Научная статья

УДК 001:002.1:930.25:004.738.5
https://doi.org/10.20913/1815-3186-2024-3-28-37

Современные технологии веб-архивирования

Н. С. Редькина

Редькина Наталья Степановна,

Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук,
ул. Восход, 15, Новосибирск, 630102, Россия,
доктор педагогических наук,
заведующий отделом научных исследований открытой науки

ORCID: *0000-0002-3486-9711*
SPIN: *9887-6329*
e-mail: *redkina@spsl.nsc.ru*

Аннотация. Идея веб-архивирования, реализованная впервые в 1996 г. как способ сохранения веб-контента для будущих исследователей, не утратила своего значения в XXI веке, что подтверждается значительным количеством созданных веб-архивов, разработкой программного обеспечения и инструментов веб-архивирования, повышением осведомленности об инициативах по сохранению интернета, внесением изменений в законодательство некоторых стран по обеспечению доступа к целостности исторических данных в цифровом виде. Целью исследования стало определение технологий веб-архивирования, способствующих сохранению веб-контента на глобальном, национальном и локальном уровнях, а также в рамках формирования широкого спектра тематических коллекций. В результате исследования определены тренды развития веб-архивов, подходы к структурированию системы веб-архивов для более эффективной организации работы с ними, а также этапы и способы реализации веб-архивирования, позволяющие выполнить полный цикл сохранения: сбор, сохранение, предоставление доступа, распространение и оценка полученных результатов. Сделан вывод о перспективах развития веб-архивов с учетом стандартов, рекомендованных Международным консорциумом по сохранению интернета (IIPC), а также современных инструментов веб-архивирования, в том числе с открытым исходным кодом, позволяющих расширять возможности и функциональность веб-архивов как источников поиска открытой информации, получения новых знаний, восстановления утраченной информации, часто имеющей большое культурное, научное, образовательное, художественное и социальное значение, а также проверки ранее опубликованных данных.

Ключевые слова: веб-архивы, библиотеки, открытый доступ, информационные ресурсы, интернет-архив, Международный консорциум по сохранению интернета

Для цитирования: Редькина Н. С. Современные технологии веб-архивирования // Библиосфера. 2024. № 3. С. 28–37. https://doi.org/10.20913/1815-3186-2024-3-28-37.

Статья поступила в редакцию 07.06.2024
Получена после доработки 28.08.2024
Принята для публикации 30.08.2024

© Н. С. Редькина, 2024

Modern Web Archiving Technologies

Natalya S. Redkina

Redkina Natalya Stepanovna,
State Public Scientific Technological Library of the Siberian Branch of the Russian Academy of Sciences,
15 Voskhod St., Novosibirsk, 630102, Russia,
Doctor of Pedagogical Sciences,
Head of the Department of Scientific Research of Open Science

ORCID: *0000-0002-3486-9711*
SPIN: *9887-6329*
e-mail: *redkina@spsl.nsc.ru*

Abstract. The idea of web archiving, pioneered in 1996 as a way to preserve web content for future researchers, has remained important in the 21st century. It is evident by the significant number of web archives, the development of web archiving software and tools, and increased awareness of initiatives to preserve the internet-resources, introducing changes in the legislation of some countries to provide access to historical web content. The purpose of the study is to identify web archiving technologies that contribute to the preservation of web content at the global, national and local levels, as well as within the framework of the formation of a wide range of thematic collections. As a result, trends in the development of web archives, approaches to structuring the web archive system for more efficient organization of work with them, as well as stages and methods of implementing web archiving, that allow one to complete the full preservation cycle: collect, save, provide access, distribute and evaluate the results obtained. A conclusion is made, that the prospects for the further development of web archives, taking into account the standards for collecting, preserving and providing long-term access to web content, recommended by the International Consortium for Internet Preservation, including modern web archiving tools (e.g. open source codes). They allow expanding capabilities and the functionality of web archives as sources of searching for open information, obtaining new knowledge, restoring lost information, as well as checking previously published data, that often have great cultural, scientific, educational, artistic and social significance.

Keywords: web-archives, libraries, open access, information resources, Internet Archive, International Internet Preservation Consortium

Received 07.06.2024
Revised 28.08.2024
Accepted 30.08.2024

Citation: Redkina N. S. Modern Web Archiving Technologies. Bibliosphere. 2024. № 3. P. 28–37. https://doi.org/10.20913/1815-3186-2024-3-28-37.

Введение

Интернет – динамично развивающийся информационно-коммуникационный ресурс. В 2024 г. интернетом пользуются более 66 % всех жителей Земли, общее количество пользователей во всем мире составляет 5,35 млрд человек, при этом почти 61 % респондентов исследования говорят, что «поиск информации» является одной из главных причин использования интернета, что делает эту мотивацию наиболее распространенной на мировом уровне[1]. В ходе опроса, проведенного компанией Netcraft в апреле 2024 года[2], насчитывалось 1 092 963 063 активных сайта, 267 934 761 домен и 12 872 291 компьютер с выходом в интернет (рис. 1). Что касается общего количества сайтов, то, по данным Hosting Tribunal, в Сети существует от 1,6 до 1,9 млрд сайтов (Chakarov, 2023), включая, однако, и неактивные, заблокированные, изменяемые и неподдерживаемые сайты, веб-контент которых может быть удален либо переработан (например, в процессе редизайна сайта), что повлекло неактуальность ссылок и появление ошибок. В решении этой проблемы помогают веб-архивы, которые сохраняют данные и являются уникальными источниками информации для пользователей. При этом отмечается, что веб-архивы чаще используют исследователи и необходимо, чтобы они применялись для обслуживания всего цифрового общества (Gomes, 2022).

