Проблемы долговременной сохранности больших данных*

А. В. Соловьев[I], Н. Б. Баканова[II]

[I]Федеральное государственное учреждение Федеральный исследовательский центр "Информатика и управление" Российской академии наук, г. Москва, Россия
[II]Федеральный Исследовательский Центр Институт Прикладной Математики им. М.В. Келдыша Российской Академии Наук, г. Москва, Россия

Аннотация. В данной статье описаны и систематизированы проблемы долговременной сохранности больших данных. Показано, что для обеспечения сохранности большого объема цифровых данных возможно использование технологий распределенных реестров. Сделана формальная постановка задачи обеспечения долговременной сохранности больших данных. Кратко описаны возможные области применения решения задачи сохранности.

Ключевые слова: цифровая экономика, долговременная сохранность, большие данные, распределенные реестры, надежность

DOI 10.14357/20718632190205

Введение

Еще до принятия в РФ программы цифровой экономики [1] стало понятно, что в ближайшее время стремительный рост цифровых данных породит серьезную проблемы с обеспечением их сохранности. Стремительная «цифровизация» экономики лишь обострила эту проблему, потому что согласно [1], «данные в цифровой форме» в одночасье стали «ключевым фактором производства во всех сферах социально-экономической деятельности».

Проблема же заключается в том, что «данных в цифровой форме» становится все больше и больше («цифровой шторм»), растет среди них и количество ценных данных, хранение которых должно продолжаться десятилетия. Казалось бы, здесь нет никакой проблемы, на рынке существует огромное количество систем хранения данных (СХД), многие из них снабжены системой многократного дублирования информации, защитой от изменения, например, создание WORM-томов (от сокращения Write Once Read Many). При этом такие данные не могут менять даже администраторы СХД. Пользователи все больше и больше доверяют облачным хранилищам и современным технологиям хранения цифровых данных.

Однако все не так просто, как кажется на первый взгляд. Дело в том, что недостаточно только надежно сохранить цифровые данные. Нужно иметь уверенность, что данные через десятилетия все также аутентичны, и, главное, что их можно интерпретировать, т.е. прочитать с помощью средств компьютерной техники. И это не так просто, ведь даже самая надежная

* Работа выполнена при частичной финансовой поддержке РФФИ в рамках научных проектов № 18-29-03085 и № 18-29-03070.

СХД (или, по крайней мере, ее цифровые носители информации), вряд ли сможет проработать без отказа в течение 50 лет (типовой срок хранения документов по личному составу, а это не самый большой срок хранения данных [2]). А это значит, что данные придется переносить на другие носители, в другие СХД и системы, а это выполняется при участии человека. В этом случае гарантировать их целостность, аутентичность и интерпретируемость крайне сложно.

Проконтролировать аутентичность цифровых данных традиционными методами, например, с помощью электронной подписи (ЭП), также не получится. Причины этого следующие: ЭП имеют очень ограниченный срок действия (не более 5 лет, а в реальности не более 1 года), цифровых данных становится очень много и их контроль с помощью ЭП невозможен из-за крайне невысокой скорости проверки.

Такая противоречивая ситуация порождает необходимость решения важной научно-технической проблемы: обеспечить сохранность большого объема цифровых данных на протяжении длительного временного срока (десятилетия).

1. Основные понятия и определения

Сохранность – как свойство большого объема цифровых данных существовать в качестве доступного и аутентичного свидетельства (доказательства) без потери семантики данных в произвольный момент времени.

Долговременная сохранность – сохранение свойства сохранности в течение 10 лет и более.

Цифровая экономика – экономическая деятельность, основой которой являются цифровые технологии и данные в цифровой форме. В узком смысле это производство электронным бизнесом и электронной коммерцией электронных товаров и услуг. В более широком смысле цифровая экономика затрагивает практически все области обычной экономики, например, банковское дело, образование, здравоохранение. В еще более широком смысле – это некоторая сверхбольшая информационная система, представляющая собой Электронное государство и/или Электронное правительство, включающее управление экономической деятельностью с помощью цифровой платформы.

Блокчейн (*blockchain*) — «выстроенная по определённым правилам непрерывная последовательная цепочка блоков (связный список), содержащих информацию. Чаще всего, копии цепочек блоков хранятся и независимо друг от друга (чрезвычайно параллельно) обрабатываются на множестве разных компьютеров» [20].

Большие данные (*big data*) — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия.

2. Краткий обзор проблемы

Нельзя сказать, что комплексность и глубина проблемы долговременной сохранности не осознана. Попытки решения проблемы активно предпринимаются. Рассмотрим подходы к долговременному хранению больших объемов разнообразных цифровых данных для подтверждения актуальности поставленной проблемы, а также определения факторов, препятствующих ее решению. Учитывая, что большие данные – это не только большой объем и многообразие, но и постоянное изменение, необходимо признать, что речь идет об организации долговременного хранения именно больших данных.

