CLEAR: a credible method to evaluate website archivability

Vangelis Banos† Yunhyong Kim‡ Seamus Ross‡ Yannis Manolopoulos†

†Aristotle University of Thessaloniki, Greece
vbanos@gmail.com manolopo@csd.auth.gr

‡University of Glasgow, United Kingdom
{yunhyong.kim,seamus.ross}@glasgow.ac.uk

ABSTRACT

Web archiving is crucial to ensure that cultural, scientific and social heritage on the web remains accessible and usable over time. A key aspect of the web archiving process is optimal data extraction from target websites. This procedure is difficult for such reasons as, website complexity, plethora of underlying technologies and ultimately the open-ended nature of the web. The purpose of this work is to establish the notion of Website Archivability (WA) and to introduce the Credible Live Evaluation of Archive Readiness (CLEAR) method to measure WA for any website. Website Archivability captures the core aspects of a website crucial in diagnosing whether it has the potentiality to be archived with completeness and accuracy. An appreciation of the archivability of a web site should provide archivists with a valuable tool when assessing the possibilities of archiving material and influence web design professionals to consider the implications of their design decisions on the likelihood could be archived. A prototype application, archiveready.com, has been established to demonstrate the viabiity of the proposed method for assessing Website Archivability.

Categories and Subject Descriptors

H.3 [**Information Storage And Retrieval**]: Online Information Services—*Web-based services*; H.3.7 [**Digital Libraries**]: [Collection]

General Terms

Web Archiving, Website Evaluation Method

Keywords

Web Archiving, Digital Preservation, Website Archivability

1. INTRODUCTION

Web archiving is the process of gathering up digital materials from the World Wide Web, ingesting it, ensuring that these materials are preserved in an archive, and making the collected materials available for future use and research [16]. Web archiving is crucial to ensure that our digital materials remain accessible over time.

Web archiving has two key aspects: organizational and technical. The organizational 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 [21]. The technical aspect of web archiving involves the procedures of web content identification, acquisition, ingest, organization, access and use [5, 25].

In this work, we are addressing two of the main challenges associated with technical aspects of web archiving, the acquisition of web content and the quality assurance (QA) performed before it is ingested into a web archive. Web content acquisition and ingest is a critical step in the process of web archiving; if the initial Submission Information Package (SIP) lacks completeness and accuracy for any reason (e.g. missing or invalid web content), the rest of the preservation processes are rendered useless. In particular, QA is vital stage in ensuring that the acquired content is complete and accurate.

The peculiarity of web archiving systems in comparison to other archiving systems, is that the SIP is preceded by an automated extraction step. Websites often contain rich information not available on their surface. While the great variety and versatility of website structures, technologies and types of content is one of the strengths of the web, it is also a serious weakness.There is no guarantee that web bots dedicated to retrieving website content (perform web crawling) can access and retrieve website content successfully [9].

Websites benefit from following established best practices, international standards and web technologies if they are to be amenable to being archived. We define the sum of the attributes that make a website amenable to being archived as *Website Archivability*. This work aims to:

- Provide mechanisms to improve the quality of web archive content (e.g. facilitate access, enhance content integrity, identify core metadata gaps).
- Expand and optimize the knowledge and practices of web archivists, supporting them in their decision making, and risk management, processes.
- Standardize the web aggregation practices of web archives, especially in relation to QA.
- Foster good practices in website development and web content authoring that make sites more amenable to harvesting, ingesting, and preserving.
- Raise awareness among web professionals regarding web preservation.

In this work, we define the *Credible Live Evaluation of Archive Readiness (CLEAR) method*, a set of metrics to quantify the level of archivability of any website. This method is designed to consolidate, extend and complement empirical web aggregation practices through the formulation of a standard process to measure if a website is archivable. The main contributions of this work are:

- the introduction of the notion of Website Archivability,
- the definition of the Credible Live Evaluation of Archive Readiness (CLEAR) method to measure Website Archivability,
- the description of ArchiveReady.com, a web application which implements the proposed method.

The concept of CLEAR emerged from our current research in web preservation in the context of the BlogForever project[1] which involves weblog harvesting and archiving. Our work revealed the need for a method to assess website archive readiness in order to support web archiving workflows.

The remainder of this paper is organized as follows: Section 2 presents work related to web archiving, content aggregation and QA, Section 3 introduces and analyses the CLEAR method, Section 4 presents archiveready.com, a prototype web application implementing it, Section 5 discusses future work and, Section 6 summarises our results.

2. RELATED WORK AND CONTEXT

The web archiving workflow includes identification, appraisal and selection, acquisition, ingest, organization and storage, description and access [16]. This section focuses explicitly on the acquision of web content and the way it is handled by web archiving projects and initiatives.

