Int J Digit Libr (2016) 17:119–141
DOI 10.1007/s00799-015-0144-4

A quantitative approach to evaluate Website Archivability using the CLEAR+ method

Vangelis Banos · Yannis Manolopoulos

Received: 28 April 2014 / Revised: 15 February 2015 / Accepted: 25 February 2015 / Published online: 12 March 2015

Abstract Website Archivability (WA) is a notion established to capture the core aspects of a website, crucial in diagnosing whether it has the potential to be archived with completeness and accuracy. In this work, aiming at measuring WA, we introduce and elaborate on all aspects of CLEAR+, an extended version of the Credible Live Evaluation Method for Archive Readiness (CLEAR) method. We use a systematic approach to evaluate WA from multiple different perspectives, which we call Website Archivability Facets. We then analyse archiveready.com, a web application we created as the reference implementation of CLEAR+, and discuss the implementation of the evaluation workflow. Finally, we conduct thorough evaluations of all aspects of WA to support the validity, the reliability and the benefits of our method using real-world web data.

Keywords Web archiving · Website Archivability · Web harvesting

1 Introduction

The number of indexed World Wide Web pages is estimated to be 1.75 billion in early 2014 according to major search engines [35].[1] The approximate numbers of Tumbler blogs (over 80 million) and WordPress websites (over 50 million) also suggest the Web as a popular channel of information exchange. For example, 3.5 billion of the WordPress webpages are being visited every month. These channels of communication are not limited to the younger generation: the average age of active users on social media networks is estimated to be around 37–45 years [12]. Volume, rate of production, and associated demographics in itself are not reasons for archiving material. However, it would be foolish to consider that all the information being produced has no value. The level of traffic, at least, suggests social value, that is, it is indicative of the trends and the narratives of our times. These observations coupled with the recognition that the World Wide Web has been used for purposes beyond personal activities shows that people have been using blogs to upload diagrams, summaries, minutes of meetings, emails, and project plans. Disasters and emergencies are being tweeted and the strong interest in archiving these suggests a need for preservation, at least for selected websites [59].

Web archiving is paramount to preserve our cultural, scientific and social heritage on the web. Is it defined as the process of gathering up digital material from the World Wide Web, ingesting it, ensuring that it is preserved in an archive, and making this collection available for future use and research [43]. Today, a range of 35–90 % of the web has at least one archived copy [2].

As all digital preservation activities, web archiving has two aspects: organisational and technical. The organisational aspect of web archiving involves the entity that is responsible for the process, its governance, funding, long-term viability and personnel responsible for the web archiving tasks [49]. The technical aspect of web archiving involves the procedures of web content identification, acquisition, ingest, organisation, access and use [18,54]. One of the main chal-

V. Banos (✉) · Y. Manolopoulos
Department of Informatics, Aristotle University,
54124 Thessaloníki, Greece
e-mail: vbanos@gmail.com

Y. Manolopoulos
e-mail: manolopo@csd.auth.gr

[1] The numbers reported in this paragraph are from the Daily Estimated Size of the World Wide Web, http://www.worldwidewebsize.com/, January 2014.

lenges of web archiving in comparison to other digital archiving activities is the process of data acquisition. Websites are becoming increasingly complex and versatile, posing challenges for web crawlers to retrieve their content with accuracy and reliability [27]. The process of web crawling is inherently complex and there are no standard methods to access the complete website data. As a result, research has shown that web archives are missing significant portions of archived websites [8]. Furthermore, the archivability of webpages varies depending on the page type, accessibility and technologies used [32]. The problem is that there is no quantitative and objective metric to decide if a website is amenable to being archived successfully. Quality Assurance (QA) checks must be performed by web archive operators and test crawls must be executed prior to archiving to evaluate the outcomes and decide on the optimal web archiving process, if it is possible to proceed with archiving at all.

To resolve this pressing issue, in a previous work we introduced the notion of Website Archivability (WA), a metric that captures the core aspects of a website crucial in diagnosing whether it has the potentiality to be archived with completeness and accuracy [6]. We also defined a first version of the Credible Live Evaluation Method for Archive Readiness (CLEAR), which is used to calculate WA for any website. In this work, we present CLEAR+, the incremental evolution of the CLEAR method. Our contributions can be summarised as follows:

1. the CLEAR+ method to measure WA, which refines and improves every aspect of the CLEAR method (Sect. 3),
2. a detailed presentation of the architecture and implementation of archiveready.com, an online system that functions as a reference implementation of the method (Sect. 4),
3. a proof that CLEAR+ method needs only to evaluate a single webpage to calculate the WA of a website, based on the assumption that webpages from the same website share the same components, standards and technologies (Sect. 5.4).

It is important to note the many benefits in using a quantitative metric to evaluate WA, as the impact in research, industry, education and applications is significant. Since the inception of the CLEAR method to calculate WA and the creation of the archiveready.com web application in 2013, we have received significant feedback regarding WA applications. Some examples are presented below.

1. Quality assurance within web engineering can exploit metrics as a way for developers to better understand the WA level of their web systems throughout the development cycle. Including WA calculation in the testing phase of their workflow, developers are able to create archivable web resources from the beginning and avoid potential problems with web archives. Currently, web developers are not aware if their websites are going to be archived correctly by web archives until it is too late. Websites are published; web archives have issues trying to archive them correctly, forcing web developers to update their websites to resolve these issues. This process could be avoided in many cases if web developers included WA calculation in their testing as issues could be identified early. Drupal[2] is one of the most popular open source CMSs; a session about Website Archivability was part of Stanford Drupal Camp 2014.[3]
2. Web archiving systems can implement WA metrics to detect issues in specific target websites and even avoid some problematic cases. For instance, web archives could calculate the WA of target websites and inform their owners regarding specific web archiving issues. As a result, subsequent crawls could be more successful as the target websites would improve over time. Web archives could also decide to avoid capturing specific websites if their WA scores were very low or some specific evaluations failed. This way, they could save precious resources and avoid dealing with problematic websites. The Internet Archive[4] has considered using WA; sessions regarding WA were part of 2013 Web Archiving meetings in Utah, USA[5] and Innsbruck, Austria.[6]
3. Benchmarking can exploit metrics as a way to explore the archivability level of websites at a high scale, such as within an organisation (e.g. university websites), a domain (e.g. .gov.uk) or within geographical areas (e.g. different EU states).[7]
4. Web crawlers can make use of WA as guidance to take informed decisions regarding web crawling.[8]
5. Students can learn from WA evaluations as they present solid information regarding website attributes, which are essential for web archiving. Some universities in the USA already discuss WA as part of their digital archiving courses.[9]

The concept of WA emerged from our research in web preservation in the context of BlogForever, a blog preservation platform [30]. We realised that automated large-scale

[2] https://drupal.org/.

[3] https://drupalcamp.stanford.edu/.

[4] http://archive.org.

[5] http://archiveitmeeting2013.wordpress.com/.

[6] http://webarchiving2013.wordpress.com/.

[7] Personal communication.

[8] https://library.columbia.edu/bts/web_resources_collection/proposal_examples.html.

[9] Personal communication.

web archiving can be greatly improved regarding performance, resources and effectiveness using a quantitative metric to evaluate target websites prior to crawling and archiving. Using the CLEAR+ method, we are able to avoid websites, which are not archivable and make better use of available resources. We believe that the increasing scale of the web and the limited resource situation of most web archives will force archivists to consider employing quantitative metrics such as CLEAR+ in their selection processes.

The rest of this document is structured as follows: Sect. 2 presents related work. Section 3 articulates the CLEAR+ method to evaluate WA. Section 4 presents archiveready.com, the reference implementation of WA evaluation. Section 5 presents the evaluations of the CLEAR+ method. Finally, our conclusions and future work are presented in Sect. 6.

2 Related work

We present the current state of the art of data extraction for web archiving purposes and highlight issues with current approaches. Next, we look into the major issues of web archiving quality assurance. Also, we present related work in the field of web metrics and, more specifically, approaches for the evaluation of certain website attributes, a work that poses similarities with WA evaluation.

2.1 Web archiving data extraction

Web content acquisition is one of the most delicate aspects of the web archiving workflow because it depends heavily on external systems: the target websites, web servers, application servers, proxies and network infrastructure between the web archive and the target website. The number of independent and dependent elements elevates the risk of harvesting with correctness and accuracy. The case of web archiving is totally different from other cases of digital archiving, where the archivist has the total control of the content to be archived.

Web content acquisition for archiving is performed using robots, also known as "crawlers", "spiders", or "bots", self-acting agents that navigate around-the-clock through the hyperlinks of the web, harvesting topical resources without human supervision [45]. Web crawler development is an active area, with many developments coming from researchers and private companies. The most popular web harvester, Heritrix, is an open source, extensible, scalable, archival quality web crawler [41] developed by the Internet Archive in partnership with a number of libraries and web archives from across the world. Heritrix is currently the main web harvesting application used by the International Internet Preservation Consortium (IIPC)[10] as well as numerous other web archiving projects. Heritrix is being continuously developed and extended to improve its capacities for intelligent and adaptive crawling [21] or capture streaming media [28]. The Heritrix crawler was originally established for crawling general webpages that do not include substantial dynamic or complex content. In early 2014, Archive-It introduced Umbra,[11] which works in conjunction with Heritrix and improves the capture of dynamic web content. This was necessary to capture websites reliant on client-side scripting to render webpages and to ensure an optimal viewing experience to users. The content used to construct dynamic webpages is not immediately delivered to a user's browser but is dynamically rendered based on user actions. Examples of this kind are Gmail and Facebook. Even though dynamic content optimises the user experience and reduces server load, these optimisations make it difficult for traditional web crawlers to discover resources that are necessary for optimal capture and display of archived content.

Research projects also develop their own crawlers aiming at addressing some of the traditional web crawlers' shortcomings. For instance, BlogForever is utilizing blog-specific technologies to preserve blogs [5]. Another case is the ArchivePress project which is based explicitly on XML feeds produced by blog platforms to detect web content [48]. ARCOMEM EC-funded research project aimed to create a new web crawler technology exploiting the social and semantic web for guided web archiving [52].

The presented outline of the development in the domain of web crawling for web archiving indicates that it is a highly active area. As websites become more sophisticated and complex, the difficulties that web crawlers face in harvesting them increase. Thus, web archives cannot be certain that they will be able to capture target websites with precision and accuracy. This leads to the development of quality assurance systems and procedures in web archives.

2.2 Web archiving quality assurance

Web content acquisition for archiving is only considered complete once the quality of the harvested material has been established. The entire web archiving workflow is often handled using special software, such as the open source software Web Curator Tool (WCT),[12] developed as a collaborative effort by the National Library of New Zealand and the British Library, at the instigation of the IIPC. WCT supports such web archiving processes as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata. Focusing on quality review, when a harvest is complete, the harvest result is saved in the digital asset store, and

[10] http://netpreserve.org.

[11] http://blog.archive-it.org/2014/03/13/introducing-archive-it-4-9-and-umbra/.

[12] http://webcurator.sourceforge.net/.

the Target Instance is saved in the Harvested state.[13] The next step for the Target Instance Owner is to Quality Review the harvest. WCT operators perform this task manually. Moreover, according to the web archiving process followed by the National Library of New Zealand, after performing the harvests, the operators review and endorse or reject the harvested material; accepted material is then deposited in the repository [47]. A report from the Web-At-Risk project provides confirmation of this process. Operators must review the content thoroughly to determine if it can be harvested at all [24].