Веб-архивы – это новая форма архивных материалов, реализуемая посредством отбора и сохранения веб-страниц или сайтов с использованием методов, инструментов и платформ с долгосрочным сохранением ресурса и обеспечением стратегий доступа к нему. Веб-архивы активно развиваются и позволяют обеспечивать доступ к сайтам, которые были удалены или изменены. Исследователи позиционирует веб-архивы как важнейшие части веб-инфраструктуры, формирующейся и формируемой потребностями и мотивацией других участников веб-пространства (Maemura, 2023a). Утверждается, что современные веб-архивы унаследовали как сильные стороны, так и ограничения

[1] Digital 2024 : global overview report // DataReportal : website. URL: https://datareportal.com/reports/digital-2024-global-overview-report?utm_source=Global_Digital_Reports&utm_medium=Partner_Article&utm_campaign=Digital_2024 (accessed 04.06.2024).

[2] April 2024 Web server survey // Netcraft : website. URL: https://www.netcraft.com/blog/april-2024-web-server-survey/ (accessed 04.06.2024).

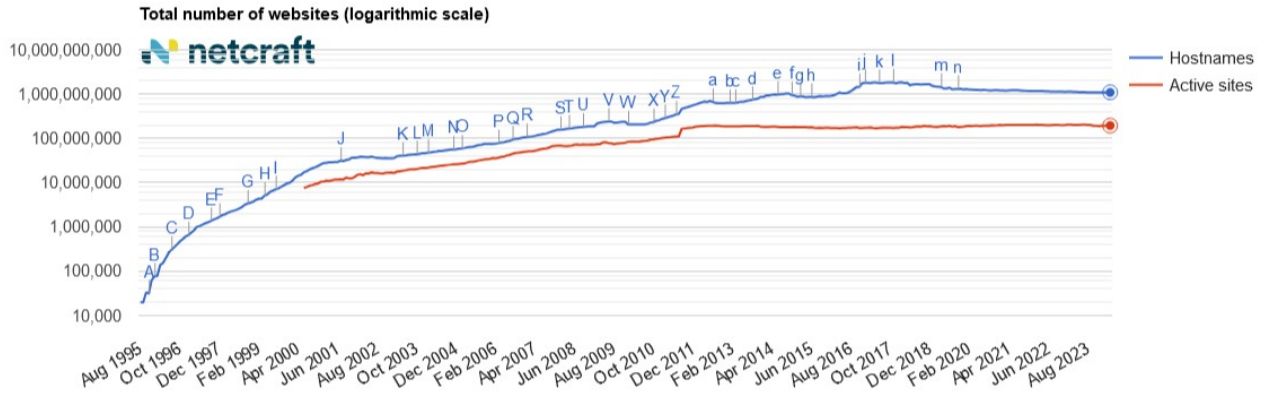

Рис. 1. Динамика развития сайтов по данным Netcraft (1995–2024 гг.)
Fig. 1 Dynamics of website development according to Netcraft (1995–2024)

от старых инфраструктур знаний, из которых они возникли (Hegarty, 2022): в первую очередь это значительно большее количество ресурсов для их подготовки (даже в небольших масштабах). Поэтому полноценное веб-архивирование, предполагающее цикл работ по сохранению веб-контента (сбор, хранение, предоставление доступа, распространение и оценку полученных результатов), – достаточно трудоемкая задача для одной или нескольких организаций. В связи с этим предлагаются разные подходы к организации веб-архивов и технологиям веб-архивирования (Балацкая, Мартиросова, 2023; Редькина, 2021; Смирнов, 2022).

Методика исследования

В исследовании были рассмотрены современные тенденции развития интернета и выявлены проблемы с сохранением веб-контента. На основании данных веб-аналитического инструмента SimilarWeb определена популярность наиболее известного веб-архива в рейтинге наиболее используемых ресурсов в мире. Далее изучены современные подходы к технологиям веб-архивирования, в первую очередь предлагаемые Международным консорциумом по сохранению интернета (IIPC), а также представленные в Сети практики некоторых организаций, и исследования, опубликованные в профессиональной печати. Полученные результаты продемонстрировали разработанность подходов к технологическим процессам создания и организации доступа к веб-архивам, что позволяет заинтересованным лицам и организациям сохранять веб-контент, развивать исследовательскую инфраструктуру и обеспечивать пользователей иногда ценнейшей, но уже недоступной для поиска в Сети информацией.

