Прикладные аспекты в информатике

Проблема интерпретации электронных документов долговременного хранения

А.В. Соловьев

Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия

Аннотация. В статье рассмотрена проблема интерпретируемости электронных документов при долговременном хранении. Показана актуальность, выполнен обзор подходов к решению проблемы. Выполнена постановка задачи обеспечения интерпретируемости электронных документов в общем виде. Определены основные сложности, связанные с использованием различных файловых форматов, разработкой формата долговременного хранения. Указаны пути дальнейших исследований, а именно: определение понятия электронного документа долговременного хранения; разработка математической модели состава информации электронного документа; оценки отчуждаемости интерпретируемости; поиск способов долговременного хранения для обеспечения интерпретируемости.

Ключевые слова: *долговременное хранение, интерпретируемость, электронный документ, метаданные, формат долговременного хранения.*

DOI: 10.14357/20790279220208

Введение

В предыдущих работах автора [1–3] было показано, что основной проблемой цифровых данных при долговременном хранении является обеспечение их сохранности, т.е. сохранение всей полноты информации как самих данных, так и метаданных, их описывающих.

В случае, когда цифровые данные преобразуются в электронный документ (далее – ЭлД), заменяющий бумажный, и дальнейшем хранении в архиве, на первый план выходит проблема интерпретируемости электронных документов.

Кроме сохранности цифровых данных и метаданных, ЭлД не должен потерять информацию о своем внешнем виде, составе расположении на листе бумаги. Информация, содержащаяся в документе, должна быть, во-первых, читаемой, т.е. ЭлД должен быть прочитан, формат, в котором он сохранен, должен интерпретироваться (раскодироваться) программно-аппаратной средой хранения.

Часто считается [4–6], что все современные форматы уже стандартизированы, поэтому проблема интерпретации ЭлД скорее надуманная. Однако, практически ни один из разработчиков форматов ЭлД не дает гарантии того, что он будет поддерживаться в течение нескольких последующих десятилетий.

Выбор формата, пригодного для долговременного хранения, является, безусловно, важным фактором для решения задачи обеспечения сохранности, что показано в работе [7]. Однако сам по себе формат не является гарантией обеспечения сохранности на длительном сроке хранения.

ЭлД должен отображаться пользователю в том виде, в котором он был создан. Т.е. должны сохраняться геометрические размеры информационных блоков документа, их взаимное расположение, шрифты, таблицы, иллюстрации и др. А это само по себе является очень нетривиальной задачей, учитывая, что разные средства интерпретации могут по-разному интерпретировать один и тот же документ даже в одном и том же формате. Так, например, разные средства работы с электронными документами, такие как Open Office, Libre Office, MS Office могут по-разному интерпретировать один и тот же документ в формате DOCX. Более того, даже разные версии MS Word по-разному интерпретируют документ в формате DOCX, в результате чего в многостраничных документах могут отличаться количество страниц. Следовательно, сохранность внешнего вида оригинального документа, особенно при длительном хранении, не гарантирована.

В работе [1] кратко описана проблема интерпретируемости цифровых данных в новых информационных условиях. Решение проблемы интерпретируемости в общем виде сводится к обеспечению возможности раскодировать формат ЭлД через десятилетия и представить его в первозданном виде на экране или при печати.

Тем самым, можно сформулировать интерпретируемость ЭлД как обеспечение возможности чтения в понятном для пользователя виде информации документа и его визуализации с помощью программно-аппаратных средств в том виде, в котором он был создан.

1. Краткий обзор проблемы интерпретируемости

Крупнейшим хранилищем электронных документов в мире является Национальный архив США (NARA[1] – National Archives and Records Administration). NARA осуществляет надзор за управлением федеральными правительственными документами и обеспечивает сохранность документов, имеющих научную, историческую и практическую ценность.

В 2005 году NARA принял решение о разработке архива электронных документов долговременного хранения ERA (Electronic Records Archives). Проект получил государственную поддержку.

Начиная с 2005 года, Государственный департамент США ежегодно передает в NARA миллионы дипломатических сообщений в электронной форме. Пентагон ежегодно передает 50 млн оцифрованных официальных документов по личному составу [8].