В РФ сравнительно недавно начали заниматься вопросом долговременной сохранности цифровых данных как таковых. В настоящее время Росархив совместно с Всероссийским научно-исследовательским институтом документоведения и архивного дела (ВНИИДАД) занимается разработкой документов, регламентирующих организацию хранения электронных документов в рекомендательном порядке [3, 4]. В частности данными рекомендациями предписывается нормализация электронных документов при длительном хранении в формат PDF/A-1. Надо сказать, что в настоящее время это пока единственный формат электронных документов, производитель которого (компания Adobe, США) гарантирует поддержку формата в течение 50 лет. Однако, по данным паспортизации архивов в РФ [5] количество электронных документов в архивах РФ растет, причем в первую очередь растет количество аудиовизуальных документов.

Безусловно, наибольший интерес с точки зрения организации долговременного хранения

большого объема цифровых данных представляет опыт **National Archives and Records Administration** (NARA). В настоящее время NARA является по факту крупнейшим в мире хранилищем данных в цифровой форме с наиболее длинной историей развития и существования [6]. Если говорить об объемах данных, то можно привести следующую иллюстрацию: Пентагон и Госдепартамент США ежегодно передают в NARA десятки миллионы электронных документов для долговременного хранения [7]. Т.е. действительно речь идет о больших данных. К заслугам NARA безусловно относится создание стандартов хранения цифровых данных [8]:

- библиографический формат MARC,
- формат хранения на оптических дисках ODISS,
- формат архивных описаний GSA6710A,
- формат хранения на магнитных лентах в кодировках ASCII и EBCDIC,
- открытый стандарт кодированного архивного описания (EAD – Encoded Archival Description).

Кроме того, разработаны бюллетень по допустимым файловым форматам, передаваемых на архивное хранение, [9] и требования для написания технических заданий для инструментов или сервисов управления цифровыми данными [10].

С 2005 года NARA разрабатывает ERA (Electronic Records Archives) для долговременного хранения цифровых документов. Однако в связи с возникшими трудностями, в 2012 году проект ERA был приостановлен [11]. Предполагается, что полнотекстовый поиск цифровых данных возможен не ранее 2021 года [12]. Причиной тому стало огромное разнообразие форматов, использование которых порождает риск не интерпретировать спустя десятилетия. Т.е. проблема интерпретируемости не рассматривалась, что и привело к значительным трудностям при создании ERA. Кроме того, выявились проблемы технологического и технического старения носителей информации, а также отсутствие понимания, что такое цифровые данные, как и в каком виде они должны сохраняться [13]. Тем не менее, NARA продолжает накапливать цифровые данные. Так в 2017 году NARA сообщала о создании новой модели хранения и доступности президентских цифровых документов, которых получено только за 2017 год более 250 ТВ [14]. Для частичного решения проблемы интерпретируемости, начиная с 2023 года [15] NARA принимает на хранение данные только в цифровом формате с описательными метаданными.

Проблемой долговременного хранения цифровых данных активно занимается также Национальный архив Великобритании. В 2017 году Национальным архивом была введена стратегия, согласно которой должно быть создано решение, позволяющее обеспечивать долговременную сохранность деловых цифровых документов, передаваемых из правительственных учреждений [16]. Следует, однако, заметить, что это не первая попытка создать подобное решение в Великобритании. Так в 2013 году была принята аналогичная стратегия [17], которая предполагала создание такого решения к 2017 году. Однако создано оно так и не было. Тем не менее, не следует сбрасывать со счетов огромный опят, накопленный в Великобритании и разработанные нормативные документы. Например [18], регламентирует порядок управления доступом, аудитом, миграцией, резервным копированием цифровых документов. Хотя модель цифрового документа в [18] не представлена, приблизительно определен состав информации, которая должна храниться.

В Германии Федеральное агентство по безопасности в информационных технологиях (Bundesamt für Sicherheit in der Informationstechnik) выпустил документ: Техническое руководство BSI TR-03125 [19]. В документе приводятся рекомендации по долговременному хранению цифровых документов, заверенных ЭП. Однако необходимо заметить, что защита данных с помощью ЭП содержит проблемы, связанные с меняющимися технологиями криптозащиты, и небольшими сроками действия ЭП. Поэтому и здесь еще комплексное решение проблемы не представлено. Решением проблемы долговременной сохранности занимаются и другие страны (Австралия, Дания, Франция (начала решение проблемы, но потом отказалась от развития долговременного хранения цифровых данных) и др.), однако формат ста-

тьи не позволяет сделать более подробный обзор. Тем более, что наиболее крупные страны по количеству цифровых данных (США, Германия, Англия) в обзоре представлены.

