МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
АРХИВНОЕ УПРАВЛЕНИЕ ТОМСКОЙ ОБЛАСТИ

ДОКУМЕНТ В СИСТЕМЕ СОЦИАЛЬНЫХ КОММУНИКАЦИЙ

Сборник материалов
III Всероссийской научно-практической конференции с международным участием
(г. Томск, 25–26 октября 2007 г.)

Томск
2008

Ж.А. Рожнёва, И.И. Корчакова

WEB-ДОКУМЕНТЫ: ПРОБЛЕМА ОПРЕДЕЛЕНИЯ И КЛАССИФИКАЦИИ

Становление и развитие постиндустриального, информационного общества сопровождается небывалым количественным ростом потоков документированной информации и появлением новых технологий документирования. В связи с этим в поле зрения документоведения наряду с традиционными документами оказались электронные документы, создаваемые с помощью электронно-вычислительной техники. В последнее время развитие и широкое внедрение во все сферы общественных отношений глобальной сети Интернет привело к формированию нового подвида электронных документов – Интернет-документов.

Информационные ресурсы и сервисы глобальной сети Интернет чрезвычайно многообразны, соответственно в ней функционируют документы разных видов. Одной из наиболее популярных служб сети является World Wide Web (WWW), которая управляет доступом к информационным ресурсам в гипертекстовом формате. Среда WWW не имеет централизованной структуры и представляет собой совокупность огромного количества web-страниц. Тематически объединенные группы web-страниц называют web-узлами или web-сайтами. Для просмотра web-страниц используются специальное программное обеспечение так называемые браузеры. Web-документы вызывают особый интерес как весьма сложные объекты, в количественном отношении уже сейчас измеряющиеся миллиардами.

Современная документоведческая наука достигла больших успехов в изучении традиционных форм документа, в последнее время активно ведется разработка концептуальных подходов к пониманию электронного документа. Однако область сетевых документов остается пока малоисследованной, в том числе проблема определения и классификации web-документов.

Понятие «web-документ» широко используется, как специалистами в области информационно-коммуникационных технологий, так и другими, включая документоведов. При этом подходы к его определению существенно различаются, что во многом обусловлено как сложностью самих web-документов, так и дискуссионностью основополагающего понятия «документ».

Напомним, что в современном документоведении в целом утвердилось представление о документе как о документированной информации,

закрепленной на материальном носителе. При этом сторонники стандартизованного подхода рассматривают в качестве документов в основном служебную, деловую документацию. Сторонники более широкого подхода предлагают в определении документа не ограничиваться рамками управления и делопроизводственной практики. Подходы к определению web-документа варьируется в рамках указанных парадигм.

Один из взглядов на определение web-документа, представлен в статьях О.И. Рыскова [1]. Под web-документом понимается документ, размещенный на web-странице, являющейся логической единицей сети. С точки зрения автора web-документы размещаются на страницах сайта в числе прочей информации (текстов, графиков, изображений). По сути, О.И. Рысков рассматривает web-документ как оцифрованный традиционный документ, размещенный в сети на определенном сайте, или как публикацию, не имеющую бумажного аналога, но отображаемую в окне браузера в традиционной для бумажного документа форме.

Данный подход, основанный на традиционном видении документа, по нашему мнению, не учитывает многообразия возможностей web-технологий. Не вдаваясь в тонкости функционирования среды WWW, следует всё же отметить, что web-страницы отличаются сложной информационной структурой. Они могут одновременно включать в себя такие разнородные объекты, как текст, графические изображения, звук, видео, интерактивные элементы и представлять их в виде взаимосвязанного целого. Традиционный аналоговый документ такими возможностями не обладает. Кроме того, выделение отдельных элементов web-страницы в качестве самостоятельных неизбежно ведет к нарушению контекста, что мешает адекватному восприятию информации.

Другую точку зрения на определение web-документа высказал С.И. Семилетов. Он считает, что развитие сети Интернет привело к возникновению нового подвида электронных документов, которые имеют программную основу. К ним относятся мультимедиа документы, в частности web-сайты [2]. Таким образом, данный автор в качестве единичного web-документа рассматривает web-сайт, что, на наш взгляд, не совсем верно.

Сайт – это сложный информационный ресурс, содержащий совокупность законченных сообщений (текстов), объединённых, как правило, общностью тематики, а иногда связанных между собой лишь косвенно. В ряде случаев сайт, представляющий собой некий ансамблевый документ, действительно выступает как единичный элемент сети. Например, в каталогах Интернет-ресурсов. Однако сам принцип гипертекстовой организации информации не позволяет оперировать сайтом как единым це-

лым. Пользователь сети перемещается посредством гиперссылок не по сайтам, а по их страницам.