Тренды развития веб-архивов

Идея веб-архивирования, реализованная впервые Национальной библиотекой Австралии, Национальной библиотекой Швеции и Internet Archive (archive.org) как способ сохранения веб-контента для будущих исследователей, продолжает свое развитие. Самый большой веб-архив в мире – Интернет-архив (archive.org), созданный в 1996 г., предлагает в 2024 г. бесплатный доступ к 866 млрд заархивированных веб-страниц, а также к книгам (44 млн), более 10 млн видеоматериалов, 4,8 млн изображений, 1 млн программ и 15 млн аудиозаписей, в том числе 255 тыс. живых концертов. Веб-архивы превратились в удобную и доступную исследовательскую инфраструктуру, интерес со стороны пользователей возрос, о чем свидетельствуют данные веб-аналитического инструмента SimilarWeb (https://www.similarweb.com), в глобальном рейтинге которого один из самых масштабных проектов Интернет-архив в марте 2024 г. занимал 181-е место с общим количеством визитов за месяц в 138 млн, а рейтинге категории «Наука и образование > Наука и образование – Другое» ему принадлежит первое место.

Ведущий сервис Internet Archive's Archive (https://archive-it.org) с 2006 г. предоставил услуги веб-архивирования более 1200 организациям из более чем 24 стран, включая библиотеки, исследовательские учреждения, социальные группы и т. д. Archive-it предоставляет инструменты, обучение и техническую поддержку для сбора и сохранения динамических веб-материалов, а также платформу для партнеров, позволяющую делиться своими коллекциями и имеющую многочисленные инструменты поиска, обнаружения и доступа. Пользователи Archive-it сохранили более 40 млрд цифровых записей, общим объемом данных, исчисляемых в петабайтах. Материалы, заархивированные через Archive-it,

Рис. 2. Топ-10 веб-архивов, распределенных по предметным областям в Archive-it в 2024 г.
Fig. 2. TOP-10 web archives distributed by subject area in Archive-it in 2024
Источник: подготовлено автором на основе анализа данных, представленных на сайте Archive-it

хранятся в некоммерческих центрах обработки данных, принадлежащих и управляемых Интернет-архивом, доступны пользователям для самостоятельной загрузки в целях дополнительного сохранения и совместного использования.

Archive-it содержит множество инструментов для создания коллекций, управления опубликованным в интернете контентом и частотой архивирования, добавления ключевых слов или метаданных в коллекции и т. д. Archive-it позволяет отследить динамику развития веб-архивов на своей платформе. Наиболее активными предметными областями среди коллекций веб-архивов в 2024 г. стали такие направления, как «Общество и культура», «Университеты и библиотеки», «Искусство и гуманитарные науки» и др. (рис. 2).

Самым популярным языком в веб-архивах Archive-it стал английский, что обусловлено большим количеством создателей (университетов, музеев, библиотек) из Великобритании и США. Анализ информации, представленной на сайте Archive-it, показал следующие данные об использовании языков:

Язык веб-архива	Количество веб-архивов, использующих язык
английский	2209
испанский	101
ирландский	66
французский	46
китайский	32
немецкий	29
исландский	13
японский	9
русский	9
тайский	9

Многие библиотеки используют сервис Archive-it для выполнения архивирования. Процесс выглядит так: веб-архиватор направляет сканер на целевой сайт → с главной страницы сайта сканер посещает его страницы в соответствии с указаниями → сканер загружает веб-контент, который находит на своем пути, включая PDF-файлы, JPEG и аудиовизуальные материалы. В результате получается копия сайта, по которой можно перемещаться, насколько это возможно, как по «живому» сайту. Сайты проходят обширный процесс тестирования, в ходе которого в параметры сканера вносятся изменения, чтобы обеспечить максимальную точность захвата. Когда окончательное архивирование завершено, каждый сайт проходит контроль качества, при котором упор делается на возможность воспроизведения, а также на сохранение документов, таких как PDF-файлы. На этом этапе добавляются метаданные на заархивированный сайт и становятся доступными.

Archive-it используют многие организации для сбора, сохранения и предоставления доступа к контенту. Библиотека Мичиганского университета (https://www.lib.umich.edu) с помощью услуг Archive-it дополняет собственно выбранные сайты, отражающие исследовательские интересы преподавателей, студентов и сотрудников университета. Archive-it сканирует и сохраняет веб-контент с помощью сканера Heritrix. Все сохраненные сайты помечены как «архивированные веб-страницы» с информацией о сканировании, чтобы их нельзя было путать с действующими сайтами, поскольку архивные версии сайтов могут оказаться неполными. Библиотека Мичиганского университета нацелена на сохранение онлайн-ресурсов в конкретных тематических областях, однако определенные типы контента сложно архивировать, например,

потоковое мультимедиа, контент на основе базы данных и контент на основе JavaScript. Кроме того, под сканирование не попадают разделы сайтов, защищенные паролем. Эти факторы оказывают влияние на полноту веб-архивов и могут содержать значительные пробелы в представленном контенте.

Таким образом, несмотря на уже более чем 25-летнюю историю веб-архивирования и развитие технологий, остаются старые проблемы и возникают новые, связанные со сбором, обработкой, долгосрочным сохранением все возрастающего количества сайтов с разнообразным наполнением и условиями доступности.

Веб-архивирование: технологии и практики

Международный консорциум по сохранению интернета определяет веб-архивирование как процесс сбора частей Всемирной паутины, сохранения коллекций в архивном формате и последующего предоставления архивов для доступа и использования (International Internet Preservation Consortium, IIPC, https://netpreserve.org).