Web content acquision 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. The number of independent and dependent elements gives harvesting a substantial risk load.

Web content acquisition for web archiving is performed using robots, also known as "spiders", "crawlers", or "bots", self-acting agents that navigate around-the-clock through the hyperlinks of the web, harvesting topical resources without human supervision [18]. The most popular web harvester, Heritrix is an open source, extensible, scalable, archival quality web crawler [15] developed by the Internet Archive[2] 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)[3] as well as numerous web archiving projects. Heritrix is being continuously developed and extended to improve its capacities for intelligent and adaptive crawling [7] or capture streaming media [10]. The Heritrix crawler was originally established for crawling general webpages that do not include substantial dynamic or complex content. In response other crawlers have been developed which aim to address some of Heritrix's shortcomings. For instance, BlogForever [2] is utilizing blog specific technologies to preserve blogs. Also, the ArchivePress project is based explicitly on XML feeds produced by blog platforms to detect web content [20].

As websites become more sophisticated and complex, the difficulties that web bots face in harvesting them increase. For instance, some web bots have limited abilities to process GIS files, dynamic web content, or streaming media [16]. To overcome these obstacles, standards have been developed to make websites more amenable to harvesting by web bots. Two examples are the Sitemaps.xml and Robots.txt protocols. The Sitemap.xml[4] protocol, 'Simple Website Footprinting', is a way to build a detailed picture of the structure and link architecture of a website [12]. Implementation of the Robots.txt protocol provide web bots with information about specific elements of a website and their access permissions [26]. Such protocols are not used universally.

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)[5], 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 the Target Instance is saved in the Harvested state [6]. The next step is for the Target Instance Owner 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 [19]. 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 [8].

Recent efforts to deploy crowdsourced 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 site, the type of content hosted, and the technical structure. However, to quote the IIPC, "it is conceivable that crowdsourcing could support targeted elements of the QA process. The comparative aspect of QA lends itself well to 'quick wins' for participants"[7].

[1] http://blogforever.eu
[2] http://archive.org
[3] http://netpreserve.org
[4] http://www.sitemaps.org/
[5] http://webcurator.sourceforge.net/
[6] http://webcurator.sourceforge.net/docs/1.5.2/Web\%20Curator\%20Tool\%20User\%20Manual\%20(WCT\%201.5.2).pdf
[7] http://www.netpreserve.org/sites/default/files/..

IIPC has also organized a Crowdsourcing 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" [8].

The previous literature shows that there is an agreement within the web archiving community that web content aggregation is challenging. QA is an essential stage in the web archiving workflow but 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[9] is to perform test crawls prior to archiving[10] 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.

Website archivability provides an approach to automating 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 reducing computational and network resource usage through not harvesting unharvestable sites and by saving on human QA of sites that could not be harvested above particular quality thresholds.

3. ARCHIVABILITY EVALUATION METHOD

The main aspects of the Credible Live Evaluation of Archive Readiness (CLEAR) method (Ver.1, as of 04/2013). After introducing the objectives of CLEAR and its key components, we provide further analysis of all its aspects.

3.1 Introduction to CLEAR

The CLEAR method proposes an approach to producing on-the-fly measurement of *Website archivability*. Website archivability is defined as the extent to which a website meets the conditions for the safe transfer of the its content to a web archive for preservation purposes. All web archives currently employ some form of crawler technology to collect the content of target websites. These all communicate through HTTP requests and responses, processes that are agnostic of the repository system of the archive. Information such as the unavailability of pages, 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. Website archivability must not be confused with website dependability, the former refers to the ability to archive a website while the latter is a system property that integrates such attributes as reliability, availability, safety, security, survivability and maintainability[1].

The main components of CLEAR are:

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

Each of the CLEAR components will be 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 five core constituent facets of archivability (Section 3.2). We further describe the website attributes (e.g. HTML elements; hyperlinks) used to examine each archivability facet (Section 3.3), and, finally, 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 (Section 3.4).

3.2 Archivability Facets

Website archivability can be measured from several different perspectives. Here, we have called these perspectives *Archivability Facets* (See Figure 1). The selection of these facets is motivated by a number of considerations. For example, whether there are verifiable guidelines to indicate what and where information is held at the target website and whether access is available and permitted (i.e. Accessibility, see Section 3.2.1); whether included information follows a common set of format and/or language specifications (i.e. Standards Compliance, see Section 3.2.2); the extent to which information is independent from external support (i.e. Cohesion, see Section 3.2.4); the level of extra information available about the content (i.e. Metadata Usage, see Section 3.2.5); and, whether server response time is below an acceptable threshold (i.e. Performance, see Section 3.2.3).