Однако в 2010-м году, в связи с трудностями создания ERA, срок ввода проекта в действие перенесли на 2015 год, затем – на 2017. В 2012 году проект был приостановлен и было объявлено об ограниченных возможностях использования ERA [9].

Причиной тому стало огромное множество компьютерных форматов, использование которых порождает риск неинтерпретируемости ЭлД спустя десятилетия. Хранение в единой кодировке ASCII, как этот делалось в NARA раньше, стало невозможно. Тем не менее, столкнувшись с проблемами обеспечения интерпретируемости ЭлД, в NARA было сделано много для систематизации полученного, хоть и отрицательного, опыта.

В 2013 году NARA разработала проект открытого стандарта кодированного архивного описания документов (EAD – Encoded Archival Description) на основе форматов XML (eXtensible Markup Language) [10]. В проекте определены классы документов, для каждого из них определен набор рекомендуемых и допустимых форматов хранения.

В 2017 году NARA выпустила документ ERM [6]. Требования, отраженные в этом документе, могут использовать сотрудники федеральных агентств при написании технических заданий для инструментов или сервисов управления ЭлД.

В 2017 году NARA сообщала о создании новой модели хранения и доступности президентских ЭлД, объем которых в год оценивается в 250 ТВ [11]. Также для частичного решения проблемы интерпретируемости, NARA приняло решение после 31 декабря 2022 г. принимать на хранение ЭлД в специальном цифровом формате с обязательными описательными метаданными [12].

Австралийский национальный архив, существующий с 1983 года, также как и NARA является передовым с точки зрения хранения электронных документов [13].

Национальным архивом Австралии выпущен стандарт PROS99/007 «Управление электронными документами», который содержал требования к управлению и хранению электронных документов в государственном секторе. Разработана методология проектирования и внедрения систем работы с документацией DIRKS, которая послужила основой международного стандарта ISO 15489-2001

[1] NARA www.archives.gov

[14]. Этот стандарт, в частности, принят в РФ как ГОСТ Р ИСО 15489-1-2007.

Согласно стандарту, «пригодным для использования является документ, который можно локализовать, найти, воспроизвести и интерпретировать» (ГОСТ Р ИСО 15489-1-2007, п.7.2.5). Хотя положения документа носят характер рекомендаций, тем не менее, стандарт определяет необходимый минимум характеристик документа, который должен быть соблюден при организации долговременного хранения.

В плане решения проблемы интерпретируемости Национальные архивы Австралии пошли по пути «нормализации документов», т.е. приведения всех документов, поступающих на длительные сроки хранения, к единому формату. Форматы [15], в которых происходит конвертация документов: формат на основе XML (eXtensible Markup Language) и RDF/XML (Resource Description Framework — модель данных, используемая для представления ресурсов т.н. семантической паутины (semantic web), представляющий документы, элементы классификации (как объекты) и связи между ними (как отношения или предикаты)). Можно утверждать, что частично решена проблема интерпретируемости данных. Частично потому, что формат XML не предполагает точного визуального отображения исходного документа, а предназначен для сохранения его информации и, возможно, метаданных ЭлД.

В 2001 году для Европейского Союза британской компанией Cornwell Affiliates plc была подготовлена спецификация Модельных требований к управлению электронными документами (MoReq) [16]. Не будучи формально стандартом, MoReq в настоящее время является им фактически. Уже в первой версии спецификации MoReq существовал раздел, посвященный долгосрочному хранению электронных документов, в котором одним из ключевых было требование предпочтительного использования открытых, документированных форматов в противовес к проприетарным (коммерческим) для снижения рисков неинтерпретируемости документов в будущем с возможностью восстановить документ с помощью опубликованного описания формата.

Из приведенного краткого обзора передового опыта организации долговременного хранения ЭлД отчетливо выявлена проблема интерпретируемости и отображения электронных документов при длительном хранении.

Из приведенного обзора можно сделать вывод о том, что пока не видно универсального решения проблемы интерпретируемости, несмотря на его активные поиски последние десять лет. Т.к. со временем количество электронных документов будет стремительно возрастать, следовательно, порядок сложности решения задачи будет стремительно увеличиваться.

