поиска. Горизонтальная ось обозначает функциональность различных методов; вертикальная ось - стоимость принятия этих методов. Идеальный метод должен располагаться в нижнем правом углу графика, когда высокая функциональность достигается при невысоких затратах. Программы поиска в веб имеют умеренную функциональность; они широко используются из-за низкой стоимости. Онлайн-каталоги на базе MARC и Z39.50 имеют значительно более высокую функциональность, но

распространены существенно меньше из-за высокой сложности. Система NCSTRL (описанная на врезке 11.4) располагается где-то посередине.

Методы достижения интероперабельности можно условно разделить на три класса:

- Большинство методов, которые ныне широко применяются для достижения интероперабельности, имеет умеренную функциональность и низкую стоимость. Главные стандарты веб - HTML, HTTP и URL - обладают этими характеристиками. Их простота привела к широкой распространенности, но она же ограничивает их функциональность.
- Некоторые высокотехнологичные сервисы обладают прекрасной функциональностью, но весьма дороги. Примеры - Z39.50 и SGML. Эти методы популярны в закрытых сообществах, где ценится их функциональность, но их цена препятствует более широкому использованию.
- Многие текущие разработки в области ЭБ представляют собой попытки найти золотую середину: значительную функциональность при умеренной цене. Примеры включают Дублинское ядро, XML, Unicode. В каждом случае разработчики уделяли особое внимание умеренным затратам на внедрение этих систем. Дублинское ядро допускает, чтобы любое поле было необязательным (optional). Unicode содержит UTF-8, который может воспринимать существующие данные в ASCII. Стоимость XML снижена за счет тесных связей с HTML и SGML.

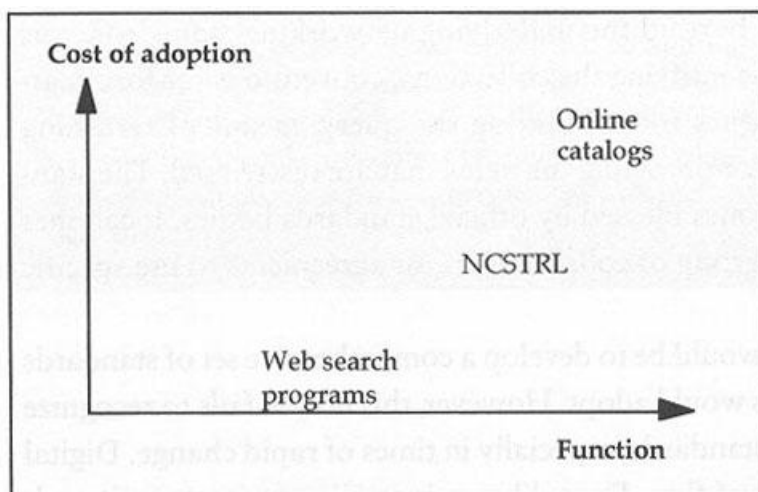


Figure 11.1
Comparison of three strategies for distributed searching.

Рисунок 11.1 приведен не в масштабе и размерности по осям "концептуальные", но он иллюстрирует фундаментальный принцип, согласно которому стоимость внедрения новой технологии является важным фактором в любом аспекте интероперабельности. Технологию никогда нельзя рассматривать изолированно, без учета организационных факторов. Когда создатели ЭБ надеются достичь совместимости с другими библиотеками, зачастую они сталкиваются с выбором между методами, наилучшими для их отдельного сообщества, и более широко используемыми стандартами, которые обладают меньшей функциональностью. Новые версии программного обеспечения иллюстрируют это противоречие. Зачастую новая версия придаст ЭБ большую функциональность, но лишь несколько пользователей будут иметь доступ к ней. Например, создатель веб-сайта может использовать лишь базовые тэги HTML, несколько самых популярных форматов и функции, поддерживаемые любой версией HTTP для того, чтобы сделать сайт, с которым сможет работать любой браузер. И наоборот, разработчик может выбрать последние версии веб-технологий, с апплетами Java, HTML-фреймами, встроенной системой обеспечения безопасности, таблицами стилей, аудио- и видео-вставками. Эти опции предоставят превосходный сервис пользователям, которые являются владельцами

наисовременнейших браузеров и имеют доступ по скоростной сети. Остальные же пользователи могут найти эти "навороты" бесполезными.

Поисковые программы для веб

Наиболее широко для распределенного поиска используются программы для веб, такие как, Infoseek, Lycos, Alta Vista и Excite. Это автоматические системы, которые индексируют материалы в Интернет. В соответствии с рисунком 11.1, они предоставляют умеренную функциональность с невысокими "барьерами" для применения: веб-сайты не должны предпринимать никаких специальных действий во время индексирования. Единственная "расплата" для пользователей - это необходимость видеть на экране рекламу. Такое сочетание функциональности и доступности делает эти программы исключительно популярными.

Большинство программ имеет сходную базовую архитектуру, хотя в деталях много различий. Заметным исключением является Yahoo, которая ведет свою историю от классификационных систем. Остальные программы состоят из двух основных частей: web crawler (веб-паук; поисковый робот), который строит индексы материалов в Интернете, и поисковой машины ("поисковик", retrieval engine).

Поисковые роботы (web crawler)

Эти роботы представляют собой индексирующие программы, которые последовательно двигаются по гиперссылкам и составляют списки страниц, которые они проходят. Они строят постоянно увеличивающийся индекс веб-страниц. Внутри системы существует список известных системе URL, чьи страницы либо уже индексировались, либо еще нет. Из этого списка выбирается URL HTML-страницы, которая еще не индексировалась. Содержимое страницы перегружается в центральный компьютер для анализа. Автоматическая программа индексирования обрабатывает страницу и создает индексные записи, которые затем добавляются к сводному индексу. Гиперссылки на другие страницы извлекаются, причем те, которых в сводных списках системы еще нет, добавляются в очередь на индексирование.

За этой простой принципиальной схемой скрываются различные проблемы и варианты их решения.

Один из вопросов - какой URL посещать следующим. В любой момент времени у робота есть миллионы страниц, которые необходимо исследовать, но нет критерия отбора следующей страницы. Такие критерии могут включать, например: какая страница индексируется в данный момент, сколько других URL на ней находится; являются ли они ссылками на новые домашние страницы или ссылками в рамках единой иерархии материалов данного сайта.

Самой большой проблемой, однако, является само индексирование. Поисковые машины используют автоматические индексирующие системы (описанные в главе 10) для создания записей, которые могут предоставляться пользователям. Последние сталкиваются при использовании автоматических индексов с самым главным - миллионами страничек, созданных тысячами людей с разными подходами к тому, как должна структурироваться информация. Обычно на веб-страничке есть лишь весьма скудные сведения для автоматического индексирования. Некоторые разработчики и издатели даже намеренно хитрят: они наполняют свои страницы терминами, которые (как они предполагают) будут чаще запрашиваться пользователями, надеясь на то, что ссылки на их сайты будут предъявляться в ответах на запросы среди первых. Без создания более четко структурированных страниц и наличия систематических метаданных качество индексных записей никогда не будет высоким. Хотя для простейших целей они вполне пригодны.

Поиск в индексе веб-страниц

Программы для поиска в веб позволяют пользователям проводить поиск в своих индексах с помощью тех методов обработки информации, которые были описаны в главе 10. Индексы организованы так, чтобы можно было эффективно выполнять поиск сразу по большому числу запросов. Поскольку индексные записи низкого качества, а пользователи, скорее всего, слабо подготовлены, поисковые программы следуют стратегии, согласно которой идентифицируются все записи, которые хотя бы приблизительно подходят под запрос и выводят их на экран, ранжируя тем или иным способом.

Большинство пользователей этих программ, скорее всего, находят их весьма полезными, но отметят ряд проблем. Алгоритмы ранжирования дают мало информации о том, на каком основании принимается решение о расположении записей в списке. В результате наивысший ранг может быть присвоен странице с минимальной ценностью для пользователя, а действительно полезные страницы будут далеко внизу списка. Программы индексирования с трудом распознают дубликаты, хотя и пытаются группировать похожие записи. А поскольку похожие записи обычно имеют сходный ранг, программы выводят длинные списки практически идентичной информации. Любопытный подход к ранжированию предложен в поисковой системе Google, которая подсчитывает гиперссылки. Google особенно эффективна при поиске вводных или обзорных материалов по конкретной теме.

Роботы исследуют веб постоянно и - вероятно - могут найти практически все. Однако важные материалы могут дожидаться индексирования месяцами после их появления в веб. С другой стороны, нельзя назвать правильным возвращение к материалам, если много индексных записей указывают на них, как на больше не существующие или перенесенные на другое место.

Врезка 11.1

Ранжирование страниц и система Google

Анализ цитирования (citation analysis) представляет собой инструмент, который часто используется в науке. Статьи, которые ссылаются друг на друга, очевидно, близки по тематике, а наиболее часто цитируемые статьи, вероятно, более важны, чем те, на которые не ссылались никогда. Lawrence Page, Sergey Brin и их коллеги из Стэнфордского университета использовали эту концепцию для веб, взяв за основу ранжирования страниц число гиперссылок. Они создали экспериментальную программу поиска, известную как Google.

В качестве примера, рассмотрим поиск по запросу "Stanford University" в различных программах поиска в веб. В Стэнфорде существует более 200 000 веб-страниц. Большая часть поисковых программ не может отделить действительно информативные и полезные страницы - от представляющих весьма относительный интерес. Все программы найдут огромное число страниц, отвечающих этому запросу, но в большинстве случаев ранжирование результатов будет неадекватным и сайты, которые большинство людей сочтут важнейшими, окажутся не в начале списка.

Когда запрос выполняется в Google, список первых 10 записей следующий:

Stanford University Homepage (www.stanford.edu/)

Stanford University Medical Center (www-med.stanford.edu/)

Stanford University Libraries & Information Resources (www-sul.stanford.edu/)

Stanford Law School (www-lland.stanford.edu/group/law/)

Stanford Graduate School of Business (www-gsb.stanford.edu/)

Stanford University School of Earth Sciences (pangea.stanford.edu/)

SUL: Copyright & Fair Use (fairuse.stanford.edu/)

Computer Graphics at Stanford University (www-graphics.stanford.edu/)

SUMMIT (Stanford University) Home Page (summit.stanford.edu/)

Stanford Medical Informatics (camis.stanford.edu/)

Большинство людей согласится, очевидно, что это хороший список.

Основной метод, которым пользуется Google, прост: более высокий ранг присваивается странице, на которую существует большее количество гиперссылок, а ссылки с таких страниц "вешат" больше, чем ссылки с других. Поскольку множество страниц в мире содержит ссылки

на Stanford Law School, эта страница имеет высокий ранг. Эта домашняя страница, в свою очередь, содержит ссылки на дюжину других страниц, включая домашнюю страницу университета, которая "набирает" свой ранг за счет ссылок на нее с "авторитетных" страниц.

Вычисления рангов требуют нестандартных, элегантных вычислений. Чтобы понять базовую концепцию, представим гигантскую матрицу, содержащую все страницы веб и идентифицирующую все страницы, связанные с ними. Сначала всем страницам присваивается одинаковый ранг. Новые изменения рассчитываются на основании числа ссылок на каждую страницу, "веса" ссылающихся страниц и пропорционально числу ссылок с каждой страницы. Эти ранги используются для следующей итерации - и так до тех пор, пока результаты двух последующих циклов не совпадут.

Реальные расчеты уточнили эту схему. В 1998 году Google содержал информацию о 25 миллионах страниц, отобранных в процессе анализа ссылок между страницами. Программа содержала "весовые факторы" для страниц без ссылок и для группы страниц, связанных только между собой. Отвергались страницы, генерируемые динамически CGI-скриптами. Тот факт, что система смогла проверить, проиндексировать и проранжировать все эти страницы всего лишь за 5 дней, ярко иллюстрирует вычислительные возможности современных стандартных рабочих станций.

Использование гиперссылок при ранжировании страниц помогает разрешить две проблемы, которые досаждают другим поисковым веб-программам. Если они не могут индексировать одновременно все страницы веб, как они должны определять очередность? Каким образом должны они ранжировать страницы ответа на простой запрос с тем, чтобы отдать приоритет наиболее полезной?

Другая проблема эффективности индексирования состоит в том, что поисковые роботы не могут индексировать материалы, к которым не имеют прямого доступа. Если веб-страница имеет ограничения доступа или является интерфейсом базы данных или коллекции ЭБ, индексирующие программы ничего не смогут узнать о тех ресурсах, которые скрываются за этим интерфейсом. Поскольку все большее число страниц превращается в Java-интерфейсы (или иные скрипты), программы индексирования теряют много высококачественной информации.

Эта слабость достаточно важна, но ее не стоит переоценивать. Практика - критерий истины. Опытные пользователи, используя комбинации инструментов и зачастую несколько поисковых инструментов веб, обычно могут найти необходимую им информацию. Программы далеки от совершенства, но они весьма хороши - а главное, они бесплатны.

Большинство поисковых сервисов выросли на базе исследовательских групп, но быстро коммерциализируются. Тот факт, что базовые сервисы и службы в Интернете бесплатные, имеет важнейшее значение как для него самого, так и для таких коммерческих компаний. Их стремление к прибыли ведет к засилью рекламы, они быстро двигаются на смежные рынки с использованием таких приемов, как лицензирование собственного программного обеспечения другим организациям, которые хотят строить индексы на своих веб-сайтах.

Менее приятные аспекты такой модели бизнеса заключаются в стремлении ограничить индексирование. Изначально индексирующие программы предназначались для обследования всего веб. Однако, по мере его роста и по мере того, как управление такими программами превращалось в бизнес, полнота ушла на второй план по сравнению с улучшением интерфейса и развитием смежных сервисов. Для построения реального высококачественного индекса Интернета и поддержания его в актуальном состоянии, требуются значительные инвестиции. Большинство компаний делает полезную работу, но если бы их мотивация к совершенству была сильнее, их индексы были бы еще лучше.

Федеративные цифровые библиотеки

Способы разрешения противоречия между функциональностью и ценой зависят от контекста. Иногда вполне допустимо выбрать простую технологию и обеспечить широкую, но

неглубокую интероперабельность. В другом случае более мудрым было бы избрать высокотехнологичные и дорогие методы. На такой выбор способны только высокомотивированные ЭБ - но именно они будут предоставлять максимальные функциональные возможности.

Термин "федеративная ЭБ" описывает группу организаций, работающих вместе на формальной или неформальной основе и поддерживающих определенный набор общих сервисов и стандартов, достигая таким образом максимальной интероперабельности между собой. Партнеры в такой федерации могут иметь различные системы и при этом поддерживать согласованный список сервисов. При этом необходимо согласовать как технические стандарты, так и политику (включая финансовые соглашения, правила защиты интеллектуальной собственности, безопасности и частной жизни).

Исследования в университете Иллинойса (Urbana-Champaign) представляют собой хороший пример, иллюстрирующий проблемы интероперабельности. За период 1994-1998 годов, являясь участниками Digital Libraries Initiative, команда исследователей из библиотеки Grainger Engineering этого университета попыталась создать федеративную библиотеку журнальных статей от различных крупных издательств. Поскольку каждый издатель предполагал использовать разметку SGML, представлялось, что это может послужить базой для такого проекта. Университет будет предоставлять ряд сервисов, например, поиск; коллекции будут находиться у издателей. Проблемы возникли из-за того, что издатели использовали SGML по-разному. Каждый из них имеет собственные определения типов документов. Университет был вынужден приложить большие усилия по пересмотру семантики DTD как для извлечения индексной информации, так и для построения согласованного пользовательского интерфейса. Это оказалось столь сложной задачей, что университет переориентировался на копирование всей информации в единую компьютерную систему с конвертацией в общую DTD. Если уважаемая высокопрофессиональная исследовательская группа университета столкнулась с такими трудностями, можно ожидать, что у других они будут не меньше...

Врезка 11.2

Федеративная библиотека научной литературы университета Иллинойса

Библиотека Grainger Engineering в университете Иллинойса является прототипом федерации коллекций научных статей. Работа началась в рамках Digital Libraries Initiative под руководством Bruce Schatz и William Mischo. К 1998 году экспериментальная коллекция состояла из 50000 журнальных статей из изданий IEEE, IEEE Computer Society, American Society of Civil Engineers, American Institute of Aeronautics and Astronautics, American Physical Society. Каждая организация предоставляла свои статьи в разметке SGML одновременно с выходом печатной версии журнала.

Этот прототип представлял собой вклад в практику концепций поиска информации в размеченных текстах, которые широко обсуждались, но мало использовались. На первом этапе (который был особенно тяжелым) предполагалось пересмотреть DTD, используемые различными издателями, каждый из которых работал с собственными DTD для представления структурных элементов своих документов. Некоторые различия были синтаксическими, например, тэг автора мог быть обозначен так: <author>, <aut>, <au>. Другие различия отражали существенные семантические отличия. Для индексирования и поиска было написано специальное программное обеспечение, которое классифицировало все тэги DTD согласно каноническим наборам. Интерфейсы ЭБ используют эти тэги с тем чтобы пользователь мог искать некий текст в данном контексте (например, в подписях к рисункам). В схеме на рисунке 11.1 использование этих наборов тэгов изображено справа сверху. Конвертирование разметки из всех коллекций в единый набор тэгов означает высокие затраты в случае присоединения к федерации новой коллекции - но расширяет функциональность.

Из-за технических трудностей, реально все документы были загружены в единое хранилище в университете Иллинойса. Тем не менее в планах на будущее значится использование в рамках федерации хранилищ индивидуальных издателей. Есть желание включить в коллекции библиографические базы данных, каталоги и другие индексы.

Даже первые результаты подтвердили плодотворность такого подхода для изучения потребностей пользователей. Предоставление пользователям более мощных методов поиска было с благодарностью отмечено потребителями - но при этом возрастали их запросы. Они указывали, что зачастую рисунки и математические выражения более информативны, чем резюме или заключения. Эти эксперименты снова показали, что у пользователей существуют большие проблемы при составлении точных поисковых запросов если отсутствует точный контроль над словарями слов, используемых в статьях, их рефератах и в поисковых системах.

Онлайновые каталоги и Z39.50

Многие библиотеки имеют онлайновые каталоги своих коллекций, которые свободно доступны через Интернет. Эти каталоги можно рассматривать как форму федерации. Каталогные записи составляются в формате MARC согласно Правилам англо-американского каталогизирования и библиотеки обмениваются ими для снижения издержек. Библиотечное сообщество разработало протокол Z39.50 (описанный на врезке 11.3) для нужд обмена и общего использования записей и распределенного поиска. В США, где Библиотека Конгресса, OCLC и Research Library Group являются наиболее активными в развитии и распространении этих стандартов, есть также много независимых реализаций Z39.50 на академических сайтах или у коммерческих поставщиков. Принадлежность к этой федерации, создававшейся многие годы, стоит больших денег, которые "образуются" в основном за счет экономии от построения и использования совместных каталогов.

Одно из принципиальных назначений Z39.50 заключается в обеспечении коммутации между серверами. Каталогная система большой библиотеки может использовать этот протокол для поиска нужной работы (ее копии или каталожной записи) в группе других библиотек. Пользователи могут пользоваться клиентом Z39.50 для поиска в нескольких каталогах последовательно или одновременно. Библиотеки и их клиенты получают значительную выгоду от такого использования общих каталогов, несмотря на то, что интероперабельность между публичными каталогами оставляет желать лучшего. Некоторые реализации Z39.50 имеют опции, отсутствующие в других версиях, однако первопричиной несовместимости обычно является то, что индивидуальные каталоги поддерживаются людьми, которые лояльны в первую очередь к своему местному сообществу. Поддержка других инструкций не является для них первоочередным приоритетом. Даже если организации совместно используют совместимые версии Z39.50, разница в организации каталогов и способах их представления для внешних пользователей остается.

Врезка 11.3 **Z39.50**

Z39.50 - это протокол, разработанный библиотечным сообществом, который позволяет одному компьютеру (клиенту) искать и обрабатывать информацию на другом (сервере). Протокол этот важен как технически, так и из-за его широкого использования в библиотечных системах. Концептуально он не связан с какой-либо особой категорией информации или типом баз данных - основное внимание при его разработке уделялось библиографическим данным. Большинство реализаций концентрируются на поиске по библиографическим данным в базах данных, состоящих из записей в формате MARC и представлении результатов пользователю.