Последняя точка зрения, широко распространенная в профессиональной среде специалистов по информационно-коммуникационным технологиям, отождествляет единичный web-документ с web-страницей. На наш взгляд, такой подход следует признать наиболее правильным. При этом, несмотря на широкое использование термина «web-документ», в специализированной литературе по сетевым технологиям преимущественное внимание уделяется техническим и технологическим аспектам. Поэтому весьма интересным представляется рассмотреть web-страницу с документоведческих позиций.

Прежде всего, анализ внешней структуры web-страницы позволяет говорить об ее сходстве с традиционными аналоговыми документами. Несмотря на всё многообразие дизайна и содержания, можно говорить о складывании определённых правил построения web-страниц. В частности, web-страница, как и любой традиционный документ, включает набор определенных элементов (заголовок, навигационная панель, содержание или контент и т.д.). При этом каждая из частей имеет свой фиксированный размер, соблюдение которого необходимо для адекватного отображения страницы на экране компьютера. Фиксированный размер и определенное положение реквизитов, как известно, является характерной чертой традиционного документа. Таким образом, во внешней структуре web-документа, выделяются элементы, которые можно сопоставить с реквизитами традиционного официального документа.

Кроме того, каждый web-документ имеет внутренние технологические характеристики, представляющие собой метаинформацию. В начале каждого кода страницы содержатся мета-теги. Они не отображаются в браузере и содержат информацию, описывающую web-документ в целом (уникальный URL-адрес страницы, формат файлов, их размер, дата создания и последнего обновления страницы, заголовок, язык на котором написан основной текст страницы, краткое её содержание). Эти данные также можно рассматривать в качестве своего рода реквизитов web-документа.

Несмотря на некоторое внешнее сходство с традиционными документами web-страницы обладают существенными особенностями. О гипертекстовой организации информации и комплексном ее характере уже говорилось. Другой отличительной чертой web-страниц является их включенность в сетевое пространство. Только в так называемом on-line режиме web-документы отображаются и функционируют в полной мере. Для них характерно также отсутствие «жесткого» форматирования.

Внешнее оформление web-страницы, которое видит конкретный пользователь, напрямую зависит от параметров используемого компьютера и настроек программы-браузера. Более того, один и тот же HTML-файл, то есть один и тот же набор данных, абсолютно по-разному отображается в браузере и текстовом или специальном web-редакторе, где можно наблюдать исходный код страницы. Web-страницу, особенно при наличии интерактивных и динамических элементов, невозможно адекватно перенести на бумажный носитель. То есть для web-документа принципиально важна цифровая форма существования.

Таким образом, следует признать, что web-страницы являются очень сложными объектами. Они отличаются не только от традиционных документов, но и от электронных документов других форматов. Тем не менее, их, несомненно, можно отнести к документированной информации и рассматривать как специфическую документную единицу информационного web-пространства.

Другой сложной задачей является классификация web-документов. Отсутствие единой классификационной схемы затрудняет поиск необходимой информации в глобальной сети. При этом создать одну универсальную классификацию применительно к web-страницам представляется весьма затруднительным или даже невозможным ввиду их чрезвычайного многообразия. Это делает неизбежным существование различных классификаций web-документов.

Ряд исследователей, изучающих электронные документы, предлагают свои классификации, которые включают Интернет-ресурсы. Большинство из них выделяют web-страницы (или web-сайты) в качестве отдельных документов и относят их к тем или иным группам. Так, Г.З. Залаев включает их в группу современных электронных документов, существующих только в цифровом виде [3. С. 62–65]. Е.В. Боброва относит web-страницы к числу Интернет-документов, создаваемых самим человеком (вручную, или с помощью каких-то программных средств) [4. С. 108–109]. Е.В. Злобин выделяет домашние странички и web-сайты в отдельную группу [5. С. 17]. Таким образом, исследовательские классификации охватывают Интернет-ресурсы в целом и не касаются собственно web-страниц.

Для классификации web-документов можно использовать специальные стандарты, созданные для описания ресурсов, в том числе электронных. Например, в межгосударственном стандарте ГОСТ 7.83–2001 электронные издания классифицируются по различным основаниям: по природе основной информации, технологии взаимодействия с пользователем, целевому назначению и т.д. [6] Подобные критерии подходят и для классификации web-страниц.