Веб-архивисты обычно используют веб-сканеры для автоматического сбора данных из-за огромных масштабов интернета[3]. Материал в веб-архивах не состоит из исходных сайтов/страниц, а представляет собой тип нового материала, возникшего из живой Сети и реконструированного с помощью вмешательства человека и технологий. Целью веб-архивирования не может быть ни сохранение оригинала, ни его полное копирование. Вместо этого он собирает фрагменты сайта, части контента, дизайн разработчика и опыт пользователей, технологии поддержки системы и контекстные ссылки (Cui et al., 2023).

Веб-архивы представляют собой цифровые коллекции, для использования которых требуются специальные программные инструменты. Так, организации, входящие в IIPC, обязуются обеспечить сохранность своих коллекций веб-архивов и сделать их доступными для будущих исследователей, историков и общественности путем реализации следующих этапов веб-архивирования.

1. **Выбор и планирование** – в зависимости от миссии архива он может быть универсальным, как Интернет-архив, ограниченным географическим регионом, как Веб-архив Великобритании, или по теме. Архивы создаются для сохранения веб-контента конкретного учреждения, коммерческой компании или национального правительства. В рамках этих полномочий веб-архивы также имеют политику по включению только определенных типов веб-страниц и материалов, которые они собирают.

2. **Согласование** – обеспечение разрешения на сбор сайтов, например, поддержка во многих европейских странах (Франция, Великобритания и др.) законодательства об обязательном экземпляре, которое распространяется на архивирование в Сети, что расширяет эту функцию по сбору цифровых материалов.

3. **Сбор данных.** Веб-архивы используют сканеры, связанные с программным обеспечением, которое собирает сайты. Сканеры могут быть как бесплатными, так и являться частью коммерческих услуг. Бесплатные сканеры:

Heritrix – расширяемый проект веб-сканера с открытым исходным кодом, разработанный Интернет-архивом, создающий файлы WARC;

HTTrack – простая в использовании утилита, позволяющая создавать снимок веб-страницы и переходить от ссылки к ссылке;

Screaming Frog – мощный и гибкий сканер сайтов, способный эффективно сканировать как небольшие, так и очень большие сайты, позволяя при этом анализировать результаты в режиме реального времени. В базовой версии веб-сканера можно сканировать до 500 URL-адресов;

Conifer – служба веб-архивирования с открытым исходным кодом, которая создает интерактивную копию любой веб-страницы, включая контент, обнаруженный в результате действий пользователя, таких как воспроизведение видео и аудио, прокрутка, нажатие кнопок и т. д.;

Wget – пакет программного обеспечения для извлечения файлов с использованием HTTP, HTTPS, FTP и FTPS – наиболее широко используемых интернет-протоколов.

Данные, полученные с URL-адресов, сохраняются как файлы WARC (Web ARChive) – это открытый стандартный формат ISO для долгосрочного сохранения веб-страниц и материалов, которые их поддерживают или сопровождают.

Масштаб и охват веб-сбора привели к координации усилий между учреждениями, а международное сообщество практиков разработало общий набор инструментов, основанный на формате WARC для записи информации. Недавние инициативы направлены на расширение использования файлов Web ARChive посредством исследовательских проектов, анализирующих веб-архивы в больших масштабах (Maemura, 2023b).

Формат файла WARC является наиболее распространенной и широко используемой стандартизированной структурой для управления долгосрочным сохранением веб-ресурсов и цифровых данных. Его предшественник, формат файла ARC, был разработан в 1996 г. и использовался Интернет-архивом как «файл-контейнер для собственных веб-ресурсов» (Ruest et al., 2022).

4. **Описание.** После сбора файлы необходимо сохранить с метаданными, которые точно

[3] Awesome web archiving // GitHub : developer platform. URL: https://github.com/iipc/awesome-web-archiving (accessed 04.06.2024).

описывают характеристики и происхождение файлов. Определено, что большинство современных коллекций веб-архивов, особенно те, которые созданы и производятся подписчиками Internet Archive Archive-It, обычно представлены с базовым набором описательных метаданных Dublin Core, что часто связано с распространением веб-архивирования среди библиотек (Ruest et al., 2022). Коллекция веб-архива включает как минимум название коллекции, указанного куратора(ов) коллекции и исходный список.

5. **Долгосрочное сохранение** – веб-архивы хранят свои файлы в аккредитованных безопасных средах с несколькими дубликатами и процессами самопроверки и восстановления.

6. **Доступ** – после сбора и описания файлы должны быть доступны запрашивающим сторонам в формате, который изначально был предназначен для этого программного обеспечения, которое может читать и представлять файлы WARC в качестве необходимых веб-страниц. Интернет-архив использует Wayback Machine, которая предоставляет доступ к миллиардам собранных сайтов.

Сканирование, в зависимости от поставленных задач, может осуществляться с разным интервалом и качеством:

- ежедневное (например, национальные веб-ресурсы собираются каждый день с использованием комбинации браузерных и обычных сканеров);
- ежемесячное;
- ежеквартальное;
- текущее («Сохранить страницу сейчас»), то есть архивирование страницы пользователями в веб-архиве в высоком качестве с помощью инструментов самостоятельного архивирования;
- высококачественное сканирование (с максимально возможным качеством и с использованием наилучшего сочетания доступных технологий);
- полное сканирование страниц;
- специальное сканирование (с выборкой страниц по заданной теме, сканируемых с различной частотой и с использованием методик широкого спектра).