Z39.50 базируется на абстрактном подходе к поиску в базах данных. Предполагается, что сервер хранит набор баз данных с поисковыми индексами. Взаимодействие строится на концепции "сессий". Клиент открывает соединение с сервером, выполняет последовательность операций и закрывает соединение. В ходе сессии и сервер, и клиент запоминают условия их взаимодействия. Важно понять, что клиент является компьютером. Применение Z39.50 для нужд конечных пользователей требует специального пользовательского интерфейса. В протоколе не содержится никаких утверждений относительно интерфейса и способа его взаимодействия с клиентом Z39.50.

Типичная сессия начинается с того, что клиент соединяется с сервером и они обмениваются исходной информацией, используя процедуру (facility) "init" (инициализация). В ходе этого обмена устанавливаются базовые соглашения, например, предпочтительный размер сообщений, возможно, условия аутентификации (хотя реальная процедура в составе стандарта не рассматривается). Затем клиент может использовать процедуру "explain" (пояснения) для выяснения на сервере списка доступных баз данных, списков полей, синтаксиса, поддерживаемых форматов и др.

Процедура "search" (поиск) позволяет клиенту направлять запросы к базе данных, например:

В базе данных "Книги" найти все записи, в которых точка доступа (access point) "заголовок" содержит значение "evangeline", а точка доступа "автор" содержит значение "longfellow".

Стандарт предоставляет несколько вариантов синтаксиса для конкретного поиска, но широко используются только Булевы операторы. Сервер выполняет запрос и строит "набор результатов" (results set). Особенностью Z39.50 является то, что сервер сохраняет эти наборы, а последующие сообщения от клиента могут содержать ссылки на них. Таким образом, клиент может модифицировать большой набор результатов последующими уточняющими запросами или запросить вывод любой записи из него без нового поиска во всей базе.

В зависимости от параметров поискового запроса, клиенту могут быть возвращены одна или большее число записей. Стандарт устанавливает несколько способов, которыми клиент обрабатывает наборы результатов, включая такие процедуры, как "sort" (сортировка) и "delete" (удаление). После завершения поиска следующий возможный шаг состоит в том, что клиент посылает запрос процедуры "present" (доставить). Этот запрос требует от сервера пересылки указанных записей из набора результатов в формате, указанном клиентом. Эта процедура имеет большое число опций для управления содержанием и форматами, а также для управления большими записями или наборами результатов.

В дополнение к базовым сервисам, Z39.50 содержит инструменты для просмотра индексов, управления доступом, управления ресурсами. Он поддерживает внешние сервисы, которые представляют большое число расширений. Это обширный и весьма гибкий стандарт.

NCSTRL и Dienst

Объединенный каталог (union catalog) представляет собой каталог, который содержит записи о материалах нескольких библиотек. Такие каталоги использовались в библиотеках задолго до появления компьютеров и позволяют выполнять распределенный поиск за счет объединения поисковой информации в одном месте (каталоге). Службы поиска в веб могут рассматриваться как объединенные каталоги веб, даже как один каталог с "сырыми" записями. Альтернативный метод распределенного поиска состоит в том, чтобы каждая коллекция имела свой поисковый индекс. Поисковая программа посылает запросы к этим отдельным индексам и объединяет результаты для представления пользователю.

Сетевая компьютерная реферативная библиотека (Networked Computer Science Technical Reference Library, NCSTRL) представляет собой федерацию коллекций ЭБ, которая весьма важна для ученых в области вычислительной техники. Она базируется на протоколе под названием Dienst. Для минимизации расходов на внедрение, этот протокол строится на нескольких технических стандартах, которые уже знакомы профессионалам, которые обычно являются продвинутыми пользователями Unix, Интернета и веб. Первая версия Dienst посылала поисковые запросы на все сервера. По мере увеличения их числа от этого подхода пришлось отказаться - если один сервер был недоступен, страдала вся система. Сейчас Dienst содержит сводный (master) индекс, который является разновидностью объединенного каталога.

Врезка 11.4

NCSTRL и модель распределенного поиска Dienst

NCSTRL представляет собой распределенную библиотеку результатов исследований в области вычислительной техники, главным образом, это технические отчеты. Сотрудничающие организации размещают свои коллекции на своих серверах. Доступ к ним осуществляется либо

по протоколу FTP, либо Dienst. Последний представляет собой распределенный библиотечный протокол, разработанный Jim Davis из Xerox Corp. и Carl Lagoze из Корнельского университета в рамках проекта "Технические отчеты по вычислительной технике", упоминавшегося в главе 4. Сначала в проекте участвовали пять университетов; к 1998 году число участников возросло до 100 из всех стран мира, причем в 43 из них функционировали Dienst-серверы. NCSTRL и Dienst предоставляют повседневные услуги и при этом являются тестовой площадкой для изучения проблем управления распределенной информацией. NCSTRL является одним из немногих примеров исследовательских групп, которые реально поддерживают ЭБ.

Архитектура Dienst разделяет сервисы ЭБ на четыре базовые категории: репозитории, индексы, коллекции и пользовательские интерфейсы. Он предоставляет открытый протокол, в котором определяются эти сервисы. Протокол поддерживает распределенный поиск в независимых коллекциях. Каждый сервер имеет индекс хранимых материалов. В ранних версиях Dienst при поиске в коллекциях пользовательский интерфейс посылал запрос на все сайты Dienst в поисках объектов, удовлетворяющих запросу; затем он ожидал ответов от всех серверов. Это была реализация распределенного поиска, наиболее близкая к идеалу, однако она столкнулась с проблемами из-за больших задержек ответов от некоторых серверов. Основной проблемой было то, что - с точки зрения пользователя - качество сервиса определялось самым плохим сервером в сети Dienst. Кстати, такой сервер в университете Карнеги-Мэллона был независимым. Если ответы не поступали, интерфейс по истечении времени ожидания продолжал работу. Даже если функционировали все серверы, зачастую наблюдались длительные задержки из-за проблем с Интернет-соединением на нескольких серверах.

Медленный поиск утомляет и раздражает пользователей. Отсутствие части коллекций представляет собой более серьезную проблему. Невозможность поиска во всех индексах означает, что исследователь может пропустить важную информацию просто из-за технических проблем на сервере.

Для разрешения подобных проблем Dienst был реконструирован. NCSTRL ныне разделен на регионы. Сначала было два региональных центра в США и четыре в Европе. В рамках этой региональной модели, сводный индекс находился на центральном сервере (Корнельский университет), а в региональных центрах находились полные копии. Все, что необходимо пользователю для поиска и обнаружения информации, находится на региональном сайте. С индивидуальными сайтами пользователь связывается только для получения конкретных хранящихся на нем материалов. Пользователь имеет возможность выбрать любой региональный центр (т.е. из-за особенностей Интернета он может иметь лучшую связь с более удаленным центром).

Исследования альтернативных подходов к распределенному поиску

Для федераций коллекций успех или неудача в равной мере определяются как техническими, так и организационными проблемами. Очевидно, что некоторые члены предоставляют сервис более высокого качества, чем другие, а уровень поддержки сильно различается. Однако качество сервиса не может определяться худшей организацией. Врезка 11.4 описывает реконструкцию системы Dienst для разрешения этой проблемы.

В каждой информационной службе заложены определенные предположения о сценариях поддержки, формах допустимых запросов, вариантах ответов. Они реализованы в таких инструментах, как поиск, просмотр, фильтры и выборка (extract). Пользователю необходима связная персональная информация; однако в распределенном мире внутренне согласованные источники информации могут различаться друг от друга. Каким образом конгломерат организаций может предоставить пользователю эффективный доступ к своим ресурсам? Каким образом службы, созданные на основе различных концепций (сценариев), могут предоставить эффективные средства поиска без поистине драконовской стандартизации? Даже в случае отдельной централизованной службы остаются технические проблемы. Они становятся гигантскими, если источники информации контролируются независимыми организациями.

Организационные проблемы столь велики, что заслоняют технические аспекты. За исключением жестких федераций, надежда на лучшее будущее связана с принятием технических соглашений параллельно с принятием стратегий, на которые постепенно могут перейти все организации. Для каждого метода должны существовать альтернативы низкого уровня (обычно, сохранение статус-кво) так, чтобы службы не прекращали продуктивную деятельность из-за отдельных систем. Так, NCSTRL считает своим базовым протоколом Dienst, но более половины ее сайтов размещают свои коллекции также и на обычных серверах с использованием FTP-протокола.

Большая часть исследований в области распределенного поиска заключается в попытках построения сводных каталогов на основе метаданных от создателей и издателей. Это было одним из мотивов создания Дублинского ядра. Необходимо объединить эти данные на компьютерной системе и сгенерировать на их основе поисковый индекс.

Другая область исследований имеет целью разработку методов сужения области поиска в самых перспективных коллекциях. Пользователи редко хотят искать во всех источниках информации в Интернете. Они хотят провести поиск в конкретных категориях, например, в каталогах монографий или индексах медицинских исследований. Возникает необходимость в кратких описаниях содержания отдельных коллекций. Это особенно важно когда доступ ограничен механизмами идентификации или оплаты. Если к источнику есть свободный доступ, внешняя программа (по крайней мере, теоретически) может генерировать статистический профиль типов материалов и используемого словаря. Если же внешние пользователи имеют доступ только через поисковый интерфейс, такой анализ невозможен.

Врезка 11.5

Архитектура Harvest

Harvest - это исследовательский проект по распределенному поиску под руководством Michael Schwartz из университета Колорадо. Хотя он окончен в 1996 году, архитектурные подходы, использованные в нем, остаются актуальными. Базовая концепция состояла в разделении основных функций центральной поисковой системы на несколько отдельных подсистем. Проект определил форматы и протоколы для коммуникации между этими подсистемами, было создано демонстрационное программное обеспечение для их реализации.

Центральной в архитектуре Harvest является концепция "сборщика" (gatherer), программы, которая собирает индексную информацию из коллекций ЭБ. Сборщики наиболее эффективны, когда инсталлированы на той же системе, что и сама коллекции. Каждый сборщик извлекает индексы из коллекций и передает в стандартном формате по стандартному протоколу программам, называемым "брокерами" (brokers). Последняя строит сводные индексы информации из разных коллекций.

Архитектура Harvest значительно более эффективна в сетях, чем традиционные методы индексирования, которыми пользуются поисковые роботы. При этом разработчики для повышения эффективности ввели кэширование и методы репликации; однако реальные преимущества - это лучший поиск и обнаружение информации. Все сборщики передают информацию с использованием определенного протокола (Summary Object Interchange Format, SOIF), но саму информацию они могут собирать по внутренним правилам отдельных коллекций. Если поисковые роботы могут обрабатывать только открытую информацию, сборщики могут получать привилегии доступа и индексировать закрытые коллекции. Они могут быть сконфигурированы под определенные базы данных и не ограничены только веб-страницами или конкретными форматами. Они могут включать словари или словники по отдельным дисциплинам. В итоге это дает большие преимущества.

Многие достоинства этой архитектуры утрачиваются, если сборщики не установлены на тех же серверах, где находятся коллекции ЭБ. С этой точки зрения Harvest наиболее эффективен для федеративных цифровых библиотек. В составе федерации каждая библиотека может поддерживать своего "сборщика" и передавать индексную информацию брокерам, которые выстраивают сводные индексы для всей библиотеки, объединяя таким образом преимущества локального индексирования с централизованным использованием этих индексов.

В 1996 году Luis Gravano из Стэнфордского университета изучал, каким образом клиент может объединять результаты из различных поисковых служб. Для этой цели он предложил протокол, названный STARTS. Он разрабатывался в рамках совместного проекта Стэнфордского университета и ряда ведущих Интернет-компаний. То, как охотно компании поддержали эту разработку показывает, что они рассматривают концепции как фундаментально важные для расширения возможностей поиска в Интернете. Относительно небольшие усилия по стандартизации могут привести к значительному улучшению этого сервиса.

В своем анализе Gravano рассматривал информацию в Интернете как большое число коллекций материалов, причем каждая организована различно и каждая со своей поисковой машиной (search engine). Фундаментальная концепция заключалась в том, чтобы позволить клиентам использовать все широкие возможности этих поисковых машин и тех коллекций, которые они обслуживают. Проблема в том, что эти программы различны, а коллекции имеют разные характеристики. Трудность даже не в том, что интерфейсы используют различный синтаксис, в результате чего запросы должны переформулироваться при передаче в другую систему. Различны сами базовые алгоритмы. Некоторые используют методы с Булевыми операторами; другие - методы с ранжированием результатов. "Поисковик", возвращающий ранжированный список, не дает указаний на то, каким образом вычислялись ранги. На самом деле, алгоритм ранжирования зачастую является коммерческим секретом. В результате, ранжированные списки из разных источников невозможно просто объединить. Ранжирование сильно зависит от набора слов, которые используются в коллекции; поэтому даже объединение результатов из двух источников, использующих одинаковый алгоритм ранжирования, сопряжено с трудностями. Протокол STARTS позволяет поисковым машинам обмениваться характеристиками их коллекций и рангами, которые они генерируют, поэтому клиентская программа может попытаться свести воедино результаты из разных источников.

После поиска

Обнаружение информации - это больше, чем поиск. Большинство людей пользуется некими комбинациями просмотра и систематического поиска. В главе 10 обсуждались потребности пользователей при поиске информации и трудности оценки эффективности поиска информации в интерактивных сессиях. Распределенные ЭБ усугубляют все эти проблемы.

Просмотр всегда был важным методом поиска информации в библиотеках. Это может быть просто осмотр полок с книгами. Более систематический подход заключается в том, чтобы начать с одного источника и дальше двигаться по тем документам, на которые они ссылаются. Многие журнальные статьи и ряд других материалов содержат списки литературы со ссылками на другие документы. Проследить эту цепочку - отличный метод поиска, но это превращается в тяжелую работу, если каждый документ представляет собой отдельный том. В компьютерных системах, с помощью гиперссылок эта концепция становится совершенной. Главный вывод заключается в том, что отслеживать цепочки ссылок существенно легче в ЭБ, а вот каталоги и индексы лучше в обычных библиотеках. Поэтому представляется, что просмотр как метод поиска информации в ЭБ имеет меньшее значение.

Если люди реализуют эвристическую комбинацию просмотра и поиска с использованием различных источников и поисковых программ, могут ли они доверять полученным результатам? Ранее в этой главе уже отмечалось, как трудно сравнивать результаты поиска в различных наборах информации или решить, являются ли два объекта, найденных в различных коллекциях, дубликатами. Серьезные пользователи ЭБ сталкиваются и с более серьезной проблемой: зачастую трудно понять, насколько полным был проведенный поиск. Пользователь крупной базы данных, такой как Medline в Национальной медицинской библиотеке, может быть уверен, что поиск охватывал все записи в индексах. Все значительно хуже при распределенном поиске в большом числе наборов данных. Каковы шансы на то, что важная информация будет утеряна по той или иной причине?

Распределенный поиск являет собой квинтэссенцию современного состояния ЭБ. С одной точки зрения, любая технология имеет серьезные недостатки. Технические стандарты отсутствуют, понимание пользователей находится в зачаточном состоянии, организационные трудности выглядят непреодолимыми. В тоже самое время огромные массивы информации доступны в Интернете, поисковые программы для веб доступны бесплатно, федеративные и коммерческие сервисы быстро расширяются. Разумно сочетая просмотр и поиск, целеустремленный пользователь обычно может найти необходимую ему информацию.

Глава 12. Модели объектов, идентификаторы и структурные метаданные

Информация поступает в различных формах и форматах, каждый из которых должен быть идентифицирован, найден и применен. Большинство из ранних разработок в области ЭБ были связаны с материалами, которые имели прямой аналог в виде физического объекта. Обычно эти материалы можно представить как компьютерный файл простой структуры. Возможности ЭБ значительно шире, они включают все, что может быть представлено в цифровом формате. Цифровая среда создает возможности для работы с новыми объектами, такими как, программное обеспечение, моделирование, анимация, видео, слайд-шоу, аудио, а также новые способы структурирования материалов. Вычислительная техника ввела свои собственные типы объектов, включая электронные таблицы, базы данных, символьную математику и гипертекст. Компьютеры и сети поддерживают непрерывно растущий поток цифровой информации - речь, музыку, видео. Цифровые объекты могут существовать во многих версиях и многократно реплицироваться.

Средства, обеспечивающие управление этой сложной средой, можно условно разделить на несколько категорий: идентификаторы для цифровых объектов; типы данных, которые специфицируют то, что представляют эти данные; структурные метаданные, которые отражают связь между цифровым объектом и его составными частями. В целом это создает "модель объекта" (object model), описание некоторой категории информации, что и позволяет компьютерным системам хранить и предоставлять доступ к комплексной информации. И поскольку информацией в нынешних ЭБ нужно будет пользоваться многие годы, причем люди будут работать на компьютерных системах, которые сейчас даже нельзя представить, интероперабельность и долговременное хранение (persistence) являются постоянными темами изучения в этой области.

Произведения, представления, публикации, экземпляры (Works, Expressions, Manifestations, and Items)

Пользователи ЭБ обычно предпочитают обращаться к объектам более высокого уровня абстрагирования, нежели файлы. Обычные слова, такие как, отчет, компьютерная программа, музыкальное произведение часто относятся ко многим цифровым объектам, которые должны рассматриваться вместе. Причем объекты могут быть в различных форматах, различаться содержанием, иметь различные ограничения на доступ к ним, но пользователь обычно рассматривает их как эквивалентные. Это требует введения концептуальной модели, которая позволила бы описывать содержание на разных уровнях абстрагирования.

В главе 1 отмечалась необходимость различения основной интеллектуальной работы и соответствующего объекта библиотеки, а также трудности, с которым связано такое различение с учетом всех возможных видов и форм представления произведения. В 1998 году Международная федерация библиотечных ассоциаций (International Federation of Library Associations) опубликовала отчет, в котором содержатся требования к библиографическим записям. В нем предложены следующие четыре уровня для описания содержания:

Произведение (work) Произведение - это основная абстрактная категория, например, "Илиада", "Пятая симфония Бетховена" или "Операционная система Unix".

Форма представления (Expression) Произведение реализуется в некоторой форме представления. "Илиада" изначально выражалась устно, в речи; затем она была записана как определенная последовательность слов. Музыкальное произведение может быть представлено письменно в виде нотации или исполнителем. Компьютерная программа, такая как Unix, имеет разные представления в виде исходного и машинного кода.

Публикация Представление является конкретной формой одной или нескольких публикаций. Текст "Илиады" был опубликован в многочисленных рукописях и книгах. Музыкальное исполнение может распространяться на компакт-дисках или транслироваться по телевидению. Программное обеспечение публикуется в виде файлов, которые могут храниться или передаваться в любой цифровой среде.

Экземпляр Когда существует много копий публикации, каждая из них - это отдельный объект (item), например, конкретный файл или экземпляр книги.

Очевидно, что эти четыре уровня содержат многочисленные нюансы. Разница между версиями, изданиями, переводами и другими вариантами - предмет разбирательства суда и не подчиняется твердым правилам. Некоторые работы - такие, как джазовые композиции (когда каждое исполнение является творчеством) или живопись (когда каждый объект является произведением) - трудно "втиснуть" в рамки такой модели. Тем не менее, эта модель является весьма плодотворной для описания сложных предметов.

Форма представления

Мультимедиа

В главах 9-11 было уделено особое внимание работам, представленным текстуально. Текстовые материалы особенно важны для ЭБ, но модели объекта обязаны поддерживать все носители и типы объектов. Некоторые нетекстовые материалы представляют собой файлы с фиксированной информацией, такие как цифровые эквиваленты знакомых объектов (например, карты, аудиозаписи, видео) и иные объекты, которые предоставляются пользователю путем прямого воспроизведения объекта хранения в ЭБ. Даже для таких материалов существует много тонкостей, если информация не текстовая. В этой главе рассмотрены три примера. Врезка 12.1 описывает проект ЭБ Alexandria по географической (geospatial) информации. Врезка 12.2 описывает, как ЭБ Infomedia объединяет несколько методов обработки информации для автоматического построения индексов и поиска в коллекциях видеофрагментов. По отдельности каждый метод не является точным, но в комбинации они позволяют достичь хороших результатов. Врезка 12.3 описывает RealAudio, способ доставки цифровых аудиозаписей пользователям. Звукозапись требует специальных методов распространения с тем, чтобы непрерывный поток данных поступал к пользователю в определенном порядке. Показано, что реально может быть передано через Интернет на сегодняшнем уровне технологии.