Из приведенного краткого обзора можно сделать следующие выводы: проблема долговременного хранения является крайне актуальной, все крупные архивы отмечают, что цифровые документы становятся проблемой, рост данных порождает риск потерять управление ими, проблема еще далека от разрешения, т.е. отсутствует универсальное тиражируемое программно-аппаратное решение обеспечения сохранности цифровых данных. Отсутствие универсального решения связано с несколькими причинами:

- отсутствует четкая системная классификация проблем долговременного хранения;
- проблема не является чисто технической, а носит междисциплинарных характер;
- не все проблемы сохранности цифровых данных изучены и систематизированы;
- отсутствует стратегия обеспечения долговременной сохранности больших данных;
- отсутствует методология контроля параметров сохранности цифровых данных.

3. Технологии распределенных реестров

Почему речь в данной статье идет о больших данных? Потому что в рамках цифровой экономики информационные системы будут охватывать целые отрасли или, по крайней мере, группы предприятий, объединенных выпуском конечной сложной и многосоставной продукции. А это означает, что данные в такой информационной системе будут накапливаться в огромном объему и разнообразии, что как раз соответствует определению больших данных.

Действительно, представим ситуацию, при которой для получения изделия $Prod_0$ в рамках сложного технологического процесса (совокупность технологических процессов $TP = \{ TP_j \}$, $j \in [1,M]$, где M – общее количество технологических процессов) объединены несколько информационных систем ($IS = \{ IS_k \}$, $k \in [1,K]$, где K – общее количество ИС) нескольких предприятий. Путь на каждом предприятии выполняется часть сложного технологического процесса TP_j для функционирования которого некоторая информационная система (ИС) IS_k накапливает данные в цифровой форме $didf_k$ (например, о произведенной продукции, логистике, составлении и выполнении планов производства, выпуске продукции, контроле качества, параметрах технологического процесса, материалах, отзывах о внедрении продукции и др.).

Тогда со временем в IS_k накапливается огромный объем данных, характеризующий всю историю производства в рамках процесса. Эти данные обладают несомненной ценностью, т.к. на их основе строятся системы поддержки принятия решений (СППР), позволяющие усовершенствовать производство продукции, улучшать экономические показатели и показатели качества. Потеря этой информации скажется на всей совокупности технологических процессов, а хранение информации для СППР должно измеряться годами, а то и десятилетиями.

Другой пример из области архивного хранения. Поскольку количество значимых цифровых документов только увеличивается, причем такие документы приобретают юридическую значимость (например, согласно №379-ФЗ), либо же представляют собой ценность, например деловая переписка по электронной почте органов власти, дипломатов и др. [7, 14], то их гарантированную сохранность можно осуществить, применив многократное резервирование. Например, с помощью создания резервных территориально-распределенных архивов и центров обработки данных (ЦОД).

Однако в рамках сложного технологического процесса (совокупность технологических процессов TP) или в распределенном архиве данные распределены по разным ИС, по узлам сети, объединяющей ИС, которые географически разделены друг с другом. Тогда можно утверждать, что цифровые данные ИС хранятся по правилам распределенных реестров. Следовательно, можно предположить, что технологии распределенных реестров (технологии блокчейн) можно использовать как одно из средств обеспечения сохранности (например, [20]). Различные ИС в приведенных примерах могут находиться в юрисдикции разных юридических лиц, обслуживаться разными IT-

структурами, тем самым, можно говорить о необходимости поддержания аутентичности хранимых цифровых данных в отсутствии единого доверенного центра. А это еще одно из условий, говорящее в пользу возможности использования технологий распределенных реестров. Конечно, данные можно защитить с помощью технологий криптозащиты, например ЭП. Однако по причине устаревания этих средств и небольших сроках их действия говорить об обеспечении с помощью них долговременной сохранности очень сложно. Также существует проблема признания доверенными сертификатов ЭП разными удостоверяющими центрами (УЦ). Кроме того, заверять каждое данное в цифровой форме с помощью ЭП – очень трудоемкий процесс, требующий довольно больших ресурсов ИС. Хранение в ИС инфраструктуры открытых ключей, сертификатов и список отзыва сертификатов тоже требует дополнительных ресурсов.

4. Обзор проблем долговременной сохранности больших данных

Прежде чем перейти к формальной постановке задачи сохранности больших данных приведем обзор проблем, возникающих при хранении больших данных.

В первую очередь следует заметить, что информационные технологии постоянно изменяются и совершенствуются. Это касается как общего программного обеспечения (операционные системы, базы данных, офисные пакеты и форматы), так и аппаратного обеспечения. Кроме того, постоянно совершенствуются и меняются средства криптографической защиты информации (СКЗИ). При этом срок технологического старения как программных, так и технических средств вычислительной техники обычно не превышает 10 лет, в то время как при долговременном хранении данные могут храниться 50 лет и более.

Тогда первой проблемой, с которой можно столкнуться при обеспечении долговременной сохранности данных – это зависимость от конкретных технических и программных средств хранения данных. Т.е. когда данные невозможно перенести на другие технические средства. Либо же когда данные хранятся в БД, управляемой некоторой СУБД в таком формате, который не поддерживается другими СУБД. Первую проблему можно сформулировать как проблему **независимости** данных.

