Нежурбеда Г. Г.

Обеспечение сохранности электронных документов в библиотеках

В последние годы количество электронных документов, поступающих в библиотеки, стремительно возрастает. Наряду с пополнением фондов электронными изданиями, поступающими на отдельных носителях информации, увеличение библиотечного фонда происходит также благодаря библиотечным программам создания электронных версий "бумажных" документов и проектам сохранения Интернет-ресурсов. С увеличением количества электронных документов возникает обеспокоенность возможностью сохранить и обеспечить эффективный надежный доступ к ним.

Целью исследования, которое предшествовало написанию статьи, было желание показать трудности применения традиционных подходов обеспечения сохранности документов к электронным документам. И, как следствие, возникает необходимость разработки специальных программ, нацеленных на развитие комплекса организационных, технологических, правовых механизмов, обеспечивающих сохранение электронных документов библиотеками. Не отрицая необходимости проведения работ в этом направлении, результатом проведенного исследования стало твердое убеждение в том, что современная библиотечная наука и практика располагают значительным теоретическим и методическим потенциалом в области обеспечения сохранности электронных документов. Следует лишь грамотно применять накопленный десятилетиями опыт.

Первостепенное значение для обеспечения процессов сохранности электронных документов имеет понимание того, что именно является объектом хранения: носители информации (дискеты, компакт-диски, магнитные ленты); файлы (определенная последовательность битов) или информация, содержащаяся в электронном документе (интеллектуальное содержание документа). Сохранение носителей информации как процесс поддержания неизменности во времени физико-химических характеристик их материальной основы не может привести к удовлетворительному результату, поскольку быстрое развитие технологий приводит к вытеснению старого аппаратного и программного обеспечения новым, не всегда поддерживающим возможность использования ранее записанной информации либо гарантирующим идентичность воспроизведения документа. По той же причине сохранение электронного документа как определенной последовательности битов не гарантирует того, что информация, которую содержит документ, будет сохранена.

Сохранение только лишь информации, которую содержит документ (даже если бы это было возможным в полном объеме), также не может привести к удовлетворительным результатам. Сама информация достаточно быстро устаревает. В случае с традиционными документами, речь может идти о хранении "книжного памятника" как документа, обладающего культурной и исторической ценностью, сущность которого реализуется в единстве опубликованного произведения и способа его материального воплощения. Под сохранностью книжных памятников понимается их сохранение в возможно полном объеме изначальных характеристик и важных в историко-культурном отношении особенностей, приобретенных в процессе бытования книги. Говоря о сохранении электронного документа, также следует помнить о необходимости сохранения в "возможно полном объеме изначальных характеристик" и оформления, и функциональных возможностей документа, например, возможности использования гиперссылок или других способов навигации по документу и между документами.

Наиболее удачная, по мнению многих экспертов, концептуальная модель сохранения цифрового объекта была предложена К. Тибодо (Kenneth Thibodeau) [1]. Термин "цифровой объект" он определил как "информационный объект любого типа информации или любого формата, который выражен в цифровой форме", что в контексте задачи цифрового сохранения соответствует понятию "электронный документ", применяемому в библиотечной терминологии.

Сохранение цифрового объекта рассматривается как сохранение (наследование) его свойств. "Все цифровые объекты — объекты с многократным наследованием, то есть свойства любого цифрового объекта унаследованы от трех классов. Каждый цифровой объект — физический объект, логический объект и концептуальный объект, и свойства на каждом из этих уровней различны".

Как физический объект, цифровой объект является "простой надписью признаков на среде", которую определяет способ регистрации информации на носителе. Например, существуют физические различия между регистрацией информации на магнитных и оптических носителях (в одном случае используется технология рельефной записи, в другом магнитной). Способ регистрации информации может изменяться и в пределах одного типа носителя информации. Например, данные на магнитном носителе могут быть записаны с различной плотностью, иметь разные размеры блока и т.д. Вопросы сохранения цифрового объекта на физическом уровне в основном связаны с физическими файлами, которые идентифицируются и управляются некоторой системой хранения. Подходы к сохранению информации на физическом уровне едины для различных видов документов — текстового файла, графического изображения или аудиофайла. Сохранение информации на физическом уровне требует периодического копирования (перемещения) информации на новые носители. При этом особенно важно использование современных, высокой степени надежности носителей информации. Немаловажное значение при выборе носителя информации имеет простота его использования. Планируя долговременное хранение носителей информации, необходимо обратить внимание и на поддержание в работоспособном состоянии соответствующего оборудования в тот же период времени.