Recent efforts to deploy crowd-sourced techniques to manage QA provides an indication of how significant the QA bottleneck is. The use of these approaches is not new, they were deployed by digitisation projects. The QA process followed by most web archives is time consuming and potentially complicated, depending on the volume of the website, the type of content hosted, and the technical structure. However, to quote IIPC, "it is conceivable that crowd-sourcing could support targeted elements of the QA process. The comparative aspect of QA lends itself well to 'quick wins' for Participants".[14]

IIPC has also organised a crowd-sourcing Workshop in its 2012 General Assembly to explore how to involve users in developing and curating web archives. QA was indicated as one of the key tasks to be assigned to users: "The process of examining the characteristics of the websites captured by web crawling software, which is *largely manual in practice*, before making a decision as to whether a website has been successfully captured to become a valid archival copy".[15].

The topic of quality assurance (QA) practices within the field of web archiving has been the topic of an extensive survey conducted by the University of North Texas [51]. The survey involved 54 institutions engaged in web archiving, which included national libraries, colleges and universities, and museums and art libraries.

Web archiving systems quality has been also the target of evaluation for projects such as the Perseus project [15] where they designed documents to enhance the performance of digital libraries and the UCLA Online Campaign Literature Archive [26] where they compared the performance of two alternative web archiving systems. Another approach worth mentioning was Project Prism which studied risk factors for webpages and websites focusing on preservation [34]. There is also the Archival Acid Test approach to evaluate web archive performance on advanced HTML and JavaScript content [33].

There is a consensus within the web archiving community that web content aggregation is challenging. Thus, QA is an essential stage in the web archiving workflow to ensure that it is done correctly. The problem is that currently, the process requires human intervention and research into automating QA is in its infancy. The solution used by web archiving initiatives such as Archive-It[16] is to perform test crawls prior to archiving[17] but these suffer from, at least, two shortcomings: (a) the test crawls require human intervention to evaluate the results, and (b) they do not fully address such challenges as deep-level metadata usage and media file format validation.

Finally, a new development in web archiving quality assurance is the use of headless web browsers to test archived websites. New versions of WCT enable testing all archived URLs with a headless browser and record any missing objects, while the Internet Archive is also following a similar process.

2.3 Web metrics

Since the inception of the web, there have been online tools to evaluate several aspects of websites. Metrics are quite useful as they try to quantify specific website characteristics. Also, metrics make it easier to address issues, communicate and educate about this topic. Metrics work with specific topics, systematic and scientific. The problem of identifying various website attributes in an automated way using data produced by software evaluation tools has been addressed in the past. The problem is that reliable and robust metrics are not an easy thing to define, especially for inherently complex entities such as websites.

W3C has created an array of tools to evaluate the validity of website markup, CSS, feeds, etc. [57]. W3C metrics are expressed as the number of errors and warnings identified by the tool after each evaluation. Search engine optimisation (SEO) tools have also been available for quite some time. Their aim was to calculate various SEO metrics, which quantify how high they will appear in search engine results. PageRank is another good example of a metric (value 1–10), which quantifies webpage popularity.

Web accessibility is also an area that has been addressed with several automated metrics. For instance, the accessibility measurement method enumerates the total number of points of failure encountered in a page and the total number of potential barriers [55]. Another interesting example of metrics application is the monitoring of web accessibility in Brazilian municipalities' websites [23].

In web engineering, Mendes et al. [40] proposed a set of metrics for web applications development. According to [16], software metrics must be: (a) simple to understand and pre-

[13] http://webcurator.sourceforge.net/docs/1.5.2/Web%20Curator%20Tool%20User%20Manual%20(WCT%201.5.2).pdf.

[14] http://www.netpreserve.org/sites/default/files/.../CompleteCrowdsourcing.pdf.

[15] http://netpreserve.org/sites/default/files/attachments/CrowdsourcingWebArchiving_WorkshopReport.pdf

[16] http://www.archive-it.org/.

[17] https://webarchive.jira.com/wiki/display/ARIH/Test+Crawls.

cisely defined, (b) objective, (c) cost effective, and, (d) informative. Measuring web application quality with WebQEM (Web Quality Evaluation Method), a global ratio is calculated for a given page based on the percentile of accessibility problems related to their respective potential problems for each barrier [44].

Web archiving is an area not yet addressed with any quantitative metric, to the best of our knowledge. In our initial work on this area, we introduced a Credible Live Evaluation of Archive Readiness (CLEAR) method to calculate Website Archivability (WA) [6]. After receiving feedback and WA started to be used in many occasions by the web archiving community and in the academic field (university courses on digital libraries and archiving), we move forward with the extended version of the CLEAR method presented in this work.

WA provides an approach to automate QA, by assessing the amenability of a website to being archived before any attempt is made to harvest it. This approach would provide considerable gains by saving computational and network resource usage through not harvesting unharvestable websites and by saving on human QA of sites that could not be harvested above particular quality thresholds.

3 The Credible Live Evaluation of Archive Readiness Plus method (CLEAR+)

We present the Credible Live Evaluation of Archive Readiness Plus method (CLEAR+) as of 08/2014. We focus on its requirements, main components, WA facets and evaluation methods. We also include a detailed example website evaluation to illustrate CLEAR+ application in a detailed manner.

3.1 General overview

The CLEAR+ method proposes an approach to produce on-the-fly measurement of WA, which is defined as the extent to which a website meets the conditions for the safe transfer of its content to a web archive for preservation purposes [6]. All web archives currently employ some form of crawler technology to collect the content of target websites. They communicate through HTTP requests and responses, processes that are agnostic of the repository system of the archive. Information, such as the unavailability of webpages and other errors, is accessible as part of this communication exchange and could be used by the web archive to support archival decisions (e.g. regarding retention, risk management, and characterisation). Here, we combine this kind of information with an evaluation of the website's compliance with recognised practices in digital curation (e.g. using adopted standards, validating formats, and assigning metadata) to generate a credible score representing the archivability of target websites. The main components of CLEAR+ are:

1. WA Facets: the factors that come into play and need to be taken into account to calculate total WA.
2. Website Attributes: the website homepage elements analysed to assess the WA Facets (e.g. the HTML markup code).
3. Evaluations: the tests executed on the website attributes (e.g. HTML code validation against W3C HTML standards) and approach used to combine the test results to calculate the WA metrics.

It is very important to highlight that WA is meant to evaluate *websites only* and is not destined to evaluate distinct webpages. This is due to the fact that many of the attributes used in the evaluation are website attributes and not attributes of a specific webpage. The correct way to use WA is to provide as input the website homepage. Furthermore, in Sect. 5 we prove that our method needs only to evaluate the home webpage to calculate the WA of the target website, based on the premise that webpages of the same website share the same components, standards and technologies.

WA must also not be confused with website dependability, since the former refers to the ability to archive a website, whereas the latter is a system property that integrates several attributes, such as reliability, availability, safety, security, survivability and maintainability [4].

In the rest of this section we will present in detail the CLEAR+ method. First, we look into the requirements of reliable high-quality metrics and how the CLEAR+ method fulfils them (Sect. 3.2). We continue with the way each of the CLEAR+ components is examined with respect to aspects of web crawler technology (e.g. hyperlink validation; performance measure) and general digital curation practices (e.g. file format validation; use of metadata) to propose four core constituent Facets of WA (Sect. 3.3). We further describe the website attributes (e.g. HTML elements; hyperlinks) are used to examine each WA Facet (Sect. 3.4), and, propose a method for combining tests on these attributes (e.g. validation of image format) to produce a quantitative measure that represents the Website's Archivability (Sect. 3.5). To illustrate the application of CLEAR+, we present an example in Sect. 3.6. Finally, we outline the development of CLEAR+ in comparison with CLEAR in Sect. 3.7.

3.2 Requirements

It is necessary for a newly introduced method and a novel metric, such as WA, to evaluate its properties. A good metric must be Quantitative, Discriminative, Fair, Scalable and Normative according to [46]. In the following, we explain how the WA metric satisfies these requirements.

1. Quantitative: WA can be measured in a quantitative score that provides a continuous range of values from perfectly archivable to completely not archivable. WA allows assessment of change over time, as well as comparison between websites or between groups of websites. For more details see the evaluation using assorted datasets in Sect. 5.2.
2. Discriminative: the metric range of values has a large discriminating power beyond simple archivable and not archivable. Discrimination power of the metric allows assessment of the rate of change. See the underlying theory and an example implementation of the metric in Sects. 3.5 and 3.6.
3. Fair: the metric is fair, taking into account all the attributes of a web resource and performing a large number of evaluations. Moreover, it also takes into account and adjusts to the size and complexity of the websites. WA is evaluated from multiple different aspects, using several WA Facets as presented in Sect. 3.3.
4. Scalable: the metric is scalable and able to conduct large-scale WA studies given the relevant resources. WA supports aggregation and second-order statistics, such as STDDEV. Also WA is calculated in an efficient way; it is relevant to the number of web resources used in a webpage. WA is calculated in real time. The scalability of the archiveready.com platform is presented in Sect. 4.1.1.
5. Normative: the metric is normative, deriving from international standards and guidelines. WA stems from established metadata standards, preservation standards, guidelines of W3C, etc. The proposed metric is based on established digital preservation practices. All WA aspects are presented in Sect. 3.3.

The WA metric has many strengths, such as objectivity, practicality and ability to conduct a large-scale assessment without many resources. Following, we focus on each WA Facet.

3.3 Website Archivability facets

WA can be measured from several different perspectives. Here, we have called these perspectives *WA Facets* (see Fig. 1). The selection of these facets is motivated by a number of considerations:

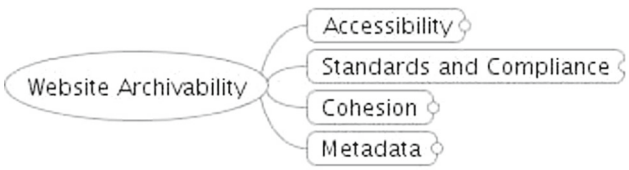

Fig. 1 WA Facets: an overview

1. whether there are verifiable guidelines to indicate what and where information is held at the target website and whether access is available and permitted by a high-performance web server (i.e. Accessibility, see Sect. 3.3.1);
2. whether included information follows a common set of format and/or language specifications (i.e. Standards Compliance, see Sect. 3.3.2);
3. the extent to which information is independent from external support (i.e. Cohesion, see Sect. 3.3.3); and,
4. the level of extra information available about the content (i.e. Metadata Usage, see Sect. 3.3.4).

Certain classes and specific types of errors create less or more obstacles to web archiving. The CLEAR+ algorithm has been enhanced to reflect the significance of each evaluation based on the following criteria:

1. High significance: critical issues which prevent web crawling or may cause highly problematic web archiving results.
2. Medium significance: issues which are not critical but may affect the quality of web archiving results.
3. Low significance: minor details which do not cause any issues when they are missing but will help web archiving when available.

Each WA Facet is computed as the weighted average of the scores of the questions associated with this Facet. The significance of each question defines its weight. The WA calculation is presented in detail in Sect. 3.5.

Finally, it must be noted that a single evaluation may impact more than one WA Facets. For instance, the presence of a Flash menu in a website has a negative impact in the Accessibility Facet because web archives cannot detect hyperlinks inside Flash and also in the Standards Compliance Facet because Flash is not an open standard.

3.3.1 F_A: Accessibility

A website is considered archivable only if web crawlers are able to visit its homepage, traverse its content and retrieve it via standard HTTP protocol requests [22]. In case a crawler cannot find the location of all web resources, then it will not be possible to retrieve the content. It is not only necessary to put resources on a website, it is also essential to provide proper references to allow crawlers to discover them and retrieve them effectively and efficiently.

Performance is also an important aspect of web archiving. The throughput of data acquisition of a web bot directly affects the number and complexity of web resources it is able to process. The faster the performance, the faster the ingestion

of web content, improves a website's archiving process. It is important to highlight that we evaluate performance using the *initial HTTP response time* and not the total transfer time because the former depends on server performance characteristics, whereas the latter depends on file size.

Example 1 A web developer is creating a website containing a Flash menu, which requires a proprietary web browser plug-in to render properly. Web crawlers cannot access the flash menu contents so they are not able to find the web resources referenced in the menu. Thus, the web archive fails to access all available website content.