Второй, связанной с первой, проблемой является зависимость от конкретного формата данных, который со временем перестает поддерживаться, а также от средств отображения данных. Такая проблема может привести к тому, что данные превратятся в бессмысленную бинарную последовательность «нулей» и «единиц», расшифровка которой станет невозможной. Вторую проблему можно сформулировать как проблему **интерпретируемости** (читаемости) данных.

Третью проблему можно охарактеризовать, как проблему сохранения **аутентичности** (неизменности) данных в процессе хранения. Такая проблема может возникнуть как по причине намеренного или случайного искажения или удаления данных вследствие несанкционированного доступа (НСД) либо же при переносе данных из одного хранилища в другое вследствие смены программно-технической платформы ИС.

Четвертая проблема связана с невозможностью существования цифровых данных вне цифровых программно-технических средств. Тогда проблему можно охарактеризовать как проблему **надежности** хранения цифровых данных, и, шире, как проблему надежности конкретных программно-технических средств хранения цифровых данных.

При хранении больших данных, когда ценность представляет собой даже не конкретное атомарное цифровое данное, а совокупность цифровых, накопленная годами может возникнуть проблема потеря **семантики** данных. Так можно определить пятую проблему. Это означает, что сами атомарные данные привязаны к некоторым метаданным, которые и определяют их смысл. Например, хранятся данные в виде «27000». Что это такое? Расстояние? Время? Деньги? Когда же данные имеют вид «Расстояние 27000 метров», то не возникает сомнений что это такое. В этом случае ценность представляют также метаданные, которые и определяют семантику хранимых данных. Потеря метаданных критическим образом скажется на возможности их использовать, даже надежно сохранив, в течение десятилетий.

Шестой проблемой можно считать проблему **устойчивости** программно-технических средств хранения к внешним воздействиям, в том числе и катастрофического характера. Программно-технические средства могут работать совершенно надежно, но могут быть выведены из строя воздействиями катастрофического характера, либо же вследствие нарушения информационной безопасности. Проблема устойчивости выделена отдельно от надежности, т.к. оценивать ее с помощью математического аппарата теории надежности не имеет смысла (вероятность будет очень маленькой). Здесь, скорее всего, должен использоваться сценарный подход: описание сценариев вероятных воздействий (например, магнитной супербури) и мер противодействия.

При использовании для обеспечения сохранности технологий распределенных реестров, по отношению к ним также можно рассматривать воздействие всех перечисленных проблем. Технологии распределенных реестров с одной стороны должны способствовать сохранению аутентичности защищаемых с помощью них данных в цифровой форме. С другой стороны как цифровые технологии, реализованные с помощью конкретных программно-технических средств [21-23], они также подвержены тем же проблемам долговременного хранения, ведь вся их информация должна сохраняться то же время, что и защищаемые с помощью них данные.

5. Постановка задачи обеспечения долговременной сохранности больших данных

Как было показано выше, срок технологического старения как программных, так и технических средств вычислительной техники обычно не превышает 10 лет, в то время как при долговременном хранении цифровые данные могут храниться 50 лет и более.

В таком случае можно представить цифровые данные как объект управления, а технические и программные средства обеспечения сохранности как нестабильную среду хранения цифровых данных.

Утверждение 1. Большие цифровые данные – это объект управления, а задача долговременной сохранности цифровых данных – это задача оптимального управления цифровыми данными в условиях параметрических возмущений цифровой программно-технической среды хранения цифровых данных [24, 25]. А для оптимального управления необходимо научится контролировать параметры среды хранения и разработать алгоритмы компенсации возмущений для стабилизации объекта управления [26].

Перейдем к математической постановке задачи. Дано:

1) Множество данных в цифровой форме $\boldsymbol{Didf} = \{ \boldsymbol{didf_i} \}$

2) Множество блоков распределенных реестров $\boldsymbol{Bch} = \{ \boldsymbol{bch_j} \}$

3) Множество параметрических возмущений среды хранения $\boldsymbol{E} = \{ \boldsymbol{\varepsilon_q} \}$:

- $\boldsymbol{\varepsilon_1}$ – нарушение независимости ***Didf***, ***Bch***
- $\boldsymbol{\varepsilon_2}$ – нарушение интерпретируемости ***Didf***, ***Bch***
- $\boldsymbol{\varepsilon_3}$ – нарушение аутентичности ***Didf***, ***Bch***
- $\boldsymbol{\varepsilon_4}$ – нарушение надежности хранения ***Didf***, ***Bch***
- $\boldsymbol{\varepsilon_5}$ – нарушение семантики ***Didf***, ***Bch***
- $\boldsymbol{\varepsilon_6}$ – нарушение устойчивости ***Didf***, ***Bch***

4) Множество информационных систем $\boldsymbol{IS} = \{ \boldsymbol{IS_k} \}$ подверженное $\boldsymbol{E}$ и осуществляющее обработку, передачу и хранение ***Didf***, ***Bch***

5) Множество требований безопасности, аутентичности, надежности, устойчивости данных $\boldsymbol{T} = \{ \boldsymbol{NTD_l} \}$

6) Исходный уровень сохранности для ***Didf***, ***Bch*** – $\boldsymbol{SV_0}$.