Логический объект в модели К. Тибодо — единица, признаваемая определенным программным обеспечением. Другими словами: цифровой информационный объект является объектом "логики" некоторого программного обеспечения. Правила, которым подчиняется логический объект, не зависят от способа регистрации информации. При помещении данных в оперативную память компьютера не имеют значения ни тип носителя информации, с которого были считаны данные, ни правила, по которым они были размещены на носителе информации. Каждый физический объект содержит один или несколько логических объектов. Например, если текстовый файл представлен только одним файлом, логический объект совпадает с физическим объектом. Если программное обеспечение использует для отображения документа специальные шрифты, хранящиеся в других файлах или специальные "инструкции" (например CSS для html-файла), то в этих случаях одному логическому объекту соответствует несколько физических объектов. Несколько логических объектов могут храниться в одном архиве (например, ZIP-, ARJ- или RAR-файл) — логические объекты могут содержать другие логические объекты.

Таким образом, чтобы сохранить цифровой объект на логическом уровне, необходимо иметь определенную информацию обо всех составляющих объект файлах, влияющих на его воспроизведение, т.е. знать их местоположение, идентификаторы, требования правильной обработки каждого типа данных, программное обеспечение, которое может производить правильную обработку.

Концептуальный объект — это объект, с которым человек имеет дело в реальном мире: конкретный документ, который был признан как значащая единица информации: книга, статья, фотография, карта. Например, текст этой статьи, подготовленный в WORDe и сохраненный в виде doc-файла, и соответствующий ему html-файл, представленный на сайте НБУВ — два различных логических представления одного и того же концептуального объекта. Каждый из этих файлов, будучи обработан соответствующим программным обеспечением, позволяет воспроизвести текст истинного объекта. Приведя подобный пример, К. Тибодо делает вывод о двух важных аспектах цифровых объектов: (первый) могут существовать разные способы представления одного и того же концептуального объекта и (второй) каждый из этих способов может сохранять необходимые характеристики первоначального концептуального объекта. На этом строится концепция цифрового сохранения. Идеальная система сохранения цифрового объекта представляет нейтральный канал для передачи информации в будущее. Этот канал не должен искажать или изменять сообщения, переданные любым способом. Парадокс цифрового сохранения состоит в том, что, стремясь сохранить цифровой объект в неизменном, подлинном состоянии, неизбежно приходится производить изменения цифрового объекта. Возникает вопрос: является ли подлинным цифровой объект после многократного изменения? Ссылаясь на К. Линча (Clifford Lynch), К. Тибодо показывает, что установление подлинности объекта, в конечном счете, лишь вопрос доверия. На основе доверия к определенному человеку, организации, некоторой системе или методу, который осуществляет контроль над передачей информации в пространстве, времени или в рамках перехода от одной технологии к другой, делается вывод о подлинности цифрового объекта.

Другой парадокс цифрового сохранения состоит в том, что чем более востребован электронный документ, тем выше вероятность его сохранения, так как чаще будут прилагаться усилия к восстановлению документа в форме, удобной для восприятия потребителем. Из библиотечной практики известно, что существуют документы, которые должны быть сохранены, несмотря на невысокий текущий уровень востребованности. Поэтому необходима разработка единообразного подхода к обеспечению сохранности электронных документов, как части культурного наследия.