Example 2 A website is archivable only if it can be fully retrieved correctly by a third party application using HTTP protocols. If a website is employing any other protocol, web crawlers will not be able to copy all its content.

Example 3 If the performance of a website is slow or web crawling is throttled using some artificial mechanism, web crawlers will have difficulties in aggregating content and they may even abort if the performance degrades below a specific threshold.

To support WA, the website should, of course, provide valid links. In addition, a set of maps, guides, and updates for links should be provided to help crawlers find all the content (see Fig. 1). These can be exposed in feeds, sitemap.xml [53], and robots.txt[18] files. Proper HTTP protocol support for Etags, Datestamps and other features should also be considered [13,25].

The Accessibility Evaluations performed are presented in detail in Table 1. For each one of the presented evaluations, a score in the range of 0–100 is calculated depending on the success of the evaluation.

3.3.2 F_S: Standards Compliance

Compliance with standards is a sine qua non theme in digital curation practices (e.g. see Digital Preservation Coalition guidelines [14]). It is recommended that for digital resources to be preserved they need to be represented in known and transparent standards. The standards themselves could be proprietary, as long as they are widely adopted and well understood with supporting tools for validation and access. Above all, the standard should support disclosure, transparency, minimal external dependencies and no legal restrictions with respect to preservation processes that might take place within the archive.[19]

Disclosure refers to the existence of complete documentation, so that, for example, file format validation processes can take place. Format validation is the process of determining whether a digital object meets the specifications for the format it purports to be. A key question in digital curation is, "I have an object purportedly of format F; is it really F? [42] Considerations of transparency and external dependencies refers to the resource's openness to basic tools (e.g. W3C HTML standard validation tool; JHOVE2 format validation tool [19]).

Example If a webpage has not been created using accepted standards, it is unlikely to be renderable by web browsers using established methods. Instead it is rendered in "Quirks mode", a custom technique to maintain compatibility with older/broken webpages. The problem is that the quirks mode is really versatile. As a result, one cannot depend on it to have a standard rendering of the website in the future. It is true that using emulators one may be able to render these websites in the future but this is rarely the case for the average user who will be accessing the web archive with his/her latest web browser.

We recommend that validation is performed for three types of content (see Table 2): webpage components (e.g. HTML and CSS), reference media content (e.g. audio, video, image, documents), HTTP protocol headers used for communication and supporting resources (e.g. robots.txt, sitemap.xml, JavaScript).

The website is checked for Standards Compliance on three levels: referenced media format (e.g. image and audio included in the webpage), webpage (e.g. HTML and CSS markup) and resource (e.g. sitemap, scripts). Each one of these are expressed using a set of specified file formats and/or languages. The languages (e.g. XML) and formats (e.g. jpeg) will be validated using tools, such as W3C HTML[20] and CSS validator,[21] JHOVE2 and/or Apache Tika[22] file format validator, Python XML validator[23] and robots.txt checker.[24]

We also have to note that we are checking the usage of QuickTime and Flash explicitly because they are the major closed standard file formats with the greatest adoption on the web, according to the HTTP Archive.[25]

3.3.3 F_C: Cohesion

Cohesion is relevant for both the efficient operation of web crawlers, and, also, the management of dependencies within digital curation (e.g. see NDIIPP comment on format dependencies [3]). If files comprising a single webpage are dis-

[18] http://www.robotstxt.org/.

[19] http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

[20] http://validator.w3.org/.

[21] http://jigsaw.w3.org/css-validator/.

[22] http://tika.apache.org/.

[23] http://code.google.com/p/pyxmlcheck/.

[24] http://tool.motoricerca.info/robots-checker.phtml.

[25] http://httparchive.org/.

Table 1 F_A: accessibility evaluations

Id	Description	Significance
A_1	Check the percentage of valid vs. invalid hyperlink and CSS urls. These urls are critical for web archives to discover all website content and render it successfully	High
A_2	Check if inline JavaScript code exists in HTML. Inline JavaScript may be used to dynamically generate content (e.g. via AJAX requests), creating obstacles for web archiving systems	High
A_3	Check if sitemap.xml exists. Sitemap.xml files are meant to include references to all the webpages of the website. This feature is critical to identify all website content with accuracy and efficiency	High
A_4	Calculate the max initial response time of all HTTP requests. The rating ranges from 100 % for initial response time less than or equal to 0.2 s and 0 % if the initial response time is more than 2 s. The limits are imposed based on Google Developers speed info[a]. The rationale is that high-performance websites facilitate faster and more efficient web archiving	High
A_5	Check if proprietary file format such as Flash and QuickTime are used. Web crawlers cannot access the proprietary files contents; so they are not able to find the web resources referenced in them. Thus, the web archive fails to access all available website content	High
A_6	Check if the robots.txt file contains any "Disallow:" rules. These rules may block web archives from retrieving parts of a website but it must be noted that not all web archives respect them	Medium
A_7	Check if the robots.txt file contains any "Sitemap:" rules. These rules may help web archives locate one or more sitemap.xml files with references to all the webpages of the website. Although not critical, this rule may help web archives identify sitemap.xml files located in non-standard locations	Medium
A_8	Check the percentage of downloadable linked media files. Valid media file links are important to enable web archives to retrieve them successfully	Medium
A_9	Check if any HTTP Caching headers (Expires, Last-modified or ETag) are set. They are important because they can be used by web crawlers to avoid retrieved not modified content, accelerating web content retrieval	Medium
A_{10}	Check if RSS or Atom feeds are referenced in the HTML source code using RSS autodiscovery. RSS function similarly to sitemap.xml files providing references to webpages in the current website. RSS feeds are not always present; thus, they can be considered as not absolutely necessary for web archiving and with low significance	Low

[a] https://developers.google.com/speed/docs/insights/Server

Table 2 F_S Standards Compliance Facet evaluations

Id	Description	Significance
S_1	Check if the HTML source code complies with the W3C standards. This is critical because invalid HTML may lead to invalid content processing and unrenderable archived web content in the future	High
S_2	Check the usage of QuickTime and Flash file formats. Digital preservation best practices are in favour of open standards; so it is considered problematic to use these types of files	High
S_3	Check the integrity and the standards of images. This is critical to detect potential problems with image formats and corruption	Medium
S_4	Check if the RSS feed format complies with W3C standards. This is important because invalid RSS feeds may prevent web crawlers from analysing them and extracting metadata or references to website content	Medium
S_5	Check if the HTTP Content-encoding or Transfer-encoding headers are set. They are important because they provide information regarding the way the content is transferred	Medium
S_6	Check if any HTTP Caching headers (Expires, Last-modified or ETag) are set. They are important because they may help web archives avoid downloading not modified content, improving their performance and efficiency	Medium
S_7	Check if the CSS referenced in the HTML source code complies with W3C standards. This is important because invalid CSS may lead to unrenderable archived web content in the future.	Medium
S_8	Check the integrity and the Standards Compliance of HTML5 Audio elements. This is important to detect a wide array of problems with audio formats and corruption	Medium
S_9	Check the integrity and the standards compliance of HTML Video elements. This is important to detect potential problems with video formats and corruption	Medium
S_{10}	Check if the HTTP Content-type header exists. This is significant because it provides information to the web archives about the content and it may potentially help to interpret it	Medium

Table 3 F_C Cohesion Facet evaluations

Id	Description	Significance
C_1	The percentage of local vs. remote images is the score of this evaluation	Medium
C_2	The percentage of local vs. remote CSS is the score of this evaluation	Medium
C_3	The percentage of local vs. remote script tags is the score of this evaluation	Medium
C_4	The percentage of local vs. remote video elements is the score of this evaluation	Medium
C_5	The percentage of local vs. remote audio elements is the score of this evaluation	Medium
C_6	The percentage of local vs. remote proprietary objects (Flash, QuickTime) is the score of this evaluation	Medium

persed across different services (e.g. different servers for images, JavaScript widgets, other resources) in different domains, the acquisition and ingest is likely to risk suffering from being neither complete nor accurate. If one of the multiple services fails, the website fails as well. Here we characterise the robustness of the website in comparison to this kind of failure as *Cohesion*. It must be noted that we use the top-level domain and not the host name to calculate Cohesion. Thus, both http://www.test.com and http://images.test.com belong to the top-level domain test.com.

Example A flash widget used in a website but hosted elsewhere may cause problems in web archiving because it may not be captured when the website is archived. More important is the case where, if the target website depends on third party websites, the future availability of which is unknown, then new kinds of problems are likely to arise.

The premise is that, keeping information associated to the same website together (e.g. using the same host for a single instantiation of the website content) would lead to a robustness of resources preserved against changes that occur outside of the website (cf. *encapsulation*[26]). Cohesion is tested at two levels:

1. examining how many domains are employed in relation to the location of referenced media content (images, video, audio, proprietary files),
2. examining how many domains are employed in relation to supporting resources (e.g. robots.txt, sitemap.xml, CSS and JavaScript files).

The level of Cohesion is measured by the extent to which material associated to the website is kept within one domain. This is measured by the proportion of content, resources, and plug-ins that are sourced internally. This can be examined through an analysis of links, on the level of referenced media content, and on the level of supporting resources (e.g. JavaScript). In addition the proportion of content relying on predefined proprietary software can be assessed and monitored. The Cohesion Facet evaluations are presented in Table 3.

One may argue that if we choose to host website files across multiple services, they could still be saved in case the website failed. This is true but our aim is to archive the website as a whole and not each independent file. Distributing the files in multiple locations increases the possibility of losing some of these files.

3.3.4 F_M: metadata usage

The adequate provision of metadata (e.g. see Digital Curation Centre Curation Reference Manual chapters on metadata [17], preservation metadata [10], archival metadata [20], and learning object metadata [9]) has been a continuing concern within digital curation (e.g. see seminal article by Lavoie [36] and insightful discussions going beyond preservation[27]). The lack of metadata impairs the archive's ability to manage, organise, retrieve and interact with content effectively. It is, widely recognised that it makes understanding the context of the material a challenge.

We will consider metadata on three levels. To avoid the dangers associated with committing to any specific metadata model, we have adopted a general view point shared across many information disciplines (e.g. philosophy, linguistics, computer sciences) based on syntax (e.g. how is it expressed), semantics (e.g. what is it about) and pragmatics (e.g. what can you do with it). There are extensive discussions on metadata classification depending on their application (e.g. see National Information Standards Organization classification [50]; discussion in Digital Curation Centre Curation Reference Manual chapter on Metadata [17]). Here we avoid these fine-grained discussions and focus on the fact that much of the metadata approaches examined in existing literature can be exposed already at the time that websites are created and disseminated.

For example, metadata such as transfer and content encoding can be included by the server in HTTP headers. The required end-user language to understand the content can be

[26] http://www.paradigm.ac.uk/workbook/preservation-strategies/selecting-other.html.

[27] http://www.activearchive.com/content/what-about-metadata.