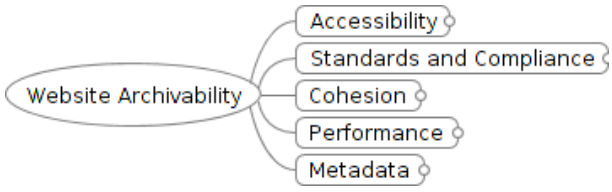

Figure 1: Archivability Facets: An Overview

3.2.1 F_A: Accessibility

A website is considered archivable only if web crawlers are able to visit its home page, traverse its content and retrieve

./CompleteCrowdsourcing.pdf

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

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

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

it via standard HTTP requests. In the case a crawler cannot find the location of all web resources, it will not be possible to retrieve the content. It is not only necessary to put resources on a web site, it is also essential to provide proper references to allow crawlers to discover them and retrieve them effectively and efficiently.

Example: a web developer is creating a website containing a javascript menu, which is generated on the fly. Web crawlers cannot understand this menu, so they are not able to find the web resources.

To support archivability, 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 Figure 2). These can be exposed in feeds, site maps, and robots.txt files. Information on whether the webpage is archived elsewhere (e.g. the Internet Archive[11]) and whether there are any errors in exporting them to the WARC format[12] could also help in determining the website accessibility.

Figure 2: Archivability Facet: Accessibility

3.2.2 F_S: Standards Compliance

Compliance with standards is a recurring theme in digital curation practices (e.g. see Digital Preservation Coalition guidelines [4]). 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[13].

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?"[14]. 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).

[11] `http://www.archive.org`

[12] Popular standard archiving format for web content. `http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml`

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

[14] `http://www.portico.org/digital-preservation/wp-content/uploads/2010/01/Portico_DLF_Fall2005.pdf`

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 pages. The problem is that the quirks mode is really versatile. As a result, you cannot depend on it to have a standard rendering of the web site in the future.

We recommend validation be performed for three types of content (see Figure 3): webpage components (e.g. HTML and CSS), reference media content (e.g. audio, video, image, documents), and supporting resources (e.g. robots.txt, sitemap.xml, javascript).

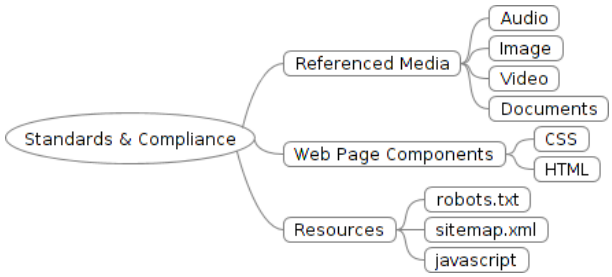

Figure 3: Archivability Facet: Compliance Standards

3.2.3 F_P: Performance

Performance is an important aspect of web archiving. The throughput of data acquisition of a web spider 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, improving a website's archiving process.

Example: if the performance of a website is slow, web spiders will have difficulty aggregating content and they may even abort if the performance degrates below a specific threshold.

While crawler performance can be adjusted and improved from within the archive, the server response time is under the control of the website creators. Website archivability is improved by optimising this response time. Depending on the size of the web archive and demands on acceptable server response time will differ. The performance is measured in relation to these needs. In a real world scenario, each archive would have a threshold indicating the maximum allowable server response time.

3.2.4 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 [17]). If files constituting a single website are dispersed across different services (e.g. different servers for images, javascript widgets, other resources), the acquisition and ingest is likely to risk suffering from neither being complete nor accurate. If one of the multiple services fails, the website fails. Here we characterise the robustness of the website in comparison to this kind of failure as *C*ohesion.

Example: images used in a website but hosted elsewhere

may cause problems in web archiving because they may not be captured when the site is archived. What is more, if the target site depends on 3rd party sites, the future availability of which is unknown, 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*[15]). Cohesion is tested on three levels:

- examining how many hosts are employed in relation to the location of referenced media content,
- examining how many hosts are employed in relation to supporting resources (e.g. robots.txt, sitemap.xml, and javascripts),
- examining the number of times proprietary software or plugins are referenced.

3.2.5 F_M: Metadata Usage

The adequate provision of metadata (e.g. see Digital Curation Centre Curation Reference Manual chapters on metadata [14], preservation metadata [23], archival metadata [27], and learning object metadata [11]) has been a continuing concern within digital curation (e.g. see seminal article by Lavoie[16] and insightful discussions going beyond preservation[17]). 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 (summarised in Figure 4). 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 NISO classification [22]; discussion in DCC Curation Reference Manual chapter on Metadata [14]). 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 indicated as part of the HTML element attribute. Descriptive information (e.g. author, keywords) that can help understand how the content is classified can be included in the HTML META element attribute and values. Metadata

