Математические структуры и моделирование
2023. № 4 (68). С. 85–92

УДК 004.021
DOI 10.24147/2222-8772.2023.4.85-92

ИСТОРИЧЕСКАЯ ИММОРТАЛИЗАЦИЯ: ЭФФЕКТИВНЫЕ МЕТОДЫ ДОЛГОСРОЧНОГО ХРАНЕНИЯ ИСТОРИЧЕСКИХ ДАННЫХ

Т.Е. Болдовская[1,3]
канд. техн. наук, доцент, e-mail: boldovskaya73@gmail.com
В.И. Гресь[2,3]
лаборант, e-mail: gresvladimir02@gmail.com
А.Е. Ветров[2,3]
лаборант, e-mail: aleksandrvetrov838@gmail.com

[1]Омский государственный технический университет, Омск, Россия
[2]Омский государственный университет им. Ф.М. Достоевского, Омск, Россия
[3]Томский государственный университет, Томск, Россия

Аннотация. Изучаются вопросы долгосрочной сохранности цифровых исторических данных и обеспечения их безопасности в цифровом мире. Проанализированы современные методы архивации электронных данных и сделан вывод о целесообразности использования микросервисной архитектуры с репликацией хранилищ.

Ключевые слова: архивация, сохранение цифровых исторических данных, контейнеризация, репликация хранилищ.

Введение

С каждым днем наш мир становится всё более цифровым, сохранение исторических данных не является исключением.

Сохранение цифровых исторических данных важно для сохранения памяти и наследия наших предков. Такие данные включают в себя различные материалы: документы, фотографии, записи и другие ценные артефакты, которые помогают нам понять формирование нашего мира и общества.

Однако с ростом объёма цифровых данных и изменением технологий, сохранение этих данных становится сложной задачей. Без правильных методов и стратегий существует риск потери важных свидетельств прошлого.

Помимо этого, сохранение цифровых исторических данных также требует доступности для будущих поколений. Главная задача не только сохранить эти данные, но и создать возможность для последующих поколений ознакомиться с ними. Вследствие этого вопрос о методах сохранения цифровых исторических данных остаётся актуальным в связи с сохранением нашей хронологии и культурного наследия, обеспечением безопасности данных в цифровом мире и связью прошлого с будущим.

1. Основные положения в архивации цифровых исторических данных

1.1. Определение

Рабочая группа, сформированная в 2007 г. Американской библиотечной ассоциацией, определила понятие электронного архивирования как сохранение цифровых данных, сочетающее в себе политику, стратегии и действия, направленные на обеспечение доступа к переформатированному и рождённому цифровому контенту независимо от проблем, связанных с провалами средств массовой информации и технологическими изменениями. Целью цифрового сохранения является точное воспроизведение аутентифицированного контента с течением времени [1].

1.2. Основные принципы долгосрочного хранения данных

В первую очередь архивация цифровых исторических данных нужна для долговременного хранения (от 100 лет) и защиты информации. Также немаловажной особенностью является то, что данные должны быть свободны от необоснованного ограничения и доступны населению, даже в том случае если финансирование проекта было прекращено [2, 3]. Эта концепция позволяет восстановить эти данные даже спустя долгие годы и без большого труда воспользоваться ими в научных целях.

1.3. Виды цифровых объектов

Важно понимать какие объекты придётся сохранять, от этого будет зависеть стратегия их архивации.

Исследователи выявляют несколько видов цифровых объектов, подлежащих сохранению [4]:

- Оцифрованные версии документов;
- Цифровые материалы, для которых отсутствует аналог в печатной форме;
- Индивидуальные объекты, такие как тексты, изображения и аудиозаписи;
- Коллективные ресурсы, такие как веб-сайты, электронные журналы;
- Наборы данных, состоящие из различных научных, компьютерных и других материалов;
- Метаданные, облегчающие долгосрочное хранение файлов и извлечение необходимой информации, включая данные о формате файла, используемом программном обеспечении и истории изменений.

1.4. Метаданные

Метаданным отводят существенное значение в организации коллекций, содержащих аудио-, видео- и рукописные материалы, которые может быть сложно идентифицировать без сопровождающих текстовых описаний.

Международный стандарт «PROMISE» определяет метаданные как информацию, которую пользователь хранилища использует для поддержки процесса сохранения цифровых данных [5].