2. Постановка задачи обеспечения интерпретируемости электронных документов

Теперь на основании проведенного обзора можно выполнить постановку задачи обеспечения интерпретируемости ЭлД и определить пути дальнейших исследований, необходимых для решения поставленной задачи.

Задача исследования формулируется следующим образом: для обеспечения долговременного хранения деловых электронных документов необходимо обеспечить их интерпретируемость (читаемость) в течение всего срока хранения.

При этом предполагается, что:

- сохранность документа на момент передачи его на длительное хранение подтверждена;
- документы не искажены;
- нет ограничений на форматы данных передаваемых на длительное хранение документов;
- аппаратно-программная среда, в которой предполагается обеспечить интерпретируемость документов, подвержена постоянному изменению (в том числе и средств интерпретации);
- нет гарантии на точное интерпретирование документа (формата документа) через десятилетия.

При долговременном хранении ЭлД возникает проблема интерпретируемости и отображения данных в новых информационных условиях, т.е. наличие возможности раскодировать хранимый формат ЭлД через десятилетия и показать документ в том или ином виде, например, отобразить на экране или распечатать.

Отсутствие стратегии в данном вопросе может, спустя десятилетия, привести к тому, что часть информации невозможно будет раскодировать из-за отсутствия (устаревания) средств интерпретации хранимых форматов данных, а также из-за потери описания хранимых форматов, в случае использования закрытых форматов представления ЭлД. Некоторым решением проблемы может быть создание конвертеров, преобразующих старые форматы ЭлД в новые, но здесь следует иметь в виду, что чем позже будет поставлена задача конвертации данных, тем сложнее будет ее решение. Кроме того, такой подход предполагает наличие постоянных высоких накладных расходов на поддержание интерпретируемости ЭлД.

Судя из приведенного выше обзора, удобно иметь единый формат долговременного хранения (или набор таких форматов), многие развитые страны (США, Австралия, Великобритания) идут именно по этому пути.

Можно кратко сформулировать требования к такому формату: простой, открытый и документированный. Даже такой набор требований существенно снизил бы вероятность «не интерпретируемости» документов, представленных для долговременного хранения в данном формате, в будущем.

Подробнее о выборе форматов долговременного хранения, их особенностях и требованиях к разработке формата долговременного хранения см. [7].

Однако есть еще ряд проблем, которые необходимо учитывать при долговременном хранении ЭлД которые могут не зависеть от формата и которые могут привести к искажению документа и путанице для пользователя:

- любой деловой ЭлД может содержать в себе данные о внесенных правках, комментарии, невидимый текст, сведения о компании и авторах правок;
- в ЭлД могут быть поля, информация которых может изменяться, что приводит к искажению всего ЭлД, например, поле с текущей датой, которая может меняться при распечатке ЭлД, иные макросы, которые могут изменить ЭлД при открытии или распечатке;
- ЭлД может содержать гиперссылки на веб-страницы или на другие связанные объекты (рисунки, схемы, другие документы). В этом случае документ может быть искажен при открытии или распечатке из-за отсутствия в нужном месте связанных объектов. Необходимо сохранять кроме ЭлД при длительном хранении еще и связанные документы.

Перечисленные проблемы позволяют сформулировать дальнейшие направления исследования:

- Определение понятия ЭлД долговременного хранения.
- Разработка математической модели состава информации ЭлД долговременного хранения.
- Разработка математической модели оценки отчуждаемости ЭлД от программно-аппаратной среды хранения.
- Разработка математической модели оценки интерпретируемости ЭлД.
- Разработка способов долговременного хранения для обеспечения интерпретируемости ЭлД.

Эти направлениям исследования будет посвящена серия статей.

Заключение

Проблема интерпретируемости электронных документов при долговременном хранении, безусловно, является актуальной, что показано в статье на основе проведенного краткого обзора подходов к решению проблемы. Однако даже существующие подходы к решению оказываются сложно реализуемыми несмотря на то, что предложены в странах с высоким развитием информационных технологий.

Это связано с отсутствием, в первую очередь, постановки задачи обеспечения интерпретируемости электронных документов, которая выполнена в общем виде в данной статье. Кроме того, в статье определены основные сложности, связанные с использованием различных файловых форматов, проблемы разработки формата долговременного хранения.