Найти:

1) Множество математических моделей $\boldsymbol{\mu} = \{ \boldsymbol{\mu_r} \}$, контроля параметров сохранности (включая независимость, интерпретируемость, аутентичность, надежность, сохранение семантики, устойчивость) при хранении, информационном обмене ***Didf***, ***Bch*** между ***IS***.

2) Множество алгоритмов $\boldsymbol{A} = \{ \boldsymbol{A_r} \}$, обеспечивающих сохранность ***Didf***, ***Bch*** на исходном уровне $\boldsymbol{SV_0}$: $\boldsymbol{A_r}(\boldsymbol{didf_i}, \boldsymbol{bch_j}) = \boldsymbol{SV_0}$.

3) Множество программно-технических решений $\boldsymbol{R} = \{ \boldsymbol{R_r} \}$, обеспечивающих сохранность ***Didf***, ***Bch*** на исходном уровне $\boldsymbol{SV_0}$: $\boldsymbol{R_r}(\boldsymbol{didf_i}, \boldsymbol{bch_j}) = \boldsymbol{SV_0}$.

Т.е. результатом решения данной задачи является технология, как совокупность математических моделей, алгоритмов и программно-

технических решений для обеспечения долговременной сохранности данных в цифровой форме.

Утверждение 2. Задача контроля сохранности данных в цифровой форме представляет собой задачу оптимального выбора по многим критериям.

Действительно, пусть сохранность некоторого $didf_i$ характеризуется функцией $\mu(t)$, значение которой представляет собой вероятность сохранности $didf_i$ на уровне SV_0 в произвольный момент времени t. Тогда можно утверждать, что задача обеспечения сохранности $didf_i$ формулируется, как задача достижения максимума функции $\mu(t)$ на произвольном временном интервале t.

Т.е.: $M = \max_{t \in [t0, \infty]} \mu(t)$, где $t0$ – момент времени создания $didf_i$ в некоторой IS_k.

Тогда $\mu(t) = \min(\varsigma_d(t)^{\omega 1}\ \zeta_d(t)^{\omega 2}\ \alpha_d(t)^{\omega 3}\ \rho_d(t)^{\omega 4}\ \sigma_d(t)^{\omega 5}\ \varphi_d(t)^{\omega 6}\ ,\ \varsigma_b(t)^{\omega 1}\ \zeta_b(t)^{\omega 2}\ \alpha_b(t)^{\omega 3}\ \rho_b(t)^{\omega 4}\ \sigma_b(t)^{\omega 5}\ \varphi_b(t)^{\omega 6})$,

где ω_i – коэффициенты важности показателей, назначаемые экспертами $\sum \omega_i = 1$, $\omega_i > 0$, i=[1,6].

$\varsigma_d(t)$, $\varsigma_b(t)$ – соответственно вероятности сохранения независимости для ***Didf*** и ***Bch*** на произвольном временном интервале t.;

$\zeta_d(t)$, $\zeta_b(t)$ – соответственно вероятности интерпретируемости для ***Didf*** и ***Bch*** на произвольном временном интервале t.;

$\alpha_d(t)$, $\alpha_b(t)$ – соответственно вероятности сохранения аутентичности для ***Didf*** и ***Bch*** на произвольном временном интервале t.;

$\rho_d(t)$, $\rho_b(t)$ – соответственно надежность хранения для ***Didf*** и ***Bch*** на произвольном временном интервале t.;

$\sigma_d(t)$, $\sigma_b(t)$ – соответственно вероятности сохранения семантики для ***Didf*** и ***Bch*** на произвольном временном интервале t.;

$\varphi_d(t)$, $\varphi_b(t)$ – соответственно вероятности сохранения устойчивости для ***Didf*** и ***Bch*** на произвольном временном интервале t.

6. Возможное практическое применение

Как было сказано выше, результатом решения задачи обеспечения сохранности больших данных должна стать технология, реализация которой позволит решить задачу сохранности большого объема данных в цифровой форме.

Возможные практические применения:

- Хранение медицинских данных (например, медицинских карт, как правило, это время жизни пациента, плюс 2-5 лет).
- Хранение цифровых данных по личному составу (50-75 лет).
- Хранение ретроспективных данных о технологических процессах, истории продаж, финансовых показателей для СППР предприятия (годы и/или десятилетия).
- Хранение данных о заработке в БД государственных и негосударственных Пенсионных фондов (до 75 лет).