Table 4 F_M Metadata Facet evaluations

Id	Description	Significance
M_1	Check if the HTTP Content-type header exists. This is significant because it provides information to the web archives about the content and may potentially help retrieve more information	Medium
M_2	Check if any HTTP Caching headers (Expires, Last-modified or ETag) are set. They are important because they provide extra information regarding the creation and last modification of web resources.	Medium
M_3	Check if the HTML meta robots noindex, nofollow, noarchive, nosnippet and noodp tags are used in the markup. If true, they instruct the web archives to avoid archiving the website. This tag is optional and usually omitted	Low
M_4	Check if the DC profile[a] is used in the HTML markup. This evaluation is optional and with low significance. If the DC profile exists, it will help the web archive obtain more information regarding the archived content. If absent, there will be no negative effect	Low
M_5	Check if the FOAF profile[b] is used in the HTML markup. This evaluation is optional and with low significance. If the FOAF profile exists, it will help the web archive obtain more information regarding the archived content. If it does not exist, it will not have any negative effect	Low
M_6	Check if the HTML meta description tag exists in the HTML source code. The meta description tag is optional with low significance. It does not affect web archiving directly but affects the information we have about the archived content	Low

[a] http://dublincore.org/documents/2008/08/04/dc-html/
[b] http://www.foaf-project.org/

indicated as part of the HTML element attribute. Descriptive information (e.g. author, keywords) that can help in understanding how the content is classified can be included in the HTML META element attribute and values. Metadata that support rendering information, such as application and generator names, can also be included in the HTML META element. The use of other well-known metadata and description schemas (e.g. Dublin Core [58]; Friend of a Friend (FOAF) [7]; Resource Description Framework (RDF) [39]) can be included to promote better interoperability. The existence of selected metadata elements can be checked as a way of increasing the probability of implementing automated extraction and refinement of metadata at harvest, ingest, or subsequent stage of repository management. The score for Metadata Usage Facet evaluations are presented in Table 4.

3.4 Attributes

We summarise what website attributes we evaluate to calculate WA. They are also presented in Fig. 2.

RSS The existence of an RSS feed allows the publication of webpage content that can be automatically syndicated or exposed. It allows web crawlers to automatically retrieve updated content, whereas the standardised format of the feeds allows access by many different applications. For example, the BBC uses feeds to let readers see when new content has been added.[28]

Robots.txt The file robots.txt indicates to a web crawler which URLs it is allowed to crawl. The use of robots.txt helps preventing the retrieval of website content that would be aligned with permissions and special rights associated to the webpage.

Sitemaps.xml The Sitemaps protocol, jointly supported by the most widely used search engines to help content creators and search engines, is an increasingly used way to unlock hidden data by making it available to search engines [53]. To implement the Sitemaps protocol, the file sitemap.xml is used to list all the website pages and their locations. The location of this sitemap, if it exists, can be indicated in the robots.txt. Regardless of its inclusion in the robots.txt file, the sitemap, if it exists, should ideally be called 'sitemap.xml' and put at the root of your web server (e.g. http://www.example.co.uk/sitemap.xml).

HTTP Headers HTTP is the protocol used to transfer content from the web server to the web archive. HTTP is very important as it contains a significant information regarding many web content aspects.

Source code and linked web resources The source code of the website (HTML, JavaScript, CSS).

Binary files The binary files included in the webpage (images, pdf, etc.).

Hyperlinks Hyperlinks comprise a net that links the web together. The hyperlinks of the website can be examined for availability as an indication of website accessibility. The lack of hyperlinks does not impact WA but the existence of missing and/or broken links should be considered problematic.

3.5 Evaluations

Combining the information discussed in Sect. 3.3 to calculate a score for WA goes through the following steps.

[28] http://www.bbc.co.uk/news/10628494.

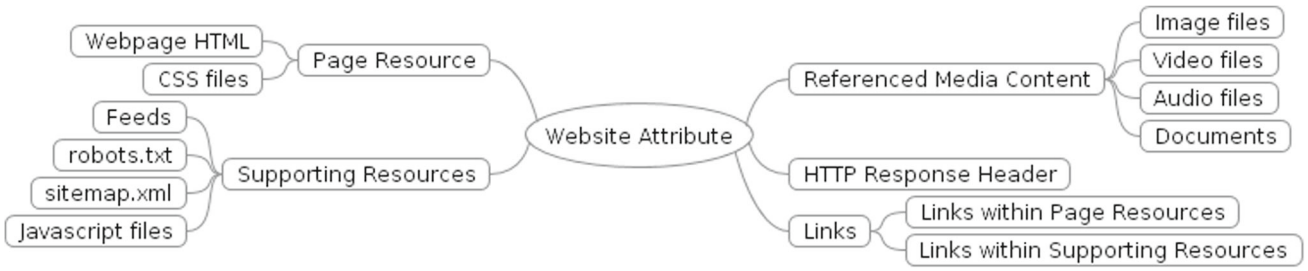

Fig. 2 Website attributes evaluated for WA

1. The WA potential with respect to each facet will be represented by an N-tuple $(x_1, \ldots, x_k, \ldots, x_N)$, where x_k equals a 0 or 1 and represents a negative or positive answer, respectively, to the binary question asked about that facet, whereas N is the total number of questions associated to that facet. For example, an example question in the case of the Standards Compliance Facet would be "I have an object purportedly of format F; is it?" [42]; if there are M files for which format validation is being carried out then there will be M binary questions of this type.
2. Not all questions are considered of equal value to the facet. Depending on their significance (low, medium and high), they have different weights $w_k = (1, 2 \text{ or } 4$, respectively). The weights follow a power law distribution where medium is twice as important as Low and High is twice as important as medium. The value of each facet is the weighted average of its coordinates:

$$F_\lambda = \sum_{k=1}^{N} \frac{\omega_k x_k}{C} \tag{1}$$

where ω_k is the weight assigned to question k and

$$C = \sum_{i=1}^{N} w_i$$

Once the rating with respect to each facet is calculated, the total measure of WA can be simply defined as:

$$\text{WA} = \sum_{\lambda \in \{A,S,C,M\}} w_\lambda F_\lambda \tag{2}$$

where F_A, F_S, F_C, F_M are WA with respect to Accessibility, Standards Compliance, Cohesion, Metadata Usage, respectively, and

$$\sum_{\lambda \in \{A,S,C,M\}} w_\lambda = 1$$

for

$$0 \leq w_\lambda \leq 1 \quad \forall \quad \lambda \in \{A, S, C, M\}$$

Table 5 WA Facet weights

Facet	Weight
F_A	(5*4) + (4*2) + (1*1) = 29
F_S	(2*4) + (8*2) = 24
F_C	6*2 = 12
F_M	(2*2) + (4*1) = 8
Total	73

Depending on the curation and preservation objectives of the web archive, the significance of each facet is likely to be different, and w_λ could be adapted to reflect this. In the simplest model, these w_λ values can be equal, i.e. w_λ=0.25 for any λ. Thus, the WA is calculated as:

$$\text{WA} = \frac{1}{4}F_A + \frac{1}{4}F_S + \frac{1}{4}F_C + \frac{1}{4}F_M \tag{3}$$

We can calculate WA by adopting a normalised model approach, i.e. by multiplying Facet Evaluations by special weights according to their specific questions (of low, medium or high significance). To this end, in Table 5 we calculate the special weights of each facet. Thus, we can evaluate a weighted WA as:

$$\text{WA}_{\text{weighted}} = \frac{29}{73}F_A + \frac{24}{73}F_S + \frac{12}{73}F_C + \frac{8}{73}F_M \tag{4}$$

Actually, accessibility will be the most central consideration in WA since, if the content cannot be found or accessed, then the website's compliance with other standards, and conditions becomes moot. In case the user needs to change the significance of each facet, it is easy to do so by assigning different values to their significance.

3.6 Example

To illustrate the application of CLEAR+, we calculate the WA rating of the website of the Aristotle University of Thessaloniki (AUTH).[29] For each WA Facet, we conduct the neces-

[29] http://www.auth.gr/ as of 10 August 2014.

sary evaluations (Tables 6, 7, 8, 9) and calculate the respective Facet values (see Eqs. 5–8) using Eq. 1.

$$F_A = \frac{\begin{array}{r}(99*4)+(0*4)+(100*4)\\ +(100*4)+(100*4)+(0*2)+(0*2)\\ +(100*2)+(100*2)+(100*1)\end{array}}{(4*5)+(2*4)+(1*1)} \approx 72\,\% \quad (5)$$

$$F_S = \frac{\begin{array}{r}(0*4)+(100*4)+(100*2)+(100*2)+\\ (100*2)+(100*2)+(54*2)+(100*2)\end{array}}{(4*2)+(2*6)} \approx 75\,\% \quad (6)$$

$$F_C = \frac{(87*2)+(90*2)+(100*2)}{3*2} \approx 92\,\% \quad (7)$$

$$F_M = \frac{(100*2)+(100*2)+(100*1)+(100*1)}{(2*2)+(1*2)} = 100\,\% \quad (8)$$

Finally, assuming the flat model approach we calculate the WA value as:

$$\text{WA} = \frac{F_A + F_C + F_S + F_M}{4} \approx 85\,\%$$

whereas, by following the normalised model approach, the weighted WA value is calculated as:

Table 6 F_A evaluation of http://auth.gr/

Id	Description	Rating (%)	Significance
A_1	121 valid and 1 invalid links	99	High
A_2	6 inline JavaScript tags	0	High
A_3	Sitemap file exists http://auth.gr/sitemap.xml	100	High
A_4	Network response time is 100 ms	100	High
A_5	No use of any proprietary file format such as Flash and QuickTime	100	High
A_6	Robots.txt file contains multiple "Disallow" rules. http://auth.gr/robots.txt	0	Medium
A_7	No sitemap.xml reference in the robots.txt file	0	Medium
A_8	16 in 16 images	100	Medium
A_9	HTTP caching headers available	100	Medium
A_{10}	One RSS feed http://auth.gr/rss.xml found using RSS autodiscovery	100	Low

Table 7 F_S evaluation http://auth.gr/

Id	Description	Rating (%)	Significance
S_1	HTML validated, multiple errors	0	High
S_2	No proprietary external objects (Flash, QuickTime)	100	High
S_3	16 well-formed images checked with JHOVE	100	Medium
S_4	RSS feed http://auth.gr/rss.xml is valid according to the W3C feed validator	100	Medium
S_5	Content encoding was clearly defined in HTTP Headers	100	Medium
S_6	HTTP Caching headers clearly defined	100	Medium
S_7	6 valid and 5 invalid CSS	54	Medium
S_8	No HTML5 audio elements	–	Medium
S_9	No HTML5 video elements	–	Medium
S_{10}	Content type. Clearly defined in HTTP Headers	100	Medium

Table 8 F_C evaluation http://auth.gr/

Id	Description	Rating (%)	Significance
C_1	14 local and 2 external images	87	Medium
C_2	10 local and 1 external CSS	90	Medium
C_3	7 local and no external scripts	100	Medium
C_4	No HTML5 audio elements	–	Medium
C_5	No HTML5 video elements	–	Medium
C_6	No proprietary objects	–	Medium

Table 9 F_M evaluation http://auth.gr/

Id	Description	Rating (%)	Significance
M_1	Content type clearly defined in HTTP Headers	100	Medium
M_2	HTTP Caching headers are set	100	Medium
M_3	No meta robots blocking	–	Low
M_4	No DC metadata	–	Low
M_5	FOAF metadata found	100	Low
M_6	HTML description meta tag found	100	Low

