https://doi.org/10.52027/18294685-vb.1.25-05

**БРИЧКОВСКИЙ ВЯЧЕСЛАВ ИВАНОВИЧ**
Национальная библиотека Беларуси

# СОВЕРШЕНСТВОВАНИЕ ПРОЦЕССОВ АРХИВИРОВАНИЯ ВЕБ-САЙТОВ ДЛЯ ОРГАНИЗАЦИИ ДОЛГОВРЕМЕННОГО СОХРАНЕНИЯ ЦИФРОВОГО КУЛЬТУРНОГО НАСЛЕДИЯ

**Բանալի բառեր՝** կայքեր, վեբ արխիվացում, կայքերի արխիվային պատճեններ, կայքերի արխիվացում, մետատվյալներ, էլեկտրոնային տեղեկատվական ռեսուրսներ։

**Keywords:** websites, web archiving, archived copies of websites, website archivability, metadata, electronic information resources.

WWW открыл беспрецедентную эру динамичного обмена знаниями. Интернету потребовалось немногим более двух десятилетий, чтобы вырасти из относительно небольшого сервиса, используемого в основном учеными, в глобальную информационную среду. Это теперь не только узел связи, но и уникальная хронология жизни двадцать первого века.

Тем не менее, сама скорость, с которой он развивается, представляет угрозу нашей цифровой культурной памяти, ее техническому наследию, эволюции и нашей социальной истории. Это также создало новые проблемы для учреждений, чья миссия заключается в документировании и сохранении современных знаний и культуры. Многие материалы, такие как научные публикации, агитационные материалы, произведения искусства, правительственные документы, переписка и новости, теперь доступны только в Интернете. Веб-страницы становятся все более динамичными, они постоянно меняются. Чтобы убедиться, что этот контент сохранится для следующего поколения, его необходимо фиксировать в режиме реального времени.

Многие бизнес-процессы, которые раньше ориентировались на физические записи, теперь все чаще используют и ссылаются на страницы и документы, хранящиеся на веб-сайтах. Некоторые документы доступны для пользователей только на сайтах и не существуют в бумажном виде.

WWW стал исключительно важным средством коммуникации для научных исследований, социальной истории и культурного наследия. Однако уникальная и ценная информация находится под угрозой, поскольку контент сайтов постоянно обновляется, заменяется или теряется.

В связи с этим актуальной является проблема веб-архивирования с целью обеспечения долговременного сохранения контента интернет-сайтов в виде доступных для дальнейшего использования архивных коллекций интернет-сайтов.

Веб-архивирование – это процесс сбора веб-сайтов и содержащейся на них информации из всемирной паутины и их сохранение в архиве. Веб-архивирование – жизненно важный процесс, позволяющий людям и организациям получать доступ к знаниям и повторно использовать их в долгосрочной перспективе, а также удовлетворять потребности в получении информации.

Все технологии и инструменты архивирования веб-контента можно разбить на три группы: веб-архивирование на стороне клиента, веб-архивирование на основе транзакций и

веб-архивирование на стороне сервера. Все 3 метода отличаются от технологии «резервного копирования» веб-сайта, которая просто позволяет восстановить сайт из сохраненных файлов в случае возникновения проблемы или необходимости миграции на другую платформу или хостинг. Указанные выше методы касаются архивирования веб-сайтов, и это означает, что пользователи могут собирать, сохранять, получать доступ и перемещаться по сайтам способами, аналогичными исходному действующему сайту [Бричковский, 2022, 299].

За последние годы исследователями был предпринят ряд попыток систематизировать сайты, используя для этого различные критерии. Рассматривая национальный сегмент WWW с точки зрения его составляющих, структуры и содержания, электронные информационные ресурсы сайтов можно разделить на 5 основных групп: – государственной официальной информации; – деловой информации (биржевой, финансовой, статистической и т.п.); – профессионально-ориентированной (научно-технической, политической, образовательной; - специализированной (библиографической, реферативной и др.); – массовой/потребительской, включающей новости, форумы, чаты, энциклопедии и т.п.

