

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В БИБЛИОТЕКАХ

УДК 02:004

DOI: 10.33186/1027-3689-2021-1-99-114

Н. С. Редькина

ГПНТБ СО РАН, Новосибирск, Россия

Мировые тенденции развития веб-архивов библиотек

Аннотация: Обоснована необходимость изучения и распространения практики библиотек мира в области веб-архивирования в целях долгосрочного сохранения информации и обеспечения её доступности в будущем. Выявлены современные технологии, используемые в веб-архивировании, а также проблемы, связанные с динамичной природой сайтов, ошибками, сложностью контента для сохранения и др. Рассмотрены примеры активно развивающихся веб-архивов библиотек мира (технологии отбора, поиска, описания, условия доступа и др.). Результаты исследования показывают, что веб-архивы отбираются для дополнения существующих цифровых коллекций библиотек по актуальным проблемам, например, по коронавирусной болезни (COVID-19), либо в интересах определённой группы пользователей. В целях сохранения культурного наследия национальные библиотеки чаще сосредоточены на сборе сайтов по доменам, отражающих принадлежность к тому или иному государству. Университетские библиотеки концентрируются на сборе веб-архивов, которые служат исследовательским или образовательным потребностям пользователей конкретного учреждения, а публичные библиотеки – на ресурсах, отражающих жизнь местного сообщества. Рассмотренный опыт может быть распространён среди других библиотек мира для развития цифровых коллекций.

Ключевые слова: веб-архивирование, национальная библиотека, научная библиотека, вузовская библиотека, публичная библиотека, культурное наследие, сохранение информации, цифровая коллекция, электронные ресурсы.

INFORMATION TECHNOLOGIES IN LIBRARIES

UDC 02:004

DOI: 10.33186/1027-3689-2021-1-99-114

Natalya S. Redkina

*State Public Scientific and Technological Library, Russian Academy of Sciences
Siberian Branch, Novosibirsk, Russia*

Global trends in library web-archives

Abstract: The need for studying and promoting web-archiving for long-term information preservation and accessibility in future is substantiated. The existing technologies of web-archiving are specified and the problems related to the web dynamic character, errors, content complexity, are revealed. Successful experience in the world libraries' web-archiving is discussed (selection, search, description technologies, access terms, etc.). The study findings demonstrate that web-archives are selected to supplement the libraries' digital collections on hot topics, like COVID-19, or to meet the demands of specific user groups. For the purpose of cultural heritage preservation, the national libraries often focus of acquiring websites by the domains in the corresponding country. The university libraries focus on acquiring web-archives that meet research and educational demands of their users; and public libraries prefer the resources of interest to their local community. The findings may be used by world libraries for developing their digital collections.

Keywords: web-archiving, national library, academic library, university library, public library, cultural heritage, information preservation, digital collection, digital resources.

Интернет – крупнейшее и постоянно увеличивающееся в объёмах хранилище культуры и знаний, источник получения ценной исследовательской информации. Согласно отчётам, публикуемым компанией *Netcraft*, в сентябре 2006 г. в интернете насчитывалось 96 854 877 сайтов [1]; в ходе анализа, проведённого в марте 2020 г., получены ответы от 1 263 025 546 сайтов (в 2018 г. – 1 805 260 010), 189 504 449 уникальных доменных имён. Одна из причин уменьшения количества сайтов – их блокировка большинством веб-браузеров из-за использования устаревших протоколов и уязвимости для атак. Сайты

блокируются, обновляются, подвергаются реструктуризации, удаляются; часть информации теряется либо становится недоступной. Следовательно, небезосновательно считается, что информация, представленная в веб-архивах, полезна для обслуживания архивами, библиотеками и другими учреждениями [2].

Для того, чтобы сохранить контент, который может представлять интерес для последующих поколений пользователей сети и исследователей, были начаты масштабные инициативы веб-архивирования под руководством таких организаций, как Интернет-архив (www.archive.org) и Международный консорциум по сохранению интернета (*International Internet Preservation Consortium, IIPC*). К ним присоединятся различные учреждения, заинтересованные в сохранении веб-ресурсов и имеющие опыт сбора и обработки информации.

Технологии веб-архивирования

По определению *IIPC* (<http://netpreserve.org>), веб-архивирование – это процессы сбора частей Всемирной паутины, сохранения коллекций в архивном формате и последующего предоставления архивов для доступа и использования. Для достижения этой цели используются инструменты, стандарты, технологии, обеспечивающие управление веб-архивами. Этапы архивирования: изначально определяются сайты (устанавливаются параметры по уровням сканирования), затем осуществляются их захват и создание метаданных (схема метаданных, автоматическое извлечение метаданных, а в некоторых случаях – анализ и усовершенствование специалистом).