Меры, традиционно применяемые библиотеками, по обеспечению сохранности документов, сопровождают целый ряд библиотечных процессов. Обычно особо выделяют требования к обеспечению сохранности на этапах поступления документов в библиотеку; обработки; хранения; во время использования; во время перемещения или транспортировки. "Условия хранения и сохранности закладываются в процессе моделирования фонда, поскольку именно тогда определяется, документы какого рода подлежат постоянному, долговременному или кратковременному пребыванию в фонде библиотеки… Подход к хранению непосредственно зависит от того, на что нацелена библиотека: хранить ли знания, заключенные в документе или хранить материальную основу документа… Политика хранения зависит непосредственно от уставных целей библиотеки. Эти цели предстоит четко сформулировать и зафиксировать в основополагающих документах" [3, с.75-76]. Если библиотека принимает на хранение электронные документы, это должно быть отражено в Уставе библиотеки, а в профиле комплектования подробно оговорено.

На этапе поступления документа в библиотеку защитную функцию выполняет штемпелевание приобретенных документов, простановка на них инвентарных номеров, наклейка экслибриса [2, с.151.]. Сегодня при приеме электронных документов на хранение в библиотеках все чаще используют право цифровой подписи, что позволяет, во-первых,

получить гарантию соответствия копии оригиналу и, во-вторых, при необходимости подтвердить право обладания библиотекой электронным документом.

Другой, более доступный способ проверки достоверности хранимого в библиотеке электронного документа дает использование алгоритма MD5, который часто применяется в программных продуктах при реализации механизма цифровой подписи. Кроме того, этот алгоритм можно использовать и для проверки электронных документов на дублетность. Описание алгоритма и его реализации размещены на многих страницах Интернета (например, [4, 5]). Результатом работы алгоритма является 128-битный ключ. По утверждению автора алгоритма этот ключ уникален для каждого файла, подобно тому, как уникальны отпечатки пальцев каждого человека [4]. Этот алгоритм уже несколько лет используется во многих библиотеках. Например, в национальной библиотеке Финляндии в рамках проекта сохранения Интернет-документов используется специализированная программа, которая копирует свободно доступные файлы с веб-серверов национальной доменной зоны .fi, определяет для каждого файла MD5-ключ, переименовывает архивную копию файла, используя MD5-ключ в качестве имени файла [6]. Это делает невозможным дублирование хранимых в архиве файлов, появляется возможность проверить аутентичность хранимого документа (повторное вычисление MD5-ключа должно давать тот же результат). Кроме того, MD5-ключ является частью номера государственной регистрации в национальной библиографии (National Bibliography Number) и используется с приставкой urn:nbn:fi-fea- как универсальное имя ресурса URN [7].

Можно назвать и другие примеры составления имен файлов для электронных документов, поступивших в библиотеку. В Национальной библиотеке Украины имени В.И.Вернадского для документов электронной библиотеки использовался такой алгоритм составления 8-ми символьного имени файла: первые 2 символа — две последние цифры года издания, 3 следующие — первые буквы фамилии, имени и отчества автора документа; заключительные 3 символа имени файла — первые буквы трех последних слов названия документа. Для документов полнотекстовой базы данных "Президент Украины: выступления, обращения, поздравления" — первые две буквы имени файла — первые буквы названия издания, следующие — год издания, месяц и число, например: gu20030827 — газета "Голос Украины" за 27 августа 2003 года. Использование мнемонических имен удобнее для пользователя, но преимущества алгоритмов, создающих гарантировано уникальное и фиксированное имя для определенного документа — очевидны.

На этапе поступления электронного документа в фонд, безусловно, должен осуществляться контроль целостности и полноты документа, а также антивирусный контроль. На этом же этапе желательно создание копии документа. При этом документ, поступивший в библиотеку, считается эталонным (контрольным), его копии — рабочими. Раздельное хранение эталонного и рабочего экземпляра электронного документа повышает надежность его сохранения. Читателям предоставляют возможность пользования только рабочим экземпляром. На этапе поступления электронного документа в фонд, библиотека может преобразовывать поступившие документы в более приемлемые, с ее точки зрения, форматы для хранения документов, например, в PDF-формат. Следуя развитию информационных технологий, все чаще для этих целей библиотеки применяют XML-формат. Сохранение документов в альтернативных форматах повышает возможность аутентичного воспроизведения документа в будущем, поскольку развитие программного обеспечения для каждой версии документа будет происходить раздельно и управление версиями документа будет осуществляться также раздельно.