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

[16] `http://www.dlib.org/dlib/april04/lavoie/04lavoie.html`

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

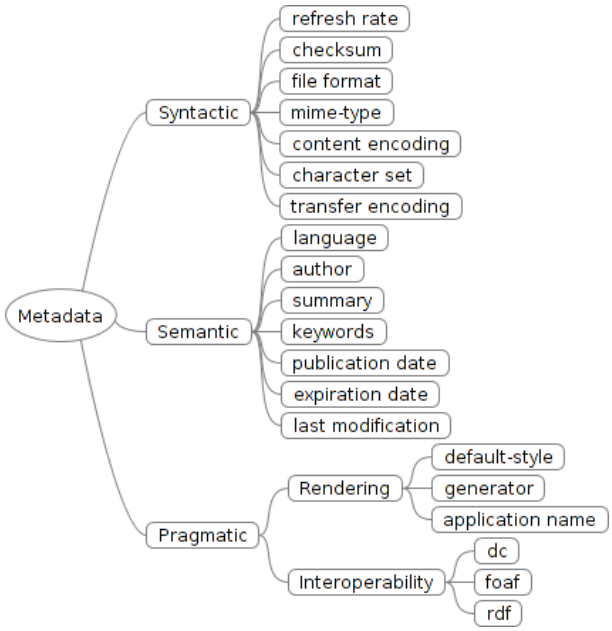

Figure 4: Archivability Facet: Metadata

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 [28]; Friend of a Friend (FOAF) [3]; Resource Description Framework (RDF) [13]) 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.

3.3 Websites Attributes

In this section, we examine the website attributes used to measure the archivability facets discussed in Section 3.2. In Figure 5, we have illustrated the components of the website that will be examined to measure the website's potential for meeting the requirements of the archivability facets.

For example, the level of **Accessibility** can be quantified on the basis of: whether or not,

- feeds exist (e.g. RSS and ATOM);
- robots.txt exists;
- sitemap.xml is mentioned in robots.txt and sitemap.xml exists at the location specified, and/or sitemap.xml is found at the root directory of the server;
- hyperlinks are valid and accessible; and,
- there are existing instantiations of the webpage elsewhere (e.g. snapshots at the Internet Archive[18]).

[18] `http://www.archive.org`

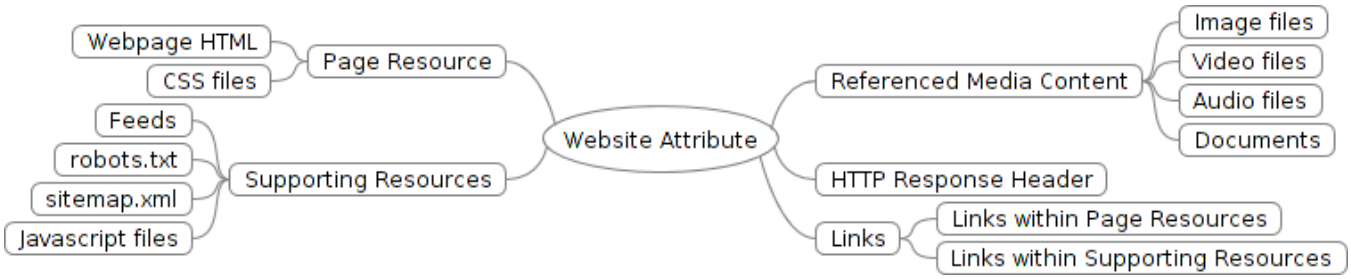

Figure 5: Website Archivability: mapping archivability facets to website attributes.

The existence of an RSS feed allows the publication of webpage content that can be automatically syndicated or exposed. It allows web crawlers automatically to retrieve updated content and the standardised format of the feeds allow access by many different applications. For example the BBC uses feeds to let readers see when new content has been added[19].

The file robots.txt[20] 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.

The Sitemaps protocol, supported jointly by the most widely used search engines to help content creators and search engines, is an increasingly widely used way unlock this hidden data by making it available to search engines [24]. To implement the Sitemaps protocol, the file sitemap.xml is used to list all the pages of the website and their location. 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`).

The hyperlinks of the website can be examined for availability as an indication of website accessibility. A website with many missing and/or broken links is not likely to be archived to any significant degree of completeness or accuracy.

The website will be 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, javascript) and formats (e.g. jpeg) will be validated using tools, such as W3C HTML[21] and CSS validator[22], JHOVE2[23] and/or Apache Tika[24] file format validator, python XML validator[25], robots.txt checker[26], ECMAScript[27] language specification.

The level of **Cohesion** is measured by the extent to which material associated to the website is kept within one host. This is measured by the proportion of content, resources, and plugins 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. javacsript). In addition the proportion of content relying on predefined proprietary software can be assessed and monitored.