2. Методы архивации электронных данных

Проблемы с хранением возникают из-за двух факторов; во-первых, электронные носители документов, как правило, менее надёжны, чем бумажные, и со временем быстрее изнашиваются; во-вторых, быстрые темпы технического прогресса приводят к регулярному обновлению систем кодирования информации, типов носителей (таких как жёсткие диски, твердотельные накопители, флэш-накопители), устройств и программного обеспечения. Следовательно, старые технологии устаревают и перестают быть частью повседневного использования. Простого сохранения носителя недостаточно для сохранения электронного документа в течение долгого времени. С течением времени информация становится нечитаемой либо из-за износа носителя, либо из-за устаревания технологии обработки. Следовательно, пассивное сохранение носителя оказывается бесполезным. Оттого становится крайне важным принимать меры путём постоянного мониторинга безопасности электронных документов, проведения анализа рисков и заблаговременного переноса документов в новые форматы носителей с возможным преобразованием. Другая потенциальная проблема возникает, когда речь заходит о поддержании согласованности в процессе пересмотра, обновления или замены информационных форматов по мере развития технологий. Каждый тип информации, такой как тексты, изображения и видео, часто требует правил преобразования. Однако новые программы и устройства, как правило, не поддерживают устаревшие форматы. Для точного считывания, интерпретации и проверки подлинности документа крайне важно сохранить не только его содержание, но и метаданные, описывающие его характеристики и жизненный цикл, включая архивное хранение.

2.1. Консервация

Сохранение цифровых данных предполагает работу с материалами в их форматах и на носителях с использованием оригинальной технологии или ранее использовавшегося аппаратного и программного обеспечения. Такой подход приводит к созданию «компьютерного музея», где материалы представлены и поддерживаются в их форматах с их функциональными возможностями. Хотя консервация считается стратегией архивирования, она может быть хорошим вариантом для определённых цифровых данных, поскольку позволяет сохранить исходные инструменты доступа (программное обеспечение). Однако при реализации этой стратегии могут возникнуть проблемы с обслуживанием и затратами, связанными с обеспечением постоянной доступности определённых типов файлов. Более того, консервация также ограничивает переносимость ресурсов, поскольку зависит от оборудования, хранящегося в определённых местах [6].

2.2. Эмуляция

Эмуляция относится к процессу репликации операционной системы с целью обеспечения совместимости с форматами данных. Включает в себя создание виртуальной машины на компьютере, которая способна имитировать функции аппаратной и программной среды. В некотором смысле эмуляция похожа на стратегии консервации, поскольку она предполагает сохранение целостности прикладной программы. Отличие в том, что основная цель эмуляции – сохранить как внешний вид, так и функциональность объекта путём дублирования его технического содержимого, что позволяет в будущем использовать либо исходный объект, либо его обновлённую версию.

2.3. Инкапсуляция

Хранение описания документа внутри объекта вместе с самим документом, известное как инкапсуляция, влечёт за собой включение описания документа в цифровом формате. Это снижает зависимость от внешних факторов, поскольку позволяет либо воспроизвести среду, либо перенести документ в новую, используя эту информацию.

3. Проектирование системы для иммортализации исторических данных

В рамках проекта «РНФ № 23-78-10119» одной из задач является создание научно обоснованной аналитической модели исторического информационного ресурса «Православный ландшафт таёжной Сибири: акторы, институты, сети», включающей выявление зависимостей между социальными, культурными и природными факторами формирования поселенческой сети методами математической статистики, которая подразумевает создание информационной системы. В ней, в свою очередь, будет храниться множество исторических данных.

Возникает вопрос долговременного хранения этих данных, чтобы даже после завершения выполнения проекта данными мог воспользоваться любой желающий.

Преимуществом является создание системы с полного нуля. Это даёт возможность изначально проектировать его с возможностью к расширению, поддержанию и сохранению данных в будущем.

3.1. Архитектура платформы

Первоначально архитектура системы планировалась таким образом, чтобы в дальнейшем не возникало трудностей внедрения новых сотрудников или передачи её третьим лицам для поддержания или совершенствования.

За основу взят подход микросервисной архитектуры. Такой подход, в отличие от стандартного «RESTFulAPI», имеет много преимуществ. Микросервисы могут быть легко добавлены, удалены или изменены без влияния на остальную систему. Это важное свойство в проекте, потому что разработка будет проходить совместно

с несколькими группами разработчиков из разных городов. Такая гибкость позволит качественнее делегировать обязанности между группами. Каждый микросервис может быть разработан, тестирован и развернут независимо от других. Благодаря этому, как было ранее сказано, можно значительно упростить работу и в начале проекта, и после его завершения, благодаря другим людям. Так как в дальнейшем предполагается интеграция со сторонними сервисами для улучшения качества технологичной среды для исторических исследований, микросервисы позволят не привязываться к определённому языку или технологии, что даст возможность даже спустя долгие годы продолжить разработку системы.