Сайты собираются поисковыми роботами, представляющими собой программное обеспечение, предназначенное для захвата веб-сайтов в определённый момент времени, и эти архивные сайты затем сохраняются и становятся доступными иногда даже после их исчезновения. Для сбора, сохранения и предоставления доступа к копиям веб-сайтов используют различные инструменты с открытым исходным кодом, такие как *Heritrix* (<https://webarchive.jira.com/wiki/spaces/Heritrix/overview>) и *Open Wayback* (<http://netpreserve.org/web-archiving/openwayback/>). Определяются три основных принципа выбора материалов, которые должны быть сохранены в современных стратегиях

веб-архивирования: тематическое, доменное и временное архивирование [3]. Эти принципы применяются и в проектах библиотек. Библиотеки устанавливают алгоритмы, подобные разработанным в Швеции, для захвата сайтов с доменом конкретной страны (например, «se» для Швеции) или тех, которые относятся к Швеции, но размещены в другой стране. Создание коллекций веб-архивов фиксируется и после появления крупного новостного события. Например, коллекция вируса Эбола в *Archive-It* была создана Национальной медицинской библиотекой (США) через несколько месяцев после вспышки Эболы в 2014 г., а архив по коронавирусной болезни (*COVID-19*) в апреле 2020 г. содержал уже более 400 ресурсов (<https://archive-it.org/collections/4887>).

Сканирование сайтов происходит по ссылкам, а сохраняются они как файлы *WARC (Web ARChive)* (предшественник – формат файлов *ARC*), что обеспечивает постоянный доступ исследователей к ним в будущем. Помимо обеспечения сохранности архивов, необходим эффективный доступ к ресурсам, основанный на потребностях практического исследователя, таких как простота использования, расширяемость и возможность повторного использования. Для достижения этих целей предлагают *ArchiveSpark*, *Warcbase*, *Archives Unleashed Toolkit* и другие платформы для распределённой обработки веб-архивов.

Автоматическая генерация метаданных приобретает всё большее значение в области поиска информации, и существует несколько инструментов с открытым исходным кодом, которые извлекают метаданные из файлов, например, *ImageMagick™*. Однако достаточно часто метаданные извлекают ручным способом, что происходит из-за нестабильности программного обеспечения и разнообразия версий файлов в коллекциях.

Некоторые веб-архивы, ранее доступные только в стенах одного учреждения, становятся открытыми и более удобными для поиска информации. Так, в библиотеке Дейтонского университета (США) коллекции веб-архивов добавили в *EBSCO Discovery Service* и предоставили к ним доступ. Локальный словарь данных был создан на основе описательных метаданных *OCLC* для веб-архивирования (2018), в результате

проиндексирован контент на единой, удобной для пользователя платформе [4]. Приведём типичные описания веб-архивов.

Пример 1

Название: Охрана реки Лакаваксен / Защита и сохранение реки Лакаваксен, её дикой природы, водосбора и природной красоты.

URL: <http://lackawaxenriver.org/>

Описание: Охрана реки Лакаваксен / Защита и сохранение реки Лакаваксен, её дикой природы, водосбора и природной красоты.

Сохранено 2 раза (ОКТ 30, 2017 и АПР 1, 2019)

Раздел: Организации водосбора

Создатель: Penn State University (Университет штата Пенсильвания (США))

Пример 2

Название: Коронавирусная болезнь 2019 (COVID-19)

URL: <http://dhhs.ne.gov/Pages/Coronavirus.aspx>

Сохранено 7 раз в период (МАР 14, 2020 и АПР 9, 2020)

Видео: 2 видео

Группа: Вспышка коронавирусной болезни (COVID-19)

Язык: английский

Создатель: Национальная медицинская библиотека (США)

Проблемы веб-архивирования и их решения

Несмотря на значительный рост проектов по веб-архивированию и увеличение числа стран, принимающих эти инициативы [5, 6], объём заархивированных ресурсов по-прежнему остаётся незначителен по сравнению с количеством информации, публикуемой в интернете. Архивирование современных веб-ресурсов остаётся непростой задачей в связи с отсутствием чёткой концепции исправления ошибок при сохранении сайтов. Проблема возникает из-за динамичной природы сайтов, к примеру, контент может контролироваться технологиями *JavaScript* (язык программирования, поддерживающий объектно-ориентированный, императивный и функциональный стили) или *Adobe Flash* (платформа для создания мультимедийного и интерактивного контента).