Предлагаются и иные подходы к веб-архивированию. Системный подход к процессу веб-сохранения предложен M. Khan и A. U. Rahman (2019). Основная идея состоит в том, чтобы разделить процесс веб-сохранения на небольшие понятные этапы и разработать пошаговый процесс веб-сохранения, который приведет к созданию хорошо организованного веб-архива. Авторами выделены такие этапы, как:

1) определение области веб-архива (сайтоцентричный, тематически ориентированный либо доменоцентричный);

2) понимание веб-структуры;

3) определение веб-ресурса (блоги, сайты социальных сетей, институциональные сайты или образовательные институциональные сайты, сайты газет или развлекательные сайты) и веб-контента (текстовое содержимое, изображения, мультимедиа-файлы);

4) определение целевой аудитории;

5) приоритизация веб-ресурсов;

6) выбор политики доступа;

7) идентификация метаданных;

8) формат хранения или структура архива;

9) механизмы распространения информации.

На каждом этапе веб-сохранения были выявлены различные методы реализации, которые можно использовать при цифровом архивировании. Потенциальная ценность предлагаемой модели заключается в том, чтобы помочь архивариусу, связанному с ней персоналу и организациям эффективно сохранять свой интеллектуальный цифровой контент для будущего использования. Более того, модель может помочь инициировать процесс сохранения веб-страниц и создать хорошо организованный веб-архив для эффективного управления архивным веб-контентом.

А. А. Смирновым предложен селективный (выборочный) подход к веб-архивированию: копирование и создание коллекций веб-документов, посвященных социально значимым событиям (Смирнов, 2022). П. А. Демидов считает, что существуют два основных типа коллекций: коллекции доменов и селективные коллекции, а архивирование на основе событий – это тип выборочного архивирования, который генерирует специальные коллекции в ответ на конкретное событие (Демидов, 2017). Автор отмечает, что у разных типов коллекций есть свои сильные и слабые стороны. В частности, коллекции доменов часто бывают неполными: файлы могут быть неполными или полными, но не отображаются должным образом или в полной мере на сайте не были захвачены. Чем масштабнее и сложнее сайт, тем более вероятно, что он будет неполным. Однако явный объем коллекций доменов означает, что отношения с другими сайтами и внешним связанным с ним контентом сохраняются лучше, чем на сайте, архивированном как часть выборочной коллекции. Селективное архивирование фокусирует ресурсы на сайтах, которые считаются особенно ценными, и позволяет осуществлять захват в пределах определенной выборки сбора (Демидов, 2017). Отмечается множество различных подходов к структурам коллекций веб-архивов. Некоторые коллекции веб-архивов поддерживают субколлекции, а на некоторые разрешено эмбарго. Часть платформ ограничивает использование сканов одной коллекцией, а другие позволяют снимкам сайтов перемещаться между

коллекциями. Разработчикам инструментов необходимо понимать структуру коллекций для удовлетворения потребностей своих пользователей (Jayanetti et al, 2022).

Н. Г. Поврозник справедливо отметил, что современные веб-архивные инициативы направлены на сохранение веба в глобальном, национальном и локальном масштабах и формирование широкого спектра тематических коллекций (Поврозник, 2020). N. Brügger классифицировал подходы к веб-архивированию на макро- и микроуровнях (Brügger, 2005). Архивирование макросети связано со сбором больших частей интернета. Обычно это делается посредством веб-архивирования доменов, при котором архивируется домен верхнего уровня страны. Веб-архивирование доменов часто подкрепляется законодательством: так обстоит дело в Соединенном Королевстве, где Веб-архив Великобритании ежегодно проводит широкое сканирование домена верхнего уровня (Bingham, Byrne, 2021). Архивирование в микросети носит более избирательный и ограниченный характер, часто используется национальными библиотеками для создания выборочных и тематических веб-архивов. Эти веб-архивы обычно создаются по определенной теме или могут использоваться для сбора онлайн-представлений о событии – например, избрании президента страны.

Выборочное веб-архивирование также позволяет национальным библиотекам работать с профильными специалистами и сообществами для создания уникальных специальных коллекций веб-архивов. Веб-архивирование доменов позволяет получить более широкий, но часто поверхностный снимок веб-сферы страны. Каждый подход к веб-архивированию создает отдельный веб-архив, ни один из которых не заменяет другой (Ryan et al., 2022). Подходы к веб-архивированию в национальных библиотеках различаются в зависимости от таких факторов, как законодательство и ресурсы. Веб-архивирование часто осуществляется сторонним поставщиком от имени национальной библиотеки или может проводиться собственными силами, как это происходит в Британской библиотеке.

Еще один пример совмещения подходов к веб-архивированию у Australian Web Archive (https://webarchive.nla.gov.au/collection, https://trove.nla.gov.au/search/advanced/category/websites), предлагающего поиск более ранних версий сайтов, созданных в прошлом, или тех, которые больше не существуют (официальный сайт Олимпийских игр в Сиднее в 2000 г.). Типы сайтов, которые включаются в архив: новостные сайты и сайты организаций, социальные или личные сайты, в том числе блоги. Сбор публикаций является частью требований к обязательному экземпляру, изложенных в Законе об авторском праве (1968 г.). В 2016 г. обязательный экземпляр был расширен за счет включения онлайн-публикаций. Содержимое собирается автоматически. При веб-архивировании используются роботизированные технологии, которые сканируют все доменное пространство .au и сами собирают веб-домены. Однако у робототехники есть ограничения. Создатели отмечают, что невозможно собрать все. Это означает, что многие заархивированные сайты являются неполными или не имеют тех функций, которые есть на исходном сайте. Иногда сайт может быть настроен таким образом, что роботы не могут автоматически собирать онлайн-материалы, а это означает, что мы не можем их собирать. Отбор материалов для веб-архива осуществляется в соответствии с Политикой развития фонда библиотеки[4] и включает разные подходы к архивированию: коллекция австралийских доменов (.au), тематическое веб-коллекционирование, выборочный сбор сайтов, социальные медиа.