Возможно классифицировать сайты и по следующим доминирующим признакам: принадлежность ресурса к определенной организационно-технологической системе, способ выделения объектов учета, структурный тип, назначение и содержание ресурса, а также его видовой состав, правовой статус, открытость, язык, степень распространения, уровень структурированности содержащейся информации Комплексный анализ предложенных классификационных схем свидетельствует о том, что эти критерии в своем подавляющем большинстве актуальны и для систематизации их «архивных аналогов». Так, базовыми характеристиками для группирования архивных копий сайтов в процессе их источниковедческой обработки могут выступать следующие параметры:

- тематическая направленность;
- характер представленной информации;
- фондообразователь/правовладелец;
- область применения и принцип взаимодействия с пользователем на этапе, когда ресурс выполнял свое первоначальное функциональное назначение;
- способ представления информации;
- уровень доступа;
- видовой состав источника и его структурирование;
- уровень актуализации и т.п.

Необходимо отметить, что для каждой группы сайтов необходимо учитывать особенности представления информации на нем при выборе политики веб-архивирования.

Для оценки и отбора на постоянное хранение сайтов можно применять как общие критерии ценности, которые распространяются на все виды документов, такие как критерии происхождения и содержания, так и специфические – обусловлены особенностями таких документов: дублирование информации в пределах одного источника, возможность ее воспроизведения, отсутствие вредоносного программного кода и т.п.

Одним из важных шагов при выборе стратегии веб-архивирования является определение функций и целевого назначения веб-сайта.

WWW – это обширный и разнообразный ресурс. состоит из множества различных типов информации, ресурсов и функций. Государственные органы и организации используют сайты для предоставления информации о направлениях своей деятельности, услугах, обсуждения своих программ сбора отзывов. Семья и друзья используют веб-платформы для того, чтобы общаться. Каждое из этих применений создает совершенно разные типы веб-контента.

Несмотря на значительный рост проектов по веб-архивированию и увеличение числа стран, принимающих эти инициативы, 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) [Редькина, 2021, 103].

Веб-контент представляет собой особую проблему для цифрового сохранения, большая часть которой связана с тем, что исходный цифровой контент разнороден и использует различные форматы представления информации. При сборе контента сайтов возникают проблемы контроля и поддержки этих форматов.

Кроме того, веб-страницы представляют собой сложные элементы, состоящие из множества файлов, скриптов и медиафайлов, которые образуют публикуемый на сайте объект. Веб-страница – это множество контекстуализированных взаимосвязей, которые необходимо сохранить с учетом контекста. Учет этих факторов в требуемом масштабе и детализации контента является одной из самых больших проблем архивирования.

Частично это может быть связано с ограничениями сбора – некоторые из наиболее проблематичных форматов, возможно, никогда не удастся собрать, – а также с нормализующим эффектом сбора, который сводит динамический контент к статическим веб-страницам.

Таким образом, необходимо решить проблему «архивируемости» веб-контента. Наряду с доступностью и производительностью сайта архивируемость является еще одним фактором, который должны принимать во внимание не только создатели сайта при публикации контента в Интернете, но и организации, занимающиеся созданием архивных копий сайтов. «Архивируемость» – это «легкость, с которой содержимое, структура, функциональность и внешнее представление веб-сайта могут быть сохранены, а затем повторно представлены с использованием современных инструментов веб-архивирования [Banos et al., 2013, 11].

Несмотря на многочисленные требования со стороны архивистов и библиотекарей к созданию устойчивых сайтов, текущая тенденция в веб-разработке к созданию приложений на React (сайты, работающие в основном на JavaScript с небольшим количеством HTML страниц), значительно усложняет проблему архивирования. Поскольку многие решения по архивированию ориентированы на HTML, растущее доминирование фрэймворков типа React ставит под угрозу эффективность применения существующих инструментов для архивирования. В разделе часто задаваемых вопросов Интернет-архива (наиболее известного проекта веб-архивирования) говорится, что архив не может обеспечить функциональность исходного (архивируемого) сайта, если сайт содержит интерактивные элементы.