The calculation of **Performance** is straightforward based on the response time of the server and can be implemented as a pass/fail test depending on a pre-set threshold of acceptability. In a archival context, it is likely that there is an acceptable performance threshold for the website if it is to be archivable given the web crawler and archival objectives.

The score for **Metadata Usage** can be assessed on the basis of whether or not,

- the <HTML> element includes a "lang" attribute specifying a value for the primary end-user language;
- the website includes element tags (i.e. <dc>, <foaf>, <rdf>), that indicate the use of Dublin Core, FOAF, and RDF (in the long-term, other elements related to initiatives such as SIOC[28], LOD[29], ORE[30] can be added as needed);
- fixity information is included ("content-md5" attribute can be used to include this in the HTTP response header);
- content mime-type identification is available ("content-type" can be used in the HTTP response header to indicate this; in cases where it is missing, this process might be refined to use JHOVE2 or Apache Tika to identify the format of content);
- character set is described (this can be exposed using "content-type" along with mime-type);
- transfer encoding is specified (this describes and compression methods in use and can be specified in the HTTP response header);

[19] `http://www.bbc.co.uk/news/10628494`
[20] `http://www.robotstxt.org/`
[21] `http://validator.w3.org/`
[22] `http://jigsaw.w3.org/css-validator/`
[23] `http://www.jhove2.org`
[24] `http://tika.apache.org/`
[25] `http://code.google.com/p/pyxmlcheck/`
[26] `http://tool.motoricerca.info/robots-checker.phtml`
[27] `http://www.ecmascript.org/`
[28] `http://sioc-project.org/`
[29] `http://linkeddata.org/`
[30] `http://www.openarchives.org/ore/1.0/datamodel`

- content encoding is specified (this can be included in the HTTP response header);
- HTML <META> element includes: "author", "description", "keywords", "default-style", "application-name", "generator" & "refresh" information.

In the case of HTTP response header, the availability of selected metadata elements and their values will be examined. In the case of more specific metadata schemas such as DC, at this stage, we envision only examining whether the schema is being used or not. At a later stage we might extend this to examine which elements are in use. The premise is that the information in the HTTP response header is essential to the archivability, whereas the elements and values associated with specific standards are considered to be desirable characteristics that would lead to richer metadata generation but are not necessarily essential.

The <META> tag attribute often embodies semantic data (e.g. authorship and keywords); however, the quality of metadata here can vary widely. Metadata harvested from this element should be used in conjunction with that derived from other components, for example, RSS feeds and Microformats[31], where these are available.

3.4 Evaluations

Combining the information discussed in Section 3.3 to calculate a score for website archivability goes through the following steps.

- The website's archivability potential with respect to each facet will be represented by an N-tuple (x_1, ..., x_k, ..., x_N) where the value of x_k is a zero or one representing a negative or positive answer, respectively, to the binary question asked about that facet, and where 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?"[32]; if there are M files for which format validation is being carried out then there will be M binary questions of this type.
- If all questions are considered to be of equal value to the facet, then the archivability with respect to the facet in questions is just the sum of all the coordinates divided by N (simplest model). If some questions are considered to be more important, then these can be assigned higher weights so that the archivability is $\sum_{k=0}^{N} \frac{\omega_k x_k}{N}$, where ω_k is the weight assigned to question k and $\sum \omega_k = 1$.
- If selected questions are grouped to represent sub-facets to be calculated at different hierarchical levels then this will also change the weighting. Ideally, this could be adjusted on the basis of the needs of the community for which the website is being archived. Some will be more interested in preservation of images, while others will be interested in text. This can be easily incorporated into the current methodology.

Once the archivability with respect to each facet is calculated, the total measure of Website Archivability can be simply defined as:

$$\sum_{\lambda \in \{A,S,C,P,M\}} w_\lambda F_\lambda$$

where F_A, F_S, F_C, F_P, F_M are archivability with respect to Accessibility, Standards Compliance, Cohesion, Performance, Metadata Usage, respectively, and $\sum_{\lambda \in \{A,S,C,P,M\}} w_\lambda = 1$ and $0 \leq w_\lambda \leq 1 (\lambda \in \{A, S, C, P, M\})$.

Depending on the curation and preservation objectives of the web archive, the weight of each facet is likely to be different, and w_λ should be assigned to reflect this. In the simplest model, these can be set to be equal so that $w_\lambda = 0.2$ for all λ. In actuality accessibility will be the most central consideration in archivability since, if the content cannot be found or accessed, then the website's compliance with other standards, and conditions become moot.

4. A WEBSITE ARCHIVABILITY EVALUATION TOOL: ARCHIVEREADY.COM

ArchiveReady, a web application located at `http://www.archiveready.com`, implements the CLEAR method for evaluating website archivability. We describe its technology stack, and website archivability evaluation workflow. To demonstrate ArchiveReady, we also present an evaluation of the iPRES2013 Conference website.