Врезка 12.1

Географическая коллекция в цифровой библиотеке Alexandria

ЭБ Alexandria, расположенная в университете Калифорнии (Санта-Барбара), возглавляемая Terence Smith, была одним из шести проектов, финансировавшихся в рамках Digital Libraries Initiative в 1994-98 гг. Коллекции Alexandria охватывают любую географическую печатную информацию, включая все виды карт, аэрофото- и спутниковую съемку, астрономические карты, базы данных и связанную текстовую информацию. Проект объединяет широкую программу исследований с практической реализацией университетской библиотеки карт.

Оказалось, что эти коллекции являются плодотворными областями исследований ЭБ. Географическая информация представляет ценность и интерес для многих отраслей: картография, планирование городов, навигация, исследования окружающей среды, сельское хозяйство, астрономия и пр. Данные поступают из многих источников: обзорные (survey) данные, цифровые фотографии, печатные карты и аналоговые изображения. ЭБ таких коллекций сталкивается с теми же проблемами, что и обычная текстовая ЭБ, но она вынуждена исследовать каждую проблему заново для того, чтобы понять, каким образом стандартные методы могут использоваться и что должно быть модифицировано.

В случае географических данных обработка информации сосредоточена на покрытии (coverage) или охвате (scope), тогда как для других категорий материалов внимание уделяется предметной области или библиографической информации (например, автор или заглавие). Покрытие означает географический регион, например, город Санта-Барбара или Тихий океан. Охват описывает виды информации, например, топографические особенности, политические связи, плотность населения. Alexandria предоставляет несколько методов для поиска подобной информации и использования для обработки.

Долгота и широта - это базовые метаданные для карт и другой географической информации. Были разработаны алгоритмические процедуры по извлечению этих данных из существующих карт. Создан большой географический справочник, состоящий из базы данных и набора процедур, которые предназначены для установления соответствий различных представлений географических объектов (например, географических наименований, особенностей рельефа, координат, почтовых кодов, данных переписей). Поисковая программа Alexandria специально адаптирована для поиска географических наименований. Исследователи добились значительного прогресса в трудной задаче по извлечению содержательных данных с использованием программ автоматической идентификации объектов на аэрофотографиях или на картах.

Исследователи в рамках ЭБ Alexandria применили высокопроизводительные методы вычислений. Один из используемых ими методов хранения информации есть и в составе пользовательских интерфейсов - т.н. Wavelets. Оцифрованные карты и другая географическая информация хранится в очень больших файлах. Указанный метод позволяет проводить послойную "разборку" (decomposition) изображения, причем на первом уровне генерируется маленькая иконка. Последующие уровни содержат все более детализированное изображение и требуют все больше данных.

В рамках проекта Alexandria было разработано несколько пользовательских интерфейсов, причем каждый базируется на опыте предыдущего. Две общих темы в этих работах - это ограничения, обусловленные маленькими размерами компьютерных экранов и медленная скорость доставки больших файлов пользователям через Интернет. Хороший дизайн позволил как-то обойти первое ограничение, хотя, безусловно, невозможно втиснуть большое изображение и совершенный поисковый интерфейс в один экран. Для улучшения производительности сделано все возможное для того, чтобы не повторять данные по сети дважды. В результате пользователь может прервать работу, а потом вернуться к тому месту, где остановился, без повторения предыдущих этапов.

Врезка 12.2

Infomedia: мультимодальный информационный поиск (multi-modal information retrieval)

Большая часть внимания в рамках проекта ЭБ Infomedia (уже упоминавшегося в главе 8) была сосредоточена на проблемах индексирования и поиска видеофрагментов без участия человека.

Infomedia - мультимодальна. Она использует компьютерные программы для анализа видеоряда, аудиоряда, заключительных титров (при наличии), изменения сцены (scene) и другой информации. Анализ каждого индивидуального аспекта дает несовершенную информацию, но их объединение оказывается удивительно эффективным.

Infomedia создается с использованием методов искусственного интеллекта, включая распознавание речи, обработку естественного языка и распознавание образов, объединяя усилия независимых исследовательских групп в нечто большее, чем простая сумма компонентов.

Первый шаг при добавлении новых материалов к коллекции состоит в разделении видеоматериала на тематические фрагменты. Компьютерная программа использует набор методов обработки образов и звука в поиске ключа - критерия того, что один фрагмент закончился, а другой начался. Например, в телевизионной программе перерыв для рекламы часто совпадает со сменой сюжета. Второй шаг - идентификация текста, ассоциированного с

каждым фрагментом. При этом выполняется распознавание звукового сопровождения, идентификация надписей и извлечение титров. Последний шаг - создание технического текста с использованием средств распознавания текстов на естественном языке. На основе этого текста генерируются индексные записи, которые загружаются в поисковую систему.

Методы открытия информации, описанные в главах 10 и 11, могут применяться к аудиопленкам и звуковому сопровождению видео (sound track), если удастся конвертировать звуки в компьютерный текст. Такая конверсия представляет собой сложную проблему, но за годы исследований был достигнут существенный прогресс, основывающийся на все увеличивающейся мощности компьютеров.

Infomedia необходимо сталкивается с труднейшими проблемами в распознавании речи. Речь в звуковом сопровождении может быть неразборчивой, в т.ч. из-за наложения музыки или звуковых эффектов, может содержать любые слова английского языка, включая сленг, а также слова из других языков. Человек даже пропускает некоторые слова. Infomedia успешно распознает от 50 до 80% слов в зависимости от характеристик конкретного видеофрагмента.

Для поиска в коллекции Infomedia пользователь может или набрать запрос или произнести его вслух и он будет обработан системой распознавания речи. Поскольку могут быть ошибки распознавания, а индекс строится на неточных данных, подходящие записи ранжируются. Алгоритм ранжирования и поиска тот же, что и в поисковой машине Lycos.

Врезка 12.3 **RealAudio**

Час звучания (с качеством компакт-диска) оцифрованного звука без компрессии требует 635 Мб дискового пространства, что создает проблемы для ЭБ. Требования к хранению сколь угодно значительных коллекций получаются просто огромными, а передача данных требует высокоскоростных сетей. Звук без компрессии с таким качеством представляет проблему даже для сетей со скоростью 155 Мбит/с. Поскольку большинство локальных сетей основаны на Ethernet со скоростью на порядок меньшей, а с использованием модемов скорость еще меньше, то абсолютно необходимы методы компрессии.

RealAudio представляет собой метод компрессии и одновременно протокол для передачи оцифрованного звука. В формате RealAudio час звучания требует около 5 Мб. Передача осуществляется по потоковому протоколу из репозитория в программу воспроизведения на пользовательском компьютере. После сигнала готовности от компьютера пользователя, репозиторий передает непрерывную последовательность пакетов со звуком. По мере прибытия на пользовательскую машину они превращаются в звук и проигрываются. Этот процесс выполняется в строго определенные интервалы времени без проверки ошибок передачи. Если пакет не прибыл в установленное время (для воспроизведения), он игнорируется, а пользователь слышит краткий разрыв воспроизведения. Процесс выглядит грубым - но если сеть достаточно хорошая, RealAudio вполне прилично передает речь по модемным линиям на 28,8 Кб. На ранних стадиях разработки, RealAudio использовался для хранения и передачи коллекции фрагментов из программ Национального Общественного радио (National Public Radio).

RealAudio пользуется стандартными методами веб, за исключением двух особенностей, которые обусловлены необходимостью передачи аудио сигнала через Интернет в реальном масштабе времени. Во-первых, пользовательский браузер должен принимать поток аудиоданных в формате RealAudio. Для этого необходимо добавить этот проигрыватель к браузеру (он доступен в Интернете). Во-вторых, для того, чтобы генерировать именно постоянный поток данных, ЭБ использует протокол UDP вместо TCP. (Поскольку системы сетевой безопасности в некоторых сетях не принимают пакеты UDP, RealAudio не может быть доступен отовсюду.)

Динамические и комплексные объекты

Многие из цифровых объектов ЭБ в настоящее время не могут быть представлены статическим файлом с данными. Рассмотрим следующие понятия:

Динамические объекты Динамические (или активные) библиотечные объекты включают компьютерные программы, апплеты Java, объекты моделирования, данные от научных сенсоров, видеоигры. Для этого типа объектов презентация для пользователя зависит от стадии исполнения компьютерной программы и других внешних факторов, так что клиент получает различные результаты (картинки) в различные моменты доступа к таким объектам.

Комплексные объекты Отдельный объект может быть собран из многих элементов, причем связанных по-разному. Это могут быть взаимодополняющие элементы (канал аудио и канал изображения видеозаписи)); альтернативные представления (спутниковая фотография в высоком и низком разрешении); смещанные (surrogates), такие как данные и метаданные. На практике такое строгое разделение не всегда удается соблюсти, например, как рассматривать маленькую иконку графического изображения, как альтернативную манифестацию или как метаданные о большом изображении?

Альтернативное представление (dissemination) Цифровые объекты могут предоставлять пользователю возможность выбора метода доступа. Например, библиотечный объект может сообщать сводку погоды в аэропорту Сан-Франциско. Когда пользователь обращается к нему, в качестве ответа он может получить количественные данные (температура, осадки, скорость и направление ветра, влажность) или фотографию затянутого облаками неба. Информация может непосредственно считываться с сенсоров (по запросу), а может поступать из регулярно обновляемых таблиц.

Базы данных Базы данных включают много различных записей, которые разные пользователи выбирают при обращении к базе. Некоторые базы данных представлены наилучшим образом именно как коллекция ЭБ с индивидуальными способами представления объектов. Другие базы, такие, как адресные книги, являются самодостаточными.

Методы управления такими объектами по-прежнему являются дискуссионными. Если веб предоставляет пользователям универсальное средство обращения к статическим файлам, то для объектов других типов таких средств пока нет. Даже терминология в этой области утверждается в дебатах. Набор соглашений, связанных с содержательным взглядом на внутреннюю структуру библиотечных материалов иногда называют "моделью документов"; однако, поскольку он приложим ко всем аспектам ЭБ, представляется, что термин "модель объекта" более точен.

Идентификация

Первая стадия построения модели объекта состоит в выборе метода идентификации материалов. Идентификатор необходим по отношению к объектам в каталогах или ссылках, для хранения или доступа к объектам. Выглядит это просто, но идентификаторы должны удовлетворять набору противоречивых критериев. Лишь немногие темы в ЭБ вызывают такие жаркие дискуссии, как имена и идентификаторы. Не существует универсального подхода к именованию. Одно противоречие связано с тем, необходимо ли включать семантическую информацию в имя. Некоторые отстаивают полностью семантические имена. Примером является стандарт Serial Item and Contribution Identifier (SICI). С помощью четко определенного набора правил, SICI идентифицирует как отдельные тома сериальных изданий или отдельные статьи. Обычно возможно сгенерировать SICI непосредственно из данных статьи или цитаты; однако это устрашающая цель, а стандарт SICI пользуется успехом только потому, что де-факто является единственным для идентифицирования заглавий сериальных изданий. Ниже приводится типичный SICI, идентифицирующий журнальную статью издательства John Wiley and Sons.

0002-8231(199601)47:1<23:NDOMII>2.0TX;2-2

Полностью семантические имена, такие, как SICI, неизбежно ограничены узкими классами информации; они обычно длинные и устрашающе выглядят вследствие высокой формализации

правил их генерации. Поскольку для более общих классов объектов (которые составлены из торговых марок и иных имен) трудно создавать семантические идентификаторы, некоторые ученые предлагают другой подход: использовать произвольные идентификаторы, не содержащие семантической информации о том, кто присвоил имя и что оно означает. Произвольные последовательности в качестве идентификаторов обычно короче, но без смысловой информации их труднее запоминать; их также труднее обрабатывать компьютерам.

На практике большинство имен - мнемонические. Они содержат информацию, которая позволяет легко их запоминать. Рассмотрим доменное имя `www.apple.com`. На первый взгляд, это семантическое имя (веб-сайт коммерческой организации под названием Apple), но это лишь догадка, основанная на некоей предыдущей информации. Префикс `www` обычно используется в именах веб-сайтов, но это просто соглашение. Существует несколько коммерческих организаций с названием Apple и приведенное имя не содержит никаких указаний на то, что этот веб-сайт поддерживается Apple Computer или иной компанией.

Другая трудность состоит в том, чтобы решить, должно ли имя относиться к производству, представлению, публикации или экземпляру. Рассмотрим международный стандарт книжных номеров (International Standard Book Numbers, ISBN), который был разработан издателями и книготорговцами для своих нужд. ISBN различает отдельные продукты, которые могут быть куплены или проданы; например, издание в твердом переплете обычно будет иметь иной ISBN, чем в мягком переплете, даже если содержание идентично. Библиотеки при этом могут найти подобные различия бесполезными. Для библиографических целей истинное различие - между версиями с различным содержанием, а не между форматами; с другой стороны, для целей управления коллекциями каждая копия является отдельной и нуждается в собственном идентификаторе.

Доменные имена и URL

В Интернете в качестве идентификаторов наиболее широко используются доменные имена и унифицированные указатели ресурсов (URL). URL расширили концепцию доменных имен в нескольких направлениях, но все расширения базируются на концепции именования местонахождения в Интернете.

Врезка 12.4

Доменные имена

Базовая цель доменных имен - идентификация компьютеров в Интернете по имени, отличному от IP-адреса. Одно из достоинств такого подхода заключается в том, что при замене компьютерной системы не нужно менять имени. Например, долгие годы доменное имя `library.dartmouth.edu` имели многие компьютеры с различными IP-адресами, но пользователи и не подозревали об этом.

С годами доменные имена стали более гибкими. Такое имя больше не должно относиться к одному конкретному компьютеру. К одному и тому же компьютеру могут относиться несколько доменных имен, или, наоборот, одно имя может означать сервис, выполняемый группой компьютеров.

Доменные имена образуют иерархию. На вершине находятся корневые имена. Один набор корневых имен базируется на типе организации, например `.com` для коммерческих, `.edu` для образовательных, `.gov` для правительственных, `.net` для сетевых, `.org` для иных организаций. Вторая серия корневых имен строится по географическому принципу; типичные примеры - `.ca` (Канада), `.jp` (Япония), `.ru` (Россия). Доменные имена организациям присваиваются в соответствии с их корневыми доменами. Типичные примеры - `cmu.edu` (университет Carnegie-Mellon), `elsevier.nl` (издательство Elsevier Science), `loc.gov` (Library of Congress, Библиотека Конгресса) `dlib.org` (D-Lib Magazine, сетевой журнал).

В США доменные имена исторически присваивались по принципу "первый пришел, первый получил". Контроль над тем, кто мог получить доменное имя и какими эти имена могли быть, был минимальным, а плата за них была очень низкой. В результате, человек, не имеющий отношения к городу Питтсбургу, мог зарегистрировать имя `pittsburg.net`. Оставляя в стороне

проблему появления неприемлемых (по разным причинам) имен, такая практика привела к возникновению споров по поводу зарегистрированных торговых марок и пр.

Врезка 12.5

Информация в URL

Набор знаков, называемый URL, жестко структурирован. Он объединяет название протокола, имя файла и опции, которые будут использованы для доступа к этому файлу. URL может содержать:

Протоколы Первая часть полного URL представляет собой имя протокола или сервиса, оканчивающегося двоеточием. Примеры: http:, mailto:, ftp:

Абсолютные и относительные URL URL может относиться к файлу по его доменному имени или по его положению относительно иного файла. Если за протоколом следует двойной прямой слэш, URL содержит полное доменное имя - например, <http://www.dlib.org/figure.jpg>. В противном случае адрес дается относительно текущей директории. Например, на HTML-странице якорь `<a href = "figure.jpg">` относится к файлу "figure.jpg" в той же самой директории.

Файлы URL идентифицирует конкретный файл на конкретной компьютерной системе. В URL <http://www.dlib.org/contents.html> "www.dlib.org" представляет собой доменное имя компьютера в Интернете, а "content.html" - файл на этом компьютере.

Порт Сервер в Интернете может предоставлять несколько сервисов одновременно. Протокол TCP позволяет указать "порт", идентифицирующий конкретный сервис, который необходимо использовать. Он указывается как двоеточие с последующими цифрами в конце доменного имени. URL <http://www.dlib.org:80/index.html> указывает на порт 80, который является портом по умолчанию для протокола HTTP и поэтому может не указываться в данном конкретном URL.

Параметры URL может содержать ряд параметров, например, "a # or a ?". Они передаются серверу после того, как получен доступ к файлу.

URL доказали свою исключительную полезность. Они позволяют создать любое количество приложений в Интернете. Однако в них содержится отложенная на будущее проблема для ЭБ - отсутствие согласованности. Пользователи ЭБ хотят иметь возможность постоянного устойчивого доступа к материалам в течение длительного времени. URL идентифицирует ресурсы указанием на расположение исходя из доменного имени. Если такое имя больше не существует или ресурсы были передвинуты в другое место, URL больше не является правильным. Знаменитый пример имел место в самом начале распространения веб. После того, как в 1994 году в Массачусетском технологическом институте был учрежден консорциум W3C, документация по веб была передана из ЦЕРНа (Швейцария) в MIT и все ссылки на ЦЕРН стали неверными. В таких случаях принято оставлять на старых местах ссылки на новое расположение ресурсов, но со временем переадресующий адрес либо ликвидируется, либо цепочка адресов становится слишком длинной. Если доменное имя ликвидировано (например, потому, что компания разорилась), все URL, базировавшиеся на таком доменном имени, теряют смысл. Для таких случаев существуют различные формы псевдонимов (aliases), которые могут использоваться с доменными именами и URL - но это скорее фокус, чем действительное решение.

Долговременно действующие имена и унифицированные имена ресурсов (Persistent Names and Uniform Resource Names (URN))

Сообщество ЭБ и издателей электронных материалов заинтересовано в долговременно действующих именах. Иногда их называют URN. Идея проста - имена должны быть глобально уникальными и существовать практически вечно, дольше чем любое нынешнее программное обеспечение и даже дольше, чем Интернет.

Долговременно действующее имя должно быть таким, что бы оно могло обозначать любой Интернет-ресурс или набор ресурсов. Одно из приложений URN - давать ссылки на текущее положение копий объекта, заданного списком URL. Другое - определять адреса электронной почты, которые не должны изменяться, если даже человек меняет работу или Интернет-провайдера. Другая возможность заключается в предоставлении открытых ключей к сервисам именования. Врезка 12.6 описывает систему handle, которая предназначена для генерации и управления долговременно действующими именами, а также то, как издатели используют ее для управления идентификаторами цифровых объектов (Digital Object Identifiers).

Компьютерные системы для разрешения (resolving) имен

Какая бы система идентификаторов не использовалась, должен существовать быстрый и эффективный метод, по которому можно было бы найти компьютер в Интернете. Это так называемое разрешение имен. В результате разрешения доменного имени выясняется (устанавливается) IP-адрес компьютерной системы с таким именем. Разрешение URN дает сведения, ассоциирующиеся с ним - обычно это URL, который указывает текущее место хранения данного ресурса.

Поскольку практически каждый компьютер в Интернете имеет соответствующее программное обеспечение, необходимое для разрешения доменных имен и обработки URL, некоторые исследовательские группы попытались использовать существующие механизмы для построения системы идентификации материалов в ЭБ. Один из подходов представлен OCLC. Это система PURL (P обозначает "долгоживущий"). PURL может быть, например, таким URL:

<http://purl.oclc.org/catalog/item1>

В этом идентификаторе purl.oclc.org является доменным именем компьютера, которое - как ожидается - должен быть долговременно существующим. На этом компьютере файл [catalog/item1](#) хранит URL на текущее расположение некоторого объекта "item1". Если объект передвигается, этот URL должен быть изменен, но PURL (внешнее имя) остается неизменным.

Врезка 12.6

Имена дескрипторов (handles) и идентификаторы цифровых объектов (DOI)

Handles - это система именования, разработанная в CNRI как часть более общей работы, предложенной Robert Kahn (CNRI) и Robert Wilensky (университет Беркли, Калифорния). Хотя эта система была разработана независимо от идей URN, они совместимы, и она может рассматриваться как первая URN-система для использования в ЭБ.

Система Handles имеет схему имен, которая позволяет независимым уполномоченным лицам создавать имена дескрипторов, причем будет уверенность в том, что эти имена являются уникальными. Реально это распределенная компьютерная система, которая хранит дескрипторы вместе с данными, которые они описывают) и набор административных процедур, обеспечивающих высокое качество информации в течение длительного времени.

