

Упражнение по архивированию. Архивирование с ArchiveReady

Выберите несколько сайтов (от 3 шт.) и проверьте их в ArchiveReady -> Check now



1. Интересно проверить в метриках ArchiveReady сайты которые пытались архивировать в задании по архивации через wpull. В ходе работы по архивации одни сайты сохранились лучше, другие хуже, другие загрузились частично без загрузки файлов, что согласно статье CLEAR: a credible method to evaluate website archivability могли бы охарактеризовать разные метрики (фасеты) из предложенных

В ходе работы с wpull мы рассматривали несколько страничек с актуальными чартами топ-популярных русскоязычных подкастах на музыкальных платформах, разберем результаты архивации за время работы ~1.5ч по одному:

- [Чарт подкастов на Яндекс-музыке](#) (прогрузка страницы и показателей, неуспешная прогрузка картинок)
- [Стартовая страница подкастов Google](#) - архивация стартовой страницы (выбор языка осуществляется в js)
- [Чарт топ-подкастов castbox.fm](#) - успешная архивация стартовой страницы с метриками, без картинок
- [Аналитика/чарт топ-подкастов Apple music с портала Chartable](#) - успешная архивация страницы
- [Чарт топ-подкастов Deezer.com на русском](#) - отсутствие файлов/ошибка
- [Стартовая страница zvuk.com](#) (ошибка: replay-web.page отображает множество файлов но ни один из них не открывается)

2. Рассмотрим каждую страницу по метрикам CLEAR в сервисе ArchiveReady.com

2.1 [Чарт подкастов на Яндекс-музыке](#):

Overall Rating	Results	
79%  One page printable HTML  EARL XML results	Web Archivability Facet	Rating
	Accessibility	62%
	Cohesion	83%
	Metadata	100%
	Standards Compliance	71%



Наблюдаем предположительно высокую общую оценку. Обращает на себя внимание 100% результат по фасету метаданных – согласно статье CLEAR, это может быть

связано с наличием одновременно тегов lang, тегов отвечающих за семантическую разметку <dc>, <foaf>, <rdf> итд, прописанный тег <META> и другими метаданными.

Кратко суммируя остальные вкладки:

- в файлах сайта ArchiveReady находит сотни (393) ошибок в HTML, других проблем в документах нет (ссылки валидные, JS-скриптов в HTML нет итд)
- во вкладке HTTP ошибок нет (“Правильно определенные HTTP-заголовки помогают веб-ботам понять контент”).
- во вкладке медиа утверждается что картинок нет, что рассматривается как положительная черта тк снижает нагрузку на архивирование (на самом деле картинки на сайте есть, вероятно они подтягиваются в js, что не позволило wpull в прошлой работе их архивировать). Из ошибок приводится долгое время ответа (“Network response time is 3000 ms”) - что относится на ArchiveReady к фасету Accessibility а не Performance.
- вкладка Sitemaps указывает на ограничения Disallow в robots.txt, а также на ненайденный sitemaps.xml.

2.2 [Стартовая страница подкастов Google](#)

Overall Rating	Results	
62%  One page printable HTML  EARL XML results	Web Archivability Facet	Rating
	Accessibility	49%
	Cohesion	0%
	Metadata	100%
	Standards Compliance	100%



Интересным кажется как два фасета Metadata и Compliance получают 100% оценку, в то время как Cohesion получает 0.

- Заметим что именно к Cohesion авторы CLEAR относят независимость от ресурсов подгружаемых в js, что соответствует нашему опыту архивации с wpull.

Рассмотрим другие вкладки для более полной картины

- HTML and CSS: единственная запись о быстрой загрузке (“Retrieved HTML document in 0.27 sec.”). Возможно не считая контента подгружаемого в JS, сама страница HTML очень пустая
- HTTP: несколько записей в пользу Standards Compliance и Metadata (“Content type was clearly defined.”, “Content encoding was clearly defined in HTTP Headers.”)
- Media: отсутствие картинок на сайте но подгружаются 140 картинок извне (Cohesion), долгое время ответа
- Sitemaps: штраф accessibility: наличие Disallow в robots.txt, отсутствие sitemaps.xml



2.3 [Чарт русскоязычных топ-подкастов castbox.fm](#)

Overall Rating	Results	
72%  One page printable HTML  EARL XML results	Web Archivability Facet	Rating
	Accessibility	58%
	Cohesion	69%
	Metadata	80%
	Standards Compliance	82%

Интересно насколько умеренные скоры по всем фасетам получает сайт по сравнению с рассмотренными выше сайтами

- HTML and CSS: здесь также, множество валидных файлов и множество ошибок в других файлах по всем типам ("There are 39 valid and 1 invalid links."), несколько ошибок в HTML и в CSS файлах – accessibility и standards compliance
- HTTP: предупреждение по Accessibility и Metadata ("HTTP caching headers are not available."), других ошибок нет
- Media: также умеренный результат ("Local images found: 7, remote images found: 21"), все проверенные картинки правильно размечены и загружены, долгое время загрузки (721мс)
- Sitemaps: accessibility: наличие Disallow в robots.txt, наличие sitemaps.xml, но отсутствие его упоминания в robots.txt

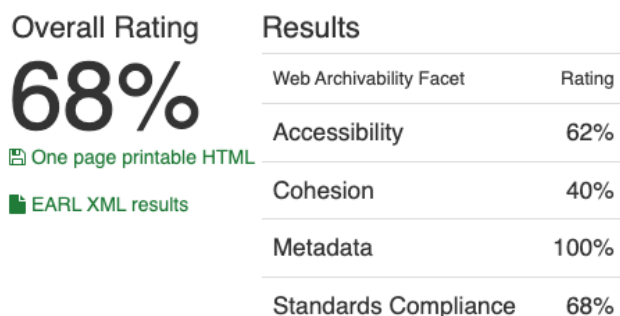
2.4 [Аналитика/чарт топ-подкастов Apple music с портала Chartable](#)