4.1 Technology Stack

ArchiveReady is a web application based on the following key components: Debian linux[33] operating system for development and production servers, Nginx web server[34] to server static web content, Python programming language[35], Gunicorn python WSGI HTTP Server for unix[36] to server dynamic content, BeautifulSoup[37] to analyse html markup and locate elements, Flask[38], a python microframework to develop web applications, Redis advanced key-value store[39] to manage job queues and temporary data, Percona Mysql RDBMS[40] to store long-term data. JSTOR/Harvard Object Validation Environment (JHOVE) [6] for object validation, Javascript libraries such as jQuery[41] and Bootstrap[42] are utilized to create a compeling user interface.

To ensure high level compatibility with W3C standards the initiative used open source web services provided by the

[31] `http://microformats.org/about`
[32] `http://www.portico.org/digital-preservation/wp-content/uploads/2010/01/Portico_DLF_Fall2005.pdf`
[33] `http://www.debian.org`
[34] `http://www.nginx.org`
[35] `http://www.python.org/`
[36] `http://gunicorn.org/`
[37] `http://www.crummy.com/software/BeautifulSoup/`
[38] `http://flask.pocoo.org/`
[39] `http://redis.io`
[40] `http://www.percona.com`
[41] `http://www.jquery.com`
[42] `http://twitter.github.com/bootstrap/`

W3C. These include: the Markup Validator[43], the Feed Validation Service[44] and the CSS Validation Service[45].

The greatest challenge in implementing ArchiveReady is performance. According to the HTTP Archive Trends, the average number of HTTP requests initiated when accessing a web page is over 90 and is expected to rise[46]. 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[47] was deployed. This approach enables the parallel execution of multiple evaluation processes, resulting in huge performance benefits when compaired to traditional web application execution model.

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 workflow 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 spider. ArchiveReady analyses only the URL submitted by the user, it does not evaluate the whole website recursively.

3. In stage 3, Website Attributes are evaluated (See Section 3.3).

4. The metrics for the Archivability Facets are calculated according to the CLEAR method and the final website archivability rating is calculated.

Note that in the current implementation, CLEAR evaluates only a single webpage based on the assumption that all website pages share the same components, standards and technologies. This issue is further discussed in Future Work.

4.3 Demonstration

To demonstrate ArchiveReady, we evaluate the website of iPRES'2013 international conference as it was available on 23 April 2013[48] and present the results in Table 1. The corresponding result is also presented in Figure 6.

[43] `http://validator.w3.org/`
[44] `http://validator.w3.org/feed/`
[45] `http://jigsaw.w3.org/css-validator/`
[46] `http://httparchive.org/trends.php`
[47] `http://python-rq.org/`
[48] `http://ipres2013.ist.utl.pt/`

Table 1: http://ipres2013.ist.utl.pt/ Website Archivability evaluation

Facet	Evaluation	Rating	Total
Accessibility	No RSS feed	50%	50%
	No robots.txt	50%	
	No sitemaps.xml	0	
	6 Valid links	100%	
Cohesion	1 external & no internal scripts	0	70%
	4 local & 1 external images	80%	
	No QuickTime or Flash objects	100%	
	1 local CSS file	100%	
Standards Compliance	1 Invalid CSS file	0	77%
	Invalid HTML	0	
	Meta description found	100%	
	No content encoding in HTTP headers	50%	
	Content type HTTP header	100%	
	Page expiration HTTP header	100%	
	Last-modified HTTP header	100%	
	No QuickTime or Flash objects	100%	
	5 images checked successfully with JHOVE	100%	
Metadata	Meta description found	100%	87%
	Content type	100%	
	No page expiration metadata	50%	
	Last-modified HTTP header	100%	
Performance	Avg network response time is 0.546ms	100%	100%
		Website Archivability	77%

5. FUTURE WORK

Future work directions stem from two facts: a) the identification of limitations which nevertheless do not refute the claim that the proposed method is significant, and b) the novelty of this work which promises to improve considerably the web archiving process.