Известно, что функцию сохранности библиотечного фонда обеспечивает его учет, задача которого состоит в том, чтобы фиксировать наличие и местонахождение каждого документа [3, с. 82]. В последние годы крупные библиотеки усиленно работают в направлении создания "электронных библиотек" или "полнотекстовых баз данных" формируя их фонд из оцифрованных "бумажных" документов и из электронных документов, переданных библиотеке на хранение. При этом каталог "электронной библиотеки" часто представляет отдельную базу данных и имеет отдельный интерфейс. В генеральный каталог библиотеки информация о документах электронной библиотеки не включается. Таким образом, генеральный каталог библиотеки не содержит информацию обо всех документах, хранимых в библиотеке. Забыта существующая практика учета копий документов, независимо от вида носителя, аналогично их оригиналу. Электронная копия бумажного документа по-другому представляет информацию, содержащуюся в документе, является другим способом ее материального воплощения, и должна учитываться отдельно.

Библиографическое описание документов в "электронной библиотеке" часто не является описанием собственно электронного документа, а формируется из описания документа, существующего на бумажном носителе и гипертекстовой ссылки, позволяющей увидеть электронный вариант этого документа. Такое пренебрежительное отношение к учету электронных документов может привести к потере информации о составе фонда электронных документов библиотеки и усложнить автоматизацию процессов управления электронными документами.

Базовым средством, широко применяемым для описания значительного класса электронных документов, служат метаданные. Существует несколько десятков систем метаданных, разработанных для разных целей. Руководство по сохранению цифрового наследия [8], вышедшее под эгидой ЮНЕСКО в 2004 году, приводит схему метаданных, необходимых для выполнения процессов сохранения цифрового объекта, разработанную Национальной библиотекой Новой Зеландии [9]. Эта схема включает такие группы элементов: описание цифрового объекта; описание процессов, связанных с объектом (включая создание); описание технических характеристик всех файлов, составляющих цифровой объект; описание изменений в метаданных.

Большое значение имеет определение единиц учета электронных документов. Чаще всего фонд электронных документов характеризуют, сообщая количество файлов и их суммарный объем. Учитывают электронные документы и в названиях. Например, в работе Маргарет Филлипс (Margaret Phillips) сообщается, что в архиве PANDORA (архив копий веб-сайтов, созданный в Национальной библиотеке Австралии) в конце 2002 года было более 3000 названий, приблизительно 14 млн файлов, что составляло 400 Гигабайт информации [10].

На этапе хранения обеспечение сохранности фонда достигается работой в двух направлениях. Первое направление — "профилактическое", превентивная защита, включающая обеспечение оптимального режима хранения, проведение дополнительных мер по стабилизации, обеспечение безопасности библиотечных фондов. Основные цели — защита документов от различного вида повреждений (механического, физического, химического и биологического свойства), охрана от хищений, предотвращение возгорания, затоплений и т.п.

Второе направление обеспечения сохранности фондов — "восстановительное" — кроме проведения собственно реставрационных работ, включает работу по замене изношенных и утраченных документов идентичными экземплярами. Известно, что реставрационные

работы требуют больших затрат труда, наличия хорошо подготовленных профессионалов и материальной базы для проведения таких работ. Поэтому они дорогостоящие и применяются только к тщательно отобранным документам.

Обеспечение сохранности электронных документов требует выполнения того же комплекса работ. Понятно, что остаются актуальными меры по защите от механических, физических, химических и биологических повреждений. Поскольку электронные носители информации занимают сравнительно небольшой объем и некоторые из них достаточно чувствительны к пыли, солнечному свету, изменению температуры и магнитным воздействиям — целесообразно создавать специальные гермозоны с поддержанием необходимых санитарно-гигиенического и температурно-влажностного режимов.