В условиях развивающейся цифровой экономики могут быть и другие применения: хранение юридически значимых документов, хранение истории взаимодействия граждан с органами власти, наблюдательные и страховые дела и др.

Заключение

В данной статье проведена систематизация проблем долговременного хранения больших объемов цифровых данных в условиях цифровой экономики. Для обеспечения сохранности больших данных возможно использование технологий распределенных реестров, что с одной стороны повышает их сохранность, с другой усложняет структуру данных и информационных систем. Сделана постановка задачи обеспечения долговременной сохранности цифровых данных как задача оптимального управления объектом хранения в условиях параметрических возмущений программно-технической среды хранения. В приведенном в начале статьи обзоре показана актуальность решения данной задачи. Результатом решения задачи должна стать технология, как совокупность математических моделей, алгоритмов и программно-технических решений для обеспечения долговременной сохранности данных в цифровой форме. Обозначен круг возможного практического применения решения задачи обеспечения долговременной сохранности. В дальнейшем планируется создать математические модели контроля параметров, состава информации большого объема цифровых данных, разработать алгоритмы обеспечения сохранности.

Литература

1. Программа «Цифровая экономика Российской Федерации». Утверждена распоряжением Правительства Российской Федерации от 28 июля 2017 г. № 1632-р. М.:2017 – 88 с.
2. Изменился срок хранения документов по личному составу [Электронный ресурс] – Режим доступа: https://its.1c.ru/news/131651 - 2019.03.29.
3. Рекомендации по комплектованию, учету и организации хранения электронных архивных документов в архивах организаций. - М.: ВНИИДАД, 2013 – 58 с.
4. Рекомендации по комплектованию, учету и организации хранения электронных архивных документов в государственных и муниципальных архивах (проект) – М.: ВНИИДАД, 2013 – 33 с.
5. Справка об итогах паспортизации федеральных архивов по состоянию на 01.01.2016 [Электронный ресурс] – Режим доступа: http://archives.ru/reporting/spravka-passportization-federal-archives-2016.shtml - 2019.03.29.
6. Рысков, О.И. Об основных направлениях деятельности зарубежных архивных органов в области исследования и нормативного регулирования работы с электронной документацией / О.И. Рысков // Секретарское дело. – 2005. - № 3. – С.76.
7. 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.
8. Рысков, О.И. Основные направления деятельности национальных архивов США и Соединенного Королевства Великобритании и Северной Ирландии в области управления электронными документами правительственных учреждений / О.И. Рысков // Отечественные архивы. – 2004. - № 3.
9. Блог Национальных Архивов США: http://blogs.archives.gov/records-express/2013/11/01/ opportunity-for-comment-transfer-guidance-bulletin/
10. Universal Electronic Records Management (ERM) Requirements. U.S. National Archives and Records Administration, 2017: https://www.archives.gov/records-mgmt/policy/universalermrequirements
11. Miller, J. NARA to suspend development of ERA starting in 2012 [Электронный ресурс] / J. Miller – Режим доступа: FederalNewsRadio.com http://www.federalnewsradio.com/?sid=2204570&nid=35
12. Lipowicz, A. NARA officials defend searchability of electronic archive [Электронный ресурс] / A. Lipowicz // Federal Computer Week.– Режим доступа: http://fcw.com/articles/2011/11/01/nara-officials-defending-searchability-of-electronic-archive.aspx
13. Carlstrom, G. Is DoD's new pay system fair? [Электронный ресурс] G. Carlstrom // FederalTimes.com – Режим доступа: http://federaltimes.com/index.php?S=3502888
14. National Archives Announces a New Model for the Preservation and Accessibility of Presidential Records. U.S. National Archives and Records Administration [Электронный ресурс] – 2017 – Режим доступа: https://www.archives.gov/press/press-releases/2017/nr17-54.
15. Draft National Archives Strategic Plan. U.S. National Archives and Records Administration [Электронный ресурс] – 2017 – Режим доступа: https://www.archives.gov/about/plans-reports/strategic-plan/draft-strategic-plan.
16. Суровцева, Н.Г. Хранение электронных документов: зарубежный опыт // Вестник культуры и искусства. – 2017. – №4(52). – С.17.-23.
17. Open Government Partnership UK National Action Plan 2013 to 2015. London SW1A 2AS – 2013 – 58 p.
18. Типовые требования к автоматизированным системам электронного документооборота. Спецификация MoReq // Office for Official Publications of the European Communities as INSAR Supplement VI, ISBN 92-894-1290-9.
19. Preservation of Evidence of Cryptographically Signed Documents // BSI Technical Guideline TR-03125 – Version 1.2 – Federal Office for Information Security – 2015 – 183 p.
20. Даниленко А.Ю., Пашкина Е.В., Пашкин М.А., Соловьев А.В. Применение технологии блокчейн в информационных системах. Часть 2. Подтверждение авторства и обеспечение целостности. // Системы высокой доступности. 2018. Т. 14. № 1. С. 9–11.
21. Melanie Swan. Blockchain: Blueprint for a New Economy. — O'Reilly Media, Inc., 2015. — 152 p. — ISBN 978-1-4919-2047-3. В русском переводе *Мелани Свон*. Блокчейн: Схема новой экономики. — Олимп-Бизнес, 2016. — 240 с. ISBN 978-5-9693-0360-7
22. Anderson, L., Holz, R., Ponomarev, A., Rimba, P., & Weber, I. (2016). New kids on the block: an analysis of modern blockchains (2016).
23. Даниленко А.Ю., Акимова Г.П. Особенности применения технологии блокчейн. // Материалы 27-й научно-технической конференции Методы и технические средства обеспечения безопасности информации 24-27 сентября 2018 года. СПб: Издательство политехнического университета. 2018. С. 73–75. ISSN 2305-994X.
24. Емельянов С.В. Системы автоматического управления с переменной структурой. – М.: Наука, 1967 г. – 336 с.
25. Емельянов С.В., Костылева Н. Е., Матич Б. Л., Миловидов Н. Н. Системное проектирование средств автоматизации. — М.: Машиностроение, 1978.
26. Емельянов С.В. Новые типы обратной связи. – М.: Наука, Физматлит, 1997. – 352 с.