Следует также упомянуть и международный стандарт ИСО 15836:2003 «Дублинское ядро», предлагающий структуру метаописаний различных ресурсов, которая считается сейчас одной из лучших. Концепция дублинского ядра предполагает, что каждый создатель сетевого ресурса должен включать в структуру web-страницы определенный набор элементов описания информации, то есть метаданных. Наибольший интерес для построения классификации представляет такой элемент метоописаний, как «тип ресурса». В дублинском ядре содержится специальный словарь определяющий типы ресурсов, который, по сути, и представляет собой базовый классификатор. Он включает 12 элементов: коллекция, структура данных, изображение, интерактивный ресурс, сервис, компьютерная программа, звук, текст, статическое изображение и др. Таким образом, тот или иной web-документ может быть отнесен к одному или нескольким указанным типам ресурсов [7].

Различные классификации web-документов разрабатываются и в самой Интернет-среде. К ним можно отнести классификации специалистов в области web-программирования и составителей каталогов сетевых ресурсов.

Как правило, разработчики web-сайтов группируют их по тематике, по организации, которая заказывает сайт или по объему материальных, интеллектуальных, временных затрат для создания ресурса. При этом большинство дизайнерских агентств предлагают схожие между собой тематические классификации web-ресурсов, которые лишь незначительно различаются между собой. Например, чаще всего выделяются домашние странички, информационные сайты, бизнес сайты, поисковые системы и каталоги, операционные сайты, образовательные ресурсы, правительственные сайты и др.

Среди классификаций, предлагаемых в каталогах Интернет-ресурсов, следует выделить как наиболее разработанную классификацию каталога Яндекс. Данная классификация является фасетной, то есть представляет собой совокупность нескольких независимых классификаций, осуществляемых одновременно по различным основаниям. Основные фасеты, используемые в каталоге, – это тема, регион, жанр, источник информации, адресат информации, сектор экономики. Значения фасетов проставляются вручную редакторами при описании ресурсов. Например, признак «тема» имеет порядка 600 значений и описывает предметную область Интернет-ресурса, «регион» определяет принадлежность ресурса к одному из 230 географических областей и т.д. [8]. Преимуществом данной классификации является то, что она предоставляет пользователю многоаспектное описание web-документов.

Подводя итог, следует еще раз подчеркнуть сложность проблемы определения web-документов и их классификации. Традиционный документ отличает двуединство его природы (информации и материального носителя). Применительно к web-документу можно говорить о единстве информационной, технологической и технической составляющих, что обусловливает трудности в его определении. Также представляется весьма сложным создание какой-либо одной исчерпывающей классификации сетевых web-документов. Более перспективным является описание web-документов по различным основаниям в рамках отдельных классификаций и дальнейшее усовершенствование методов поиска документов в сети на основе комплекса этих классификаций.

Примечания

1. ***Рысков О.И.*** Web-документ // Делопроизводство. 2004. № 7; ***Рысков О.И.*** Web-документ // Служба кадров. 2004. № 10.
2. ***Семилетов С.И.*** Формирование коллекций из Интернет-документов и проблемы авторского права // Информационный бюллетень «Вестник архивиста». 2003. № 5. С. 258–274.
3. ***Залаев Г.З.*** Анализ и классификация электронных документов // Вестник архивиста. 1999. № 2/3.
4. ***Боброва Е.В.*** Интернет-документ как объект архивного хранения // Информационный бюллетень Ассоциации «История и компьютер». М., 2000. № 26/27.
5. ***Злобин Е.В.*** О некоторых проблемах классификации и описания электронных документов как исторического источника // Круг идей: электронные ресурсы исторической информатики: Труды VIII конф. Ассоциации «История и компьютер». Москва; Барнаул: Изд-во Алт. ун-та, 2003.
6. ГОСТ 7.83–2001. Электронные издания. Основные виды и выходные сведения. (Система стандартов по информации, библиотечному и издательскому делу).
7. ***Манцивода А.В.*** Система метаописаний Dublin Core // TeaCODE.com. Исследовательская группа ЦНИТ ИГУ. Электрон. дан. Иркутск, 2002–2005. – Режим доступа: http://teacode.com/concept/eor/dc.html#contr.
8. ***Браславский П.И., Вовк Е.А., Маслов М.Ю.*** Фасетная организация интернет-каталога и автоматическая жанровая классификация документов // Яндекс. Company. – Электрон. дан. 1997–2007. – Режим доступа: http://company.yandex.ru/articles/article8.html.