Архивируемость веб-сайта охватывает основные аспекты веб-сайта, имеющие решающее значение для диагностики того, может ли он быть заархивирован с приемлемой полнотой и точностью. Оценка возможности архивирования веб-сайта должна предоставить архивистам ценный инструмент при оценке возможностей архивирования его контента.

Архивируемость можно рассматривать как степень, в которой веб-сайт соответствует условиям безопасной передачи его содержимого в веб-архив в целях сохранения.

При веб-архивировании в настоящее время используются технологии сканирования для сбора контента целевых веб-сайтов. Все они взаимодействуют через HTTP-запросы и ответы.

Такая информация, как недоступность страниц и другие ошибки имеются в протоколах этого обмена сообщениями, и может использоваться для повышения качества сканирования, управления рисками при создании архивных копий сайтов.

Целесообразно объединить такого рода информацию с оценкой соответствия веб-сайта признанным практикам цифрового курирования (например, использование общепринятых стандартов, проверка форматов и использование метаданных) для генерирования оценки, отражающей архивируемость веб-сайта. Архивируемость веб-сайта не следует путать с надежностью веб-сайта. Первое относится к способности архивировать контент веб-сайта, в то время как последнее является системным свойством, которое объединяет такие атрибуты, как надежность, доступность, безопасность, живучесть и ремонтопригодность.

При оценке показателя архивируемости необходимо принимать во внимание следующие моменты:

- Аспекты архивации: факторы, которые необходимо учитывать при расчете показателя архивируемости веб-сайта (например, соответствие стандартам).
- Атрибуты веб-сайта: элементы веб-сайта, проанализированные для оценки архивируемости (например, HTML-разметка, код скриптов).
- Оценки: тесты, выполненные для атрибутов веб-сайта (например, проверка HTML-кода на соответствие требованиям W3C).
- Выбор аспектов архивации связан с рядом следующих соображений:
- Например, существуют ли достоверные рекомендации, указывающие на то, что информация доступна и доступ разрешен (т. е. доступность).
- Соответствует ли информация общему набор спецификаций формата и/или языка (т. е. соответствие стандартам).
- Степень, в которой сайт независим от внешних сайтов (т. е. связность).
- Уровень доступной дополнительной информации о содержании (т. е. об использовании метаданных).
- Является ли время отклика сервера ниже допустимого порога (т. е. производительность).

Веб-сайт считается архивируемым только в том случае, если сканеры могут посещать домашнюю страницу, просматривать ее содержимое, извлекать контент и метаданные его с помощью стандартных HTTP-запросов.

В случае, если сканер не может найти расположение всех веб-элементов, будет невозможно извлечь контент и другую информацию.

На веб-сайте необходимо не только размещать ресурсы, но и обеспечивать надежность ссылок, чтобы сканеры могли их обнаруживать и извлекать информацию. Пример: веб-разработчик создает веб-сайт, содержащий меню javascript, которое генерируется на лету. Сканеры не могут понять это меню, поэтому они не могут найти веб-ресурсы, связанные с этим меню. Для поддержки возможности архивирования веб-сайт, конечно же, должен содержать действительные ссылки. Кроме того, должен быть предоставлен набор карт, руководств и обновлений для ссылок, чтобы помочь сканерам найти весь контент. Они могут отображаться в карте сайта и файле robots.txt.

Производительность является важным аспектом веб-архивирования. Пропускная способность сбора данных сканером напрямую влияет на количество и сложность веб-ресурсов, которые он может обрабатывать. Чем выше производительность, тем быстрее захватывается веб-контент, улучшая процесс архивирования веб-сайта. Пример: если

производительность веб-сайта низкая, то сканеру будет трудно агрегировать контент, и он может даже перестать работать, если производительность упадет ниже определенного порога.