Многие веб-архивы являются результатом сотрудничества с другими учреждениями. Налаживание партнерских отношений повышает эффективность сбора веб-контента.

Технологии самостоятельного веб-архивирования

При необходимости заархивировать собственные сайты, не используя какие-либо службы веб-архивирования, разработаны технологии самостоятельного сканирования и сохранения. Поскольку это собственные материалы, относящиеся к конкретным проектам, первые два этапа из упомянутых выше выполняются самостоятельно. Далее, как и в случае с веб-архивами, используется веб-сканер для создания файлов WARC собственных сайтов. Затем рекомендуется заархивировать их в исследовательском хранилище.

Для самостоятельного архивирования веб-контента разработаны различные инструменты. В частности, Webrecorder (https://webrecorder.net), который предоставляет набор инструментов и пакетов с открытым исходным кодом для захвата интерактивных сайтов и их последующего воспроизведения с максимально возможной точностью. ArchiveWeb.page (https://archiveweb.page) – инструмент от Webrecorder, позволяет превратить браузер в полнофункциональную интерактивную систему веб-архивирования, архивировать веб-контент с помощью

[4] Collection development policy // National Library of Australia : website. URL: https://www.nla.gov.au/about-us/corporate-documents/policy-and-planning/collection-development-policy (accessed 04.06.2024).

браузера и сохранять ее в соответствующем стандартном формате WARC40. Отмечается, что проект Webrecorder стал прорывом в веб-архивировании, поскольку он позволяет любому пользователю или небольшому учреждению создавать свои собственные веб-архивы выбранной информации, в соответствии со стандартными форматами, которые дают возможность повторно использовать и обеспечивать совместимость данных, хранящихся в веб-архивах (Gomes, 2022). Все разработанное программное обеспечение доступно в виде бесплатных проектов с открытым исходным кодом.

В дополнение к этому существуют службы, которые позволяют любому человеку архивировать веб-страницу по URL-адресу. Это системы архивирования веб-ссылок по требованию (цитируемые веб-страницы и сайты или другие виды цифровых объектов, доступных в интернете), которые могут использовать авторы, редакторы, издатели научных статей и книг, чтобы гарантировать, что цитируемые веб-материалы останутся доступными читателям в будущем. Если цитируемые веб-ссылки в журнальных статьях, книгах и так далее не архивируются, пользователи в дальнейшем могут столкнуться с ошибкой «404. Файл не найден» при нажатии на цитируемый URL-адрес. Каждая заархивированная страница получает уникальную ссылку (например, цифровой идентификатор объекта), которая направляет читателей к ее исходной версии в открытом доступе. Эти услуги удовлетворяют потребности пользователей, например, ученые сохраняют веб-страницы, цитируемые в их работах (Costa et al., 2017). Такими инструментами являются, например, Perma.cc (https://perma.cc) и Archive.is (http://archive.is). Perma.cc прост, удобен в использовании, создан и поддерживается библиотеками. На сайте Perma.cc приводятся данные о том, что более 50 % ссылок, цитируемых в решениях Верховного суда, больше не указывают на нужную страницу, а примерно 70 % цитируемых ссылок в академических юридических журналах и 20 % всех статей в области науки, технологий и медицины страдают от «испорченных» ссылок. Archive.is сохраняет текст и графическую копию страницы с большей точностью и также предоставляет короткую и надежную ссылку на неизменяемую запись любой веб-страницы. Создателями отмечается, что это может быть полезно, если необходимо сделать «снимок» страницы, которая может вскоре измениться: прайс-лист, предложение о работе, список недвижимости, пост в блоге и др. Сохраненные страницы не будут иметь активных элементов и скриптов, за счет чего обеспечивают безопасность, поскольку не содержат всплывающих окон или вредоносных программ.

Поиск в веб-архивах

Поиск в веб-архивах – одна из актуальных и до сих пор не окончательно решенных проблем. В частности, полнотекстовый поиск является весьма востребованной функцией веб-архивов. Однако многие веб-архивы включают только поиск по URL. К примеру, Wayback Machine Интернет-архива предоставляет поисковую систему, но она основана на метаданных, таких как заголовок страницы и домен, а не на полнотекстовом индексировании. Некоторые веб-архивы, такие как Португальский веб-архив, Британский веб-архив и коллекции в Archive-It, поддерживают полнотекстовый поиск с учетом возможностей расширенного поиска, выдачи требуемого формата файла (HTML, PDF, MS WORD и др.), диапазона даты сканирования, ограничений по полям: «Содержит все», «Точная фраза», «Не содержит». Однако возникает другая проблема. Для веб-архивов, включающих функцию полнотекстового поиска, несколько версий одной и той же веб-страницы, соответствующих поисковому запросу, отображаются индивидуально без перечисления изменений или группируются вместе таким образом, что изменения скрываются. Но предлагаются системы поиска текста изменений, которые позволяют пользователям находить их на веб-страницах (Frew et al., 2023).