Соловьев Александр Владимирович. Главный научный сотрудник ИСА ФИЦ ИУ РАН. Доктор технических наук. Количество печатных работ: 88. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, математическое моделирование, долговременное хранение электронных документов. E-mail: soloviev@isa.ru

Баканова Нина Борисовна. Зав. сектором, Институт Прикладной Математики им. М.В. Келдыша Российской Академии Наук». Доктор технических наук. Количество печатных работ: более 40. Область научных интересов: системный анализ, управление и обработка информации, проектирование информационных систем, поддержка принятия решений, проблемно-ориентированные системы, экспертные системы. E-mail: nina@keldysh.ru

Problems of long-term safety of big data

A.V. Solovyev[I], N.B. Bakanova[II]

[I] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia
[II] M.V. Keldysh, Institute of Applied Mathematics of the Russian Academy of Sciences, Moscow, Russia

Abstract. This article describes and systematizes the problems of long-term preservation of big data. It is shown that to ensure the safety of large amounts of digital data, it is possible to use distributed registry technologies. A formal formulation of the problem of ensuring the long-term safety of big data has been made. The possible areas of application of the solution of the preservation problem are briefly described.

Keywords: digital economy, long-term preservation, big data, distributed registries, reliability

DOI 10.14357/20718632190205

References

1. Program "Digital Economy of the Russian Federation". Approved by the order of the Government of the Russian Federation at July 28, 2017. № 1632-r. M.:2017 – 88 p.
2. Izmenilsya srok hraneniya dokumentov po lichnomu sostavu [Changed the storage period of documents on the staff], [Electronic resource] – Access mode: https://its.1c.ru/news/131651 - 2019.03.29.
3. Rekomendacii po komplektovaniyu, uchetu i organizacii hraneniya elektronnih arhivnih dokumentov v arhivah organizacii [Recommendations for the acquisition, accounting and organization of storage of electronic archival documents in the archives of organizations]. - M.: VNIIDAD, 2013 – 58 p.
4. Rekomendacii po komplektovaniyu, uchetu i organizacii hraneniya elektronnih arhivnih dokumentov v gosudarstvennih i municipal'nih arhivah [Recommendations for the acquisition, accounting and organization of storage of electronic archival documents in state and municipal archives]. – M.: VNIIDAD, 2013 – 33 p.
5. Spravka ob itogah pasportizacii federal'nih arhivov po sostoyaniyu na 01.01.2016 [Certificate of certification results of federal archives as of 01/01/2016], [Electronic resource] – Access mode: http://archives.ru/reporting/spravka-passportization-federal-archives-2016.shtml - 2019.03.29.
6. Ryskov, O.I. Ob osnovnih napravleniyah deyatelnosti zarubezhnih arhivnih organov v oblasti issledovaniya I normativnogo regulirovaniya raboti s elektronnoi dokumentaciey [On the main activities of foreign archival bodies in the field of research and regulatory work with electronic documentation] // Sekretarskoye delo [Secretarial business]. – 2005. - № 3. – P.76.
7. Afanasyeva, L.P. Avtomatizirovannie arhivniye tehnologuu [Automated Archive Technologies] // Federal'noye agentstvo po obrazovaniyu. Gosudarstvennoye Obrazovatelnoye uchrezhdeniye vysshego professional'nogo obrazovaniya Rossiyskiy Gosudarstvenniy Gumanitarniy universitet [Federal Agency for Education. State Educational Institution of Higher Professional Education Russian State University for the Humanities] – M.: 2005. – P.114.
8. Ryskov, O.I. Osnovniye napravleniya deyatelnosti nacional'nih arhivov USA I Soedinennogo Korolevstva Velikobritanii I Severnoy Irlandii v oblasti upravleniya elektonnimi dokumentami pravitelstvennih uchrezhdeniy [The main activities of the national archives of the United States and the United Kingdom of Great Britain and Northern Ireland in the field of management of electronic documents of government agencies] // Otechestvenniye Arhivy [Russian archives]. – 2004. - № 3.