На основе анализа выборки 10 тыс. веб-страниц на наличие ошибок в содержимом *Webis Web Archive 2017* авторы работы [7] определили, что около 10% страниц подвержены ошибкам в содержании, а также выявили пять различных классов ошибок, встречающихся при веб-архивировании: сообщения об ошибках (4,5% страниц), всплывающие окна (3,9%), страницы, которые состоят в основном из рекламы (1,1%), *CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart)* – тип теста «вызов-ответ», используемый в вычислениях для определения, является ли пользователь человеком или компьютером) (0,8%) и индикаторов загрузки (0,5%) [Там же]. Ещё одна проблема связана с достоверностью архивированных ресурсов, которые могут быть изменены несанкционированным способом. Фальсификация веб-контента является распространённым преступлением, при котором контент изменяется злонамеренными пользователями и действиями [8].

Анализ материалов ежегодных семинаров по веб-архивации и цифровым библиотекам (*WADL*) позволил определить ещё ряд проблем, которые возникают при проведении веб-архивирования: увеличение масштабов онлайн-информации; сложность контента, как с технической, так и с интеллектуальной точки зрения; эфемерная природа веб-контента; меняющиеся технологии. Актуальны для обсуждения и поиска возможных решений вопросы, связанные со сканированием динамического и мобильного контента, архивированием в социальных сетях, разработкой стандартов, протоколов и инструментов архивирования, использованием метаданных, описанием, классификацией архивных ресурсов, функциональной совместимостью систем веб-архивирования, этикой и др.

Целью семинара в 2019 г. стало изучение интеграции веб-архивов и электронных библиотек, т.е. рассмотрение аспектов, охватывающих весь жизненный цикл цифровых ресурсов (создание, загрузка, публикация в интернете, сканирование/сбор, сжатие, форматирование, хранение, сохранение, анализ, индексирование, поддержка доступа и т.д.), а также процессы архивирования и инструменты для «нетрадиционных» ресурсов, таких как научные и правительственные данные, трёхмерные объекты и цифровое онлайн-искусство [9].

Проекты веб-архивов библиотек мира

В целях сохранения культурного, научного и исторического наследия ряд библиотек мира реализует проекты по включению в БД собственной генерации заархивированных веб-ресурсов либо создание коллекций веб-архивов. Результаты исследования 2017 г. [10] показывают, что наблюдается значительный рост инициатив организаций разных организационно-правовых форм, принимающих участие в проектах по созданию веб-архивов. Вместе с тем подчёркивается, что объём архивированных данных по сравнению с объёмом данных, которые публикуются в интернете, небольшой.

Библиотеки регулярно пересматривают свои стратегии развития коллекций в свете меняющихся моделей сайтов и технологий, а также постановлений правительств. К примеру, Британская библиотека благодаря постановлению Правительства Великобритании получила возможность собирать цифровые публикации и веб-сайты в Великобритании в целях законного хранения [11]. Библиотеки делятся опытом создания веб-архивов, а также проводят обучение специалистов. К примеру, в Национальной библиотеке *Széchenyi* (Венгрия) разработаны тренинги по веб-архивированию для специалистов в области цифровых библиотек Европы [12].

В 2009 г. был проведён опрос среди 16 национальных библиотек с целью прояснить, как национальные библиотеки осуществляют свою деятельность по архивированию через интернет [13]. Результаты показали, что национальные библиотеки предполагают следующее: а) выгоды от их инициатив превышают общие затраты, б) затраты, налагаемые на библиотеки, превышают затраты, налагаемые на заинтересованные стороны, в) все они прилагают усилия для реагирования на риски различными способами (например, законодательство, политика заключения договоров и отказ от участия), хотя существуют компромиссные отношения с точки зрения затрат на ведение переговоров, объёма доступа, размера и объёма веб-архива. Активны были библиотеки Новой Зеландии, Великобритании, Канады, Кореи и других стран [14–17].