Вот два типичных handles:

`hdl:cnri.dlib/magazine`

`hdl:loc.music/musdi.139`

Каждый из них состоит из трех частей. Первая указывает на тип - `hdl:`, следующая (`cnri.dlib` и `loc.music`) называется naming authority и назначается по иерархическому принципу, причем части `cnri` и `loc` - центральным уполномоченным органом, остальное (`dlib`, `music`) - на месте. Части после слэшей могут представлять собой любую последовательность знаков, уникальную для данного naming authority.

Разрешение имен дескрипторов осуществляется путем посылки в компьютерную систему и получения в качестве ответа хранимых там данных. В системе существует центральный компьютер, известный в качестве "глобального регистра имен дескрипторов". Однако, организации могут организовывать собственные службы handle и предоставлять услуги по разрешению. Существует единственное требование - naming authorities высшего уровня назначаются централизованно, а остальные должны быть зарегистрированы в центральном

компьютере. Для повышения производительности и надежности каждый такой сервис может поддерживаться группой компьютеров, а данные - автоматически дублироваться. Предоставляется также кэш-сервис, а также плагины (plug-ins) для веб-браузеров для разрешения имен дескрипторов.

В 1996 году по инициативе Ассоциации американских издателей (Association of American Publishers) дескрипторы были приняты в качестве метода идентификации электронных опубликованных материалов. Эти идентификаторы называют Идентификаторами цифровых объектов (Digital Object Identifiers, DOI). Эта инициатива привела к созданию международного фонда для дальнейшего развития концепции DOI. Цифровые naming authorities присвоены издательствам - например, 10.1006 означает Academic Press. Вот пример DOI книги, изданной Academic Press: doi10.1006/0121585328. Использование для naming authorities чисел отражает желание минимизировать семантическую информацию. Издательства часто реорганизуются, сливаются или переносят работы в другие издательства. Поскольку DOI позволяют такие изменения, они не должны содержать имя исходного издательства.

Концепция PURL добавила интересный момент в управление именами. Другие системы поименования должны иметь единый координационный центр имен для больших сообществ (вплоть до всего мира). Это можно рассматривать как "глобальный" подход. PURL - это более низкий уровень. Поскольку каждый PURL-сервер функционирует самостоятельно, нет необходимости координировать размещение имен между ними. Имена могут повторяться или использоваться с различной семантикой, в зависимости от решения, принимаемого на данном сервере. И наоборот, когда издательства разработали DOI, они выставили единую систему имен, гарантированно глобально уникальных.

Структурные метаданные и объектные модели

Типы данных

Типы данных - это структурные метаданные, используемые для описания различных возможных объектов в ЭБ. В главе 2 обсуждалась важность стандарта MIME для определения типов файлов при обмене посредством электронной почты или через веб. MIME является собой замечательный пример достаточно гибкого стандарта для того, чтобы удовлетворять нужды большого числа приложений, но в тоже время достаточно простого для реализации в компьютерных системах.

Объектная модель для веб включает файлы данных MIME-типа, связанные гиперссылками. Гиперссылки определяют связи между файлами; типы данных указывают пользователю интерфейсу как обрабатывать каждый файл для представления пользователю.

Комплексные объекты (Complex Objects)

Многие материалы в ЭБ сложнее, чем файлы, которые могут быть представлены простыми MIME-типами. Например, они могут состоять из нескольких элементов с различными типами данных (например, графическое изображение в тексте, или отдельно аудио- и видеотреки), могут быть связаны с другими материалами связями типа часть/целое или "последовательность". Цифровой текст может состоять из страниц, глав, индексов, иллюстраций и других элементов. Статья в онлайн-периодическом издании может храниться в компьютерной системе в виде нескольких файлов, содержащих текст и графику со сложными связями между ними. Поскольку цифровые материалы легко изменить, различные их версии могут создаваться постоянно.

Врезка 12.7 MIME

Спецификация, называемая MIME (Multipurpose Internet Mail Extension) была разработана Nathaniel Borenstein и Ned Freed для электронной почты, но оказалась полезной для широкого ряда Интернет-приложений. Это один из простых, но гибких строительных блоков, которые привели к успеху веб.

Полная MIME спецификация весьма сложна из-за того, что должна удовлетворять различными системам электронной почты, но для ЭБ важно некое ядро, концепция которого в MIME называется "тип содержимого" (Content-Type). MIME-описание состоит из трех частей, как, например: text/plain; charset = "US-ASCII".

MIME-тип состоит из собственно типа ("text" в примере выше), подтипа ("plain") и одного или нескольких параметров. Приведенный выше пример определяет гладкий текст в кодировке ASCII. Наиболее часто используются следующие типы данных: text/plain, text/html, image/gif, image/jpeg, image/tiff, audio/basic, audio/wav, video/mpeg, video/quickdraw.

"Тип приложения" представляет собой тип данных для указания каким приложением должны обрабатываться эти данные. MIME-типы для файлов в форматах PDF, Microsoft Word, PowerPoint - это, соответственно, application/pdf, application/msword и application/ppt. (Поскольку файлы Word могут содержать не только текстовую информацию и требуют специальных компьютерных программ для своей интерпретации, application/msword не рассматривается как формат.)

Приведенные примеры были формально составлены из зарегистрированных MIME-типов. Однако, возможно также создавать неофициальные типы и подтипы с помощью имен, начинающихся с "x-". Пример такого рода - audio/x-pn-realaudio.

Успех MIME являет собой показательный урок на тему превращения хорошей идеи в широко распространенную систему. Возможности MIME позволяют ей добиваться совместимости с большинством систем. Существующие почтовые системы в Интернет не требуют изменений для обработки сообщений MIME. Процессы проверки версии MIME, регистрации новых типов и подтипов, внесения изменений в систему были разработаны таким образом, чтобы естественно "стыковаться" со стандартными процедурами Интернета. Еще более важно, что MIME не пытается решать все проблемы типов данных и не привязан жестко ни к одному особенному приложению. Эта спецификация предоставляет гибкий набор сервисов, которые могут использоваться в различных контекстах. Величайший триумф спецификация MIME пережила при создании веб, хотя разрабатывалась она еще до его появления.

Структурные типы

Для пользователя объект представляется единым, а его внутреннее представление для него не имеет значения. Библиография или индексы будут восприниматься именно как такие отдельные единые объекты ЭБ. При этом внутреннее представление в ЭБ может быть более сложной конструкцией. Для представления различных компонентов и связей между ними используются структурные метаданные. Выбор структурных метаданных для конкретной категории материалов создает объектную модель.

Различные категории объектов требуют различных объектных моделей - например, текст в разметке SGML, объекты веб, компьютерные программы, оцифрованный звук. Для каждой категории правила и соглашения описывают способы организации такой информации в набор цифровых объектов. Например, представление цифровых звукозаписей описывается конкретными правилами. Для каждой категории правила описывают представление материала в библиотеке, группировку компонентов как группы цифровых объектов, внутреннюю структуру каждого компонента, ассоциированные метаданные, соглашения по наименованиям цифровых объектов. Такие категории различаются по структурному типу.

Объектные модели для компьютерных программ являются стандартной обязательной частью computing уже много лет. Большие компьютерные программы состоят из многих файлов с программами или данными, со сложной структурой и взаимосвязями. Эти связи описываются в отдельной структуре данных, которая используется при компилировании и

построении программы. Для программ под Unix такая процедура называется "сделать файл" (make file).

Структурные типы могут различаться по жанрам (genres). В поисковых задачах метаописание по жанрам является вполне обычным. (Например, некоторые жанры популярной музыки - джаз, блюз, реп, рок.) Жанры - это естественный и полезный способ описания материалов для поиска и других библиотечных целей, но для управления распределенными ЭБ требуется другая категоризация. Хотя художественный фильм, документальная лента и учебное видео явно представляют различные жанры, их цифровые эквиваленты могут кодироваться и обрабатываться абсолютно одинаково; они имеют один и тот же структурный тип. И наоборот, два текста могут быть одного жанра, но если один представлен в разметке SGML, а другой - в формате PDF, то они имеют различные структурные типы и модели объектов.

Врезка 12.8 описывает объектную модель для отсканированного изображения. Эта модель была создана для представления цифровых фотографий, но тот же структурный тип может быть использован для любых битовых изображений. Например, карта, афиша, техническая диаграмма представляют различное содержание, но с точки зрения хранения и обработки в компьютере они представляют собой одну структуру. В настоящее время можно полагать, что даже сложные коллекции ЭБ могут быть представлены небольшим числом структурных типов. Менее десяти структурных типов оказалось достаточно для адекватного описания всех категорий материалов, которые конвертировались в Библиотеке Конгресса. Среди них оцифрованные изображения, наборы изображений страниц, наборы изображений страниц со связанным SGML текстом, оцифрованные звукозаписи, оцифрованные видеозаписи.

Объектные модели, обеспечивающие интероперабельность (Object Models for Interoperability)

Объектные модели медленно эволюционируют. То, каким образом различные части должны группироваться вместе, трудно определить несколькими жесткими правилами. Решение зависит от контекста, специфики объектов, типа содержания, а иногда - и от реального содержания. После появления книгопечатания структура книг формировалась десятилетиями, прежде чем пришла к своей нынешней форме с титульной страницей, главами, рисунками и индексами. Неудивительно поэтому, что для ЭБ разработано лишь несколько соглашений по объектным моделям и структурным метаданным. Структурные метаданные все еще на стадии экспериментов с собственными спецификациями в каждой отдельной ЭБ, причем последние подвергаются изменениям по мере обретения опыта.

Это может показаться проблемой интероперабельности, поскольку программа-клиент не будет знать, какие метаданные использовались для хранения цифровых объектов в экспериментальном репозитории. Однако, клиентам и не нужно знать внутренние детали хранения объектов; они должны знать лишь функции, которые предоставляются хранилищем.

Рассмотрим хранение печатных материалов, которые конвертированы в цифровые форматы. Один объект в коллекции ЭБ состоит из набора изображений страниц. Если материал был также конвертирован в SGML, то должна присутствовать разметка SGML, DTD, таблица стилей и связанная информация. В репозитории структурные метаданные определяют связи между компонентами, но программе-клиенту эти детали не нужны. В главе 8 мы обсуждали пользовательские интерфейсы, которые обрабатывают последовательности изображений страниц. Их функции предусматривают демонстрацию указанной страницы и переход к странице с определенным номером. Пользовательский интерфейс, который "знает" о функциях, поддерживающих этот структурный тип информации сможет предоставлять этот материал пользователю без деталей его хранения.

Доставка (Dissemination)

В ЭБ форма хранения и форма доставки информации пользователю обычно не совпадает. В модели доступа для веб информация копируется с сервера на пользовательский компьютер, где

и выводится на дисплей. Такая обработка обычно предполагает конвертирование данных из файла в экранное изображение с использованием подходящих цветов и шрифтов, которое затем выводится на дисплей с окнами, меню и "иконками".

Получение хранящейся информации и представление пользователю ЭБ может быть значительно сложнее, чем в описанной веб-модели. По меньшей мере, общая архитектура должна разрешать обработку хранящейся информации непосредственно на сервере перед пересылкой клиенту; или программой-клиентом перед предоставлением пользователю. Данные не обязательно должны храниться непосредственно на сервере, т.е. программы по извлечению нужных данных из хранящейся информации и передаче пользователю выполняются на многих серверах. Эта информация может быть полностью отформатирована сервером, но зачастую она передается в виде файла в формате HTML (или ином промежуточном формате), который может распознаваться пользовательским компьютером.

Основные способы распространения включают прямое взаимодействие между клиентом и хранящимися цифровыми объектами, а также непрерывные потоки данных. Когда пользователь взаимодействует с информацией, его доступ не является единичным дискретным событием. Хорошо известный пример такого рода - видеоигры; тоже верно и для любых интерактивных материалов, как например моделирование физических ситуаций. Доступ к такой информации представляет собой серию взаимодействий, которые управляются пользователями и структурными метаданными объектов.

Часто программы-клиенты могут выбирать способ распространения. Выбор может определяться оборудованием - например, пользователь в медленной сети иногда выбирает текстовые версии веб-страниц. Это может определяться программным обеспечением на пользовательском компьютере - если текст, к примеру, доступен и в HTML, и в PostScript, пользователь, у которого отсутствует программа для просмотра PostScript, выберет HTML. Причины могут быть и не техническими, вроде пользовательских привычек и предпочтений - например, пользователь может предпочесть короткую версию видеопрограммы полной версии.

Одной из целей объектных моделей является предоставление пользователю набора методов доставки. Некоторые исследовательские проекты изучали варианты предоставления пользовательскому компьютеру возможностей автоматически получать информацию о доступном наборе способов и выбирать один наиболее полезный в данный момент времени. Обычно, предусматривается способ "по умолчанию" - то, что пользователь получит, если не выберет сам. Другая опция - короткое резюме; зачастую это просто строка текста (возможно, автор и название), но это может быть и что-то более наглядное (маленькая "иконка" изображения или фрагмент видео). Какой бы метод не рассматривался, цель резюме - предоставить пользователю некоторую информацию, которая помогла бы ему идентифицировать и описать объект. Это все еще область для исследований, на практике существует лишь несколько хороших примеров таких систем.

Глава 13. Репозитории и архивы (Repositories and Archives)

В этой главе рассматриваются методы хранения цифровых материалов в репозиториях и их долгосрочное архивирование. Рассказывается также о протоколах, которые предоставляют доступ к материалам в репозиториях. Может показаться странным, что такие важные темы затрагиваются в конце книги, поскольку долговременное хранение является одной из центральных задач ЭБ. На это есть свои причины. На всем протяжении книги особо подчеркивалось, выделялось то, что сегодня реально уже существует. Исследовательские темы затрагивались только в необходимых случаях, но большей частью обсуждались системы, которые используются в библиотеках сейчас. Темы данной главы менее конкретны и определены. Помимо вездесущих веб-серверов, согласие относительно хранилищ ЭБ практически отсутствует. Область цифрового архивирования - новая. Ее потребности начинают осознаваться, но специфические методы находятся в зародышевом состоянии.

Репозитории

Репозиторий - это любая компьютерная система, чья основная функция заключается в хранении цифровых материалов для использования в библиотеке. Репозитории - это книжные полки ЭБ. Они могут быть гигантскими или крошечными, хранить миллионы цифровых объектов или только один объект. В некоторых контекстах подвижный "мобильный агент" (mobile agent), содержащий лишь несколько цифровых объектов может рассматриваться как хранилище, хотя большинство репозиторий находятся на компьютерных системах, которые хранят информацию в файловых системах или базах данных и представляют ее миру посредством четко определенных интерфейсов.

Веб-серверы

В настоящее время наиболее распространенным способом доступа к коллекциям являются веб-серверы. Некоторые фирмы поставляют прекрасные серверы. Главные различия между их продуктами связаны прежде всего с составом программного обеспечения, включая электронную почту, программы индексации, системы безопасности, механизмы электронных платежей и другие сетевые сервисы.

Врезка 13.1

Веб-серверы

Веб-сервер - это компьютерная программа, чья задача заключается в хранении файлов и ответе на запросы, поступающие по HTTP и близким протоколам. Веб-сервер функционирует на компьютере, подключенном к Интернету. Компьютер быть может полностью задействован для этой цели или выполнять одновременно и другие приложения. Для маленького веб-сайта это может быть простой персональный компьютер.

Сердце веб-сервера - это процесс под названием httpd. Буква в конце означает слово "демон" (daemon). Демон - это программа, которая выполняется непрерывно, но большую часть времени просто ожидает событий (сообщений для обработки). Протокол HTTP функционирует поверх протокола TCP, транспортного протокола Интернета. TCP предоставляет каждому компьютеру несколько "портов". Веб-сервер связан с одним из таких портов, обычно это порт 80, хотя можно использовать и другие. Когда сообщение поступает на этот порт, оно передается демону. Последний запускает процесс обработки данного конкретного сообщения, а сам продолжает следить за поступлением новых сообщений. Таким образом, одновременно могут

обрабатываться несколько сообщений, а демон свободен от необходимости наблюдать за деталями этих процессов.

Реальный процесс, который выполняет веб-сервер, управляется HTTP-протоколом. Ранние версии этих программ, в сущности, выполняли просто команду `get` (получить). Эта команда получает от клиента сообщение, содержащее URL, который указывает на файл, хранящийся на сервере. Сервер находит файл и возвращает его в качестве ответа на запрос клиента. После этого HTTP-соединение для данного конкретного обмена посланиями прекращается.

По мере того, как веб-сайты становились больше, веб-серверы становились сложнее. В дополнение к полному набору команд HTTP, они поддерживают CGI-скрипты и другие расширения. Одно из требований к веб-серверам (и к веб-браузерам) заключается в сохранении поддержки предыдущих версий HTTP-протокола. Они должны быть готовы к получению сообщений от любой версии протокола и правильно обрабатывать их. Затем, веб-серверы имеют постоянно усложняющуюся систему обеспечения безопасности. Версия 1.1 протокола включает также "сохраняемые" (persistent) соединения, которые позволяют обрабатываться несколько HTTP-команд через одно TCP-соединение.

Самые крупные веб-сайты столь загружены, что одного компьютера недостаточно. Есть несколько методов разделения нагрузки. Наиболее очевидный - просто продублировать данные на нескольких идентичных серверах. Это возможно, когда число запросов велико, а объем данных относительно невелик. Для балансирования нагрузки применяется метод под названием "DNS round robin", который представляет собой расширение системы доменных имен, позволяющее использовать одно доменное имя для группы компьютеров с различными IP-адресами. Например, доменное имя `www.cnn.com` относится к группе компьютеров, на каждом из которых работает веб-сайт Cable News Network. Когда пользователь обращается на этот сайт, система доменных имен выбирает один из компьютеров для обслуживания запроса. Дублирование веб-сайтов становится неудобным, если объем данных велик или быстро изменяется. Примером являются системы поиска в веб. Некоторые из них используют отдельные компьютеры для выполнения запроса, "сборки" страницы, которая будет возвращена пользователю в качестве ответа и для добавления рекламы.

В случае ЭБ веб-серверы обладают умеренной функциональностью при низкой стоимости. Такие характеристики привели к широкому использованию этих программ и достижению базового уровня интероперабельности. Сам по себе веб столь успешен именно благодаря простоте, а веб-серверы являются частью этого успеха. К сожалению, некоторые упрощения и допущения в таких программах неудобны для разработчиков ЭБ. Веб-серверы поддерживают только одну объектную модель: иерархическая файловая система, в которой информация организована в отдельных файлах. Ее обработке внутренне присуща забывчивость (stateless); каждое сообщение получается, обрабатывается и забывается.

Современные репозитории (Advanced Repositories)

Несмотря на широкое применение веб-серверов, в качестве репозитория для ЭБ используются другие системы промышленной обработки данных. В коммерческой обработке данных стандартом для управления большими объемами являются реляционные базы данных. Они основываются на модели объекта, которая состоит из таблиц данных и связей между ними. Эти связи позволяют объединять данные из разных таблиц различными способами. Таблицы и поля данных в реляционной базе данных определяются схемой и словарем данных. Эти базы данных прекрасно подходят для управления большими объемами данных с жестко определенной структурой. Многие крупные издательства размещают свои коллекции в реляционных базах данных, а веб-сервер используется в качестве интерфейса между коллекциями и пользователем.

Каталоги и индексы для ЭБ обычно создаются с использованием коммерческих поисковых систем. Такие системы имеют индексы, которые связывают их с объектами. Обычно они базируются на гибкой и изощренной модели индексирования информации - но на примитивной модели реального содержания. Многие из них начинались как полнотекстовые системы,

потому их сильные стороны связаны с поиском в больших объемах текста. Некоторые системы имеют такие сервисы, как ранжирование по степени релевантности и поиск по полям - их создатели надеются поднять таким способом их функциональность, а значит и объем продаж.

Реляционные базы данных и коммерческие поисковые системы представляют удобные инструменты для таких задач, как загрузка данных, формально-логический контроль (validating), манипулирование, защита. Управление доступом очень жесткое, предоставляются сервисы, весьма важные в бизнес-приложениях (например, audit trails). Наблюдается общая тенденция добавлять системам управления базами данных возможности полнотекстового доступа, а для поисковых систем - некоторые аспекты модели реляционной базы данных. Эти дополнительные возможности могут быть полезными, но компании пока еще только разрабатывают системы, которые объединили бы сильные стороны обоих подходов.

Хотя некоторые ЭБ успешно используют реляционные базы данных, реляционная модель данных недостаточно гибка для разработки разнообразных моделей объектов. Среди лидирующих ЭБ существует общее понимание того, что необходимы более современные репозитории. Вот возможный набор требований к такому репозиторию:

Скрытие информации Внутренняя организация репозитория должна быть скрыта от компьютеров-клиентов. Должны быть доступными такие действия, как реорганизация коллекций, изменение ее внутреннего представления, перенос на различные компьютеры без всяких внешних проявлений.