The method as currently implemented treats all website archivability facets equally in calculating the total Archivability Score. This may not be the optimal approach as in different organisational and policy contexts the objectives of web archiving might put greater or lesser emphasis on the individual Archivability Facets. The ability to weight the various individual Archivability Facet Scores in calculating the total Archivability Score is a feature which users will find valuable. For instance Metadata breadth and depth might be critical for a particular web archiving research task and

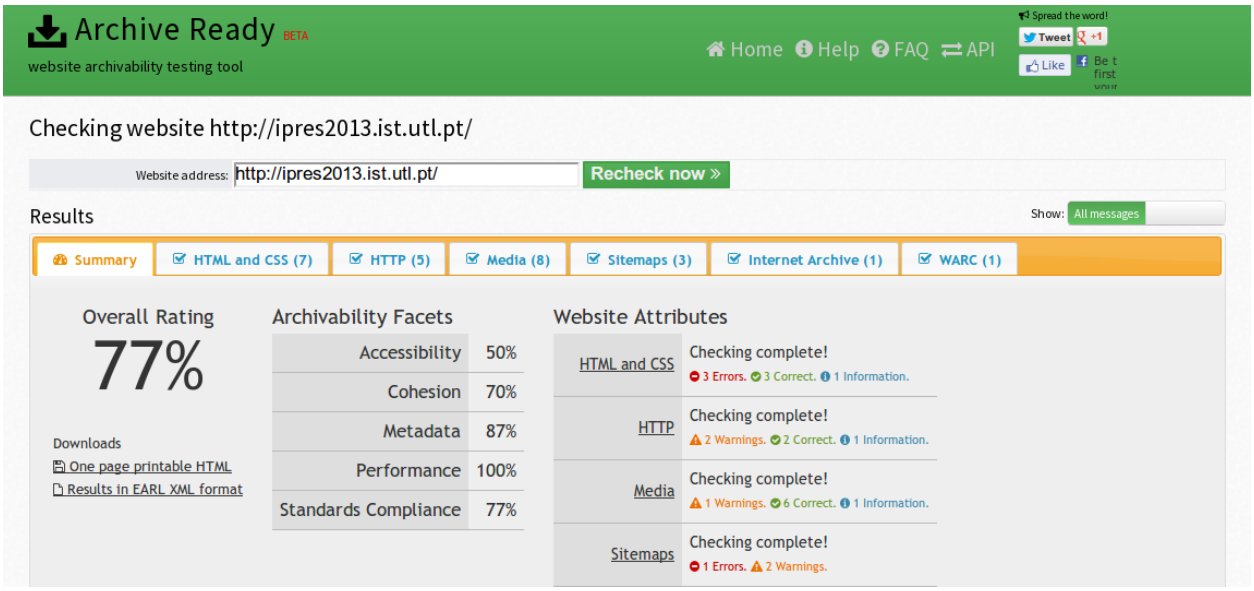

Figure 6: Evaluating iPRES2013 Website Archivability using ArchiveReady

therefore in establishing the archivability score for a particular site the user may which to instantiate this thinking in calculating the overall score. A next step will be to introduce a mechanism to allow the user to weight each Archivability Facet to reflect specific objectives.

Currently, CLEAR evaluates only a single website page based on the assumption that webpages from the same website share the same components and standards. To achieve a more objective evaluation, it would be better to perform sampling using sitemap.xml and RSS referenced pages to increase the breadth of the target website content to be evaluated.

There are some open questions that could lead to further refinement of the website archivability concept:

- Is it correct to consider archivability to be directly proportional to the number of binary questions answered positively? Are there points in the archivability curve that move at a faster/slower rate?
- Evidence from other archiving projects demonstrates that certain classes and specific types of errors create lesser or greater obstacles to website acquisition and ingest than others. The website archivabliity tool needs to be enhanced to reflect this differential valuing of error classes and types.
- Recognising that the different classes and types of errors do not have a purely summative combinatorial impact on archivability of a website this research in its next stage must identify the optimal way to reflect this weighting to enable comparisons across websites.
- Currently the system is envisaged as being used to guide the process of archiving websites, but a further extension would support its use by developers to assist them in design and implementation.

One way to address these concerns might be to apply an approach similar to normalized discounted cummulative gain (NDCG) in information retrieval[49]: for example, a user can rank the questions/errors to prioritise them for each facet. The basic archivability score can be adjusted to penalise the outcome when the website does not meet the higher ranked criteria. Further experimentation with the tool will lead to a richer understanding of new directions in automation in web archiving.

6. CONCLUSIONS

Our main aims were to improve web archive quality by establishing standards and tools to enhance content aggregation. Moreover, our aims were to help web archive operators improve their content ingestion workflows and also raise awareness among web professionals regarding web archiving.

To this end, we introduced the *Credible Live Evaluation of Archive Readiness (CLEAR)* method, a set of metrics to quantify the level of *Website Archivability* based on established web archiving standards, digital preservation principles and good practices. Also, one of the authors of this paper developed a web application implementing this method, ArchiveReady.com. This approach, provided the authors with an environment to test the concept of Archivability Facets and offered a method for web archive operators to evaluate target websites before content harvesting and ingestion, thus avoiding invalid harvests, erroneous web archives and unnecessary wasted resources which could be used elsewhere. ArchiveReady provides web professionals with an easy but thorough tool to evaluate their websites and improve their archivability. This achieved the twin goals of on the one hand instantiating methods to improve website archiving and on the other raising awareness of the challenges to web archiving among a broader audience.