Как же изменилась ситуация десять лет спустя? Результаты анализа 25 веб-архивов библиотек мира свидетельствуют о значительном увеличении инициативных проектов среди библиотек разных видов, расширении поисковых возможностей внутри коллекций и условий

доступа пользователей. Рассмотрим примеры (технологии отбора, поиска, доступности и др.) активно развивающихся веб-архивов библиотек:

Национальная библиотека Германии (http://www.dnb.de/EN/Netzpublikationen/Webarchiv/webarchiv_node.html). Доступ к заархивированным ресурсам возможен через поиск на сайте, а полные тексты доступны в здании библиотеки;

Национальная и университетская библиотека Исландии (<https://vefsafn.is>). В веб-архив собираются копии исландских веб-сайтов с 2004 г. в соответствии с законом об обязательном депонировании от 2002 г. Коллекция ограничена доменами «.is» и отобранным набором исландских веб-сайтов в других доменах верхнего уровня. Сбор ресурсов осуществляется с помощью веб-сканера *Heritrix*, разработанного Интернет-архивом. Доступ к коллекции открыт для всех через индивидуальную установку *OpenWayback*;

Национальная библиотека Великобритании (<https://www.webarchive.org.uk/ukwa/>). Архив содержит сайты, которые отражают различные аспекты жизни на всей территории Великобритании. Поиск осуществляется по названию веб-сайта, полному тексту или URL-адресу, а также по тематике, коллекциям или по алфавиту;

Национальная библиотека Уэльса (Великобритания) (<https://www.library.wales/collections/activities/conservation/web-archive-wales>). Проект начат в июне 2003 г. и всё ещё действует. Библиотека разработала свою политику и стратегию цифрового сохранения, в которой отражены цели и задачи создания веб-архива, условия отбора, описание сайтов, использование метаданных, программного обеспечения, возможности сотрудничества с другими организациями, частными лицами и организациями, работающими в области сохранения цифровых данных;

Национальная библиотека Ирландии (Великобритания) (<https://archive-it.org/home/nli>) создаёт ирландский веб-архив. Сайты собираются поисковым роботом. Национальная библиотека архивирует эти материалы в интересах общества, делает их доступными для решения различных исследовательских задач;

Национальная библиотека Швеции (<https://www.kb.se/om-oss>) начала работу по веб-архивированию в 1997 г. Коллекция включает как веб-серверы, расположенные под шведским доменом верхнего уровня «.se», так и серверы, расположенные в других местах, но с идентифика-

цией ресурсов по шведской геолокации. Сбор сведений производится два раза в год; для некоторых ресурсов, к примеру, для коллекции газет, ведётся ежедневное архивирование. Архив открыт для всех, но только в здании библиотеки;

Национальная библиотека Австралии. Библиотека участвует в проекте *PANDORA* (<http://pandora.nla.gov.au>), цель которого – сохранение и доступ к сетевым документным ресурсам страны. Коллекция формируется в сотрудничестве с другими австралийскими библиотеками и организациями. В феврале 2020 г. архив включал: 62 959 заархивированных названий, 210 090 заархивированных экземпляров (отдельный снимок или копия заголовка, который был добавлен в архив, каждая новая копия заголовка называется «архивным экземпляром»), 818 895 257 файлов. Полный текст заархивированного контента каталогизируется и отражается в Национальной библиографической базе данных, записи доступны для поиска через единую поисковую службу поиска библиотеки (*Trove*). Коллекцию *PANDORA* можно просмотреть в тематических и алфавитных списках на домашней странице *PANDORA*. Поисковые сервисы, такие как *Yahoo!* и *Google*, индексируют архив до уровня выбранных отдельных заголовков, но не позволяют полностью раскрыть содержимое коллекции;

Библиотека Конгресса США (<https://www.loc.gov/collections/?fa=original-format:archived+web+site>). Архивы доступны в разделе «Цифровые коллекции» и представлены как тематические коллекции («Афганистан», «Выборы в Индии в 2009 г.», «Веб-архив отдела рукописей» и др.);

Национальная библиотека Новой Зеландии. Новозеландский веб-архив (<https://natlib.govt.nz/collections/a-z/new-zealand-web-archive>) включает веб-сайты с 1999 г., в том числе те, которых уже нет в сети. Эта избранная коллекция сайтов о Новой Зеландии в начале 2020 г. содержала более 5 600 уникальных названий сайтов и более 33 500 веб-ресурсов. Поиск в веб-архиве осуществляется с использованием каталога Национальной библиотеки Новой Зеландии и фильтра «Тип ресурса»; архивные веб-сайты обозначаются значком со словом «веб-сайт». Отметим, что веб-архив не содержит сайты, которые требуют регистрации. При этом некоторые виды ресурсов, например, музыка и видео, которые являются встроенными файлами с внешних сайтов

(*YouTube*, *Vimeo* или *Soundcloud*), или для работы с ними требуется внешний медиаплеер, не архивируются, но могут быть доступны в других специализированных коллекциях библиотеки.