Модели объектов Хранилища должны поддерживать гибкий набор моделей объектов лишь с небольшими ограничениями для данных, метаданных, внешних и внутренних взаимосвязей. Новые категории информации не должны приводить к фундаментальным изменениям других аспектов ЭБ.

Открытые протоколы и форматы Клиенты должны общаться с репозиториями с помощью четко определенных протоколов, типов данных и форматов. Архитектура репозитория должна допускать изменения протоколов (по мере их эволюции). Это, в частности, относится к управлению доступом. Репозиторий должен поддерживать широкий набор политик доступа, которые были бы применимы на любом уровне, для любых групп пользователей.

Надежность и производительность Репозиторий должен обеспечивать хранение очень больших объемов данных, быть абсолютно надежным и производительно работать.

Метаданные в репозиториях

В репозиториях находятся и данные, и метаданные. Последние можно рассматривать как классы описательных, структурных и административных. Для различения элементов цифровых объектов могут потребоваться идентификаторы. Хранение метаданных требует гибкости, определяемой, в частности, следующими особенностями:

- Описательные данные часто хранятся в каталогах и индексах, управляемых извне репозитория. Эти каталоги и индексы могут содержаться в отдельных репозиториях и охватывать материалы многих независимых ЭБ. Для ассоциирования метаданных с соответствующими данными используют идентификаторы.
- Структурные и административные метаданные часто хранятся с собственно цифровыми объектами. Например, эти метаданные могут "встраиваться" в объект.
- Некоторые метаданные относятся к группе объектов. Административные данные для управления доступом могут использоваться как для всего репозитория или для отдельных коллекций в его составе. "Подсказки для поиска" (finding aids) пригодны для многих объектов.
- Метаданные могут храниться как отдельные цифровые объекты со ссылками на них от объектов, с которыми они связаны. Некоторые метаданные вообще физически не хранятся - а генерируются в случае необходимости.

В каждой ЭБ есть своя собственная концепция выбора и определения метаданных, хотя проблема обмена метаданными является центральной в достижении интероперабельности. Концепция Warwick, описанная на врезке 13.2 представляет собой попытку сбалансировать эту чреватую хаосом ситуацию.

Врезка 13.2

Концепция Warwick

Концепция Warwick ведет свое происхождение от некоторых идей, высказанных на рабочей встрече в университете Warwick (Англия) в 1996 году. Основная цель - организация метаданных. К одному цифровому объекту может прилагаться огромный массив метаданных, включая такие описательные метаданные, как каталоги MARC, метаданные управления доступом, структурные метаданные, а также идентификаторы. Участники рабочей встречи предложили организовать метаданные в "пакеты" (package) - например, один пакет для Дублинского ядра, другой для географических данных. Такая специализация облегчает достижение интероперабельности: если клиент и репозиторий способны обрабатывать пакеты метаданных одного типа, они способны успешно взаимодействовать, даже если другие пакеты у них разные.

Carl Lagoze из Корнельского университета и Ron Daniel из Национальной лаборатории в Los Alamos подхватили эту идею и разработали элегантный взгляд на все компоненты ЭБ. Первое положение состояло в том, что различие между данными и метаданными зачастую совсем неочевидно. Аналогично, является ли "Содержание" книги частью содержания книги или это метаданные о содержании? В концепции Warwick такие различия неважны. Все подразделяется на пакеты и при этом не делается разницы между данными и метаданными. Следующее положение - не каждый пакет должен храниться явным образом как часть цифрового объекта. Описательные метаданные зачастую хранятся отдельно в виде каталожных записей или индексов. Определения и условия, приложимые ко многим цифровым объектам, зачастую лучше хранить в отдельных записях "политики" (policy record), а не встраивать в каждый отдельный объект. Такое разделение может быть обеспечено для счет введения косвенных пакетов. Пакет хранится в наиболее удобном месте, а ссылка на него хранится в репозитории. Это может быть простой указатель на местонахождение либо ссылка на компьютерную программу, результатом работы которой будет создание пакета согласно требованию.

Цифровые объекты одного и того же структурного типа обычно будут состоять из конкретных групп пакетов. Это формирует модель объекта, обеспечивающую интероперабельность между клиентами и репозиториями для данного типа цифрового объекта. Концепция Warwick в явном виде еще не применялась для крупных систем, но идея выглядит многообещающей. Разделение информации на четко определенные пакеты упрощает спецификацию цифровых объектов и повышает гибкость и интероперабельность.

Протоколы, обеспечивающие интероперабельность

Для достижения интероперабельности требуются протоколы, которые клиенты могут использовать для отправки сообщений в репозитории, а последние - для отправки информации клиентам. На базовом уровне необходимы функции, которые размещают информацию в репозитории и предоставляют доступ. Реализация эффективных систем требует, чтобы клиент мог получать представление о структуре цифровых объектов. Различные типы объектов требуют различных методов доступа, а управление доступом может потребовать идентификации или "переговоров" между клиентом и хранилищем. Кроме того, клиенты могут пожелать провести поиск в индексах репозитория.

В настоящий момент самым распространенным протоколом ЭБ является HTTP, протокол доступа в веб. Другим широко применяемым протоколом является Z39.50.

Объектно-ориентированное программирование и распределенные объекты

Одно из направлений исследований - создание простейшего протокола для репозитория, который бы поддерживал все необходимые функции. Если протокол простой, информация о

сложных композиционных типах объектов должна содержаться в самих объектах. (Такой подход был назван SODA, smart object, dumb archives - умный объект, глупый архив.)

Некоторые современные проекты разрабатывают архитектуры, использующие компьютерную концепцию распределенных объектов (distributed objects). В этом контексте слово "объект" имеет исключительно техническое значение, которое отличается от значения этого слова в выражениях "электронный объект" и "библиотечный объект", широко используемых в данной книге. В современной вычислительной технике под объектом понимается независимый фрагмент компьютерного кода с ассоциированными данными, который может быть по-разному использован в разных контекстах. Данные и методы их обработки скрыты внутри объекта, так что все внутренние детали спрятаны. Все, что окружающий мир знает о классах объектов - это открытый интерфейс, состоящий из методов (т.е. операций над объектами) и экземпляров данных (instance data). Результат каждого метода может изменяться от класса к классу. Например, в ЭБ метод "render" может по-разному интерпретироваться для различных классов объектов.

После десятилетий развития, такие языки объектно-ориентированного программирования, как C++ и Java стали общепринятыми для построения самых эффективных компьютерных систем. Движущей силой развития объектно-ориентированного программирования является сложность области. Такое программирование позволяет разрабатывать и тестировать компоненты программы независимо, при этом не требуется их адаптация под различные версии компьютерных систем. Активным сторонником объектно-ориентированного программирования является фирма Microsoft. Различные версии ее объектно-ориентированной среды называются OLE, COM, DCOM, Active-X. Все это варианты одной ключевой концепции.

Распределенные объекты представляют идею объектов в сетевом окружении. Базовая концепция заключается в том, что объект, исполняемый на одном компьютере должен иметь возможность взаимодействовать с объектом, исполняемым на другом, через свой открытый интерфейс, определенный в терминах методов и данных. Лидирующие компании-разработчики по созданию программного обеспечения (исключение составляет Microsoft) разработали стандарт для распределенных объектов, известный как CORBA. CORBA предоставляет разработчикам распределенных компьютерных систем практически тоже самое разнообразие возможностей программирования, которые объектно-ориентированные языки дают для одиночного компьютера.

Врезка 13.3 **HTTP**

Метод доступа HTTP get (получить) является командой от клиента серверу вернуть в качестве ответа информацию, обозначенную URL, приведенным в сообщении. Если URL относится к процессу, который генерирует данные, именно эти данные подлежат передаче клиенту.

Ответ на команду get начинается с трехзначного цифрового кода состояния. Некоторые из этих кодов пользователи могут видеть в браузерах, когда встречаются сообщения об ошибках - например, 404, которое выдается, если ресурс, указанный в URL, не найден. Если ошибки нет, код состояния сопровождается технической информацией (используемой главным образом в кэш или прокси-сервером) и метаданными о теле ответа (body of the response). Метаданные указывают клиенту на длину типа данных, язык, кодировку, хэш и дату. Клиент использует эти метаданные для обработки тела ответа (представляющее собой последнюю часть сообщения, которое обычно является файлом, указанным в URL).

Два других метода HTTP тесно связаны с get. Метод head запрашивает те же данные, что и get, за исключением того, что собственно тело сообщения не пересылается. Это полезно при проверке правильности гиперссылок, доступности, проверке после модификации, поскольку не требует пересылки больших файлов. Метод post используется для расширения возможностей увеличения объема информации, передаваемой от клиента к серверу. Метод post обычно применяется клиентом при передаче блока данных, таких, например, как формы HTML, которые затем могут обрабатываться CGI-скриптами или иными приложениями на сервере.

Главное назначение HTTP состоит в поиске и получении информации с сервера, но этот протокол может быть использован для изменения информации непосредственно на сервере. Метод put обычно используется для сохранения конкретной информации по указанному URL; метод delete используется для уничтожения информации. (Эти методы используются редко. Обычный способ поместить информацию на веб-сервер состоит в использовании специальных программ, которые манипулируют данными веб-сервера непосредственно, а не с помощью HTML-сообщений, отправленных откуда-то снаружи.)

Большинство изменений в HTTP новых версий направлено на сохранение совместимости разных версий и улучшение производительности в Интернет. HTTP способен распознавать, что многие сообщения обрабатываются прокси и кэшами. Его последние версии включают набор данных и сервисов для поддержки этих "посредников". Есть также специальные методы, например options (которые позволяют клиенту запрашивать информацию о доступных опциях передачи данных) и trace (используется для диагностики и тестирования).

HTTP совершенствуется уже много лет, но все еще остается простым протоколом. Разработчики выполнили важную работу по противостоянию возрастающему давлению желающих "улучшить" протокол, добавляя все новые и новые возможности, и при этом постоянно внедряя практические улучшения для повышения производительности. Не найти двух людей, которые бы согласились в том, какой набор сервисов должен предоставлять протокол, но совершенно очевидно, что HTTP представляет собой пример успешного проекта Интернета.

Ключевой элемент в CORBA - это брокер объектных запросов (Object Request Broker, ORB). После того, как ORB добавлен к программе-приложению, он устанавливает клиент-серверные связи между объектами. Используя ORB, клиент может прозрачно вызывать объект, который может быть как на той же машине, так и где-то в сети. ORB перехватывает вызов, находит объект, удовлетворяющий этому запросу, передает параметры этому объекту, вызывает его методы (methods) и возвращает результаты. Клиент не должен беспокоиться о том, где находится объект, его язык программирования, операционная система и другие системные аспекты, которые не являются частью интерфейса объекта. Таким образом, ORB обеспечивает интероперабельность между приложениями на разных машинах в неоднородной распределенной среде.

Соккрытие данных (Data Hiding)

Объекты в смысле программирования и цифровые объекты в библиотечном смысле совершенно различны, но они имеют общие черты. Термин "сокрытие данных" пришел из объектно-ориентированных языков программирования, но равным образом приложим к цифровым объектам. Когда клиент обращается к информации в репозиторий, ему необходимо знать интерфейс, предоставляемый репозиторием, но не нужно знать, каким образом в нем хранится информация. Для веб-сервера интерфейсом является протокол (HTTP), схема адресов (URL) и набор форматов и типов данных. В других репозиториях он может выражаться в терминах объектно-ориентированного программирования. То, что пользователь воспринимает как единичный объект, может в репозитории находиться в виде сложного набора файлов, как в виде записей в таблицах, так и в виде активных объектов, исполняемых по запросу.

Соккрытие внутренней структуры и предоставление полного доступа только через четко определенный интерфейс упрощает достижение интероперабельности. Выгода клиентов состоит в том, что не нужно знать внутреннюю организацию репозитория. Два репозитория могут хранить одну и ту же информацию различными способами. В одном случае звуковой и видеоряд оцифрованного фильма могут храниться как два отдельных объекта, а в другом - как один. Клиентская программа должна иметь возможность послать запрос на воспроизведение независимо от этих внутренних различий. Выгода репозитория в том, что их внутренняя организация остается их сугубо внутренним делом. То, что пользователь видит отдельным объектом, может фактически являться страницей в HTML-формате со связанными изображениями и апплетами Java. С использованием концепции сокрытия данных становится

возможным перенос файлов изображений в другое место или переход на новую версию Java, причем прозрачным методом для окружения.

JSTOR и Национальная программа ЭБ в Библиотеке Конгресса предоставляют маленькие "иконки", которые являются уменьшенными копиями больших изображений. В Библиотеке Конгресса было решено создавать "иконки" в первую очередь и хранить их как отдельные данные. JSTOR не хранит "иконок", вместо этого они генерируются по запросу из хранящихся больших изображений. Каждый подход имеет свои преимущества. Эти внутренние детали должны быть скрыты от пользовательского интерфейса и могут позднее изменяться. Внешним системам важно знать, что хранилище может предоставлять такие "иконки", но им не надо знать, каким образом они генерируются.

Унаследованные системы (Legacy Systems)

Попытки достичь интероперабельности обычно базируются на согласовании стандартов. Стратегия заключается в том, чтобы вынудить все организации принять набор технических и организационных стандартов. Для сложных систем вроде ЭБ, это практически невозможно осуществить. Необходимы отдельные стандарты для сетей, типов данных, методов идентификации, безопасности, поиска и обработки, сообщений об ошибках, организации оплаты. Каждый из этих стандартов будет состоять из нескольких частей, описывая синтаксис, семантику, процедуры обработки ошибок, расширения и пр. Если бы удалось согласовать полный набор стандартов и если бы каждая организация полностью внедрила их у себя - то был бы достигнут замечательный уровень интероперабельности!

На практике скорость стандартизации всегда отстает от скорости развития технологии. Ни одна организация никогда не успеет внедрить все стандарты в свои системы до того, как она уже будет вынуждена начать изменения в этих системах в связи с новыми требованиями или проблемами. Таким образом, в отношении интероперабельности фундаментальный вопрос стоит так: каким образом системы разных поколений могут успешно работать вместе. Более старым системам иногда дают ироническое имя "унаследованные системы" - но многие такие системы отлично работают. Планы на будущее всегда должны предусматривать совместимость и возможность использования существующих систем.

Врезка 13.4

Система InfoBus в Стэнфордском университете

Один из проектов, поддержанных в рамках Digital Libraries Initiative в Стэнфордском университете, был проект InfoBus под руководством Hector Carcia-Molina, Terry Winograd и Andreas Paepcke. Он посвящен проблемам интероперабельности существующих систем. При этом InfoBus воспринимает эти системы такими, какие они есть, не пытаясь определять новые стандарты или модифицировать что-либо. Базовый подход проекта представляет т.н. прокси библиотечного сервиса (Library Service Proxies), который является набором объектов CORBA для предоставления онлайн-сервисов. Эти прокси взаимодействуют с существующими сервисами любыми способами коммуникации, доступным этим сервисам, и преобразуют их сообщения в стандартный интерфейс, определенный в терминах метода CORBA. Например, клиент с поисковым интерфейсом Z39.50 может обратиться к онлайн-поисковой службе, такой как Dialog. Это потребует двух прокси: одного для перевода из поискового протокола Z39.50 в модель InfoBus, а другого - для перевода из интерфейса Dialog в модель InfoBus. Используя эти пары прокси, клиент может искать информацию в Dialog несмотря на различные интерфейсы.

Наверное, наиболее интересным инструментом InfoBus является поддержка Z39.50. Команда из Стэнфорда разработала прокси, который позволяет клиентам Z39.50 взаимодействовать с поисковыми сервисами, не поддерживающими Z39.50. Пользователи могут формулировать запросы, обращаясь к этому прокси, причем пользовательский интерфейс разработан для коммуникации с сервером Z39.50. Прокси отправляет поисковые запросы через InfoBus к любому доступному источнику, даже если последний не поддерживает Z39.50. Прокси

конвертирует результаты поиска в формат, понятный клиенту Z39.50. Параллельно с этим исследованием, разработчики из университета Мичигана реализовали другой прокси, который делает доступными через InfoBus все серверы Z39.50. В рамках проекта Стэнфорда также разработан прокси-сервер для HTTP, веб-поисковых систем (Lycos, WebCrawler, AltaVista) и для других веб-служб (включая ConText и инструмент объединения документов фирмы Oracle (document-summarization tool)).

Архивирование

Архивы поставляют "сырье" для истории. В США Администрация национальных архивов и записей (National Archives and Records Administration) ставит своей целью хранение записей "до конца республики" (until the end of the republic). По крайней мере, архивы должны быть готовы хранить библиотечную информацию дольше, чем могут существовать нынешние компьютерные системы и дольше, чем сроки, на которые рассчитаны нынешние магнитные или электронные носители информации. Цифровое архивирование - дело исключительно сложное. Легко констатировать ситуацию, но значительно труднее ее разрешить на практике. Основные результаты работ в области цифрового архивирования представлены в отчете группы по исследованию архивирования цифровой информации (Task Force on Archiving of Digital Information), описанном на врезке 13.5.

Обычное архивирование различает "консервацию" (conservation) отдельных объектов и "сохранение" (preservation), которое позволяет сохранить содержание, даже если оригинал утрачен или разрушен. В ЭБ соответствующие технологии называют "обновление" (регенерация, refreshing), имеющее своей целью сохранение точной последовательности битов, и "преобразование" (миграция, migration), которая сохраняет содержание на семантическом уровне, но это уже не исходная "законсервированная" последовательность битов. Впервые это различие было проведено Task Force on Archiving of Digital Information, которая рекомендовала миграцию в качестве базовой техники цифрового архивирования.

И обновление, и преобразование требуют регулярных усилий. Деловые записи сохраняются длительное время потому, что группа людей получает за это зарплату. Это их работа и они уделяют внимание взаимосвязанным проблемам безопасности, резервного архивирования и долговременной доступности данных. Издатели также начали понимать, что их цифровая информация представляет собой способ получения доходов в течение десятилетий. К сожалению, в ЭБ нет никого, кто бы отвечал за сохранность информации за рамками текущего использования. Некоторые данные, которые сегодня выглядят бесполезными, могут быть находкой спустя века. Архивирование таких данных не имеет высокого приоритета для большинства и это будет первое, что сократят при уменьшении бюджета.

Врезка 13.5

Группа по архивированию цифровой информации (Task Force on Archiving of Digital Information)

Группа была создана в конце 1994 года Комиссией по сохранению и доступу (Commission of Preservation and Access) и Исследовательской группой по библиотекам (Research Libraries Group) для изучения проблем цифрового хранения. Она функционировала под руководством John Garrent (ныне работает в CNRI) и Donald Waters из Йельского университета. Отчет 1995 года был первым всеобъемлющим взглядом на эту проблему с юридической, экономической, организационной и технической точек зрения. Он иллюстрировал опасности абсолютизации технократического подхода и приводил некоторые устрашающие примеры. Например, в докладе отмечалось, что больше не существует компьютерных программ для анализа данных, собранных в рамках проекта по анализу использования земель штата Нью-Йорк и сохранения природных ресурсов, выполнявшегося в 60-е годы. Рабочая группа предложила создать национальную систему архивов в качестве базы долговременного хранения. Однако, несмотря на то, что такие системы весьма желательны и даже необходимы, признавалось, что это не является панацеей. Ответственность за сохранение информации лежит

на людях и организациях, которые ею управляют и им необходимо рассматривать себя как часть большей системы.

Архивирование, отмечается в отчете, - это больше, чем простое копирование битов с одного стареющего носителя на другой; обязательно должны предусматриваться методы интерпретации и обработки этих битов. С этой точки зрения, миграция информации с одного формата в другой или с одного типа компьютерной системы на другую, очевидно, будет доминирующей формой цифрового архивирования.

Рассмотренные в отчете юридические аспекты, связанные с архивированием, особенно интересны. Большая часть информации, которая должна архивироваться, принадлежит организациям, которые рассматривают ее как интеллектуальную собственность по финансовым или иным причинам. Естественно, они сомневаются по поводу передачи копий в архивы. С другой стороны, архивы должны проявлять настойчивость в поиске и сохранении для общества информации, которая может быть утрачена (например, в связи с уходом компании из бизнеса).