Связность также важна для эффективной работы сканера, а также для управления зависимостями между элементами веб-архива. Если файлы, составляющие один веб-сайт, рассредоточены по разным службам (например, разные серверы для изображения, виджетов javascript, других ресурсов), захват контента может быть неполным, если одна из нескольких служб выходит из строя. Например, изображения, используемые на веб-сайте, но размещенные в другом месте, могут вызвать проблемы при веб-архивировании, поскольку они могут не быть захвачены, когда сторонний сайт не является частью создаваемого архива. Тем более, если целевой сайт зависит от сторонних сайтов, доступность в будущем которых неизвестна, могут возникнуть новые проблемы. Связность проверяется на нескольких уровнях:

- анализ того, сколько хостов используется для размещения контента сайтов;
- анализ того, сколько хостов задействовано в отношении к вспомогательным ресурсам (например, robots.txt, sitemap.xml, и javascripts).

Адекватное предоставление метаданных, относящихся к контенту, также является важным фактором для архивирования. Отсутствие метаданных ухудшает возможности эффективно извлекать контент сайтов. Например, метаданные, такие как передача и кодирование контента, могут быть включены в заголовки HTTP. Требуемый язык конечного пользователя для понимания содержимого может указываться как часть атрибута элемента HTML. Информация, которая может помочь понимать, как контент классифицируется, может быть включена в атрибут и значения элемента META.

Ниже приведены некоторые рекомендации, которые позволяют повысить показатели «архивируемости» сайта.

- Необходимо поддерживать стабильные связи. Стабильные ссылки, поддерживаемые либо с помощью перенаправлений, либо без изменения веб-адресов, обеспечивают постоянное удобство использования.
- Использовать общепризнанные и надежные форматы данных. Наряду с разметкой форматы файлов могут устареть; даже если контент сохраняется в долгосрочном хранилище, в будущем может не быть возможности интерпретировать его. Следует отдавать предпочтение открытым форматам или, по крайней мере, тем, которые можно прочитать с помощью программного обеспечения с открытым исходным кодом.
- Необходимо использовать http GET вместо http POST.
- Желательно предоставить каждому уникальному ресурсу (страницам, изображениям и файлам) на веб-сайте собственный статический URL-адрес: URI.
- Необходимо разрешить просмотр и поиск всех ресурсов, находящихся в общественном достоянии, т. е. размещать их на «внешнем интерфейсе» сайта, доступном через http.
- Необходимо избегать проприетарных форматов для важного контента, особенно для домашней страницы. Не рекомендуется создавать домашние страницы, в значительной степени полагающиеся на изображения или анимацию, такую как Flash, но если такие страницы имеются, то должны быть также альтернативные текстовые версии HTML.
- Необходимо иметь карту сайта и, если возможно, карту сайта в формате XML, в которой перечислены страницы сайта, их относительная важность и частота их обновления. Создание карты сайта обеспечивает сканирование всего содержимого веб-сайта (некоторые страницы могут не обнаруживаться сканером, например, страницы, использующие навигацию Flash или JavaScript).

- Необходимо использовать файл robots.txt, чтобы предотвратить доступ к областям сайта, которые могут вызвать проблемы при сканировании, например, базы данных, включая онлайн-каталоги, функции календаря и т.п.

Веб-архивирование является сложным процессом, успех которого зависит также от решения организационных вопросов. Такой процесс требует разработки политики, рабочих процессов и систем, предполагающих значительное выделение ресурсов и обеспечение его устойчивости. Веб-архивирование необходимо рассматривать как долгосрочный проект, зависящий от поставленных целей, наличия финансовых и человеческих ресурсов. Анализ международного опыта показывает, что такие проекты реализуются в течение длительного периода времени. По сути, серьезный проект веб-архивирования требует разработки политики, определения механизмов и процедур сбора данных, разработки процедур извлечения/создания метаданных, решения вопросов проверки качества собранной информации, организации хранения данных, реализации сервисов поиска и доступа к данным, включая индексацию очень больших объемов данных.

Создание веб-архива также связано с рядом юридических вопросов, включая: законодательство об авторском праве, законодательство о защите персональных, правовые рамки, обеспечивающие целостность контента (чтобы его можно было использовать), положения о недопущении недобровольного захвата незаконного контента, разграничение зон ответственности для национального сегмента веб-пространства т. п. Также должны быть решены вопросы права доступа к веб-архиву после его создания. Важными являются проблемы периодичности и глубины архивирования, например, сколько раз в год контент будет сканироваться.