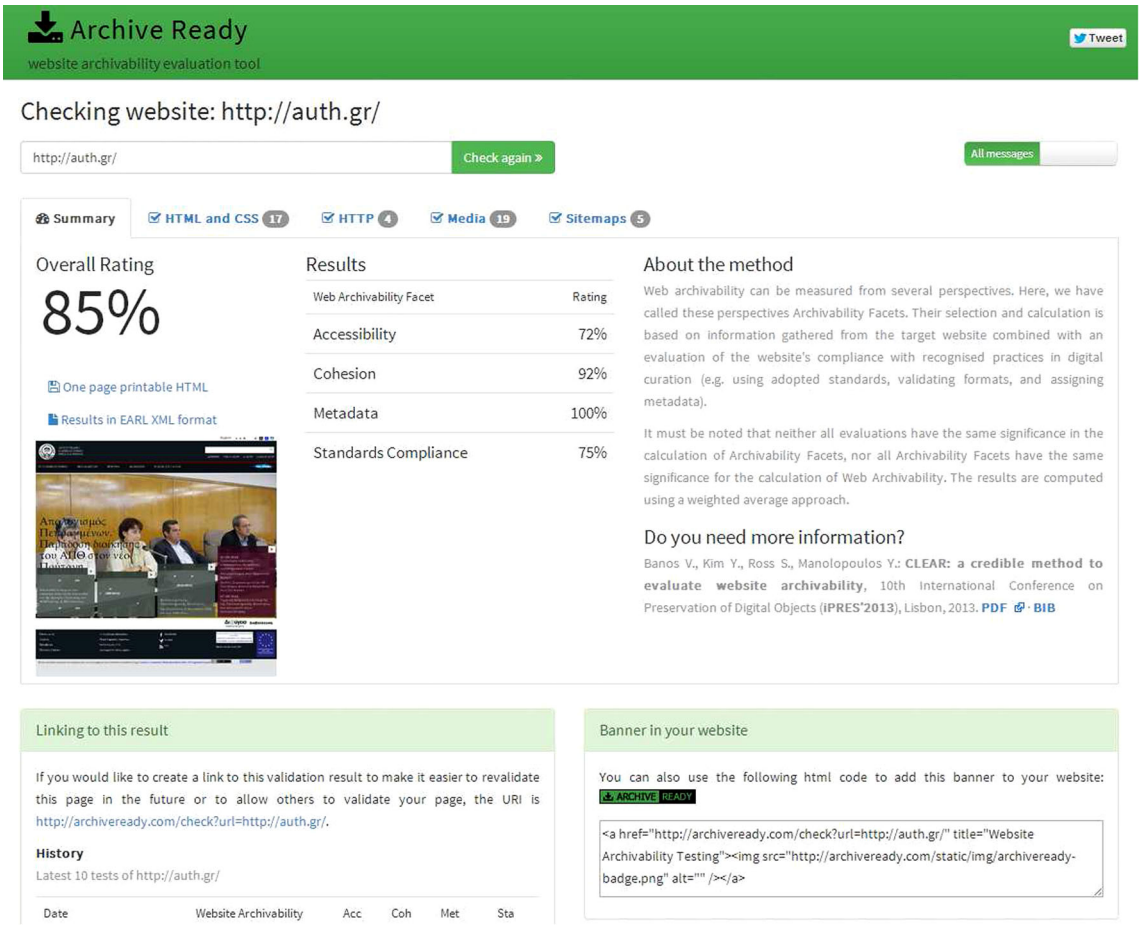

Fig. 3 Evaluating http://auth.gr/ WA using ArchiveReady

$$\text{WA}_{\text{weighted}} = \frac{29}{61}F_A + \frac{20}{61}F_S + \frac{6}{61}F_C + \frac{6}{61}F_M \approx 78\ \%$$

A screenshot of the http://archiveready.com/ web application session we use to evaluate http://auth.gr/ is presented in Fig. 3.

3.7 The evolution from CLEAR to CLEAR+

Finally, we conclude this section with the developments of the method since the first incarnation of CLEAR (Ver.1 of 04/2013) [6]. We experimented in practice with the CLEAR method for a considerable time, running a live online system which is also presented in detail in Sect. 4. We conducted multiple evaluations and received feedback from academics and the web archiving industry professionals. This process resulted in the identification of many issues such as missing evaluations and overestimated or underestimated criteria. The algorithmic and technical improvements of our method can be summarised as follows:

1. Each website attribute evaluation has a different significance, depending on its effect to web archiving, as presented in Sect. 3.3.

2. The Performance Facet has been integrated in the Accessibility and its importance has been downgraded significantly. This is a result of the fact that website performance in tests has been consistently high, regardless of their other characteristics. Thus, Performance Facet rating was always 100 % or near 100 %, distorting the general WA evaluation.
3. Weighted arithmetic mean is implemented to calculate WA Facets instead of simple mean. All evaluations have been assigned a low, medium or high significance indicator, which affects the calculation of all WA Facets. The significance has been defined based on the initial experience with WA evaluations from the first year of archiveready.com operation.
4. Certain evaluations have been removed from the method as they were considered irrelevant. For example, the check that archived versions of the target website are present in the Internet Archive or not should be part of the assessment.
5. On a technical level, all aspects of the reference implementation of the Website Archivability Evaluation Tool http://archiveready.com have been improved. The software has also the new capability of analysing dynamic websites using a headless web browser, as presented in Sect. 4. Thus, its operation has become more accurate and valid than the previous version.

In the following, we present the reference implementation of the CLEAR+ method.

4 System architecture

Here, we present ArchiveReady,[30] a WA evaluation system that implements CLEAR+ as a web application. We describe the system architecture, design decisions, WA evaluation workflow and Application Programming Interfaces (APIs) available for interoperability purposes.

4.1 System

ArchiveReady is a web application based on the following key components:

1. Debian linux[31] operating system for development and production servers,
2. Nginx web server[32] to server static web content,
3. Python programming language,[33]
4. Gunicorn Python WSGI HTTP Server for unix[34] to server dynamic content,
5. BeautifulSoup[35] to analyse HTML markup and locate elements,
6. Flask,[36] a Python microframework to develop web applications,
7. Redis advanced key-value store[37] to manage job queues and temporary data,
8. Mariadb Mysql RDBMS[38] to store long-term data.
9. PhantomJS,[39] a headless WebKit scriptable with a JavaScript API. It has fast and native support for various web standards: DOM handling, CSS selector, JSON, Canvas, and SVG.
10. JSTOR/Harvard Object Validation Environment (JHOVE) [19] for media file validation,
11. JavaScript and CSS libraries such as jQuery[40] and Bootstrap[41] are utilised to create a compelling user interface,
12. W3C HTML Markup Validation Service[42] and CSS Validation Service[43] APIs for web resources evaluation.

An overview of the system architecture is presented in Fig. 4. During the design and implementation of the platform, we took some important decisions, which influenced greatly all aspects of development.

We choose Python to implement ArchiveReady since it is ideal for rapid application development and has many modern features. Moreover, it is supported by a big user community and has a wide range of modules. Using these assets, we were able to successfully implement many important features such as RSS feed validation (feedvalidator module), XML parsing, validation and analysis (lxml module), HTTP communication (python-requests module) and asynchronous job queues (python-rq module).

We use PhantomJS to access websites which use Javascipt, AJAX and other web technologies, which are difficult to handle with HTML processing. Using PhantomJS, we can perform JavaScript rendering when processing website. Therefore, we can extract dynamic content and even support AJAX-generated content in addition to traditional HTML-only websites.

[30] http://www.archiveready.com.
[31] http://www.debian.org.
[32] http://www.nginx.org.
[33] http://www.python.org/.
[34] http://gunicorn.org/.
[35] http://www.crummy.com/software/BeautifulSoup/.
[36] http://flask.pocoo.org/.
[37] http://redis.io.
[38] http://www.mariadb.com.
[39] http://phantomjs.org/.
[40] http://www.jquery.com.
[41] http://twitter.github.com/bootstrap/.
[42] http://validator.w3.org/.
[43] http://jigsaw.w3.org/css-validator/.

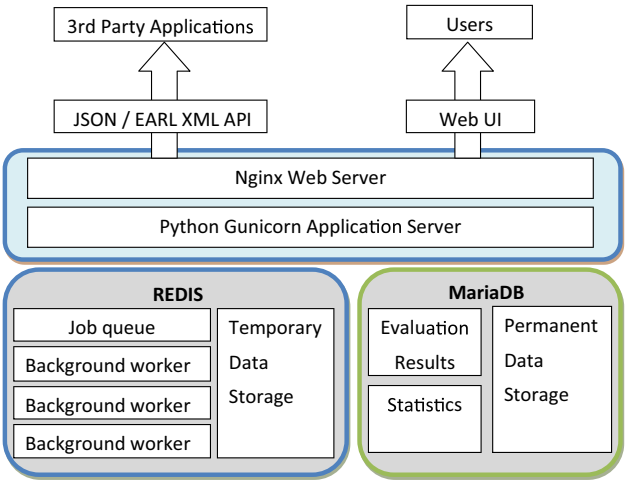

Fig. 4 The architecture of the archiveready.com system

We select Redis to store temporary data into memory because of its performance and its ability to support many data structures. Redis is an advanced key-value store, since keys can contain strings, hashes, lists, sets and sorted sets. These features make it ideal for holding volatile information, such as intermediate evaluation results and other temporary data about website evaluations. Redis is also critical for the implementation of asynchronous job queues as described in Sect. 4.1.1.

We use MariaDB to store data permanently for all evaluations. Such data are the final evaluation results and user preferences. We use JHOVE [19], an established digital preservation tool to evaluate the files included in websites for their correctness. We evaluate HTML markup, CSS and RSS correctness using W3C validator tools. We also use Python exceptions to track problems when analysing webpages and try to locate webpages, which cause problems to web client software.

4.1.1 Scalability

One of the greatest challenges in implementing ArchiveReady is performance, scalability and responsiveness. A web service must be able to evaluate multiple websites in parallel, while maintaining a responsive Web UI and API. To achieve this goal, we implement asynchronous job queues in the following manner:

1. ArchiveReady tasks are separated into two groups: real time and asynchronous. Real-time commands are processed as soon as they are received by the user as in any common web application.
2. Asynchronous tasks are processed in a different way. When a user or third party application initiates a new evaluation task, the web application server maps the task into multiple individual atomic subtasks, which are inserted in the asynchronous job queue of the system, which is stored in a Redis List.
3. Background workers equal to the number of server CPU cores are constantly monitoring the job queue for new tasks. As soon as they identify them, they begin processing them one by one and store the results in MariaDB database.
4. When all subtasks of a given task are finished, the web application server process is notified to present the results to the user. While the background processes are working, the application server is free to reply to any requests regarding new website evaluations without any delay.

The presented evaluation processing logic has many important benefits. Tasks are separated into multiple individual atomic evaluations. This makes the system very robust. An exception or any other system error in any individual evaluation does not interfere with the general system operation. More important is the fact that the platform is highly scalable as it is possible for the asynchronous job queues to scale not only vertically depending on the number of available server CPU cores, but also horizontally, as multiple servers can be configured to share the same asynchronous job queue and mysql database.

To ensure high-level compatibility with W3C standards the initiative used open source web services provided by the W3C. These include: the Markup Validator,[44] the Feed Validation Service[45] and the CSS Validation Service.[46]

According to the HTTP Archive Trends, the average number of HTTP requests initiated when accessing a webpage is over 90 and is expected to rise.[47] In response to this performance context, ArchiveReady has to be capable of performing a very large number of HTTP requests, process the data and present the outcomes to the user in real time. This is not possible with a single process for each user, the typical approach in web applications. To resolve this blocking issue, an asynchronous job queue system based on Redis for queue management and the Python RQ library[48] was deployed. This approach enables the parallel execution of multiple evaluation processes, resulting in huge performance benefits when compared to traditional web application execution model.

Its operation can be summarised as follows:

1. As soon as a user submits a website for evaluation, the master process maps the work into multiple individual jobs, which are inserted in the parallel job queues in the background.

[44] http://validator.w3.org/.

[45] http://validator.w3.org/feed/.

[46] http://jigsaw.w3.org/css-validator/.

[47] http://httparchive.org/trends.php.

[48] http://python-rq.org/.

2. Background worker processes are notified and begin processing the individual jobs in parallel. The level of parallelism is configurable, 16 parallel processes are the current setup.
3. As soon as a job is finished, the results are sent to the master process.
4. When all jobs are finished, the master process calculates the WA and presents the final results to the user.

Using the presented approach, we are able to practically eliminate the evaluation time.