В 2017 г. стало известно, что Президентская библиотека им. Б. Н. Ельцина первой в России начала формировать архив сайтов российских организаций и учреждений в целях предотвращения безвозвратного исчезновения сетевых ресурсов и обеспечения их сохранения [18]. Это направление деятельности Президентской библиотеки им. Б. Н. Ельцина обусловлено стремительным развитием и изменением цифрового пространства. Формирование подобных веб-архивов, по мнению инициаторов проекта, даст возможность пользователям исследовать историю российского интернета, поднимать исчезнувшие сайты, использовать размещённую на них информацию в качестве доказательной базы, делать ссылки на архивные копии и мн. др.

Крупнейшие библиотеки мира занимаются веб-архивированием, собирают, каталогизируют, индексируют и предоставляют в доступ информационные ресурсы. К ним присоединяются и небольшие публичные библиотеки [19], заинтересованные в сохранении локального культурного, научного и исторического наследия, а также библиотеки вузов; появляются и общие проекты.

Совместная деятельность библиотек представлена в рамках проекта веб-архива «Конец срока» (*End of Term, EOT*, <https://cdlib.org/services/pad/webarchiving/end-of-term-web-archive/>), задачей которого стало сохранение веб-сайтов правительства США в конце президентских сроков (2008, 2012 и 2016 гг.). *EOT* – это действительно совместный проект, реализуемый учреждениями-партнёрами, такими как Библиотека Университета Калифорнии, Интернет-архив, Библиотека Конгресса США, Стэнфордский университет.

Университетские библиотеки расширяют сферу своей деятельности по разработке коллекций, включая в них архивные коллекции свободно доступных интернет-ресурсов. В Библиотеке Колумбийского университета (США) эта работа стала возможной благодаря разработанной программе, финансируемой университетом с 2013 г., и поддержке Фонда Эндрю У. Меллона. Примечательно, что эта деятельность

основана на сотрудничестве с другими научными библиотеками. В 2012 г. библиотека стала членом *IIPC*.

Программа сбора веб-ресурсов предусматривает архивирование отдельных веб-сайтов в тематических областях, соответствующих направлениям деятельности университета, веб-сайтов, созданных филиалами Колумбийского университета, а также организаций или частных лиц, чьи документы или записи хранятся в физических архивах библиотеки. Коллекции архивных сайтов (<https://library.columbia.edu/collections/web-archives.html>) включают, например, веб-архив портала «Права человека», «Библиотеку редких книг и рукописей», «Университетский архив». В сотрудничестве с партнёрами ведётся работа по созданию новых тематических веб-архивов: «Веб-архив современных композиторов», «Веб-архив архитектуры, урбанизма и устойчивого развития», «Глобальный веб-комикс», веб-архив Восточной Европы и бывшего Советского Союза и др.

Библиотеки Университета штата Пенсильвания (США, <https://guides.libraries.psu.edu/c.php?g=653349&p=4584008>) также активно развивают тематические веб-архивы. Так, среди коллекций веб-архивов, курируемых библиотеками, имеются веб-архивы сайтов, связанных с темой сланцев и природного газа, по водным ресурсам, водосборным бассейнам в Пенсильвании и др.

Основная цель инициативы веб-архивирования библиотек Государственного университета штата Северная Каролина (США) состоит в том, чтобы сохранить богатое культурное наследие университета в том виде, в каком оно представлено на его веб-сайтах. Эта цель отражена в миссии библиотек по сбору, сохранению и распространению университетских документов.

В сотрудничестве с Интернет-архивом, который предоставляет инструменты и технологии, необходимые для сканирования, сбора, хранения и предоставления доступа к веб-сайтам, созданы коллекции веб-архивов (<https://www.lib.ncsu.edu/web-archiving>): по архитектуре и дизайну (представлены веб-сайты ведущих архитектурных и дизайнерских компаний региона, а также веб-сайты отмеченных наградами архитекторов); «Ландшафтная архитектура» (содержит веб-сайты известных ландшафтных архитекторов, фирм ландшафтной архитектуры, а также веб-сайты, посвящённые крупным проектам ландшафтной архи-

тектуры), «Джеймс Б. Хант-младший» (включаются статьи, написанные о библиотеке). Помимо коллекционирования веб-сайтов штата Северная Каролина, осуществляется сбор сайтов, которые связаны с направлениями исследований университета, например, по инновациям в сельском хозяйстве, компьютерному моделированию.