Этот отчет остается наиболее полным исследованием данной проблемы. Хотя финансовые расчеты сейчас выглядят устаревшими, анализ вариантов и доступных выборов по-прежнему представляет серьезный интерес.

Хранение (Storage)

В прошлом "выживание" физических объектов зависело главным образом от физической долговечности материалов, из которых они изготавливались. Были ли это церковные записи, правительственные документы или сокровища, как Rosetta Stone, Dead Sea Scrolls (рукописи Мертвого моря), Domesday Book, Gutenberg Bibles (библии Гутенберга) - выжили те, что создавались на прочных материалах, главным образом, высококачественной бумаге.

Ни один из нынешних цифровых носителей не может гарантированно служить столь же долго. Некоторые (например, магнитная лента,) имеют удручающе короткий срок службы. Компакт-диски более стабильны, но предсказать срок из жизни пока невозможно. Если об этом не позаботиться, вся цифровая информация может быть утрачена за несколько десятилетий. Врезка 13.6 описывает некоторые методы, используемые ныне для хранения цифровых материалов. Обращает на себя внимание то, что основной упор делается на минимизацию стоимости оборудования и на короткое время обработки - а отнюдь не на долговечность.

Врезка 13.6

Хранение цифровой информации

ЭБ используют устройства для хранения огромных объемов. Страница ASCII-текста содержит лишь несколько тысяч знаков, однако для сканирования цветного изображения площадью с квадратный дюйм требуется свыше мегабайта (миллион байтов!). Час оцифрованного звука на компакт диске занимает свыше 600 Мб, а минута видео - до гигабайта (без сжатия). Для снижения требований к хранению, большие объекты (почти все изображения, звук и видео) подвергаются сжатию. Идея проста, хотя реализация весьма сложна. Цифровая информация содержит повторы. Страница изображения имеет участки белого цвета, для которых нет необходимости кодировать каждый пиксель отдельно. Поскольку последовательные кадры видео различаются незначительно, проще записывать разницу между ними, чем кодировать каждый кадр отдельно.

Сжатие без потерь (lossless compression) удаляет повторяющуюся информацию полностью обратимым способом - оригинал может быть реконструирован точно и полностью. Компрессия с потерями (lossy) не может быть реконструирована в связи с утратой части информации. Для некоторых приложений сжатие без потерь обязательно. В физических экспериментах одна точка на изображении может быть критически важна, так что любые его изменения могут поставить под сомнение надежность эксперимента. Однако, для большинства приложений некоторые потери вполне приемлемы. Компрессия JPEG для графических изображений и MPEG для видео представляют собой методы с потерями, которые предназначены для создания изображений с приемлемым для человеческого глаза качеством.

Методы сжатия значительно уменьшают объем хранимых данных, однако файлы по-прежнему весьма велики. После компрессии монохромная страница сканированного текста занимает свыше 50000 байт. MPEG компрессия цифрового видео уменьшает объем с 20-30 Мб на секунду изображения до 10 Мб на минуту. Поскольку ЭБ могут хранить миллионы таких объектов, емкость устройств хранения весьма важна.

Идеальный носитель для хранения в ЭБ должен обеспечивать огромную емкость для данных при низкой стоимости; обеспечивать быстрый доступ и чтение информации; быть надежным и долговечным.

Вращающиеся магнитные диски являются стандартной средой хранения в современных компьютерных системах. Их размеры - от нескольких миллионов байт до тысяч гигабайт. (Гигабайт - это 1000 миллионов байт). Диски достаточно быстры для большинства приложений ЭБ, поскольку данные могут считываться с диска быстрее, чем передаваться по сети. Когда данные считываются с диска, задержка в связи с позиционированием головок составляет лишь около 15 миллисекунд; затем данные считываются большими блоками (обычно порядка 25 Мб/с). Такие характеристики производительности удовлетворительны для приложений ЭБ, которые обычно читают большие блоки данных одновременно.

Снижение стоимости дисков представляет собой одно из замечательных достижений технологии. Стоимость магнитных дисков падает в цене даже быстрее, чем цена полупроводников. В 1998 году цена дисков составляла несколько сотен долларов за гигабайт. Технология развивается столь стремительно, что ЭБ уверенно могут планировать, что через 10 лет их стоимость будет не более 5% нынешней (с определенной долей уверенности - не более 1%).

Слабое место дисков - их ненадежность. Данные на дисках легко утратить из-за выхода из строя механических деталей или из-за искажения вследствие программного сбоя. Для защиты от таких потерь обычной практикой является копирование данных на другие носители, обычно на магнитную ленту. На дисках обычно содержатся свободные сектора, которые автоматически задействуются для автоматической коррекции проблем. Для долговременного хранения ни диски, ни магнитная лента не годятся. Данные кодируются в тонкой магнитной пленке на некоей поверхности - рано или поздно эта пленка разрушится. Диски превосходно подходят для текущих операций, но не для архивирования.

Большие коллекции ЭБ иногда хранятся в иерархических устройствах хранения. Обычно это три уровня: магнитные диски, оптические диски и магнитная лента. Магнитные диски постоянно в работе, так что информация может быть считана в доли секунды. Оптические диски представляют собой дешевый способ хранения огромных массивов данных, но обычно они недоступны в режиме он-лайн. Перед использованием компакт-диска робот должен его найти, захватить и переместить в соответствующее устройство - а это медленный процесс. Магнитные ленты также используются для офф-лайн-ового хранения и загружаются роботами.

С точки зрения компьютера, иерархическое хранение представляется как единая согласованная файловая система. Менее часто использованные данные переносятся с быстрого, но дорого магнитного диска на медленный, но более дешевый носитель. В условиях быстрого снижения цены на диски, промежуточный уровень хранения становится дискуссионным. Магнитные диски и ленты имеют вполне определенное назначение и оба эти носителя по-прежнему необходимы, а вот промежуточный уровень хранения в будущем может уже не понадобиться.

Репликация и обновление (Replication and Refreshing)

Репликация - это один из часто используемых методов обработки данных. Существование важных данных в виде единственной копии на одном компьютере крайне ненадежно. Компьютер может выйти из строя, данные могут быть уничтожены из-за ошибки программного обеспечения, некомпетентный или нечестный пользователь может удалить данные; компьютер может быть уничтожен из-за пожара, наводнения или других стихийных бедствий. По этим причинам, компьютерные центры всегда делают архивные копии всех данных и хранят их в безопасном месте. Хорошие организации идут еще на один шаг впереди и периодически

консолидируют важные записи для долговременного хранения. Один из подходов состоит в сохранении копий финансовых и юридических записей в уменьшенном виде (микрофиши или микрофильмы), поскольку эти носители для архивных целей исключительно долговечны.

Поскольку все типы носителей для цифровой информации имеют короткий срок службы, ЭБ должны планировать периодическое обновление своих коллекций. Каждые несколько лет данные должны переноситься на новые носители. С финансовой точки зрения это не пустая трата денег. Следующие несколько десятилетий компьютерное оборудование будет падать в цене с одновременным увеличением возможностей и производительности. Оборудование, которое понадобится для переноса сегодняшних данных через десять лет будет стоить лишь несколько процентов от сегодняшней цены, а роботы помогут минимизировать трудозатраты. Как это часто бывает в случае ЭБ, вопрос скорее организационный: будут ли библиотеки и издательства систематически выполнять эти работы?

Сохранение содержания путем преобразования (Preserving Content by Migration)

Даже если допустить, что биты систематически обновляются при переносе на другой носитель и техническая потребность в сохранении исходных данных установлена, сама проблема еще только начинается. Цифровая информация бесполезна без сведений о форматах, протоколах и метаданных. Древние манускрипты могут быть прочитаны, поскольку язык и способ письма меняется медленно. Для интерпретации старых документов требуется большой опыт, но подобный опыт сохраняется в поколениях, так что и студент может расшифровать старые материалы.

Компьютерные форматы меняются постоянно. Файл, записанный в формате десятилетней давности, уже может быть трудно прочесть. В мире уже нет компьютера, который мог бы выполнить программы для компьютеров, которые были весьма распространены лишь совсем недавно. Некоторые форматы очень просты. Например, если в будущем археолог столкнется с файлом с ASCII-текстом, даже если все знания об этом формате будут утрачены, код настолько прост, что текст, вероятно, будет прочтен. Но ASCII - скорее исключение. Другие форматы весьма сложны. Трудно допустить, чтобы кто-то смог расшифровать MPEG-архивированный файл без информации о математическом алгоритме или понять большую компьютерную программу лишь по машинному коду.

Таким образом, кроме хранения исходных данных, цифровые архивы должны хранить информацию о методах получения данных, о типе, структуре и формате. Если для интерпретации данных необходима компьютерная программа, то сохраняться должна эта программа и устройство для ее выполнения. Либо эти данные должны мигрировать в другую форму. В ближней перспективе, для этих целей можно хранить старые компьютерные системы, но компьютеры весьма недолговечны. Рано или поздно он сломается, запасные части будут недоступны - и любые программы, которые исполнялись на таком компьютере, будут бесполезны. Поэтому преобразование содержания представляется необходимым.

Преобразование является стандартной технологией при обработке данных уже десятилетия. Пенсионные фонды, например, хранят записи о финансовых транзакциях многие годы. Администрация социального обеспечения (Social Security Administration) США хранит записи об уплате налогов всего работоспособного населения в течение всего периода их работы. Эти записи находятся в компьютерах, но сами компьютерные системы периодически заменяются. Устанавливаются новые компьютеры, обновляется программное обеспечение. После этих изменений данные мигрируют с компьютера на компьютера, из базы данных в базу данных. Базовый принцип преобразования состоит в том, что меняться могут форматы и структура, но семантика содержимого должна сохраняться.

Другой иногда предлагаемый метод - эмуляция. Идея состоит в создании полной имитации деталей компьютерной среды (операционной системы), необходимой для исполнения программы. В результате, в любой момент времени в будущем может быть построен эмулятор, который будет вести себя точно так же, как и исходный компьютер. С точки зрения некоторых специальных нюансов, это неоднозначное предположение. Например, возможно создать такую

спецификацию для программ, обрабатывающих простые форматы изображений, как JPEG. Во всех других обстоятельствах такая эмуляция - не более, чем химера. Даже простая операционная среда слишком сложна для точной и полной спецификации. Точная комбинация синтаксиса, семантики и специальных правил не поддается формализации - это скорее комбинация "эзотерических" аспектов системы, причем именно они зачастую критически важны для корректного исполнения программ.

Цифровая археология

Общества проходят через периоды испытаний, включая депрессии, войны, политическую нестабильность - и при этом сохранности архивных материалов придается минимальное значение. Физические артефакты могут оставаться забытыми на века в хранилищах - и при этом могут быть возвращены в оборот. Цифровая информация в этом смысле менее благодарный объект. Врезка 13.7 описывает как в один из таких периодов испытаний, во время коллапса ГДР, ситуация чуть не дошла до утраты государственных архивов. Процесс извлечения информации из поврежденных, фрагментированных и архаичных источников данных называется цифровой археологией.

Создание цифровых библиотек с учетом архивных потребностей (Creating Digital Libraries with Archiving in Mind)

Поскольку в цифровом архивировании так много подводных камней, встает вопрос - что мы можем сегодня сделать для увеличения вероятности того, что будущие археологи смогут расшифровать создаваемые нами биты? Некоторые простые меры позволяют существенно продвинуться в этом направлении. Первая - это использование для хранения информации распространенных форматов. Это увеличивает шанс, что после его устаревания, новые форматы будут поддерживать конверсию из такого формата. Например, HTML и PDF так широко используются в промышленности, что программы для их просмотра обязательно будут доступны еще многие годы.

Одно из интересных предложений - создать архив, содержащий определения форматов, стандартов метаданных, протоколов и других строительных блоков для ЭБ. Этот архив должен вестись на надежных носителях (бумаге или микрофильмах) и все должно быть описано простым текстом. Если сохранятся форматы и схемы кодировки, большая часть информации может быть расшифрована. Будущие цифровые археологи вынуждены будут потратить много сил на создание интерпретаторов, которые могут воспроизводить давно устаревшие форматы или наборы инструкций - но принципиально это возможно. Современные компьютерные форматы весьма сложны. Хотя цифровые археологи, возможно, смогут восстановить структуру и архитектуру ранних компьютеров фирмы IBM по распечатке памяти (дампу памяти, memory dump), они будут бессильны перед более сложными материалами, если только не будут сохранены спецификации.

Один из наиболее важных способов, которым ЭБ могут поддержать архивирование - это селекция. Сохраняться должно отнюдь не все подряд. Большая часть информации создавалась для краткосрочного использования, причем ее большая часть эфемерна и не имеет ценности. Издатели всегда вынуждены были выбирать, что публиковать - а что нет; даже самые крупные библиотеки собирают лишь часть издаваемых в мире материалов. ЭБ управляют коллекциями информации. Критически важной частью такого управления является отбор того, что собирать, что хранить, что сохранять на будущее, - а что уничтожить.

Врезка 13.7

Цифровая археология в Германии

В 1998 году статья в New York Times рассказала о тех проблемах, с которыми сталкиваются архивисты в цифровом мире, если данные не поддерживаются непрерывно с момента их создания.

В 1989 году, после падения Берлинской стены и объединения Германии, цифровые записи ГДР были разупорядочены. Германские федеральные архивы получили грудку перепутанных перфокарт, магнитных дисков и лент - остатки архивных записей коммунистического государства. Большая часть носителей была в плохом состоянии, данные - в недокументированных форматах, а компьютерные центры, в которых они поддерживались были либо ликвидированы, либо приватизированы. С тех пор небольшая группа немецких архивистов пытается реконструировать эти записи ГДР. Сами себя они называют "цифровыми археологами".

Первая проблема, с которой они столкнулись - считывание данных с носителей. Данные даже на лучших магнитных пленках обладают небольшим сроком жизни. Эти же пленки были в плохом состоянии, поэтому прочесть их можно было только один раз. Во многих случаях данные хранились в русских компьютерных системах, несовместимых с другими компьютерами. Хотя архивисты имели несколько русских компьютеров, около 30% данных были нечитаемыми.

После того, как данные копировались на другие носители, проблема все еще была далека от разрешения. Для экономии места многие данные архивировались редкими и недокументированными методами. Важная база данных коммунистических функционеров иллюстрирует только некоторые проблемы. Поскольку сама система управления базой данных и компьютеры, на которых генерировались эти данные, принадлежали к клону IBM, сама реконструкция базы данных не представляла трудности. Однако, интерпретация данных без документации была исключительно сложна. У архивистов было одно преимущество - они могли расспросить некоторых из тех людей, которые создавали эти базы данных, и использовать этот опыт для интерпретации и сохранения большей части этой информации.

Michael Wettengel, глава немецких архивистов, резюмировал ситуацию достаточно четко: "Компьютерные технологии созданы для обработки информации, но не для долговременного хранения."

Глава 14. Цифровые библиотеки и электронные издательства сегодня

ЭБ и электронные издательства уже существуют. Это не академические концепции и не утопическая мечта. В этой, заключительной главе ставилась задача попытаться заглянуть вперед, хотя история вычислительной техники показывает, сколь бессмысленными могут быть такие предсказания. Более выгодным оказывается жить настоящим.

Это не призыв игнорировать будущее. Работы, которые сегодня лишь в процессе создания, завтра уже будут в производстве. Не все новые проекты относятся к основному направлению развития, но понимание того, что происходит сегодня поможет нам понять, что потенциально ждет нас завтра. Эта книга писалась в течение 12 месяцев в 1997 и 1998 годах. Даже за этот короткий период времени ЭБ развивались стремительно. Эта глава анализирует некоторые современные тенденции и претендует не на предсказание будущего, но на осмысление настоящего.

Одна из причин того, насколько трудно делать какие-то предсказания, заключается в том, что в настоящее время технологии сами по себе более зрелые, чем то, для чего они создавались. Прошлые предсказания базовых технологий обычно оказывались точными. С каждым годом полупроводниковые и магнитные устройства становятся все меньше, дешевле и быстрее, а их емкость возрастала. Есть основания ожидать, что эти тенденции сохранятся на ближайшие пять-десять лет. Фонды уже вкладывают средства в высокоскоростные сети, которые будут доступны уже через несколько лет. Но предсказания о новых приложениях, которые будут разработаны, всегда были значительно менее точны. Такие обычные приложения, как электронные таблицы, настольные издательства и веб-браузеры появились без предупреждения. Даже если сравнимого по размерам прорыва и не случится, предсказания о том, какие именно приложения будут использоваться, какие социальные эффекты будут иметь новые технологии, практически невозможны. Есть лишь одно надежное правило: если ученый муж начинает со слов "Это неизбежно...", такое утверждение неизбежно неправильное.

Одно из возможных интерпретаций нынешней ситуации с ЭБ состоит в том, что они находятся в конце начальной фазы развития - и наступает новая фаза. Первая часть может рассматриваться как движение от традиционных издательств и библиотечных коллекций к цифровым сетям. Онлайн-газеты, электронные версии научных журналов, конверсия исторических материалов - все они относятся к этому процессу. Главным образом, они использовали новые технологии для улучшения имеющихся типов информации. Если это заключение правильно, следующая фаза будет связана с новыми типами коллекций и сервисов, не имеющих аналогий в традиционных формах. Формы воплощения этих коллекций практически невозможно предугадать.

Близорукий взгляд на цифровые библиотеки (Myopic view...)

Когда кто-то оглядывается назад, на определенный период развития, многие тенденции и ключевые события видны весьма отчетливо. Те из нас, кто работает в ЭБ, могут иметь только близорукий взгляд на эту область - но вот некоторые наблюдения на тему о том, как ЭБ выглядят с точки зрения постороннего наблюдателя.

Контраст между маленьким катером и супертанкером представляется полезной метафорой. Катер может быстро набирать скорость и менять направление, но обладает небольшой движущей силой. Ранние проекты ЭБ, такие как Mesquig в университете Карнеги-Меллона, были как такой катер. Был быстрый прогресс - но как только иссякал первоначальный энтузиазм или заканчивался грант, то такие проекты быстро теряли свою движущую силу. Официальные библиотеки, издательства, коммерческие корпорации - они как супертанкеры.

Они движутся более целенаправленно и медленно изменяют направление - но уж если они выбрали новое направление, то движутся в этом направлении неуклонно и постоянно. В области ЭБ и электронных издательств для достижения успеха требуется внимание к тысячам деталей, которые в итоге и трансформируют хорошую идею в практический сервис. По мере того, как Интернет и веб достигают зрелости, многие организации делают долговременные инвестиции в библиотечные коллекции, электронные издательства и онлайн-сервисы. Супертанкеры меняют направление.

В 1997-98 годах достигли зрелости многие идеи, работа над которыми велась в предыдущие годы. Онлайн-версии газет достигли высокого качества, а ряд из них стал конкурировать с печатными версиями. Большинство научных публикаций как коммерческих, так и некоммерческих издательств стали доступны он-лайн. Это было важное время и для конверсионных проектов в библиотеках, например, JSTOR и в Библиотеке Конгресса; объем материалов, доступных в рамках этих проектов, резко возрос. Подход Dublin Core к метаданным и использование Digital Object Identifiers для электронных публикаций обрели движущую силу и - как представляется - переходят из стадии экспериментов в повседневную практику.

С технической точки зрения, на рынке появились продукты для автоматической репликации (mirroring) и кэширования, становятся доступными инструменты обеспечения безопасности веб, а язык программирования Java используется все шире. Эти разработки начинались несколько лет назад и сейчас еще сохраняют исследовательский характер, - но в сумме они представляют собой значительный прогресс.

В США стремительно расширяется электронная коммерция в Интернете. Покупка книг, авиабилетов, акций и автомобилей через Интернет-транзакции становится общепринятой. Служба внутренних доходов (Internal Revenue Service) предлагает людям платить налоги он-лайн. Конгресс принял обновленную версию закона об авторских правах. Для инновационных инвестиций существуют фонды, а рынки открыты для абсолютно новых продуктов. Акции Интернет-компаний очень популярны у игроков на рынке акций. Персональные компьютеры ценой до 1000 долларов активно продаются, приближая онлайн-информацию ко все большему числу людей.

Ничто не совершенно, и пессимисты могут указать на некоторые тревожные события. В 1997 году уровень "мусора" среди электронной почты достиг раздражающего уровня. Ведущие производители выпустили несовместимые версии Java. Политика США в области шифрования продолжала "эмулировать страуса". Это кратковременные проблемы. Можно надеяться, что к моменту выхода этой книги, проблема с электронной почтой будет снята, а Java стандартизована. Появились даже проблески надежды в политике США в области шифрования.