В статье также определены пути дальнейших исследований, а именно: определение понятия электронного документа долговременного хранения; разработка математической модели состава информации электронного документа долговременного хранения; разработка математической модели оценки отчуждаемости электронного документа от программно-аппаратной среды хранения; разработка математической модели оценки интерпретируемости электронных документов; разработка способов долговременного хранения для обеспечения интерпретируемости электронных документов.

В ходе дальнейших исследований планируется подготовить серию статей для описания решения проблемы интерпретируемости электронных документов для обеспечения долговременной сохранности.

Литература

1. *Solovyev A.V.* Long-Term Digital Documents Storage Technology // Lecture Notes in Electrical Engineering. 2020. Vol.641. P. 901-911.
2. *Solovyev A.V.* Authentication control algorithm for long-term keeping of digital data // IOP Conference Series: Materials Science and Engineering (MSE), vol.862(5), 052080. 2020.
3. *Solovyev A.V.* Digital media inventory algorithm for long-term digital keeping problem // IOP Conference Series: Materials Science and Engineering (MSE). 2020. Vol. 919(5). 052003.
4. Open Government Partnership UK National Action Plan 2013 to 2015. London. SW1A 2AS. 2013. 58 p.
5. *Pitman N., Shipman A.* A manager's guide to the long-term preservation of electronic documents. London. BIP 0089 BSI. 2008. 110 p.

6. Universal Electronic Records Management (ERM) Requirements. U.S. National Archives and Records Administration. 2017. URL: https://www.archives.gov/records-mgmt/policy/universalermrequirements (2022/01/24).
7. *Соловьев А.В.* Решение проблемы интерпретации цифровых данных долговременного хранения / А.В. Соловьев // Труды ИСА РАН. 2021. Том 71. Вып. 2. С. 43-49.
8. *Афанасьева Л.П.* Автоматизированные архивные технологии / Л.П. Афанасьева // Федеральное агентство по образованию. Государственное образовательное учреждение высшего профессионального образования Российский государственный гуманитарный университет. М. 2005. С. 114.
9. *Miller J.* NARA to suspend development of ERA starting in 2012 [Electronic resource] / J. Miller – Access mode: FederalNewsRadio.com http://www.federalnewsradio.com/?sid=2204570&nid=35 (2022/01/25).
10. US National Archives Blog [Electronic resource]: http://blogs.archives.gov/records-express/2013/11/01/opportunity-for-comment-transfer-guidance-bulletin/ (2022/01/26).
11. National Archives Announces a New Model for the Preservation and Accessibility of Presidential Records. U.S. National Archives and Records Administration [Electronic resource]. 2017. Access mode: https://www.archives.gov/press/press-releases/2017/nr17-54 (2022/01/27).
12. Draft National Archives Strategic Plan. U.S. National Archives and Records Administration [Electronic resource] – 2017 – Access mode: https://www.archives.gov/about/plans-reports/strategic-plan/draft-strategic-plan (2022/01/28).
13. *Рысков О.И.* Управление документацией в Австралии / О.И. Рысков // Отечественные архивы. 2005. № 2. С. 82.
14. *Митченко О.Ю.* Применение международного стандарта ИСО 15489-2001 при создании системы управления документацией в организации / О.Ю.Митченко // Секретарское дело. 2004. № 6.
15. Блог голландской Национальной коалиции по обеспечению сохранности электронных материалов [Электронный ресурс] – Режим доступа: http://www.ncdd.nl/blog/?p=2860 (2022/01/29).
16. Typical requirements for automated electronic document management systems. Specification MoReq // Office for Official Publications of the European Communities as INSAR Supplement VI.