Университет Индианы (США) помимо архивирования своих веб-сайтов (<https://archive-it.org/collections/219>) проводит аналогичную работу по сохранению информации, публикуемой в аккаунтах социальных сетей университета (<https://archive-it.org/collections/8920>) с целью предоставления доступа к информации, созданной администрацией учреждения, школами, департаментами, отделами обслуживания, институтами, центрами, программами, а также преподавателями, студентами и выпускниками. Архив включает записи, созданные и управляемые университетским сообществом в *Facebook*, *Flickr*, *Google+*, *Instagram*, *Pinterest*, *Twitter*, *Viewpoints* и *YouTube*. Ресурсы архивируются в зависимости от технических возможностей, например, сайт защищён паролем, заблокирован robots.txt или недоступен для автоматизированных систем Интернет-архива. Инструмент поиска, используемый для предоставления полнотекстового доступа к коллекциям веб-архива библиотеки, работает на основе поисковой системы с открытым исходным кодом *Nutch*.

Большинство из рассмотренных выше библиотек являются членами Международного консорциума по сохранению интернета (*IIPC*), в который входят организации из более чем 45 стран, включая национальные, университетские и региональные библиотеки и архивы. Как показал анализ, веб-архивы собираются для дополнения существующих коллекций и служат различным целям. Национальные библиотеки часто сосредоточены на сборе своих национальных доменов для сохранения культурного наследия. Университеты концентрируются на сборе веб-архивов, которые служат исследовательским или образовательным потребностям пользователей.

Выводы

Веб-архивы становятся неотъемлемой частью цифровых коллекций библиотек разных видов. Архивирование сети считается необходимым условием для сохранения мировых онлайн-ресурсов для будущих поколений, возможностью для исследования истории виртуального мира и онлайн-событий в реальном мире. Для решения этих задач библиотеки используют доступные программные средства с открытым исходным кодом, определяют периоды архивирования, приоритеты выбора сайтов (культурные, образовательные, научные, социальные и политические темы для публичных библиотек либо, что характерно для национальных библиотек, – доменные имена, для вузов – тематика исследований и сайты и социальные аккаунты вузов). Библиотеки активно сотрудничают друг с другом и другими организациями для развития проектов, разработки совместных сервисов поиска, индексирования и сохранения веб-контента. Учитывая активность разработок в области веб-архивирования и наличие удачных примеров, представляется, что данная тенденция имеет будущее и в библиотечной практике, несмотря на некоторые проблемы, с которыми сталкиваются библиотеки в процессе этой деятельности (увеличение объёмов информации, эфемерная природа веб-контента, меняющиеся технологии и др.).

СПИСОК ИСТОЧНИКОВ

1. **January** 2020 Web Server Survey: Total number of websites (logarithmic scale). – URL: <https://news.netcraft.com/> (дата обращения: 03.04.2020).
2. **Deschamps R.**, The cost of a WARC: Analyzing web archives in the cloud / R. Deschamps, S. Fritz, J. Lin, I. Milligan, N. Ruest // Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2019 – June. – 2019. – Article № 8791179. – P. 261–264. – DOI: 10.1109/JCDL.2019.00043.
3. **Finnemann N. O.** Web archivert / N. O. Finnemann // Knowledge Organization. – 2019. – Vol. 46. – № 1. – P. 47–70. – DOI: 10.5771/0943-7444-2019-1-47.
4. **Beis C. A.** Accessing Web Archives: Integrating an Archive-It Collection into EBSCO Discovery Service / C. A. Beis, K. N. Harris, S. L. Shreffler // Journal of Web Librarianship. – 2019. – Vol. 13. – № 3. – P. 246–259. – DOI: 10.1080/19322909.2019.1625844.