На следующие несколько лет можно ожидать поступательного развития в духе 1997-98 годов. Резюмировать их можно так: большое число энергичных индивидуумов исследует возможности Интернета по созданию и предложению новых продуктов и сервисов.

Люди

С близорукой точки зрения легко увидеть индивидуальную активность. Значительно труднее увидеть те подспудные изменения, частями которого она является. Наиболее фундаментальная в долгосрочном плане тенденция может быть труднее всего измерима в краткосрочной перспективе. Как люди будут усваивать изменения? Многие люди пишут для веб. Графические дизайнеры создают онлайн-материалы. Люди читают эти материалы. Кто эти люди? Что бы они делали в ином случае?

Привычки определенно изменились. Посетите университетский кампус или другую организацию, интенсивно использующую информацию - и вы увидите, что люди проводят несколько часов в неделю перед компьютером, обращаясь к онлайн-информации. Дома часть людей проводит в веб примерно столько же времени, сколько раньше они тратили на просмотр телепрограмм. На работе люди стали читать больше - или они постепенно заменяют онлайн-информацией традиционные формы активности, вроде посещения библиотеки? Оглядываясь на прошедшее десятилетие, эти перемены могут показаться обычными. Однако

сегодня мы можем только строить гипотезы или экстраполировать наши незначительные данные. Вот некоторые догадки, основывающиеся на личных наблюдениях и частных обсуждениях.

Вспышка онлайн-информации привлекла в область ЭБ большое число новых людей. Люди, которые считали библиотечное дело скучным, а издательское педантичным, оказались вовлеченными в создание и дизайн онлайн-материалов. Привнесенные этими новичками энтузиазм и энергия повлияли на эти старые профессии сильнее, чем можно было ожидать. Хотя каждая группа имеет своих луддитов [движение ремесленников, громивших первые станки и механизмы], многие люди довольны открывающимися перед ними новыми возможностями.

С появлением веб количество профессионалов Интернета было так невелико, что любой человек с минимальными навыками мог рассчитывать на высокую зарплату. В настоящее время, хотя настоящие профессионалы по-прежнему в недостатке, средний уровень навыков весьма высок. Признаком таких изменений является увеличение числа программ по обучению новым технологиям менеджеров среднего уровня. В США каждый колледж предлагает курсы Интернета и веб. Компании, выпускающие учебные программы для компьютеров, переживают бум. Специализированные программы для ЭБ идут нарасхват.

В 1997 году студент Корнелла, которого попросили найти информацию в университетской библиотеке, ответил "Можно я воспользуюсь веб? Я не бываю в библиотеке." Совсем недавно преподаватель университета Калифорнии в Беркли заметил, что термин "цифровая библиотека" тавтологичен. Она заметила, что для студентов библиотекой является Интернет. Не будут ли они в будущем рассматривать прекрасную обычную библиотеку в Беркли как физический эрзац реальной вещи? Что это - отдельные мнения привилегированного меньшинства или они отражают взгляды нового поколения пользователей библиотек? Никто не знает. Данные фрагментарны и зачастую противоречивы. Следующая статистика была собрана из различных источников и должна рассматриваться с изрядной долей здорового скептицизма.

Опрос, проведенный в Питтсбурге в 1996 году показал, что 56% людей от 18 до 24 лет пользуются Интернетом - против 7% среди тех, кому больше 55. Другой опрос среди всех жителей США в 1997 году показал, что 61% подростков (66% ребят и 56% девочек) пользовались веб. В 1996 году 72% детей от 8 до 12 лет каждый месяц проводили некоторое время за компьютером.

В 1998 году сеть, которую многие молодые люди начинают использовать как свою библиотеку, насчитывала около 5,3 миллиона компьютеров. Тщательное исследование, проведенное в конце 1996 года, дало оценку в миллион наименований веб-сайтов и примерно 450000 серверов (host machines). Из этих наименований примерно 300000 выглядели достаточно стабильными. На публичных серверах было около 80 миллионов HTML-страниц. Двумя годами позже число веб-страниц оценивалось в 320 миллионов. Каким бы не было реальное число, любой согласится с тем, что их много и что число их быстро увеличивается.

Пессимист может истолковать эти цифры как указание на то, что американские молодые люди "растворились" в онлайн-информации, полностью восприняли ее, люди среднего возраста пытаются адекватно приспособиться к ней, а люди более старшие, - которые разрабатывают планы и распоряжаются ресурсами - безнадежно устарели. Однако, наблюдения показывают, что подобный анализ абсолютно неправилен. Тот факт, что множество крупных организаций инвестируют огромные средства в цифровую информацию, указывает на то, что, по крайней мере, часть более пожилых лидеров воспринимают изменения, новый мир.

Организации

Несмотря на то, что средства в создание ЭБ и электронных издательств вкладывают многие организации, нельзя угадать, какие типы организаций окажутся более успешными. В некоторых условиях размер может быть преимуществом, однако маленькие, проворные компании также процветают.

Статья в журнале Economist в 1995 году описывала контроль над Интернетом в терминах гигантов и карликов. Гиганты - крупные компании, такие как телефонные и медийные корпорации. Карлики - частные лица; по отдельности они мало что могут, но суммарно они вносят большой вклад в формирование Интернета, причем зачастую в направлении, противоположном интересам гигантов.

"Сообщество карликов" в особенности успешно отстаивает открытый характер Интернета. За последние пять лет и ЭБ и электронные издательства консолидировались в крупные организационные единицы. В библиотеках это выглядит как формирование консорциумов; в издательствах - как ряд крупных слияний. Тем не менее, даже в таких условиях энергия карликов не ослабевает. В настоящий момент можно говорить о том, что и гиганты, и карлики имеют право на существование и развитие. А некоторые карлики, вроде Yahoo и Amazon.com становятся гигантами.

Коллекции и доступ

Два фактора потенциально исключительно важны для будущего ЭБ - стоимость доступа через Интернет к хорошо управляемым коллекциям и бизнес-модель, которая будет создана для ЭБ.

Пока материалы размещаются в онлайн-доступе в огромных количествах и нет никаких признаков замедления роста веб. Число сервисов, сайтов явно возрастает. Наилучшими примерами качества являются сайты газет и новостных агентств - но они также ярко высвечивают уязвимость цифровой информации. Онлайн-новостные службы, поддерживаемые Sydney Morning Herald, New York Times, CNN, предоставляют актуальную, хорошо структурированную информацию бесплатно. Общее число их читателей, очевидно, превышает число читателей всех американских газет. В качестве источника текущей информации они превосходны. К сожалению, они эфемерны. Информация непрерывно меняется, а в конце дня большая ее часть просто исчезает. Обычные библиотеки собирают газеты и хранят их веками. Эти веб-сайты никто не коллекционирует и не архивирует.

Трудно сказать, какая научная дисциплина в наибольшей мере отражена в библиотечных материалах, доступных ныне в ЭБ. Велика доля текущей научной и технической литературы, а также правительственной информации. Деловая и медицинская информация менее заметна. Публичные библиотеки предоставляют доступ к текущей информации, такой как газеты, различные расписания, информация о поиске работы, формы налоговых деклараций. Они обычно доступны он-лайн бесплатно.

Многие коллекции предоставляют в цифровой форме только текущую информацию, а не исторические материалы - хотя существует много проектов по конверсии традиционных материалов в цифровые форматы и размещение в ЭБ. Библиотеки конвертируют свои исторические коллекции, а издательства - ретроспективы журналов. Текущие проекты такого рода планируют увеличить скорость конверсии, появляются новые проекты. Несколько проектов уже конвертировали свыше миллиона страниц каждый. Первый план, согласно которому планировалось конвертировать миллиард страниц, опубликован в 1998 году.

Наибольшая лагуна - это коммерческие развлечения. Некоторые основные производители развлекательной продукции - фильмов, ТВ, радио, романов, журналов - экспериментировали со способами использования Интернета, но с минимальными последствиями. Большей частью это обусловлено техническими ограничениями Интернета. Большинство людей получают картинку значительно лучшего качества по кабельному ТВ, чем это возможно по сети или на рядовом дисплее компьютера. Другой фактор заключается в том, что скорость изменения диктуется практикой бизнеса. Развлечения - это огромный бизнес, который еще не нашел способов использования Интернета с прибылью.

Похоже, свободный доступ станет постоянной характеристикой ЭБ, хотя существует всего лишь несколько успешных качественных бизнес-моделей такого рода. Некоторые веб-сайты получают существенные доходы от рекламы, но большинство финансируется за счет внешних источников. ЭБ и электронные издательства требуют подготовленного персонала для создания,

организации и управления информацией - а такие работники дороги. Безусловно, будет достигнут некий экономический баланс, когда некоторые коллекции будут открытыми, а за некоторые придется платить пользователям. Каким именно будет такой баланс, пока предсказать трудно.

Технология

Веб-технология, которая инициировала столь большие изменения и рост, стала вполне зрелой. В середине 90-х годов веб развивался столь быстро, что появилось выражение "год веб" для коротких периодов времени, за которые происходило так же много изменений, как за полный календарный год. По мере созревания веб, скорость изменений технологии стала замедляться до величины, характерной для вычислительной техники. Каждый год приносит частичные изменения, за несколько лет эти изменения становятся весьма значительными - но лихорадочного роста больше нет. Это не значит, что развитие веб закончено - до этого еще далеко. Число веб-сайтов растет очень быстро. Загруженные сайты сообщают, что объем предоставляемых услуг за год удваивается. Качество графики и стандартны обслуживания непрерывно улучшаются. Для ЭБ и электронных издательств некоторые развивающиеся технологии выглядят особенно многообещающими: "долгоживущие" имена, такие как Digital Object Identifiers; XML-разметка; Resource Description Framework; Unicode. Успех этих технологий будет зависеть от капризов рынка. Широкое распространение некоторых (или всех) могло бы принести большие преимущества ЭБ.

Производительность собственно Интернета растет впечатляющими темпами. За последние годы было несколько правительственных и коммерческих инициатив, имевших целью увеличение производительности, надежности и охват территории в течение ближайших нескольких лет. Мы не можем предсказать, как ЭБ будут использовать эти возможности, но они открывают прекрасные перспективы.

Исследования и разработки

ЭБ в настоящее время превращаются в самостоятельную область исследований, со своей инфраструктурой рабочих встреч и конференций. Было предпринято несколько попыток учредить печатные журналы, посвященные ЭБ. Более важно то, что свыше 1000 людей считают это своей работой, проводят исследования в этой области. Примеры такого рода проектов, поддержанных NSF и DARPA, приводились на протяжении всей этой книги. Многие из самых недавних финансировались именно как исследования ЭБ.

Врезка 14.1

Рабочая встреча в Санта-Фэ

В марте 1997 года NSF поддержал проведение рабочей встречи в Санта-Фэ, Нью-Мексика, для обсуждения будущих исследований в области ЭБ. Это была часть планового процесса, который позже привел к объявлению о главной исследовательской программе по ЭБ: вторая фаза Digital Libraries Initiative. Рабочая встреча стала для ученых возможностью изложить свое видение развития ЭБ и направлений исследований.

Большинство людей на встрече были участниками проектов в рамках первой фазы Digital Libraries Initiative и других федеральных исследовательских проектов. Конечно, они были заинтересованы в продолжении своих исследований, но они не просто рекомендовали продолжить те же программы. В ранних проектах создавались экспериментальные коллекции, которые использовались для исследований, главным образом технических. Некоторые участники рабочей встречи полагали, что эффективным путем использования правительственных финансовых ресурсов было бы создание больших ЭБ. Многие люди

согласились с тем, то исследования по архивированию отстают. Большинство дискуссий, однако, было вокруг того, как сделать существующие коллекции более удобными в использовании и доступными. Участниками рабочей встречи были руководители исследований, для которых проблемой является не только поиск информации, но и избавление от излишка информации. Временами дискуссия концентрировалась на том, что центральной проблемой ЭБ является перегруженность информацией. Каким образом можно использовать автоматические методы фильтрации, извлечения и консолидации информации? Дискуссия затрагивала методы, которыми отдельные пользователи могут управлять своими частными библиотеками, а группы - продолжать свою совместную работу с использованием Интернет. Каким образом ЭБ - управляемые коллекции информации - могут управляться наиболее удобным для пользователей способом?

Обсуждались социальные, экономические и юридические аспекты. Как водится, в этих вопросах основной проблемой было формулирование согласованной исследовательской стратегии. Никто не отрицает важности этих областей - но остается скептицизм относительно того, каким образом они могут быть охвачены в крупных исследовательских проектах.

Технические специалисты указали на то, что ЭБ постепенно становятся основными потребителями суперкомпьютеров. В 1986 году, когда был создан национальный суперкомпьютерный центр, магистрали (backbone) Интернета работали на скорости 56 Кб/с, чего было достаточно для всех пользователей. Сегодня такая скорость доступна через недорогой модем. Сегодняшние ноутбуки имеют производительность суперкомпьютеров 12-летней давности. Через 12 лет маленький компьютер будет иметь производительность сегодняшних суперкомпьютеров.

Через несколько лет постепенный быстрый рост приведет к качественным, фундаментальным изменениям. Участников рабочей встречи в Санта-Фэ просили высказать их предположения о развитии ЭБ. Целью было формирование базиса для формулирования новых направлений исследований в этой области.

Заключительные замечания

В процессе написания этой книги я просмотрел сотни источников. Большая часть была оригинальными материалами - отчетами, написанными исследователями или создателями ЭБ о своей работе. Другие источники - выставка в Агентстве по авторским правам США; неопубликованная книга; по некоторым темам я посылал электронные письма друзьям. Для всего остального источником был Интернет. Многие нужные материалы просто не существуют в обычных форматах. С точки зрения ЭБ, Интернет - уже библиотека.

Мечта о библиотеках будущего объединяет все лучшее из того, что есть в обычных, традиционных методах и что могут предложить новые онлайн-технологии. В некоторых ночных кошмарах объединяются наихудшие аспекты этих подходов... В первые годы XX века филантропия Andrew Carnegie открыла в США эру публичных библиотек. Сейчас развивается новая форма библиотеки. Будем надеяться, что ЭБ смогут снискать такое же внимание и уважение, будут служить тем же глубоким потребностям, которые длительное время ассоциировались с лучшими библиотеками и издательствами.

Глоссарий

Область ЭБ абсорбирует термины из различных областей, включая вычислительную технику, библиотеки, издательское дело и юриспруденция. Данный глоссарий дает короткие определения различных терминов в их современном значении в контексте ЭБ. В других контекстах некоторые термины могут иметь иное значение.

AACR2 (Anglo-American Cataloging Rules, англо-американские правила каталогизирования) Набор правил, которые описывают содержание библиотечных каталожных записей.

abstracting and indexing services (реферативные и индексные сервисы) Вторичные информационные сервисы, которые предоставляют возможности поиска образовательной и научной информации, в частности, в отдельных журнальных статьях.

access management (управление доступом) Контроль за доступом к материалам ЭБ. Иногда употребляются выражения terms and conditions (термины и условия [доступа]) или rights management (управление правами).

ACM Digital Library ЭБ журналов и материалов конференций, публикуемых Association for Computing Machinery.

Alexandria Digital Library ЭБ географической информации в университете в Санта-Барбаре (Калифорния).

American Memory and National Digital Library Program ("Американская Память и программа Национальной Цифровой Библиотеки) ЭБ Библиотеки Конгресса США, содержащая материалы, конвертированные из уникальных исторических документов, связанных с американской историей.

applet (апплет) Маленькая компьютерная программа, которая может передаваться с сервера на компьютер-клиент и выполняться на нем.

Art and Architecture Thesaurus Тезаурус по искусству, архитектуре, декоративному искусству и материальной культуре, проект Фонда J.Paul Getty.

artifact (артефакт) Физический объект в библиотеке, архиве, музее.

ASCII (American Standart Code for Information Interchange, Американский стандартный код для обмена информацией) Схема кодировки, в которой знаки представляются наборами из семи или восьми битов; печатный (printable) ASCII - часть полного набора ASCII.

authentication (аутентификация, идентификация) Проверка пользователя, компьютера или цифрового объекта с целью убедиться, что он - именно то, что объявлено.

authorization (авторизация) Выдача разрешения пользователю или компьютеру-клиенту на доступ к конкретной информации или выполнение допустимых действий.

automatic indexing (автоматическое индексирование) Создание каталожных или индексных записей с использованием только компьютерных программ.

Boolean searching (поиск с использованием Булевых операторов) Метод поиска информации, в котором поисковый запрос состоит из последовательности поисковых терминов, объединенных операторами, такими как "and", "or", "not".

browser (браузер) Пользовательский интерфейс общего назначения, используемый в веб и в других онлайн-информационных службах. Другое название - веб-браузер.

browse (просматривать) Исследовать информацию путем изучения способа организации коллекций или путем просмотра списков - в отличие от прямого поиска.

cache (кэш) Временное хранилище, которое используется для копий недавно запрошенной информации или данных, которые - как ожидается - будут частью использоваться.

California Digital Library ЭБ, которая обслуживает девять кампусов университета Калифорнии.

catalog (каталог) Коллекция библиографических записей, созданных согласно установленному набору правил.

classification (классификация) Организация библиотечных материалов по иерархии предметных категорий.

client (клиент) Компьютер, который выступает как пользователь, в т.ч. персональный компьютер или любой иной, обращающийся к серверу с такими функциями.

CGI (Common Gateway Interface) Интерфейс программирования, позволяющий веб-браузеру выступать в качестве интерфейса информационных сервисов, отличных от веб-сайтов. (A programming interface that enable a web browser to be an interface to information services other than web sites.)

Chemical Abstracts Вторичный информационный сервис по химии.

CNI (Coalition for Networked Information, Коалиция по сетевой информации) Содружество Association for Research Libraries (Ассоциации исследовательских библиотек) и Educase для развития сетевой научной информации.

complex object (комплексный, сложный объект) Библиотечный объект из многих взаимосвязанных элементов или цифровых объектов.

compression (сжатие, компрессия) Уменьшение размера цифровых материалов за счет удаления повторов или аппроксимации; различают сжатие без потерь (полностью обратимый процесс) и сжатие с потерями (часть информации теряется при аппроксимации, необратимый процесс).

computational linguistics (компьютерная лингвистика) Раздел обработки естественного языка, посвященный грамматике и лингвистике.

controlled vocabulary (контролируемый словарь) Набор предметных терминов и правил их использования для описания материалов в индексировании и обработке информации.

conversion (конверсия) Трансформация (перенос) информации с одного носителя на другой, в т.ч. из бумажной формы в цифровую.

CORBA Стандарт распределенных вычислений, в котором объект на одном компьютере обращается к брокеру объектных запросов (Object Request Broker, ORB) для взаимодействия с объектом на другом компьютере.

CORE Проект по конвертации химических журналов в цифровую форму., выполнявшийся в 1991-95 гг. Корнельским университетом, Bellcore, OCLC и American Chemical Society.

Cryptolope Безопасный контейнер, используемый для купли-продажи секретной информации через Интернет, разработанный фирмой IBM.

CSS (Cascading Style Sheets, каскадная таблица стилей) Система таблиц стилей для использования с HTML, основа для XLS.

CSTR (Computer Science Technical Reports project) Проект CNRI и пяти университетов, выполнявшийся в 1992-96 гг. и финансировавшийся DARPA.

data type (тип данных) Структурные метаданные, связанные с цифровыми данными и указывающие цифровой формат или приложение, используемое для обработки этих данных.

DES (Data Encryption Standard, стандарт шифрования данных) Метод шифрования с закрытым ключом.

Dewey Decimal Classification (Десятичная девеевская классификация) Схема классификации библиотечных материалов, которая использует цифровые коды для указания предметных областей.

desktop metaphor (метафора рабочего стола) Концепция пользовательского интерфейса персональных компьютеров, которая представляет информацию в виде файлов и папок на рабочем столе.

Dienst Архитектура сервисов ЭБ и открытый протокол для предоставления таких сервисов, разработанный в Корнельском университете; используется в NCSTRL.

digital archeology (цифровая археология) Процесс обработки разрушенной, фрагментарной цифровой информации, в т.ч. из архаичных источников (носителей) данных.