Рассмотрим поисковые возможности на примере веб-архива Новой Зеландии (New Zealand Web Archive, https://natlib.govt.nz/collections/a-z/new-zealand-web-archive). Он представляет собой коллекцию архивных сайтов Новой Зеландии и Тихоокеанского региона. В процессе использования веб-архива можно увидеть визуальную историю того, как сайты менялись с течением времени. Этот архив включает 900 уникальных названий сайтов и более 47 000 страниц сайтов, собранных с 1999 г., включая уже не существующие в Сети. Сайты архивируются для долгосрочного хранения и исследовательских целей. Большинство сайтов в веб-архиве собираются через определенные промежутки времени.

Поиск элементов в веб-архиве с помощью Каталога национальной библиотеки производится по ключевым словам. В результатах имеется возможность использовать фильтр и выполнить поиск по полям «Тип ресурса», «Сайты». Архивированные сайты обозначаются значком «сайт». При нажатии вкладки «Интернет-доступ» предоставляется возможность увидеть все версии сайта и получить доступ к большинству контента, перейдя по ссылкам меню, страницы и даты. Карты сайта также могут быть полезным способом навигации, если ссылки меню не работают. Отметим, что некоторые формы навигации не работают на заархивированных сайтах. Например, для окон поиска обычно требуется

поисковая система, и эту функцию невозможно реализовать на архивных сайтах. Любое содержимое сайта, требующее входа в систему или регистрации, как правило, не архивируется. Это происходит по техническим или юридическим причинам. Такой контент может включать форумы, информационные бюллетени, социальные сети и комментарии.

Музыка и видео, которые являются встроенными файлами с внешних сайтов, например, YouTube, Vimeo или Soundcloud, или контент, для которого требуется внешний медиаплеер, не архивируются при сборе в интернете, но могут быть доступны в другом месте коллекции. Так, веб-архив Новой Зеландии включает в себя широкий спектр сайтов, многие из которых уже недоступны в Сети: музыкальные сайты о новозеландских группах и музыкантах, лейблах, гидах по концертам, форумах, интернет-журналах, сайтах фестивалей, а также музыкальных клубах и обществах.

В веб-архиве Новой Зеландии имеется возможность найти следующие коллекции: «*Спорт и отдых*», охватывающая широкий спектр сайтов традиционных спортивных и связанных со спортом организаций, клубов и ассоциаций, а также малоизвестные сайты о спорте и отдыхе, такие как гонки в ваннах, и целый блог, посвященный поэзии крикета; «*Искусство и культура*», представляющая собой сайты, выбранные по темам искусства, гуманитарных наук и культуры (сайты и блоги художников, сайты критиков, форумы, организации и ассоциации, конференции, симпозиумы, фестивали и награды, а также музеи, галереи и пр.); «*Среда*», включающая сайты, посвященные изменению климата, охране окружающей среды, устойчивому развитию и стихийным бедствиям, таким как землетрясение в Кайкоуре в 2016 г. и др.; «*Известные сайты*» и т. д. Поиск по коллекциям и внутри них представляется перспективным в условиях увеличения потоков веб-контента.

Заключение

Инициативы по веб-архивированию со стороны библиотек и иных организаций направлены на сохранение цифрового культурного, исторического и научного наследия, фиксирование изменений содержания веб-страниц. Масштаб и охват веб-архивирования привели к координации усилий между учреждениями, а международное сообщество практиков разработало общий набор инструментов. Крупномасштабные веб-архивы дают возможность анализа закономерностей и тенденций общества. Технологии, поддерживающие коллекции веб-архивов, становятся более развитыми, а методы анализа данных и поиска – более совершенными. Получают развитие различные коллекции веб-архивов, методы веб-архивирования становятся доступными отдельным пользователям, имеются успешные примеры реализации создания веб-архивов. Мы надеемся, что раскрытие современных подходов к веб-архивированию будет принято во внимание заинтересованными сторонами и веб-архивы, сохраняющие исторические данные и являющиеся уникальными источниками данных, продолжат свое совершенствование как важнейшие ресурсы открытого доступа, способствующие развитию мирового информационного рынка.

Статья подготовлена по плану НИР ГПНТБ СО РАН, проект «Разработка модели функционирования научной библиотеки в информационной экосистеме открытой науки», № 122041100150-3

Автор прочитал и одобрил окончательный вариант рукописи.

Конфликт интересов

Н. С. Редькина входит в редакционную коллегию журнала «Библиосфера», но не имеет никакого отношения к решению редколлегии опубликовать эту статью. Статья прошла принятую в журнале процедуру рецензирования. Об иных конфликтах интересов автор не заявлял.

Список источников / References

Балацкая Н. М., Мартиросова М. Б. Краеведческий веб-архив в структуре информационных ресурсов библиотеки: модель и возможности реализации. Санкт-Петербург, 2023. 208 с. [Balatskaya NM and Martirosova MB (2023) Local history web archive in the structure of library information resources: model and implementation possibilities. Saint Petersburg. (In Russ.)].