4.2 Workflow

ArchiveReady is a web application providing two types of interaction: web interface and web service. With the exception of presentation of outcomes (HTML for the former and JSON for the latter) both are identical. The evaluation workflow of a target website can be summarised as follows:

1. ArchiveReady receives a target URL and performs an HTTP request to retrieve the webpage hypertext.
2. After analysing it, multiple HTTP connections are initiated in parallel to retrieve all web resources referenced in the target webpage, imitating a web crawler. ArchiveReady analyses only the URL submitted by the user, it does not evaluate the whole website recursively, as we have proven that the WA analysis of a single website page is a good proxy of the whole website WA rating.
3. In stage 3, Website Attributes (see Sect. 3.4) are evaluated. In more detail:
   (a) HTML and CSS analysis and validation.
   (b) HTTP response headers analysis and validation.
   (c) Media files (images, video, audio, other objects) retrieval, analysis and validation.
   (d) Sitemap.xml and Robots.txt retrieval, analysis and validation.
   (e) RSS feeds detection, retrieval, analysis and validation.
   (f) Website Performance evaluation. The sum of all network transfer activity is recorded by the system and in the end, after the completion of all network transfers, the average transfer time is calculated. There are fast and slow evaluations; fast are performed instantly at the application server, whereas slow evaluations are performed asynchronously using a job queue as presented in Sect. 4.1.1.
4. The metrics for the WA Facets are calculated according to the CLEAR+ method and the final WA rating is calculated.

Note that in the current implementation, CLEAR+ evaluates only a single webpage based on the assumption that all its webpages share the same components, standards and technologies. This is validated in Sect. 5.4.

In addition to the CLEAR+ method application, ArchiveReady is performing some additional procedures, which are practical and may provide useful insight to users.

1. ArchiveReady checks the Internet Archive to identify if the target website is already archived there and provides a link to the specific webpage.
2. It generates a WARC file [29] and makes it available for download. Thus, the user is able to see how his/her website would be when encoded in WARC, which is the most common web archive storage format.

4.3 Interoperability and APIs

ArchiveReady is operating not only as a web application for users visiting the website archiveready.com/ but also as a web service, which is available for integration into third party applications. Its interface is quite simple; by accessing archiveready.com/api?url=http://auth.gr/ via HTTP and a JSON document will be retrieved with the full results of the WA evaluation results on the target URL as presented in Listing 1.

```
{"test":{
    "website_archivability": 91,
    "Metadata":100
    "Standards_Compliance":73,
    "Accessibility":88,
    "Cohesion":71,
  },
  "url": "http://auth.gr/",
  "messages":[
    {"title":"Invalid CSS http://
       dididownload.com/wp-content
       /themes/didinew/style.css.
       Located 8 errors, 78
       warnings.",
    "attribute":"html",
    "facets":["
       Standards_Compliance"],
    "level":0,
    "significance":"LOW",
    "message":"Webpages which do
       not conform with Web
       Standards have a lower
       possibility to be preserved
        correctly",
    "ref":"http://jigsaw.w3.org/
       css-validator/validator?uri
       =http://dididownload.com/wp
       -content/themes/didinew/
       style.css&warning=0&profile
       =css3"},
    ....
  ]
}
```

Listing 1 ArchiveReady API JSON output

The JSON output can be easily used by third party programs. In fact, all evaluations in Sect. 5 were conducted this way.

Another significant interoperability feature of the archiveready.com platform is to output Evaluation and Report Language (EARL) XML [1], which is the W3C standard for expressing test results. EARL XML enables users to assert WA evaluation results for any website in a flexible way.

5 Evaluation

5.1 Methodology and limits

Our evaluation has two aims. The first is to prove the validity of the WA metric by experimenting on assorted datasets and by expert evaluation. The second is to validate our claim that it is only necessary to evaluate a single webpage from a website to calculate a good approximation of its WA value.

In our experiments, we use Debian GNU/Linux 7.3, Python 2.7.6 and an Intel Core i7-3820, 3.60 GHz processor. The Git repository for this work[49] contains the necessary data, scripts and instructions to reproduce all the evaluation experiments presented here.

WA is a new concept and even though our method has solid foundations, there are still open issues regarding the evaluation of all WA Facets and the definition of a dataset of websites to be used as a Gold Standard:

1. The presented situation regarding standards compliance raises issues regarding the accuracy of the Accessibility Facet (F_A) evaluation. Web crawlers try to mitigate the errors they encounter in web resources with various levels of success, affecting their capability to access all website content. Their success depends on the sophistication of their error mitigation algorithms. On the contrary, the F_A rating of websites having such errors will be definitely low. For instance, a web crawler may access a sitemap.xml which contains invalid XML. If it uses a strict XML parser, it will fail to parse it and retrieve its URLs to proceed with web crawling. On the contrary, if it uses a relaxed XML parser, it will be able to retrieve a large number of its URLs and it will access more website content. In any case, the F_A rating will suffer.
2. The tools we have at our disposal are limited and cannot cope with the latest developments on the web. For instance, web browser vendors are free to implement extensions to the CSS specifications that, in most cases, are proprietary to their browser.[50] The official W3C CSS Standard[51] is evolving to include some of these new extensions but the process has an inherent delay. As a result, the state of the art W3C CSS validator we use in our system to validate target website CSS may return false-negative results. This problem is also apparent in all W3C standards validators. As a result, Standards Compliance (F_S) evaluation is not always accurate. It must be noted though that W3C validators are improving on a steady rate and any improvement would be utilised automatically by our system as we are using the W3C validators as web services. Another aspect of this issue is that experts evaluating the live as well as the archived version of a website depend mainly on their web browsers to evaluate the website quality using mostly visual information. The problem is that HTML documents which are not following W3C standards may appear correctly to the viewer even if they contain serious errors because the web browser is operating in "Quirks Mode" [11] and has particular algorithms to mitigate such problems. Thus, a website may appear correctly in a current browser but it may not do the same in a future browser because the error mitigation algorithms are not standard and depend on the web browser vendor. As a result, it is possible that experts evaluating a website may report that it has been archived correctly but the F_S evaluation results may not be equally good.
3. The Cohesion (F_C) of a website does not directly affect its archiving unless one or more servers hosting its resources become unreachable during the time of archiving. The possibility of encountering such a case when running a WA experiment is very low. Thus, it is very difficult to measure it in an automated way.
4. Metadata (F_M) are a major concert for digital curation, as discussed in Sect. 3.3.4. Nevertheless, the lack of metadata in a web archive does not have any direct impact on the user; archived websites may appear correctly although some of their resources may lack correct metadata. This deficiency may become significant in the future, when the web archivists would need to render or process some "legacy" web resources and they would not have the correct information to do so. Thus, it is also challenging to evaluate this Facet automatically.
5. The granularity of specific evaluations could be improved in the future to improve the accuracy of the method. Currently, the evaluations can be grouped based on their output score into binary (100 %/0 stands for successful/-failed evaluation) and relative percentage evaluations (for instance, if 9 out of 10 hyperlinks are valid, the relevant evaluation score is 90 %). There are some binary evaluations though which may be defined better as a relative percentage. For example, we have A_2: Check if inline

[49] https://github.com/vbanos/web-archivability-journal-paper-data-2014.

[50] http://reference.sitepoint.com/css/vendorspecific.

[51] http://www.w3.org/Style/CSS/.

JavaScript code exists in HTML. We are certain that inline JavaScript code is causing problems to web archiving so we assign a 100 % score if no inline JavaScript code is present and 0 % in the opposite case. Ideally, we should assign a relative percentage score based on multiple parameters, such as the specific number of inline JavaScript files, filesizes, type of inline code, complexity and other JavaScript-specific details. The same also applies for many evaluations such as S_1: HTML standards compliance, S_4: RSS feed standards compliance, S_7: CSS standards compliance and A_6: Robots.txt "Disallow:" rules.

With these concerns in mind, we consider several possible methods to perform the evaluation. First, we could survey domain experts. We could ask web archivists working in IIPC Member Organisations to judge websites. However, this method is impractical because we would need to spend significant time and resources to evaluate a considerable number of websites. A second alternative method would be to devise datasets based on thematic or domain classifications. For instance, websites of similar organisations from around the world. A third alternative would be to perform manual checking of the way a set of websites is archived in a web archive and evaluate all their data, attributes and behaviours in comparison with the original website. We choose to implement both the second and the third method.

5.2 Experimentation with assorted datasets

To study WA with real-world data, we conduct an experiment to see if high-quality websites, according to some general standards, have better WA than low-quality websites. We devise a number of assorted datasets with websites of varying themes, as presented in Table 10. We evaluate their WA using the `ArchiveReady.com` API (Sect. 4.3) and finally, we analyse the results.

We define three datasets of websites (D_1, D_2, D_3) with certain characteristics: (a) they belong to important educational, government or scientific organisations from all around the world. (b) They are developed and maintained by dedicated personnel and/or special IT companies. (c) They are used by a large number of people and are considered very important for the operation of the organisation they belong to. We also choose to create a dataset (D_4) of manually selected spam websites which have the following characteristics: (a) they are created automatically by website generators in large numbers. (b) Their content is generated automatically. (c) They are neither maintained nor evaluated for their quality at all. (d) They have relatively very few visitors.

It is important to highlight that a number of websites from all these datasets could not be evaluated by our system for various technical reasons. This means that these websites may also pose the same problems to web archiving systems. The reasons for these complications may be one or more of the following:

- The websites do not support web crawlers and deny sending content to them. This may be due to security settings or technical incompetence. In any case, web archives would not be able to archive these websites.
- The websites were not available during the evaluation time.
- The websites returned some kind of problematic data which resulted in the abnormal termination of the `ArchiveReady` API during the evaluation.

It is worth mentioning that D_4, the list of manually selected spam had the most problematic websites: 42 out of 120 could not be evaluated at all. In comparison, 8 out of 94 IIPC websites could not be evaluated (D_1), 13 out of 200 (D_2) and 16 out of 450 (D_3).

We conduct the WA evaluation using a python script, which uses the `ArchiveReady.com` API and record the outcomes in a file. We calculate the results of the WA distribution for all four datasets and present them in Fig. 5. Also, we calculate the average, median, min, max and standard deviation functions on these datasets and present the results in Table 11 and depict them in Fig. 6 using boxplots.

Table 10 Description of assorted datasets

Id	Description	Raw data	Clean data
D_1	A set of websites from a pool of international web standards organisations, national libraries, IIPC members and other high-profile organisations in these fields	94	86
D_2	The first 200 of the top universities according to the Academic Ranking of World Universities [37], also known as the "Shanghai list"	200	187
D_3	A list of government organisation websites from around the world	450	434
D_4	A list of manually selected spam websites from the top 1 million websites published by Alexa	120	78

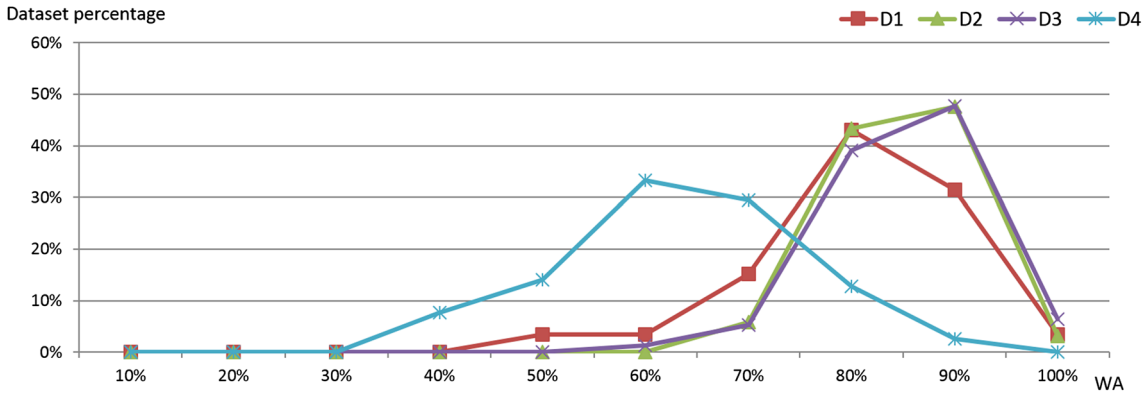

Fig. 5 WA distribution for assorted datasets

Table 11 Comparison of WA statistics for assorted datasets

Function	D_1	D_2	D_3	D_4
Average (WA)	75.87	80.08	80.75	58.37
Median (WA)	77.5	81	81	58.75
Min (WA)	41.75	56	54	33.25
Max (WA)	93.25	96	96	84.25
StDev (WA)	10.16	6.11	7.06	11.63