Для Беларуси открытыми вопросами продолжают оставаться:

- низкий уровень мотивации правообладателей по их передаче на постоянное хранение;
- проблемы авторского права и связанные с этим трудности в отношении получение или копирование архивами отдельных электронных информационных ресурсов в процессе инициативного документирования;
- отсутствие нормативно-правового определения и унифицированного подхода к критериям их отбора для передачи на архивное хранение;
- соответствие программно-технического обеспечения архивной отрасли задачам по воспроизведению информации в долгосрочной перспективе в условиях технического прогресса и т.д. Заметно тормозит темпы формирования в архивах этой составляющей длительный и затяжной процесс экспертизы ценности сайтов, что обусловлено сложностью их внутренней структуры, объемами, видовым составом.

Барьерами на пути создания и использования веб-архивов остаются:

- неподготовленность части пользователей к работе с новыми технологиями и нетрадиционными формами документов архивных фондов, в том числе, психологическая;
- относительно короткий промежуток существования этого вида исторических доказательств, что приводит отсутствие четкой методики и опыта его источниковедческого изучения;
- дискуссии вокруг проблем подлинности, первичности архивных электронных информационных ресурсов, поскольку современные технологии позволяют создать копии, совершенно подобные оригиналам;

- низкий уровень осознания обществом объективных отличий содержания локальных копий, принятых на хранение в архив от ее «аналога» в сети Интернет, если таковой продолжает функционировать.

Серьезной проблемой, с которой сталкивается практически каждое учреждение при запуске и создании программ веб-архивирования, является нехватка персонала. Веб-архивирование (особенно описание и контроль качества) требует значительного количества рабочего времени персонала.

Проведенный анализ показал, что многие библиотеки и архивы вкладывают значительные средства в разработку и внедрение технической инфраструктуры для поддержки крупномасштабных решений веб-архивирования. Постоянно растущая международная сеть архивного сообщества продолжает активно разрабатывать новые инструменты и методы для улучшения существующих возможностей, а также для устранения неизбежной потери доступа к контенту, вызванной эфемерной природой веб-контента и способов его использования.

## ԱՄՓՈՓՈՒՄ

Վեբ արխիվացման գործընթացը կարևոր նշանակություն ունի, քանի որ այն հնարավորություն է տալիս մարդկանց և կազմակերպություններին երկարաժամկետ հեռանկարում հասանելիություն ունենալ տեղեկատվությանը և օգտագործել այն։ Վեբ արխիվացման տեխնոլոգիայի ամենակարևոր խնդիրները վերաբերում են կանոնակարգման և կազմակերպչական խնդիրներին, ինչպես նաև իսկությունը և ամբողջականությունը, փաստաթղթավորումը և որակի վերահսկողությունը ապահովելու ասպեկտներին։ Հոդվածում առաջարկվում են այս խնդիրների լուծման որոշ մոտեցումներ։ Մշակվել են վեբ կայքերի արխիվացման հնարավորությունը գնահատելու չափանիշներ, որոնց կիրառումը նպաստում է վեբ կայքերի արխիվային պատճենների ստեղծման որակի բարելավմանը։

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

**1.** Бричковский, В. И. Использование технологий архивирования веб-сайтов для организации долговременного сохранения цифрового культурного наследия // Развитие информатизации и государственной системы научно-технической информации: РИНТИ-2022: XXI Международная конференция, 17 ноября 2022 г, Минск, ОИПИ НАН Беларуси, 2022, С. 298–301.

**2.** Редькина, Н. С. Мировые тенденции развития веб-архивов библиотек // Научные и технические библиотеки, 2021, Т. 1, №. 1, Москва, ГПНТБ России, С. 99–114.

**3.** Banos V. et al. CLEAR: a credible method to evaluate website archivability // Proceedings of iPRES, 2013, Lisbon, Portugal, iPRES 2013, 2013, P. 9-19.