Демидов П. А. Способы веб-архивирования в современном архивном деле // Развитие науки и техники: механизм выбора и реализации приоритетов : сб. ст. Междунар. науч.-практ. конф. (25 дек. 2017 г., Омск). Омск ; Уфа, 2017. Ч. 6. С. 69–72 [Demidov PA (2017) Methods of web archiving in modern archival work. *Razvitie nauki i tekhniki: mekhanizm vybora i realizatsii prioritetov: sb. st. Mezhdunar. nauch.-prakt. konf. (25 dek. 2017 g., Omsk)*. Omsk; Ufa, pt. 6, pp. 69–72. (In Russ.)].

Поврозник Н. Г. Веб-архив как источник для изучения современной истории // Исторические исследования в контексте науки о данных: информационные ресурсы, аналитические методы и цифровые технологии. Москва, 2020. С. 401–407 [Povroznik NG (2020) Web archive as a source for studying modern history. *Istoricheskiye issledovaniya v kontekste nauki o dannykh: informatsionnyye resursy, analiticheskiye metody i tsifrovyye tekhnologii*. Moscow, pp. 401–407. (In Russ.)].

Редькина Н. С. Мировые тенденции развития веб-архивов библиотек // Научные и технические библиотеки. 2021. № 1. С. 99–114 [Redkina NS (2021) Global trends in library web-archives. *Nauchnye i tekhnicheskie biblioteki* 1: 99–114. (In Russ.)]. DOI: https://doi.org/10.33186/1027-3689-2021-1-99-114.

Смирнов А. А. Проблемы отечественного и зарубежного веб-архивирования в библиотеках. Веб-архивирование как область деятельности // Научные и технические библиотеки. 2022. № 12. С. 104–123 [Smirnov AA (2022) The problems of national and foreign web-archiving in libraries. Web-archiving as a functional area. *Nauchnye i tekhnicheskie biblioteki* 12: 104–123. (In Russ.)]. DOI: https://doi.org/10.33186/1027-3689-2022-12-104-123.

Bingham NJ and Byrne H (2021) Archival strategies for contemporary collecting in a world of big data: challenges and opportunities with curating the UK web archive. *Big Data & Society* 8 (1). DOI: https://doi.org/10.1177/20539517219904.

Brügger N (2005) Archiving websites: general considerations and strategies. Århus, Denmark: Centre for Internet Reseach. URL: https://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf (accessed 04.06.2024).

Chakarov R (2023) How many websites are there? How many are active in 2023? *WebTribunal: website*. URL: https://webtribunal.net/blog/how-many-websites (accessed 04.06.2024).

Costa M, Gomes D and Silva MJ (2017) The evolution of web archiving. *International Journal on Digital Libraries* 18 (3): 191–205. DOI: https://doi.org/10.1007/s00799-016-0171-9.

Cui C, Pinfield S, Cox A and Hopfgartner F (2023) Participatory web archiving: multifaceted challenges. *Information for a better world: normality, virtuality, physicality, inclusivity: proc. of the 18th Intern. conf., iConference 2023, virtual event, March 13–17, 2023*. Springer, pt. 1, pp. 79–87. DOI: https://doi.org/10.1007/978-3-031-28035-1_7.

Frew L, Nelson ML, Weigle MC (2023) Making changes in webpages discoverable: a change-text search interface for web archives. *2023 ACM/IEEE Joint conference on digital libraries (JCDL): proceedings: Santa Fe, NM, USA, 26–30 June 2023*. Los Alamitos [et al.], pp. 71–81. DOI: https://doi.org/10.1109/JCDL57899.2023.00021.

Gomes D (2022) Web archives as research infrastructure for digital societies: the case study of Arquivo.pt. *Archeion* 123: 46–85. DOI: https://doi.org/10.4467/26581264ARC.22.012.16665.

Hegarty K (2022) The invention of the archived web: tracing the influence of library frameworks on web archiving infrastructure. *Internet Histories* 6 (4): 432–451. DOI: https://doi.org/10.1080/24701475.2022.2103988.

Jayanetti HR, Jones SM, Klein M, Osbourne A, Koerbin P, Nelson ML and Weigle MC (2022) Creating structure in web archives with collections: different concepts from web archivists. *arXiv: website*. DOI: https://doi.org/10.48550/arXiv.2209.08649.

Khan M and Rahman AU (2019) A systematic approach towards web preservation. *Information Technology and Libraries* 38 (1): 71–90. DOI: https://doi.org/10.6017/ital.v38i1.10181.

Maemura E (2023a) Sorting URLs out: seeing the web through infrastructural inversion of archival crawling. *Internet Histories* 7 (4): 386–401. DOI: https://doi.org/10.1080/24701475.2023.2258697.

Maemura E (2023b). All WARC and no playback: the materialities of data-centered web archives research. *Big Data & Society* 10 (1). DOI: https://doi.org/10.1177/20539517231163172.

Ruest N, Fritz S and Milligan I (2022). Creating order from the mess: web archive derivative datasets and notebooks. *Archives and Records* 43 (3): 316–331. DOI: https://doi.org/10.1080/23257962.2022.2100336.

Ryan M, Keating D and Finegan J (2022) Managing and accessing web archives: Irish practitioners' perspectives. *AI & Society* 37 (3): 975–984. DOI: https://doi.org/10.1007/s00146-021-01364-0.