Соловьев Александр Владимирович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Главный научный сотрудник, доктор технических наук. Количество печатных работ: 130. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, математическое моделирование, долговременное хранение электронных документов. E-mail: soloviev@isa.ru

The problem of interpretation of electronic documents for long-term storage

A.V. Solovyev

Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

Abstract. The article considers the problem of interpretability of electronic documents during long-term storage. The urgency of the problem is shown, a review of approaches to solving the problem is made. The statement of the problem of ensuring the interpretability of electronic documents in a general form is carried out. The main problems associated with the use of various file formats, the problems of developing a long-term storage format are identified. The ways of further research are determined, namely: definition of the concept of an electronic document for long-term storage; development of a mathematical model of the composition of information in an electronic document; development of a mathematical model for assessing alienability; development of a mathematical model for assessing interpretability; developing methods for long-term storage to ensure interpretability.
Keywords: *long-term storage, interpretability, electronic document, metadata, long-term storage data format*
DOI: 10.14357/20790279220208

References

1. *Solovyev A.V.* 2020. Long-Term Digital Documents Storage Technology. Lecture Notes in Electrical Engineering. 641: 901-911.
2. *Solovyev A.V.* 2020. Authentication control algorithm for long-term keeping of digital data. IOP Conference Series: Materials Science and Engineering (MSE). 862(5): 052080.
3. *Solovyev A.V.* 2020. Digital media inventory algorithm for long-term digital keeping problem. IOP Conference Series: Materials Science and Engineering (MSE). 919(5): 052003.
4. Open Government Partnership UK National Action Plan. 2013. London. SW1A 2AS. 58 p.
5. *Pitman N., and Shipman A.* 2008. A manager's guide to the long-term preservation of electronic documents. London. BIP 0089 BSI. 110 p.
6. Universal Electronic Records Management (ERM) Requirements. 2017. U.S. National Archives and Records Administration. 2017. Available at: https://www.archives.gov/records-mgmt/policy/universalermrequirements (accessed January 24, 2022).
7. *Solovyev A.V.* 2021. Resheniye problemy interpretatsii tsifrovykh dannykh dolgovremennogo khraneniya [Solving the problem of interpreting digital data for long-term keeping]. Trudy ISA RAN [Proceedings of the ISA RAS]. 71(2): 43-49.
8. *Afanasyeva L.P.* 2005. Avtomatizirovannyye arkhivnyye tekhnologii [Automated archival technologies]. Federal'noye agentstvo po obrazovaniyu.GosudarstvennoyeObrazovatel'noye uchrezhdeniye vysshego professional'nogo obrazovaniya Rossiyskiy Gosudarstvennyy Gumanitarnyy universitet [Federal Agency for Education. State Educational Institution of Higher Professional Education Russian State University for the Humanities]. p. 114.
9. *Miller J.* 2012. NARA to suspend development of ERA starting in 2012. FederalNewsRadio.com. Available at: http://www.federalnewsradio.com/?sid=2204570&nid=35 (accessed January 25, 2022).
10. US National Archives Blog. 2013. Available at: http://blogs.archives.gov/records-express/2013/11/01/opportunity-for-comment-transfer-guidance-bulletin/ (accessed January 26, 2022).
11. National Archives Announces a New Model for the Preservation and Accessibility of Presidential Records. 2017. U.S. National Archives and Records Administration. Available at: https://www.archives.gov/press/press-releases/2017/nr17-54 (accessed January 27, 2022).
12. Draft National Archives Strategic Plan. 2017. U.S. National Archives and Records Administration. Available at: https://www.archives.gov/about/plans-reports/strategic-plan/draft-strategic-plan (accessed January 28, 2022).
13. *Ryskov O.I.* 2005. Upravleniye dokumentatsiyey v Avstralii [Records management in Australia]. Otechestvennyye arkhivy [Domestic archives]. 2: 82.
14. *Mitchenko O.Yu.* 2004. Primeneniye mezhdunarodnogo standarta ISO 15489-2001 pri sozdanii sistemy upravleniya dokumentatsiyey v organizatsii [Application of the international standard ISO 15489-2001 when creating a document management system in an organization]. Sekretarskoye delo [Secretarial business]. 6.
15. Blog of the Dutch National Electronic Preservation Coalition. Available at: http://www.ncdd.nl/blog/?p=2860 (accessed January 29, 2022).
16. Typical requirements for automated electronic document management systems. Specification MoReq. Office for Official Publications of the European Communities as INSAR Supplement VI.

Solovyev A.V. Chief Researcher, Doctor of Technical Sciences. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia. E-mail: soloviev@isa.ru