[49] `http://en.wikipedia.org/wiki/Discounted_cumulative_gain\#Normalized_DCG`

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission Framework Programme 7 (FP7), BlogForever project, grant agreement No.269963.

8. REFERENCES

[1] A. Avižienis, J.-C. Laprie, and B. Randell. Fundamental concepts of computer system dependability. In *Proceedings of IARP/IEEE-RAS Workshop on Robot Dependability: Technological Challenge of Dependable, Robots in Human Environments*, 2001.

[2] V. Banos, N. Baltas, and Y. Manolopoulos. Trends in blog preservation. In *Proceedings of the 14th International Conference on Enterprise Information Systems (ICEIS)*, Wroclaw, Poland, 2012.

[3] D. Brickley and L. Miller. Foaf vocabulary specification 0.98. *Namespace Document*, 9, 2010.

[4] D. P. Coalition. Institutional strategies - standards and best practice guidelines. `http://www.dpconline.org/advice/preservationhandbook/institutional-strategies/standards-and-best-practice-guidelines`, 2012. [Online; accessed 18-April-2013].

[5] D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. The sharc framework for data quality in web archiving. *The VLDB Journal*, 20(2):183–207, 2011.

[6] M. Donnelly. Jstor/harvard object validation environment (jhove). *Digital Curation Centre Case Studies and Interviews*, 2006.

[7] M. Faheem and P. Senellart. Intelligent and adaptive crawling of web applications for web archiving. In *Proceedings of the 21st International Conference Companion on World Wide Web (WWW)*, pages 127–132, Lyon, France, 2012.

[8] V. D. Glenn. Preserving government and political information: The web-at-risk project. *First Monday*, 12(7), 2007.

[9] Y. He, D. Xin, V. Ganti, S. Rajaraman, and N. Shah. Crawling deep web entity pages. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 355–364, Rome, Italy, 2013.

[10] H. Hockx-Yu, L. Crawford, R. Coram, and S. Johnson. Capturing and replaying streaming media in a web archive-a british library case study, 2010.

[11] U. o. S. Lorna Campbell. Learning object metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/learning-object-metadata`, 2007. [Online; accessed 18-April-2013].

[12] S. Mansfield-Devine. Simple website footprinting. *Network Security*, 2009(4):7–9, 2009.

[13] B. McBride et al. The resource description framework (rdf) and its vocabulary description language rdfs. *Handbook on Ontologies*, pages 51–66, 2004.

[14] D. Michael Day. Metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/metadata`, 2005. [Online; accessed 18-April-2013].

[15] G. Mohr, M. Stack, I. Rnitovic, D. Avery, and M. Kimpton. Introduction to heritrix. In *Proceedings of the 4th International Web Archiving Workshop (IWAW)*, Vienna, Austria, 2004.

[16] J. Niu. An overview of web archiving. *D-Lib Magazine*, 18(3):2, 2012.

[17] L. of Congress. Sustainability of digital formats planning for library of congress collections: External dependencies. `http://www.digitalpreservation.gov/formats/sustain/sustain.shtml\#external`, 2013. [Online; accessed 18-April-2013].

[18] G. Pant, P. Srinivasan, and F. Menczer. Crawling the web. In *Web Dynamics*, pages 153–177. Springer, 2004.

[19] G. Paynter, S. Joe, V. Lala, and G. Lee. A year of selective web archiving with the web curator tool at the national library of new zealand. *D-Lib Magazine*, 14(5):2, 2008.

[20] M. Pennock and R. Davis. 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.

[21] M. Pennock and B. Kelly. Archiving web site resources: a records management view. In *Proceedings of the 15th International Conference on World Wide Web (WWW)*, pages 987–988, Edinburgh, UK, 2006.

[22] N. Press. Understanding metadata. *National Information Standards*, 20, 2004.

[23] F. C. f. L. A. Priscilla Caplan, Digital Library Services. Preservation metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/preservation-metadata`, 2006. [Online; accessed 18-April-2013].

[24] U. Schonfeld and N. Shivakumar. Sitemaps: above and beyond the crawl of duty. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*, pages 991–1000, Madrid, Spain, 2009.

[25] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW)*, pages 19–26, Madrid, Spain, 2009.

[26] Y. Sun, Z. Zhuang, and C. L. Giles. A large-scale study of robots. txt. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 1123–1124, Banf, Canada, 2007.

[27] W. D. . M. van Ballegooie. Archival metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/archival-metadata`, 2006. [Online; accessed 18-April-2013].

[28] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin core metadata for resource discovery. *Internet Engineering Task Force RFC*, 2413:222, 1998.