Проект предполагает использование из всех частей Российской Федерации независимо от часового пояса. Это, в свою очередь, требует высокой доступности. Для её обеспечения все микросервисы оборачиваются в специальные кластеры, которые можно множить и оркестрировать по отдельным сервисам в случае их отказа.

3.2. Длительное сохранение данных

Для каждого документа ведётся сбор и сохранение метаданных, чтобы гарантировать, что они могут быть воспроизведены и проиндексированы поисковыми системами в будущем.

На рисунке 1 представлен пример метаданных для работы «Толкование на Апокалипсис», созданных в Национальном исследовательском Томском государственном университете, на базе которого реализуется данный проект.

Заглавие	**Толкование на Апокалипсис**
Автор	Андрей (архиепископ Кесарии Каппадокийской)
Предмет	
Публикация	[Б. м.
Описание	Рукопись: 1910–1920-е гг.
-	Бумага: русская машинная со штемпелем с надписью "Писчая бумага", производилась в 1913 г.
-	Почерк: полуустав, кириллическая цифровая фолиация (писцовая)
-	Орнаментика и украшения: заголовки, инициалы красными чернилами; заставки растительного орнамент 12, 13, 16, 19, 23, 26, 28 , 35, 44, 46, 54, 57, 60, 65, 67, 69, 73, 79, 80, 83, 85, 87, 90, 95, 98); концовки 50 об.) и растительного орнамента (л. 101)
-	Переплет: 1910–1920-е гг., ткань, картон; на верхней крышке переплета – вырезка из конфетной короб "Апокалипсис"
-	Записи: читательские, 4 закладки из бересты с записями
-	Состояние: удовлетворительное
-	История бытования: переплет рукописи украшен в томско-чулымском скиту староверов-странников (посл
-	Содержание: толкование на Апокалипсис Андрея, архиепископа Кесарии Каппадокийской пользуется о представления о существовании 8 царств (последнее из которых – царство Антихриста) и методы экзегез
Отношения	Связи документа
Коллекция(ии)	
Ссылки	
Язык	rus

Рис. 1. Пример метаданных для работы «Толкование на Апокалипсис»

Также вся информация о документах будет храниться внутри самой системы в базе данных, что удовлетворяет требованию инкапсуляции.

3.3. Контейнеризация

Вместо того, чтобы полагаться на операционную систему, установленную на главном компьютере, контейнеры «Docker» содержат свою минимальную операци-

онную систему, специально разработанную для запуска приложения. Эта *автономная* среда включает свою файловую систему, процессы и ресурсы. Все необходимые библиотеки, зависимости и исполняемые файлы, которые необходимы для правильной работы приложения, также включены в контейнер, что устраняет любые конфликты между версиями и гарантирует работу приложения в контролируемой и предсказуемой среде.

Контейнер уже содержит всё необходимое для бесперебойной работы приложения. Сюда входят версии библиотек, настройки среды выполнения и другие компоненты. В результате воспроизведение среды выполнения в другой системе становится простым.

Docker использует файлы конфигурации, такие как Dockerfile, чтобы указать, как создавать и запускать контейнер. Они предлагают декларативный подход для определения зависимостей и настроек, упрощающий процесс воспроизведения среды при одновременном повышении надёжности.

Благодаря этому подходу реализуется сразу два принципа: консервация и эмуляция.