Digital Libraries Initiative (Инициатива по ЭБ) Исследовательская программа по ЭБ. Фаза 1, с 1994 по 1998 год включала 6 университетских проектов и финансировалась NSF, DARPA и NASA. Фаза 2 началась в 1998 году.

digital object (цифровой объект) Единица хранения в ЭБ, состоящая из данных, метаданных и идентификатора.

digital signature (цифровая подпись) Криптографический код, состоящий из hash и применяемый для индикации того, что данные не изменялись; зашифровывается открытым ключом создателя подписи.

dissemination (распространение) Перенос цифрового объекта из репозитория клиенту. (The transfer from the stored form of a digital object in a repository to a client.)

distributed computing (распределенные вычисления) Вычислительные системы, в которых пользовательские сервисы предоставляются группой взаимодействующих в сети компьютеров.

D-Lib Magazine Онлайн-ежемесячник, посвященный исследованиям и разработкам в области ЭБ.

DLITE Экспериментальный пользовательский интерфейс, используемый в системе InfoBus Стенфордского университета.

document (документ) Цифровой объект, который является аналогом физического документа, в частности, текстовых материалов; модель документа - разновидность модели объекта для документов.

domain name (доменное имя) Имя компьютера в Интернете; служба доменных имен (domain name service, DNS) превращает доменные имена в IP-адреса.

DOI (Digital Object Identifier) (Идентификатор цифрового объекта) Идентификатор, используемый издательствами для обозначения материалов, опубликованных в электронной форме; частная форма handle.

DSSSL (Document Style Semantics and Specification Language) Общая система таблиц стилей для SGML.

DTD (Document Type Definition) (Определитель типа документа) Спецификация разметки классов документов, определенная в рамках SGML.

Dublin Core (Дублинское ядро) Простой набор элементов метаданных, используемый в ЭБ, главным образом для описания цифровых объектов и управления коллекциями, а также для обмена метаданными.

dynamic object (динамический объект) Цифровой объект, для которого распространение и представление пользователю зависит от выполнения компьютерной программы или иной внешней активности.

EAD (Encoded Archival Description) Разновидность DTD, используемая для кодировки электронных версий finding aids ("помощников в поиске") для архивных материалов.

electronic journal (электронный журнал) Онлайн-публикация, организованная по образцу традиционного печатного журнала; либо онлайн-версия печатного журнала, либо только электронная версия.

eLib Британская программа разработок, в основном сосредоточенная на электронных публикациях. (A British programme of innovations, around the theme of electronic publications.)

emulation (эмуляция) Воспроизведение компьютерной системы для выполнения программ и использования данных из более ранних компьютеров, которые более недоступны. [Создание на современном компьютере программными методами виртуальной модели устаревшего компьютера, на котором "выполняется" программное обеспечение таких компьютеров.]

encryption (шифрование) Методы кодирования информации для обеспечения секретности; зашифрованные данные выглядят как случайная последовательность знаков; обратный процесс, расшифровка, требует знания цифрового ключа.

entities, elements (сущности, элементы) В языке разметки сущности являются базовыми единицами информации (включая знаковые сущности, character entities); последовательности сущностей образуют элементы, которые, в свою очередь, образуют структурные единицы.

expression (выражение) Реализация работы путем выражения абстрактной концепции в реальных словах, звуках, образах и пр.

fair use (честное использование) Концепция законодательства об авторских правах, позволяющая ограниченное использование материалов без получения разрешения владельца прав, например для обучения или использования в обзорах.

federated digital library (федеративная цифровая библиотека) Группа ЭБ, поддерживающих общие стандарты и службы, обеспечивая таким образом интероперабельность и согласованный сервис для пользователей.

field, subfield (поле, подполе) Отдельные элементы информации в структурированной записи, например, в каталоге или записи базы данных.

fielded searching (поиск по полю) Метод поиска в текстовых материалах, включая каталоги, в котором поисковые термины сопоставляются с содержанием указанных полей.

finding aid ("помощник для поиска") Текстовый документ, описывающий объекты библиотеки, архива или музея.

firewall (брандмауэр) Компьютерная система, которая отслеживает данные, проходящие между сегментами сети; используется для обеспечения безопасности в частных сетях, устанавливается обычно в точке подключения к Интернету.

first sale (первая продажа) Концепция авторского права, позволяющая покупателю книги или иного объекта передавать ее кому-либо еще без получения разрешения от держателя прав.

FTP (File Transfer Protocol) (протокол передачи файлов) Протокол, используемый для передачи файлов между компьютерами в Интернете.

full text searching (полнотекстовый поиск) Метод поиска в текстовых материалах, в котором поисковые термины сопоставляются с содержанием текста.

gatherer ("сборщик") Программа, которая автоматически собирает индексную информацию из коллекций ЭБ.

gazetteer База данных, используемая для перевода между различными представлениями географических ссылок, таких как название места или географические координаты.

genre (жанр) Класс или категория объекта с точки зрения интеллектуального содержания.

geospatial information (географическая информация) Информация, относящаяся к географии, включая местоположение.

GIF Формат для хранения сжатых изображений.

Google Поисковая программа для веб, которая ранжирует веб-страницы в списки с учетом веса гиперсвязей, указывающих на данную страницу.

gopher Протокол (до появления веб) для построения ЭБ, сейчас в значительной степени устарел.

handle Система глобально-уникальных имен Интернет-ресурсов и компьютерная система для управления ими, разработанные CNRI; разновидность URN.

Harvest ("урожай") Исследовательский проект по разработке архитектуры распределенного поиска, включая протоколы и форматы.

hash (хэш) Короткое значение, вычисленное из цифровых данных, которое служит для отличия их от других данных.

High Wire Press Издательское предприятие библиотеки Стэнфордского университета, предоставляет электронные версии журналов, издаваемых образовательными и профессиональными обществами.

hit (удар) 1. Входящий запрос на веб-сервер или иную компьютерную систему. 2. В обработке информации: документ, обнаруженный по запросу.

home page (домашняя страница) Начальная, вводная страница коллекции информации в веб.

HTML (Hypertext Markup Language) Простой язык разметки и форматирования текста, с гиперссылками между объектами; используется в веб.

HTTP (Hypertext Transfer Protocol) Базовый протокол веб, используется для коммуникации между браузерами и веб-сайтами.

hyperlink (гиперссылка) Сетевая ссылка между объектами ЭБ или веб-сайта.

ICPSR (International Consortium for Political and Social Science Research) Архив данных по социальным наукам в университете Мичигана.

identifier (идентификатор) Последовательность знаков, которая обозначает (идентифицирует) конкретные ресурсы в ЭБ или сети.

IETF (Internet Engineering Task Force) Сообщество, координирующее техническое развитие Интернета, включая стандарты.

InfoBus Подход к интероперабельности с использованием прокси в качестве интерфейсов между существующими системами, разработанный в Стэнфордском университете.

information discovery ("открытие" информации) Общий термин, относящийся ко всем стратегиям и методам поиска информации в ЭБ.

information retrieval (информационный поиск) Поиск в массиве информации объектов, которые удовлетворяют поисковому запросу.

Informedia Поисковая программа и ЭБ сегментов видео, находящаяся в университете Карнеги-Меллона.

Inspec сервис индексирования по физике, инженерным наукам, вычислительной технике и близким областям.

Internet Международная сеть, состоящая из независимо управляемых сетей, функционирующая на основе протоколов TCP/IP и общей системы имен. Наследник сети ARPAnet.

interoperability (интероперабельность) Задача построения согласованных сервисов для пользователей из технически различных и независимо управляемых компонентов.

inverted file (инвертированный файл) Список слов в наборе документов и их расположение в этих документах; инвертированный список - список расположения [в документах] данных слов.

item Отдельная порция материала в ЭБ; единичный пример или копия манифестации. (A specific piece of material in a digital library; a single instance or copy of a manifestation.)

Java Язык программирования, используемый для написания мобильного кода, в частности, для пользовательских интерфейсов, разработанный фирмой Sun Microsystems.

JavaScript Язык скриптов, используемый для встраивания в состав веб-страниц исполняемых инструкций.

JPEG Формат для хранения сжатых изображений.

JSTOR Служба подписки, инициированная фондом Andrew W. Mellon для конвертации ретроспективных выпусков важных журналов и предоставления их академическим библиотекам.

key (ключ) Цифровой код, используемый при шифровании или дешифровании посланий. Шифрование с закрытым ключом (private-key) использует один секретный ключ. Шифрование двойным (открытым) ключом использует два ключа, один из которых закрытый, а другой - открытый.

legacy system ("унаследованная система") Существующая система, обычно компьютерная, которая должна быть встроена в создающуюся новую систему.

lexicon (лексикон) Инструмент лингвистики, содержащий информацию о морфологических вариациях и грамматическом использовании слов.

Lexis Юридическая информационная служба, пример первого предоставления полнотекстовой информации он-лайн.

Los Alamos E-Print Archives Общедоступный сайт для быстрого распространения исследовательских статей по физике и смежным дисциплинам.

manifestation (манифестация) Форма, использованная для выражения содержания работы, в т.ч. представление в цифровой форме.

MARC (Machine-Readable Cataloging) (машиночитаемое каталогизирование) Формат, используемый в библиотеках для хранения и обмена каталожными записями.

markup language (язык разметки) Особые коды, включенные в документ и описывающие его структуру и(или) формат.

Medline Индексная служба для поиска информации в медицине и смежных дисциплинах, предоставляемая Национальной медицинской библиотекой США.

MELVYL Совместная цифровая библиотечная система академический организаций штата Калифорния; часть ЭБ Калифорнии.

Memex Концепция онлайн-библиотеки, предложенная Vannevar Bush в 1945 году.

Mercury Экспериментальный проект по ЭБ для предоставления онлайн-доступа к научным журналам в университете Карнеги-Меллона в 1987-93 годах.

MeSH (Medical Subject Headings) Набор предметных терминов и соответствующий тезаурус, используемый для описания медицинских исследований, поддерживаемый Национальной медицинской библиотекой США.

metadata (метаданные) Данные о других данных, обычно подразделяемые на описательные метаданные (библиографическая информация), структурные метаданные (форматы и структура) и административные метаданные (для управления информацией).

migration (миграция) Метод сохранения цифрового содержания, при котором информация сохраняется, а старые форматы и внутренние структуры замещаются на более современные.

MIME (Internet Media Type) Схема для указания типов данных цифровых материалов.

mirror (зеркало) Компьютерная система, содержащая копию информации, хранящейся в другой системе.

mobile code (мобильный код) Компьютерные программы (или части программ), которые передаются по сети и выполняются на удаленных компьютерах.

morphology (морфология) Грамматические и иные варианты слов, произведенных от одного корня или основы (stem).

Mosaic Первый широко использовавшийся веб-браузер, разработанный в университете Иллинойса.

MPEG Семейство форматов для сжатия и хранения оцифрованного звука и видеоизображения.

multimedia (мультимедиа) Комбинация нескольких типов среды в одном цифровом объекте или коллекции, например, графика, аудио и видео.

natural-language processing (обработка естественного языка) Использование компьютеров для интерпретации и манипулирования словами, как частью языка. (Use of computers to interpret and manipulate words as part of a language.)

NCSTRL (Networked Computer Science Technical Reports Library) Международная распределенная библиотека по вычислительной технике и сервисам, находящаяся в Корнельском университете.

Netlib ЭБ математического программного обеспечения и связанных коллекций.

NSF (National Science Foundation) (Национальный научный фонд) Американское правительственное агентство для поддержки научных и инженерных исследований, включая ЭБ.

object (объект) Технический компьютерный термин обозначающий независимый фрагмент компьютерного кода со своими данными. Отсюда - объектно-ориентированное программирование и распределенные объекты, когда объекты соединены через сеть.

object model (модель объекта) Описание структурных связей между компонентами библиотечного объекта, включая метаданные.

OCLC (Online Computer Library System) Организация, предоставляющая библиотекам - наряду с другими сервисами, - утилиты для совместного ведения каталожных записей.

OPAC (online public access catalog) Онлайн-библиотечный каталог, используемый клиентами библиотеки.

open access (открытый доступ) Свободный доступ пользователей к ресурсам, без требований по идентификации или оплате.

optical character recognition (оптическое распознавание символов) Автоматическая конверсия текста из цифрового изображения в компьютерный текст.

Pad++ Экспериментальный пользовательский интерфейс для доступа к большим коллекциям информации, основанный на семантическом zooming.

page description language (язык описания страниц) Система для кодировки документов, которая позволяет точно описать ее внешний вид после подготовки к выводу на печать или на дисплей.

PDF (Portable Document Format) Язык описания страниц, разработанный Adobe Corporation для хранения и представления изображений страниц.

peer review (рецензирование) Процедура, согласно которой статьи, присланные в академические журналы, рецензируются другими учеными перед принятием для публикации.

Perseus ЭБ исходных классических текстов (и смежных дисциплин), снабженных гиперссылками; функционирует в университете Tufts.

policy (политика) Правила, сформулированные менеджером (системным администратором) ЭБ, в которых предусмотрен порядок авторизации пользователей и доступ к конкретным материалам.

port (порт) Используемый протоколом TCP метод указания на программу, выполняемую на сервере, которая должна обрабатывать поступающие из Интернета сообщения.

PostScript Язык программирования для создания графической "выдачи" (output) для распечатки, используемый в качестве языка описания страниц.

precision (точность) Процент материалов, найденных в процессе поиска информации, удовлетворяющих данному запросу.

presentation profile (профиль представления, презентации) Указания, ассоциированные с цифровым объектом, относительно того, каким именно образом он может быть представлен пользователю.

protocol (протокол) Набор правил, который описывает последовательность сообщений, пересылаемых по сети с указанием как синтаксиса, так и семантики.

проxy (прокси) Компьютер, который функционирует как мост между двумя компьютерными системами, использующими различные стандарты, форматы или протоколы.

publish (публикация) Предоставление информации в открытый доступ.

PURL (Persistent URL) ("постоянный адрес") Метод присвоения постоянных идентификаторов с использованием стандартных веб-протоколов, разработанный OCLC.

query (запрос) Текстовая последовательность, возможно, структурированная, которая используется для поиска информации; задача, заключающаяся в поиске объектов, удовлетворяющим поисковым терминам в запросе.

ranked searching (ранжированный поиск) Методы поиска информации, которые возвращают списки документов, ранжированные по степени соответствия запросу.

RDF (Resource Description Framework) Метод уточнения синтаксиса метаданных, используемый при обмене метаданными.

Real Audio Формат и протокол для сжатия и хранения оцифрованного звука и передачи по сети для воспроизведения в реальном масштабе времени.

recall Процент соответствующих запросу объектов в материале, реально найденных в ходе поиска информации.

refresh (обновление, регенерация) Изготовление точной копии данных со старого носителя на новый для долговременного хранения.

render (представление) Трансформация цифровой информации из формы, в которой она хранилась в компьютере, в форму, в которой она будет выведена на экран (или по-другому представлена пользователю).

replication (дублирование, репликация) Изготовление копии цифрового материала для резервного архивирования, улучшения производительности, надежности или сохранения.

repository (репозиторий) Компьютерная система, используемая для хранения коллекций ЭБ и предоставления их пользователям.

RSA encryption Метод шифрования двойным ключом (dual-key).

scanning (сканирование) Метод конверсии, при котором физический объект, например, печатная страница, превращается в набор пикселей (графическое цифровое изображение).

search term (поисковый термин) Отдельный термин в поисковом запросе, обычно слово или короткая фраза.

secondary information (вторичная информация) Информационные источники, которые описывают другую (первичную) информацию, например, каталоги, индексы, рефераты; используется для управления коллекциями и поиска информации.

security (безопасность) Методы и практика сохранения целостности компьютерных систем, сервисов и коллекций ЭБ.

server (сервер) Компьютер в сети, который, в отличие от клиента, хранит коллекции и предоставляет услуги.

SGML (Standart Generalized Markup Language) Система для создания языков разметки, которые представляют структуру документа.

SICI (Serial Item and Contribution Identifier) Идентификатор томов сериальных изданий или статей в сериальных изданиях.

speech recognition (распознавание речи) Автоматическая конверсия произносимых слов в компьютерный текст.

STARTS Экспериментальный протокол для распределенного поиска, который позволяет клиенту комбинировать результаты от нескольких поисковых программ.

stemming Сокращение морфологических вариантов слова к общей основе (в процессе поиска информации).

stop word Слово, которое настолько распространено, что обычно игнорируется при поиске информации. Набор таких слов называют стоп-списком.

structural type (структурный тип) Метаданные, указывающие на структурную категорию цифрового объекта.

style sheet (таблица стилей) Набор правил, который определяет разметку документа при представлении (внешнего вида документа пользователю). (A set of rules that specify how markup in a document translates into the appearance of the document when rendered.)

subscription (подписка) В ЭБ - оплата частным лицом или организацией доступа к конкретным коллекциям или сервисам, обычно на фиксированный период времени, например, на год.

subsequent use (последующее использование) Использование материалов после того, как они выходят из под контроля ЭБ (поступают на компьютер пользователя).

tag (тэг) Специальная последовательность знаков в размеченном тексте, указывающая на структуру или формат.

TCP/IP Базовые протоколы Интернета. IP использует цифровые IP-адреса для объединения сегментов сети; TCP предоставляет возможность надежной доставки посланий между компьютерами в сети.

TEI (Text Encoding Initiative) Проект по представлению текстов с гуманитарной направленностью в цифровой форме. Также DTD, используемый в этой программе.

TeX Метод кодировки текста, который точно описывает внешний вид печатной копии документа; особенно удобен для математической нотации. LaTeX является версией TeX.

thesaurus (тезаурус) Лингвистический инструмент, который связывает слова с их значениями.

Ticer Summer School (летняя школа в Ticer) Программа университета в Tilburg (Голландия) по образованию опытных библиотечных работников в вопросах ЭБ.

transliteration (транслитерация) Последовательная замена знаков одного алфавита (набора знаков) знаками другого алфавита (набора).

TREC (Text Retrieval Conferences) Ежегодные конференции, на которых рассматриваются методы обработки текстов в рамках стандартных коллекций и задач.

Tulip Эксперимент, в рамках которого издательство Elsevier Science сканировало материалы из научных журналов; также группа университетов, размещавших эти материалы на своих локальных компьютерах.

UDP (User Datagram Protocol) Интернет-протокол, который передает пакеты данных без проверки ошибок.

Unicode 16-битный код для представления знаков из большинства известных на земле алфавитов. UTF-8 представляет собой альтернативную кодировку, в котором один или несколько 8-битных байтов представляют каждый знак Unicode.

union catalog (сводный каталог) Единый каталог, содержащий записи о материалах из нескольких коллекций или библиотек.

URL (Uniform Resource Locator) Ссылка на ресурс в Интернете, определяющая протокол, компьютер и файл на этом компьютере, а также параметры. Полный URL указывает расположение как доменное имя или IP-адрес; относительный URL указывает положение относительно текущего файла.

URN (Uniform Resource Name) Имя Интернет-ресурса, независимое от его расположения.

WAIS Ранняя версия Z39.50, используемая в ЭБ до веб, сейчас в значительной версии устарела.

Warwick Wramework Общая модель, которая описывает различные части комплексного объекта, включая различные категории метаданных.

watermark (водяной знак) Код, включенный в цифровой материал с целью установления владельца; может быть как видимым, так и не видимым для пользователя.

web crawler Программа индексирования в веб, которая строит индексы путем последовательного перехода по гиперссылкам с одной веб-страницы на другую.

webmaster (веб-мастер) Человек, управляющий веб-сайтом.

web search services (поисковые службы веб) Коммерческие сервисы, которые позволяют выполнять поиск в веб, включая Yahoo, AltaVista, Excite, Lycos, Infoseek и др.

web site (веб-сайт) Коллекция информации в веб, обычно хранимая на веб-сервере.

Westlaw Юридическая информационная служба издательства West.

World Wide Web (web) (веб) Связанный гиперссылками набор информационных источников в Интернете и технологии, которые при этом используются, включая HTML, HTTP, URL и MIME.

World Wide Web Consortium (W3C) Международный консорциум, базирующийся в MIT (Массачусетский технологический институт), координирующий техническую разработку веб.

work (работа) Абстрагированное интеллектуальное содержание некоторых материалов в ЭБ. (The underlying intellectual abstraction behind some material in a digital library.)

Xerox Digital Property Rights Language Синтаксис и правила для формулирования прав, условий и оплаты за цифровые работы.

XSL (eXtensible Style Language) Язык таблиц стилей для использования с XML, производный от CSS.

XML (eXtensible Markup Language) Упрощенная версия SGML, предназначенная для использования с онлайн-информацией.

Z39.50 Протокол, позволяющий компьютеру проводить поиск информации в удаленной системе, создавать наборы результатов для дальнейших манипуляций и обрабатывать информацию; используется главным образом для библиографической информации.

Индекс