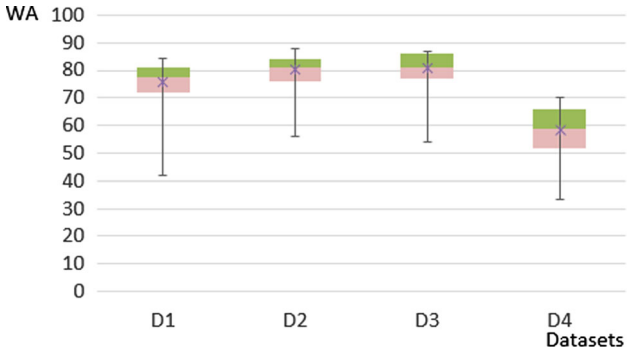

Fig. 6 WA statistics for assorted datasets box plot

From these results we can observe the following. In the websites of datasets D_1, D_2 and D_3, which are considered high quality, the distribution of WA values is leaning towards high values as illustrated in Fig. 6. This is also evident from the statistics presented in Table 11. The average WA values are 75.87, 80.08 and 80.75. The median WA values are also similar. On the contrary, D_4 websites, which are characterised as low quality, have remarkably lower WA values as shown in Table 11 and in Fig. 6. The average WA value is 58.37 and the median value is 58.75. Thus, lower quality websites are prone to issues, which make them difficult to be archived. Finally, the standard deviation values are in all cases quite low. As the WA range is [0...100], standard deviation values of approximately 10 or less indicate that our results are strongly consistent, for both lower and higher WA values.

To conclude, this experiment indicates that higher quality websites have higher WA than lower quality websites. This outcome is confirmed not only by the WA score itself but also by another indicator which was revealed during the experiment, the percentage of completed WA evaluations for each data set.

5.3 Evaluation by experts

To evaluate the validity of our metrics, a reference standard has to be employed for the evaluation. It is important to note that this task requires careful and thorough investigation, as it has been already elaborated in existing works [31,56]. With the contribution of three post-doc researchers and PhD candidates in informatics from the Delab laboratory[52] of the Department of Informatics at Aristotle University who assist us as experts, we conduct the following experiment.

We use the first 200 websites of the top universities according to the Academic Ranking of World Universities of 2013 as a dataset (D_2 from Sect. 5.2). We review the way that they are archived in the Internet Archive and rank their web archiving with a scale of 0 to 10. We select to use the Internet Archive because it is the most popular web archiving service, to the best of our knowledge.

More specifically, for each website we conduct the following evaluation:

1. We visit http://archive.org, enter the URL on the Wayback Machine and open the latest snapshot of the website.
2. We visit the original website.
3. We evaluate the two instances of the website and assign a score from 0 to 10 depending on the following criteria:

[52] http://delab.csd.auth.gr/.

Table 12 Correlation between WA, WA Facets and Experts rating

	F_A	F_C	F_M	F_S	WA	Exp.
F_A	1.000					
F_C	0.060	1.000				
F_M	0.217	−0.096	1.000			
F_S	0.069	0.060	0.019	1.000		
WA	0.652	0.398	0.582	0.514	1.000	
Exp.	0.384	0.263	0.282	0.179	0.516	1.000

(a) Compare the views of the homepage and try to find visual differences and things missing in the archived version (3.33 points).
(b) Inspect dynamic menus or other moving elements in the archived version (3.33 points).
(c) Visit random website hyperlinks to evaluate if they are also captured successfully (3.33 points).

After analysing all websites, we conduct WA evaluation for the same websites with a Python script which is using the archiveready.com API (Sect. 4.3). We record the outcomes in a file and calculate the Pearson's correlation coefficient for WA, WA Facets and expert scores. We present the results in Table 12.

From these results, we observe that the correlation between WA and Experts rating is 0.516, which is quite significant taken into consideration the discussion about the limits presented in Sect. 5.1. It is also important to highlight the lack of correlation between different WA Facets. The correlation indicators between $F_A - F_C$, $F_A - F_S$, $F_C - F_M$, $F_C - F_S$ and $F_M - F_S$ are very close to zero, ranging from −0.096 to 0.069. There is only a very small correlation in the case of $F_A - F_M$, 0.217. Practically, there is no correlation between different WA Facets, confirming the validity and strength of the CLEAR+ method. WA Facets are different perspectives of WA, if there was any correlation of the WA Facets, this would mean that their differences would not be so significant. This experiment confirms that WA Facets are totally independent.

Finally, we conduct One-Way Analysis of Variance (ANOVA) [38], to calculate the F value = 397.628 and the P value = 2.191e−54. These indicators are very positive and show that our results are statistically significant.

5.4 WA variance in the same website

We argue that the CLEAR+ method needs only to evaluate the WA value of a single webpage based on the assumption that webpages from the same website share the same components, standards and technologies. We also claim that the website homepage has a representative WA score. This is important because it would be common for the users of the CLEAR+ method to evaluate the homepage of a website and we have to confirm that it has a representative WA value. Following, we conduct the following experiment:

1. We use the Alexa top 1 million websites dataset[53] and we select 1000 random websites.
2. We retrieve 10 random webpages from each website to use as a test sample. To this end, we decided to use their RSS feeds.
3. We perform RSS feeds auto-detection and we finally identify 783 websites which are suited for our experiment.
4. We evaluate the WA for 10 individual webpages for each website and record the results in a file.
5. We calculate the WA average ($WA_{average}$) and standard deviation ($StDev(WA_{average})$) for each website.
6. We calculate and store the WA of the homepage for each website ($WA_{homepage}$) as an extra variable.

We plot the variables $WA_{average}$, $StDev(WA_{average})$ and $WA_{homepage}$ for each website in a descending order by $WA_{average}$ in Fig 7. The x-axis represents each evaluated website, whereas the y-axis represents WA. The red cross (+) markers which appear in a seemingly continuous line starting from the top left and ending at the centre right of the diagram represent the $WA_{average}$ values for each website. The blue star (*) markers which appear around the red markers represent the $WA_{homepage}$ values. The green square (■) markers at the bottom of the diagram represent $StDev(WA_{average})$. From the outcomes of our evaluation we draw the following conclusions:

1. While average WA for the webpages of the same website may vary significantly from 50 to 100 %, the WA standard deviation does not behave in the same manner. The WA standard deviation is extremely low. More specifically, its average is 0.964 points in the 0–100 WA scale and its median is 0.5. Its maximum value is 13.69 but this is an outlier; the second biggest value is 6.88. This means that WA values are consistent for webpages of the same website.
2. The WA standard deviation for webpages of the same website does not depend on average WA of the website. As depicted in Fig. 7, regardless of the $WA_{average}$ value, $StDev(WA_{average})$ value remains very low.
3. The WA of the homepage is near the average WA for most websites. Figure 7 indicates the $WA_{homepage}$ values are always around $WA_{average}$ values with very few outliers. The average absolute difference between $WA_{average}$

[53] http://s3.amazonaws.com/alexa-static/top-1m.csv.zip.

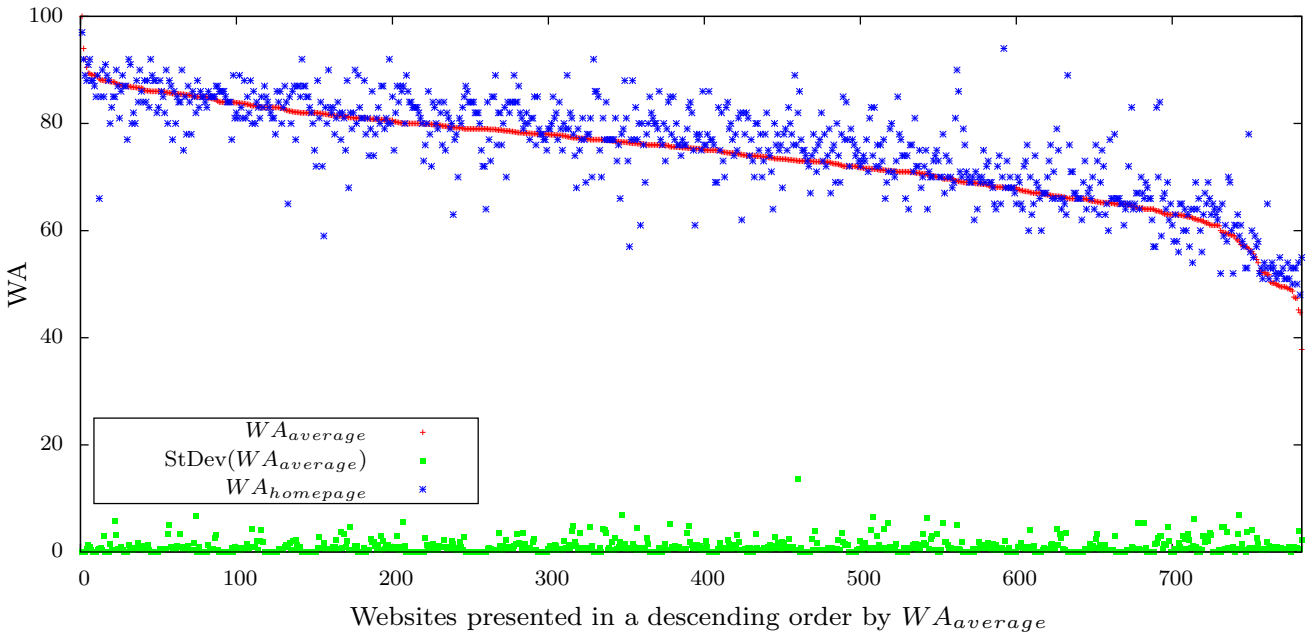

Fig. 7 WA average rating and standard deviation values, as well as the homepage WA for a set of 783 random websites

and $WA_{homepage}$ for all websites is 3.87 and its standard deviation is 3.76. The minimum value is obviously 0 and the maximum is 25.9.

4. Although $WA_{homepage}$ is near $WA_{average}$, we observed that its value is usually higher. Out of the 783 websites, in 510 cases $WA_{homepage}$ is higher, in 35 is it exactly equal and in 238 it is lower that $WA_{average}$. Even though the difference is quite small, it is notable.

Our conclusion is that our initial assumptions are valid, the variance of WA for the webpages of the same website is remarkably small. Moreover, the homepage WA is quite similar to the average, with a small bias towards higher WA values, which is quite interesting. A valid explanation regarding this phenomenon is that website owners spend more resources on the homepage than any other page because it is the most visited part of the website. Overall, we can confirm that it is justified to evaluate WA using only the website homepage.

6 Conclusions and future work

In this article, we presented our extended work towards the foundation of a quantitative method to evaluate WA. The Credible Live Evaluation of Archive Readiness Plus (CLEAR+) method to evaluate Website Archivability has been elaborated in great detail, the key Facets of WA have been defined and the method of their calculating has been explained in theory and practice.

In addition, we presented the ArchiveReady system, which is the reference implementation of CLEAR+. We overviewed all aspects of the system, including design decisions, technologies, workflows and interoperability APIs. We believe that it is quite important to explain how the reference implementation of CLEAR+ works because transparency raises the confidence for the method.

A critical part of this work is also the experimental evaluation. First, we performed experimental WA evaluations of assorted datasets and observed the behaviour of our metrics. Then, we conducted a manual characterisation of websites to create a reference standard and we identified correlations with WA. Both evaluations provided very positive results, which support that the CLEAR+ can be used to identify whether a website has the potential to be archived with correctness and accuracy. We also experimentally proved that CLEAR+ method needs only to evaluate a single webpage to calculate the WA of a website, based on the assumption that webpages from the same website share the same components, standards and technologies.