Overall Rating	Results	
58%  One page printable HTML  EARL XML results	Web Archivability Facet	Rating
	Accessibility	32%
	Cohesion	31%
	Metadata	80%
	Standards Compliance	88%

Интересно наблюдать низкий общий скор хотя по работе с wpull этот ресурс скорее казался одним из наиболее позитивных примером

- HTML and CSS: множество нерабочих ссылок, ошибок в нескольких HTML и CSS файлах
- HTTP: предупреждение по Accessibility и Metadata ("HTTP caching headers are not available."), других ошибок нет
- Media: большинство картинок с внешних ресурсов, некоторые картинке не удовлетворяют фасету Standards Compliance
- Sitemaps: accessibility: наличие Disallow в robots.txt, отсутствие sitemaps.xml

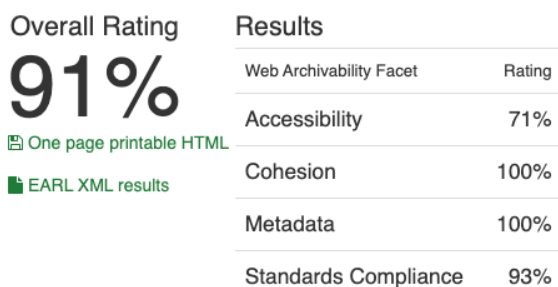
2.5 [Чарт топ-подкастов Deezer.com на русском](#)



Снова интересно наблюдать высокий рейтинг притом как с помощью wpull страницу не удалось загрузить за несколько попыток

- HTML and CSS: по несколько ошибок каждого типа (в HTML, в CSS, в ссылках)
- HTTP: без предупреждений и ошибок
- Media: долгое время ответа (3000мс), других ошибок нет
- Sitemaps: accessibility: наличие Disallow в robots.txt, но присутствует sitemaps.xml

2.6 [Стартовая страница zvuk.com](#)



Согласно метрикам CLEAR и ArchiveReady мы нашли самый архивируемый сайт, (wpull также загрузил множество файлов в архив но все же ни один из них не открывается в replay-web.page)

- HTML and CSS: по несколько ошибок каждого типа (в HTML, в CSS, в inline-js)
- HTTP: без предупреждений и ошибок
- Media: долгое время ответа (670мс), других ошибок нет
- Sitemaps: accessibility: наличие Disallow в robots.txt, но присутствует sitemaps.xml

3. Выводы и сравнение с результатами архивирования wpull

Результаты оказались интересными и во многом неожиданными.

Некоторые параметры в некоторых сайтах кажутся интуитивными и объяснимыми (напр. низкий Cohesion в Google подкастах, интуитивно объяснимый js-ориентированностью сайта).

Другие результаты оказались неожиданными (сайт Deezer показывает высокий скор при том что wpull совсем не был в состоянии создать архив сайта, а сервис Zvuk получил максимальную оценку по CLEAR, притом что страницы созданного wpull:ом архива не открываются).

Напрашивается вывод, что оценка архивации должна также зависеть от постановки задачи к архиватору и критерия качества архива (что считать успешно собранным архивом, а что менее успешным). В нашем случае работы с wpull мы ставили возможно не самую популярную задачу а именно сбор аналитики по конкретным параметрам (именно подкасты, именно на русском, необязательно с прогруженными картинками но желательно с работающими ссылками и обязательно с загруженными показателями). Для других задач (сбор контента сайта как такового) представляется что результат оценки сайта по метрикам CLEAR мог бы сильнее коррелировать с оценкой опытом создания архива самостоятельно.

Объясните, как этот сервис может быть полезен в жизненном цикле веб-архивирования. Опишите результат.

Представляется что подобный сервис и система оценки CLEAR могли бы найти множество применений в контексте веб-архивирования.

- При формировании задачи архивирования можно обращаться к ArchiveReady и уточнять задачу с продуктовой стороны (например, можно выяснить что интересующий сайт не подлежит архивации вовсе, или что точно не удастся сохранить картинки и другие графические элементы что может противоречить исходной задаче и как следствие скорректировать ее)

- Сформировав задачу, далее при выборе экземпляров страниц подлежащих архивированию можно ориентироваться на ArchiveReady в первом приближении. Как мы видим, результаты ArchiveReady не всегда соответствуют ожиданиям от работоспособности получившегося архива, и тем более ожиданиям от результатов поставленной задачи (не все сайты с высокими оценками получилось архивировать вовсе, и тем более не все архивы с высокими оценками получились удачными с точки зрения задачи по сбору аналитики)

- Создав архив, можно также использовать ArchiveReady для сверки результатов. Например если картинки не прогрузились в архиве, с помощью ArchiveReady можно просмотреть, связано ли это с тем что загрузить картинки невозможно (напр если они прогружаются со сторонних ресурсов), или дело во времени загрузки или в технических ошибках при архивировании

- Для устранения ошибок можно также ориентироваться на ArchiveReady. Например, если архив загрузился, но страницы внутри не открываются, хотя получают высокий рейтинг в ArchiveReady - возможно это связано с техническими проблемами во время архивирования. В то же время, если архив не начал загружаться вовсе, хотя ArchiveReady показывает высокий результат – возможно это связано с специальным

блоком на загрузку контента со стороны разработчиков сайта или с другими причинами в меньшей степени зависящих от методов архивации