5. **Costa M.** The evolution of web archiving / M. Costa, D. Gomes, M. J. Silva // *Int. J. Digit. Libr.* – 2017. – Vol. 18. – P. 191. – DOI: 10.1007/s00799-016-0171-9.
6. **Gomes D.** A Survey on Web Archiving Initiative / D. Gomes, J. Miranda, M. Costa // *Proceedings of TPDFL.* – Berlin : Springer, 2011. – P. 408–420.
7. **Kiesel J.** A dataset for content error detection in web archives / J. Kiesel, F. Hubricht, B. Stein, M. Potthast // *Proceedings of the 18th Joint Conference on Digital Libraries (JCDL '19).* – IEEE Press, 2019. – P. 349–350. – DOI: <https://doi.org/10.1109/JCDL.2019.00065>.
8. **Kachhawa R., Nikhil Kumar Singh, Deepak Singh Tomar.** A Novel Approach to Detect Web Page Tampering / R. Kachhawa, N. K. Singh, D. S. Tomar // *International Journal of Computer Science and Information Technologies.* – 2014. – Vol. 5. – № 3. – P. 4604–4607.
9. **Klein M.** Workshop on web archiving and digital libraries (WADL) / M. Klein, Z. Xie, E. A. Fox // *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2019 – June.* – 2019. – Article № 8791196. – P. 455–456. – DOI: 10.1109/JCDL.2019.00041.
10. **Отчёт** о научно-исследовательской работе «Сервис-ориентированные информационные технологии в задачах эффективной организации библиотечных процессов и поддержки научных коммуникаций». Номер в ИСГЗ ФАНО 0334-2016-0002. – Новосибирск, 2018 г. – Режим доступа: <http://www.spsl.nsc.ru/download/nir/2018-NIR-0334-2016-0002.pdf> (дата обращения: 03.04.2020).
11. **Brazier C.** Great libraries? Good libraries? Digital collection development and what it means for our great research collections / C. Brazier // *Digital Information Strategies: From Applications and Content to Libraries and People.* – 2016. – P. 41–56. – DOI: 10.1016/B978-0-08-100251-3.00003-2.
12. **Németh M., Drotos L.** A blended learning-based curriculum on Web archiving in the national Széchényi library / M. Németh, L. Drótos // *Digital Library Perspectives.* – 2019. – Vol. 35. – № 2. – P. 97–114. – DOI: 10.1108/DLP-03-2019-0012.
13. **Shiozaki R.** Role and justification of web archiving by national libraries: A questionnaire survey / R. Shiozaki, T. Eisenschitz // *Journal of Librarianship and Information Science.* – 2009. – Vol. 41. – № 2. – P. 90–107. – DOI: 10.1177/0961000609102831.
14. **Paynter G.** A year of selective web archiving with the web curator at the National Library of New Zealand / G. Paynter, S. Joe, V. Lala, G. Lee // *D-Lib Magazine.* – 2008. – Vol. 14. – № 5–6. – DOI: 10.1045/may2008-paynter.
15. **Lilleniit R.** Archiving the Canadian web: Experiences at library and archives Canada / R. Lilleniit // *Serials Librarian.* – 2007. – Vol. 53. – № 1–2. – P. 139–149. – DOI: 10.1300/J123v53n01_11.
16. **Choi K. H.** A web archiving system of the national library of Korea: OASIS / K. H. Choi, D. J. Jeon // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4312.* – LNCS, 2006. – P. 313–322.
17. **Senserini A.** Archiving and accessing web pages: The Goddard library web capture project / A. Senserini, R. B. Allen, G. Hodge, N. Anderson, Jr. D. Smith // *D-Lib Magazine.* – 2004. – Vol. 10. – № 11. – DOI: 10.1045/november2004-hodgee.

18. **Президентская** библиотека первой в стране начала формирование сетевого архива. – Режим доступа: <https://www.prilib.ru/news/689438> (дата обращения: 03.04.2020).

19. **Milbrodt N.** Curators Needed: How Public Libraries Are Bringing Community Members into Their Web Archiving Practice / N. Milbrodt, E. Ward, K.-R. Blumenthal // LIS Scholarship Archive. – 2018. – April 24. – URL: <https://osf.io/preprints/lissa/pr9hc/> (дата обращения: 03.04.2020). – DOI: 10.31229/osf.io/pr9hc.

REFERENCES

1. **January** 2020 Web Server Survey: Total number of websites (logarithmic scale). – URL: <https://news.netcraft.com/>.

2. **Deschamps R.** The cost of a WARC: Analyzing web archives in the cloud / R. Deschamps, S. Fritz, J. Lin, I. Milligan, N. Ruest // Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2019 – June. – 2019. – Article № 8791179. – P. 261–264. – DOI: 10.1109/JCDL.2019.00043.

3. **Finnemann N. O.** Web archivēt / N. O. Finnemann // Knowledge Organization. – 2019. – Vol. 46. – № 1. – P. 47–70. – DOI: 10.5771/0943-7444-2019-1-47.

4. **Beis, C. A.** Accessing Web Archives: Integrating an Archive-It Collection into EBSCO Discovery Service / C. A. Beis, K. N. Harris, S. L. Shreffler // Journal of Web Librarianship. – 2019. – Vol. 13. – № 3. – P. 246–259. – DOI: 10.1080/19322909.2019.1625844.

5. **Costa M.** The evolution of web archiving / M. Costa, D. Gomes, M. J. Silva // Int. J. Digit. Libr. – 2017. – Vol. 18. – P. 191. – DOI: 10.1007/s00799-016-0171-9.