CLEAR+ is an improvement over CLEAR for many reasons, as presented in detail in Sect. 3.7. The evaluations of the website attributes have been overhauled, resulting in the identification of many issues, such as missing evaluations and overestimated or underestimated criteria. Important improvements include: (a) the removal of irrelevant evaluations, such as checking the presence of a website in the Internet Archive, (b) the addition of evaluations such as the check for "Disallow:" instructions in robots.txt, and (c) the improvement of methods, such as the extraction of content from dynamic websites using a headless web browser software component.

Introducing a new metric to quantify the previously unquantifiable notion of WA is not an easy task. We believe that we have captured the core aspects of a website crucial in diagnosing whether it has the potential to be archived with correctness and accuracy with the CLEAR+ method and the WA metric. Our future efforts will be towards three directions: (a) further CLEAR+ method development, (b) dissemination to larger web-related audiences, and, (c) exploring application in web archiving and web development.

The development of the CLEAR+ method will continue to overcome the method limitations we presented in Sect. 5.1. We will also try to extend the evaluations over more website attributes, such as catching JavaScript execution and maybe also automated interaction with the page (random clicking, scrolling down, etc).

Besides method development, it is also critical to communicate the notion of WA and the method to evaluate it in larger web-related audiences, where we hope it will have important impact. We also plan to explore applications of the method in web archiving, web development and related education activities. Towards this direction, we are planning to implement plug-ins for popular CMS to enable web professionals to integrate WA evaluations in their systems.

Acknowledgments We would like to thank our colleagues: Panagiotis Symeonidis, Georgia Latsiou and Konstantinos Mokos, for their assistance in Sect. 5.3, Evaluation by experts. We would like to thank the anonymous reviewers for their valuable input, which helped us to significantly improve this manuscript. In particular, their feedback was critical to improve Sect. 3 on the CLEAR+ method and Sect. 5 on the experimental evaluation.

References

1. Abou-Zahra, S., Squillace, M.: Evaluation and report language (earl) 1.0 schema. http://www.w3.org/TR/EARL10-Schema/ (2006). Accessed 22 Dec 2014
2. Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How much of the web is archived? In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, pp. 133–136. ACM (2011)
3. Arms, C., Fleischhauer, C., Murray, K.: Sustainability of digital formats planning for Library of Congress collections: external dependencies. http://www.digitalpreservation.gov/formats/sustain/sustain.shtml#external (2013). Accessed 22 Dec 2014
4. Avižienis, A., Laprie, J.C., Randell, B.: Fundamental concepts of computer system dependability. In: Proceedings of the IARP/IEEE-RAS Workshop on Robot Dependability: Technological Challenge of Dependable, Robots in Human Environments (2001)
5. Banos, V., Baltas, N., Manolopoulos, Y.: Trends in blog preservation. In: Proceedings of the 14th International Conference on Enterprise Information Systems (ICEIS). Wroclaw, Poland (2012)
6. Banos, V., Kim, Y., Ross, S., Manolopoulos, Y.: CLEAR: a credible method to evaluate website archivability. In: Proceedings of the 10th International Conference on Preservation of Digital Objects (IPRES). Lisbon, Portugal (2013)
7. Brickley, D., Miller, L.: FOAF vocabulary specification 0.98. Namespace Document 9 (2010)
8. Brunelle, J.F., Kelly, M., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: Not all mementos are created equal: Measuring the impact of missing resources. In: 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL), pp. 321–330. IEEE (2014)
9. Campbell, L.: Learning object metadata, curation reference manual. http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/learning-object-metadata (2007). Accessed 22 Dec 2014
10. Caplan, P.: Preservation metadata, curation reference manual. http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/preservation-metadata (2006). Accessed 22 Dec 2014
11. Center, M.D.: Mozilla's quirks mode. 2007 (2008)
12. Charron, C., Favier, J., Li, C., Joseph, J., Neurauter, M., Cohen, S., McHarg, T., Kolko, J.: Social computing: how networks erode institutional power, and what to do about it. Forrester Customer Report (2006)
13. Clausen, L.: Concerning etags and datestamps. In: 4th International Web Archiving Workshop (IWAW04). Citeseer (2004)
14. Coalition, D.P.: Institutional strategies—standards and best practice guidelines. http://www.dpconline.org/advice/preservationhandbook/institutional-strategies/standards-and-best-practice-guidelines (2012). Accessed 22 Dec 2014
15. Crane, G.: Designing documents to enhance the performance of digital libraries. Time, space, people and a digital library on London. D-Lib Mag. **6**(7/8) (2000)
16. Daskalantonakis, M.: A practical view of software measurement and implementation experiences within motorola. IEEE Trans. Softw. Eng. **18**(11), 998–1010 (1992)
17. Day, M.: Metadata, curation reference manual. http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/metadata (2005). Accessed 22 Dec 2014
18. Denev, D., Mazeika, A., Spaniol, M., Weikum, G.: The SHARC framework for data quality in web archiving. VLDB J. **20**(2), 183–207 (2011)
19. Donnelly, M.: JSTOR/Harvard Object Validation Environment (JHOVE). Digital Curation Centre Case Studies and Interviews (2006)
20. Duff, W., van Ballegooie, M.: Archival metadata, curation reference manual. http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/archival-metadata (2006). Accessed 22 Dec 2014
21. Faheem, M., Senellart, P.: Intelligent and adaptive crawling of web applications for web archiving. In: Proceedings of the 21st International Conference Companion on World Wide Web (WWW), pp. 127–132. Lyon, France (2012)
22. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext transfer protocol-http/1.1. http://tools.ietf.org/html/rfc2616 (1999). Accessed 22 Dec 2014
23. Freire, A.P., Bittar, T.J., Fortes, R.P.: An approach based on metrics for monitoring web accessibility in Brazilian municipalities web sites. In: Proceedings of the 23rd ACM Symposium on Applied Computing (SAC), pp. 2421–2425. Fortaleza, Brazil (2008)
24. Glenn, V.D.: Preserving government and political information: the web-at-risk project. First Monday **12**(7) (2007)
25. Gomes, D., Silva, M.J.: Modelling information persistence on the web. In: Proceedings of the 6th International Conference on Web Engineering, pp. 193–200. ACM (2006)
26. Gray, G., Martin, S.: Choosing a sustainable web archiving method: a comparison of capture quality. D-Lib Mag. **19**(5), 2 (2013)
27. He, Y., Xin, D., Ganti, V., Rajaraman, S., Shah, N.: Crawling deep web entity pages. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), pp. 355–364. Rome, Italy (2013)

28. Hockx-Yu, H., Crawford, L., Coram, R., Johnson, S.: Capturing and replaying streaming media in a web archive—a British Library case study. In: Proceedings of the 7th International Conference on Preservation of Digital Objects (iPres). Vienna, Austria (2010)
29. ISO: 28500: 2009 information and documentation-WARC file format. International Organization for Standardization (2009)
30. Kasioumis, N., Banos, V., Kalb, H.: Towards building a blog preservation platform. World Wide Web **17**(4), 799–825 (2013)
31. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. In: Foundations and Trends in Information Retrieval, vol. 3. Now Publishers Inc., Hanover (2009)
32. Kelly, M., Brunelle, J.F., Weigle, M.C., Nelson, M.L.: On the change in archivability of websites over time. In: Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 35–47. Valletta, Malta (2013)
33. Kelly, M., Nelson, M.L., Weigle, M.C.: The archival acid test: evaluating archive performance on advanced html and javascript. In: 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL), pp. 25–28. IEEE (2014)
34. Kenney, A.R., McGovern, N., Botticelli, P., Entlich, R., Lagoze, C., Payette, S.: Preservation risk management for web resources. D-Lib Mag **8**(1) (2002)
35. de Kunder, M.: Geschatte grootte van het geïndexeerde world wide web. Tilburg University, p. 63 (2008)
36. Lavoie, B.F.: Implementing metadata in digital preservation systems: the premis activity. D-Lib Mag. **10**(4) (2004)
37. Liu, N.C., Cheng, Y.: The academic ranking of world universities. High. Educ. Eur. **30**(2), 127–136 (2005)
38. Lowry, R.: Concepts and Applications of Inferential Statistics. Lowry, Richard (1998)
39. McBride, B., et al.: The resource description framework (RDF) and its vocabulary description language RDFS. In: Handbook on Ontologies, pp. 51–66. Springer, New York (2004)
40. Mendes, E., Mosley, N., Counsell, S.: Web metrics-estimating design and authoring effort. IEEE Multimed. **8**(1), 50–57 (2001)
41. Mohr, G., Stack, M., Rnitovic, I., Avery, D., Kimpton, M.: Introduction to heritrix. In: Proceedings of the 4th International Web Archiving Workshop (IWAW). Vienna, Austria (2004)
42. Morrissey, S., Meyer, J., Bhattarai, S., Kurdikar, S., Ling, J., Stoeffler, M., Thanneeru, U.: Portico: A case study in the use of xml for the long-term preservation of digital artifacts. In: International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Montréal, Canada (2010)
43. Niu, J.: An overview of web archiving. D-Lib Mag. **18**(3), 2 (2012)
44. Olsina, L., Rossi, G.: Measuring web application quality with WebQEM. IEEE Multimed. **9**(4), 20–29 (2002)
45. Pant, G., Srinivasan, P., Menczer, F.: Crawling the web. In: Web Dynamics: Adapting to Change in Content, Size, Topology and Use, pp. 153–177. Springer, New York (2004)
46. Parmanto, B., Zeng, X.: Metric for web accessibility evaluation. J. Am. Soc. Inf. Sci. Technol. **56**(13), 1394–1404 (2005)
47. Paynter, G., Joe, S., Lala, V., Lee, G.: A year of selective web archiving with the web curator tool at the National Library of New Zealand. D-Lib Mag. **14**(5), 2 (2008)
48. Pennock, M., Davis, R.: ArchivePress: a really simple solution to archiving blog content. In: Proceedings of the 6th International Conference on Preservation of Digital Objects (IPres). San Francisco, CA (2009)
49. Pennock, M., Kelly, B.: Archiving web site resources: a records management view. In: Proceedings of the 15th International Conference on World Wide Web (WWW), pp. 987–988. Edinburgh, UK (2006)
50. Press, N.: Understanding metadata. National Information Standards 20 (2004)
51. Reyes Ayala, B., Phillips, M.E., Ko, L.: Current quality assurance practices in web archiving. http://digital.library.unt.edu/ark:/67531/metadc333026/ (2013). Accessed 22 Dec 2014
52. Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., Senellart, P.: Exploiting the social and semantic web for guided web archiving. In: Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 426–432. Paphos, Cyprus (2012)
53. Schonfeld, U., Shivakumar, N.: Sitemaps: above and beyond the crawl of duty. In: Proceedings of the 18th International Conference on World Wide Web (WWW), pp. 991–1000. Madrid, Spain (2009)
54. Spaniol, M., Denev, D., Mazeika, A., Weikum, G., Senellart, P.: Data quality in web archiving. In: Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW), pp. 19–26. Madrid, Spain (2009)
55. Sullivan, T., Matson, R.: Barriers to use: usability and content accessibility on the web's most popular sites. In: Proceedings on the ACM Conference on Universal Usability (CUU), pp. 139–144 (2000)
56. Voorhees, E., Harman, D.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (2005)
57. W3C: W3C HTML validation service (2001)
58. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. Internet Eng. Task Force RFC **2413**, 222 (1998)
59. Yang, S., Chitturi, K., Wilson, G., Magdy, M., Fox, E.A.: A study of automation from seed URL generation to focused web archive development: the CTRnet context. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 341–342. Washington, DC (2012)