9. US National Archives Blog [Electronic resource] – Access mode: http://blogs.archives.gov/records-express/2013/11/01/opportunity-for-comment-transfer-guidance-bulletin/– 2019/03/22.
10. Universal Electronic Records Management (ERM) Requirements. U.S. National Archives and Records Administration [Electronic resource], 2017, Access mode: https://www.archives.gov/records-mgmt/policy/universalermrequirements
11. 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
12. Lipowicz, A. NARA officials defend searchability of electronic archive [Electronic resource] / A. Lipowicz // Federal Computer Week.– Access mode: http://fcw.com/articles/2011/11/01/nara-officials-defending-searchability-of-electronic-archive.aspx
13. Carlstrom, G. Is DoD's new pay system fair? [Electronic resource] G. Carlstrom // FederalTimes.com – Access mode: http://federaltimes.com/index.php?S=3502888
14. 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.
15. 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
16. Suvorovtseva, N.G. Hraneniye elektronnih dolumentov: zarubezhniy opyt [Storage of electronic documents: foreign experience] // Vestnik kultury I iskusstva [Bulletin of culture and art]. – 2017. – №4(52). – P.17.-23.
17. Open Government Partnership UK National Action Plan 2013 to 2015. London SW1A 2AS – 2013 – 58 p.
18. Typical requirements for automated electronic document management systems. Specification MoReq // Office for Official Publications of the European Communities as INSAR Supplement VI, ISBN 92-894-1290-9.
19. Preservation of Evidence of Cryptographically Signed Documents // BSI Technical Guideline TR-03125 – Version 1.2 – Federal Office for Information Security – 2015 – 183 p.
20. Danilenko A.Yu., Pashkina E.V., Pashkin M.A., Solovyev A.V. Primeneniye tehnologii blockchain v informacionnih systemah. Chast 2. Podtverzhdeniye avtorstva I obespecheniye celostnosti [The use of blockchain technology in information systems. Part 2. Confirmation of authorship and integrity.] // Systemy visokoy dostupnosti [High Availability Systems]. 2018. T. 14. № 1. P. 9–11.
21. Melanie Swan. Blockchain: Blueprint for a New Economy. — O'Reilly Media, Inc., 2015. — 152 p. — ISBN 978-1-4919-2047-3.
22. Anderson, L., Holz, R., Ponomarev, A., Rimba, P., & Weber, I. (2016). New kids on the block: an analysis of modern blockchains (2016).
23. Danilenko A.Yu., Akimova G.P. Osobennosti primeneniya tehnologii blockchain [Features of blockchain technology] // Materialy 27-th nauchno-tehnicheskoy konferencii "Metody I tehnicheskiye sredstva obespecheniya bezopasnosti informacii 24-27 sentybrya 2018 [Materials of the 27th Scientific and Technical Conference. Methods and Technical Means for Ensuring the Security of Information September 24-27, 2018]. S-Pb: Izdatel'stvo politehnicheskogo universiteta [Publishing house of the Polytechnic University]. 2018. P. 73–75. ISSN 2305-994X.
24. Emelyanov S.V. Sistemy avtomaticheskogo upravleniya s peremennoy strukturoy [Automatic control systems with variable structure]. – M.: Science, 1967 – 336 p.
25. Emelyanov S.V., Kostileva N.E., Matich B.L., Milovidov N.N. Sistemnoye proektirovaniye sredstv avtomatizacii [System design automation]. – M.: Mashinostroyeniye [Mechanical engineering], 1978.
26. Emelyanov S.V. Noviye tipy obratnoy sv'azi [New types of feedback]. – M.: Science, Fizmatlit, 1997 – 352p.

Solovyev A.V. Chief Researcher, Department 94 ISA FRC CSC RAS. Moscow, prosp. 60-let Oktyabrya, 9. Doctor of Technical Sciences. Number of publications: 88. Area of scientific interests: system analysis, database management systems, reliability theory, mathematical modeling, electronic document management, electronic archive, long-term storage of electronic documents. E-mail: soloviev@isa.ru

Bakanova N.B. Head of sector, M.V. Keldysh, Institute of Applied Mathematics of the Russian Academy of Sciences, 125047, Moscow, Miusskaya Square, Building 4. Doctor of Technical Sciences. Number of publications: 40. Area of scientific interests: system analysis, management and information processing, information systems design, decision-making support, problem-oriented systems, expert systems, E-mail: nina@keldysh.ru