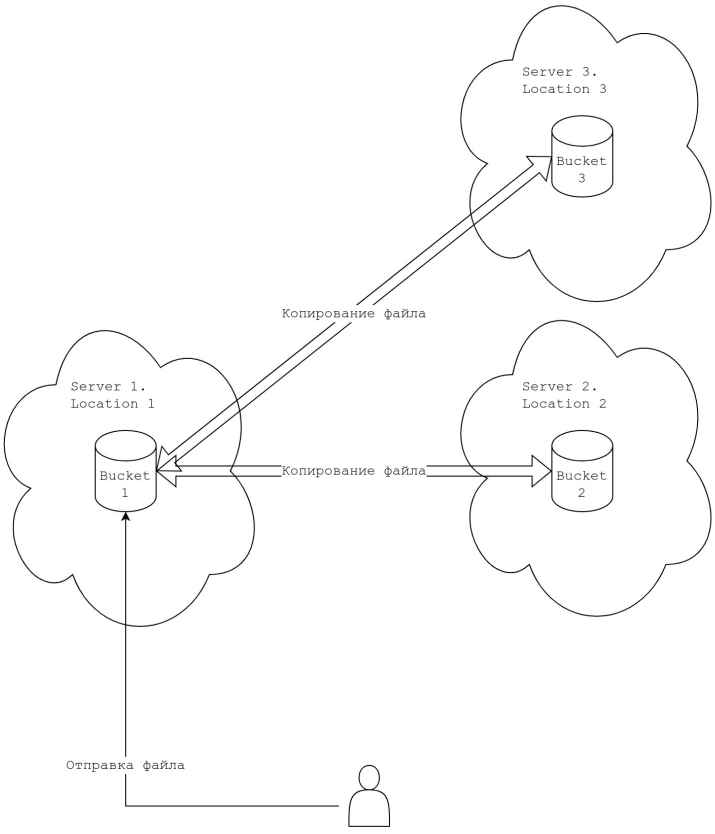

Рис. 2. Пример загрузки файла и его последующее копирование в другие хранилища

3.4. Репликация хранилищ

Репликация данных через хранилища данных позволяет создавать резервные копии в разных местах (см. рис. 2). В случае, если один из узлов или серверов, хранящих данные, становится недоступным, например, из-за неисправности или технического обслуживания, пользователи и система по-прежнему могут получать доступ к данным из реплик. Это повышает надёжность систем и к тому же уменьшает влияние сбоев на доступность данных.

Более того, репликация данных служит средством обеспечения целостности и безопасности. В случае, если одна реплика скомпрометирована из-за атак или случайной потери данных, другие реплики могут сохранять копии. Это повышает уровень безопасности данных и обеспечивает восстановление после негативных инцидентов без ущерба для важной информации.

Как и остальная инфраструктура в системе, хранилища будут помещены в контейнер, и на них будут располагаться правила, описанные в предыдущей секции.

4. Заключение

Исходя из проведённого исследования существующих правил и систем долгосрочного архивирования исторических данных, для разрабатываемого проекта перспективнее всего будет использование микросервисной архитектуры и контейнеризации системы, а также использование репликаций хранилищ, что, несомненно, скажется на доступности и сохранности исторических данных.

5. Благодарности

Данная работа была выполнена при поддержке гранта № 23-78-10119, выделенного Российским научным фондом.

Литература

1. Martyniak C. Definitions of Digital Preservation // American Library Association. 2007. Vol. 4.
2. ООН. Хартия о сохранении цифрового наследия. 2003. URL: `https://www.un.org/ru/documents/decl_conv/conventions/digital_heritage_charter.shtml` (дата обращения: 10.01.2010).
3. Digital Preservation at UW–Madison // University of Wisconsin–Madison Libraries. 2019.
4. Hazarika R. Digital Preservation in Academics Libraries // International Journal of Library and Information Studies. 2020. No. 10. P. 6.
5. Pringle A. The Role of Metadata in Digital Preservation // Library and Information Science Graduate Student Posters. 2018. P. 2.
6. The State of the Art and Practice in Digital Preservation // Journal of Research of the National Institute of Standards and Technology. 2002.
7. Толкование на Апокалипсис / Андрей Кесарийский. [Б. м.], 1910–1920 гг. URL: `http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000204036` (дата обращения: 10.01.2010).

HISTORICAL IMMORTALISATION: EFFICIENT METHODS FOR LONG-TERM STORAGE OF HISTORICAL DATA

T.E. Boldovskaya[1,2]
Ph.D. (Techn.), Associate Professor, e-mail: boldovskaya73@gmail.com
V.I. Gres[2,3]
Laboratory Assistant, e-mail: gresvladimir02@gmail.com
A.E. Vetrov[2,3]
Laboratory Assistant, e-mail: aleksandrvetrov838@gmail.com

[1]Omsk State Technical University, Omsk, Russia
[2]Tomsk State University, Tomsk, Russia
[3]Dostoevsky Omsk State University, Omsk, Russia

Abstract. The article is devoted to the study of the issues of long-term preservation of digital historical data and ensuring their security in the digital world. Modern methods of archiving electronic data are analyzed and a conclusion is made about the expediency of using a microservice architecture with storage replication.

Keywords: archiving, preservation of digital historical data, containerization, storage replication.

Дата поступления в редакцию: 25.11.2023