6. **Gomes D.** A Survey on Web Archiving Initiative / D. Gomes, J. Miranda, M. Costa // Proceedings of TPD. – Berlin : Springer, 2011. – P. 408–420.

7. **Kiesel J.** A dataset for content error detection in web archives / J. Kiesel, F. Hubricht, B. Stein, M. Potthast // Proceedings of the 18th Joint Conference on Digital Libraries (JCDL '19). – IEEE Press, 2019. – P. 349–350. – DOI: <https://doi.org/10.1109/JCDL.2019.00065>.

8. **Kachhawa R., Nikhil Kumar Singh, Deepak Singh Tomar.** A Novel Approach to Detect Web Page Tampering / R. Kachhawa, N. K. Singh, D. S. Tomar // International Journal of Computer Science and Information Technologies. – 2014. – Vol. 5. – № 3. – P. 4604–4607.

9. **Klein M.** Workshop on web archiving and digital libraries (WADL) / M. Klein, Z. Xie, E. A. Fox // Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2019 – June. – 2019. – Article № 8791196. – P. 455–456. – DOI: 10.1109/JCDL.2019.00041.

10. **Otchet** o nauchno-issledovatel'skoy rabote «Servis-orientirovannyye informatsionnyye tehnologii v zadachah effektivnoy organizatsii bibliotечnykh protsessov i podderzhki nauchnykh kommunikatsiy». Nomer v ISGZ FANO 0334-2016-0002. – Novosibirsk, 2018 g. – URL: <http://www.spsl.nsc.ru/download/nir/2018-NIR-0334-2016-0002.pdf>.

11. **Brazier C.** Great libraries? Good libraries? Digital collection development and what it means for our great research collections / C. Brazier // *Digital Information Strategies: From Applications and Content to Libraries and People*. – 2016. – P. 41–56. – DOI: 10.1016/B978-0-08-100251-3.00003-2.

12. **Németh M., Drotos L.** A blended learning-based curriculum on Web archiving in the national Széchényi library / M. Németh, L. Drótos // *Digital Library Perspectives*. – 2019. – Vol. 35. – № 2. – P. 97–114. – DOI: 10.1108/DLP-03-2019-0012.

13. **Shiozaki R.** Role and justification of web archiving by national libraries: A questionnaire survey / R. Shiozaki, T. Eisenschitz // *Journal of Librarianship and Information Science*. – 2009. – Vol. 41. – № 2. – P. 90–107. – DOI: 10.1177/0961000609102831.

14. **Paynter G.** A year of selective web archiving with the web curator at the National Library of New Zealand / G. Paynter, S. Joe, V. Lala, G. Lee // *D-Lib Magazine*. – 2008. – Vol. 14. – № 5–6. – DOI: 10.1045/may2008-paynter.

15. **Lilleniit R.** Archiving the Canadian web: Experiences at library and archives Canada / R. Lilleniit // *Serials Librarian*. – 2007. – Vol. 53. – № 1–2. – P. 139–149. – DOI: 10.1300/J123v53n01_11.

16. **Choi K. H.** A web archiving system of the national library of Korea: OASIS / K. H. Choi, D. J. Jeon // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4312. – LNCS, 2006. – P. 313–322.

17. **Senserini A.** Archiving and accessing web pages: The Goddard library web capture project / A. Senserini, R. B. Allen, G. Hodge, N. Anderson, Jr. D. Smith // *D-Lib Magazine*. – 2004. – Vol. 10. – № 11. – DOI: 10.1045/november2004-hodge.

18. **Prezidentskaya** biblioteka pervoy v strane nachala formirovanie setevogo arhiva. – URL: <https://www.prlib.ru/news/689438>.

19. **Milbrodt N.** Curators Needed: How Public Libraries Are Bringing Community Members into Their Web Archiving Practice / N. Milbrodt, E. Ward, K.-R. Blumenthal // *LIS Scholarship Archive*. – 2018. – April 24. – URL: <https://osf.io/preprints/lissa/pr9hc/>. – DOI: 10.31229/osf.io/pr9hc.

Информация об авторе / Information about the author

Редькина Наталья Степановна – доктор пед. наук, заместитель директора по научной работе ГПНТБ СО РАН, Новосибирск, Россия
to@spsl.nsc.ru

Natalya S. Redkina – Dr. Sc. (Pedagogy), Deputy Director for Research, State Public Scientific and Technological Library, Russian Academy of Sciences Siberian Branch, Novosibirsk, Russia
to@spsl.nsc.ru