Реставрационные работы для электронных документов — их перевод в на современные носители информации, а при необходимости, — преобразование документов в более современные форматы. Для случаев перехода к следующему поколению компьютеров, операционных систем, новым версиям прикладного программного обеспечения, не позволяющего аутентично воспроизводить хранимый "оригинал" документа, рассматриваются возможности разработки программного обеспечения, эмулирующего работу устаревшего.

Повышению уровня сохранности фонда электронных документов способствуют и применение специальных технологических решений. Например, использование иерархической структуры хранилища электронных документов. В соответствии с этой технологией, для каждого документа, в зависимости от частоты спроса и степени важности, определяется уровень хранения, ассоциированный с конкретными запоминающими устройствами. Верхний уровень — быстродействующие общедоступные устройства (винчестеры), нижний — отдельно хранящиеся накопители (дискеты, магнитные ленты, CD-ROMы, DVD). Промежуточный уровень образуют устройства, непосредственно подключенные к серверам и позволяющие оперативно подключать необходимые накопители. При поступлении запроса на воспроизведение электронного документа, этот документ перемещают на верхний уровень иерархии. Если ресурса накопителя верхнего уровня не хватает, часть документов верхнего уровня, к которым не было обращения наибольшее время, перемещают на уровень ниже. Таким образом, на более быстродействующих общедоступных устройствах будут находиться документы повышенного спроса. Сегодня различные компании предлагают готовые решения по организации хранения электронной информации. К сожалению, для большинства библиотек постсоветского пространства они слишком дороги. Обеспечение сохранности документов на этапе использования заключается, прежде всего, в управлении доступом к электронным документам с помощью организационных и технологических мер, предотвращающих несанкционированное обращение к документам.

Повышению сохранности электронных документов в библиотеках способствуют четкое распределение обязанностей каждого из сотрудников, участвующих в процессах комплектования, обработки, хранения, обеспечения использования электронных документов; а также — разработка и принятие соответствующих нормативных документов и методических рекомендаций; создание и развитие необходимой технологической инфраструктуры; подготовка и обучение персонала для работы с электронными документами. Существует необходимость информирования и активного содействия создателям электронных документов со стороны библиотек в использовании методов, которые помогут долгосрочному сохранению документов, основанных на стандартах и лучших современных разработках.

Література

1. *Thibodeau K.* Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years [Electronic recource] // The State of Digital Preservation: An International Perspective: Conference Proceedings (July 2002). — Way of access: http://www.clir.org/pubs/reports/pub107/thibodeau.html . — Title from the screen.
2. *Столяров Ю. Н.* Библиотечный фонд. — М., 1991. — 271 с.
3. *Столяров Ю. Н.* Как сохранить библиотечный фонд. — М., 2001. — 256 с.
4. RFC 1321. The MD5 Message-Digest Algorithm [Electronic recource] / Rivest R. — 1992. — Way of access: http://www.ietf.org/rfc/rfc1321.txt. — Title from the screen.
5. MD5 Homepage: (unofficial) [Electronic recource] . — Way of access: http://userpages.umbc.edu/~mabzug1/cs/md5/md5.html. — Title from the screen.
6. *Lounamaa K.* EVA: The Acquisition and Archiving of Electronic Network Publications In Finland [Electronic recource]. — Way of access: http://%20www.ercim.org/publication/ws-proceedings/DELOS6/eva.rtf. — Title from the screen.
7. RFC 3188. Using National Bibliography Numbers as Uniform Resource Names [Electronic recource] / Hakala J. — 2001. — Way of access: http://www.ietf.org/rfc/rfc3188.txt. — Title from the screen.
8. Guidelines for the preservation of digital heritage [Electronic recource] / UNESCO. — 2003. — Way of access: http://unesdoc.unesco.org/images/0013/001300/130071e.pdf. — Title from the screen.
9. Metadata Standards Framework — Preservation Metadata [Electronic recource] / National Library of New Zealand. — 2002. — Way of access: http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf. — Title from the screen.
10. *Phillips M.* Collecting and Preserving Australian Online Publications [Electronic recource]. — 2002. — Way of access: http://www.nla.gov.au/nla/staffpaper/2002/phillips2.html. — Title from the screen.