

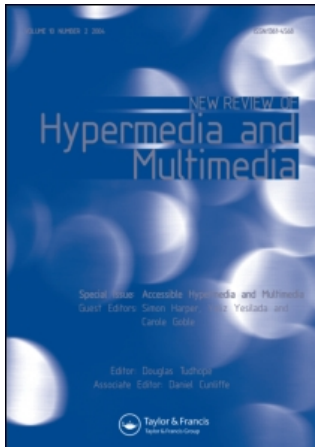
This article was downloaded by: [Lopes, Rui]

On: 17 December 2010

Access details: Access Details: [subscription number 929754479]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## New Review of Hypermedia and Multimedia

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t713599880>

## Macroscopic characterisations of Web accessibility

Rui Lopes<sup>a</sup>; Luis Carriço<sup>a</sup>

<sup>a</sup> LaSIGE Research Lab, University of Lisbon, Lisboa, Portugal

First published on: 17 November 2010

**To cite this Article** Lopes, Rui and Carriço, Luis(2010) 'Macroscopic characterisations of Web accessibility', New Review of Hypermedia and Multimedia, 16: 3, 221 — 243, First published on: 17 November 2010 (iFirst)

**To link to this Article:** DOI: 10.1080/13614568.2010.534185

**URL:** <http://dx.doi.org/10.1080/13614568.2010.534185>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# Macroscopic characterisations of Web accessibility

RUI LOPES\* and LUIS CARRIÇO

LaSIGE Research Lab, University of Lisbon, Lisboa, Portugal

(Received 15 August 2010; final version received 19 October 2010)

The Web Science framework poses fundamental questions on the analysis of the Web, by focusing on how microscopic properties (e.g. at the level of a Web page or Web site) emerge into macroscopic properties and phenomena. One research topic on the analysis of the Web is *Web accessibility evaluation*, which centres on understanding how accessible a Web page is for people with disabilities. However, when framing Web accessibility evaluation on Web Science, we have found that existing research stays at the microscopic level.

This article presents an experimental study on framing Web accessibility evaluation into Web Science's goals. This study resulted in novel accessibility properties of the Web not found at microscopic levels, as well as of Web accessibility evaluation processes themselves. We observed at large scale some of the empirical knowledge on how accessibility is perceived by designers and developers, such as the disparity of interpretations of accessibility evaluation tools *warnings*. We also found a direct relation between accessibility quality and Web page complexity. We provide a set of guidelines for designing Web pages, education on Web accessibility, as well as on the computational limits of large-scale Web accessibility evaluations.

**Keywords:** Web Science; Web accessibility; Characterisation; Automated evaluation

## 1. Introduction

Since its inception, the Web has been growing both in size and complexity. It is argued that this happens due to its decentralised properties (Jacobs and Walsh 2004): anyone can contribute to the Web without a central authority dictating what is allowed to be published or not. Consequently, unskilled people became massive Web publishers. This means that the Web is perceived as a *living organism*, constantly evolving in different directions regarding its size and quality, while retaining fundamental properties, as thoroughly discussed in the Web Science framework by Berners-Lee *et al.* (2006).

At the same time, the Web is being experienced by an ever-growing number of people interacting with Web sites. This leads to a critical mass of users that turns apparent its diversity. Minority groups become more representative and, consequently, the (lack of) adequacy of Web sites to them becomes relevant.

---

\*Corresponding author. Email: rlopes@di.fc.ul.pt

One of these groups concerns people with disabilities. This group includes those with physical disabilities such as upper limb impairment, sensorial disabilities such as blindness, or even ageing groups, who typically have mild, yet multiple disabilities. Due to the richness and openness of Web technologies, there is no implied quality control to ensure that all published Web pages are accessible to any or all of these user groups.

The base premise of this article centres on the fact that Web accessibility evaluation research is mostly focused at the analysis of small sets of Web pages. Consequently, this poses limitations on analysing the deployment of accessibility in the Web, such as understanding how the behaviour of developers and designers shape the Web's accessibility. We focused on a sufficiently large (and representative) portion of the Web that entails its diversity in Web documents, in order to characterise Web accessibility on the Web. To enable this study, we devised an experimental analysis over a large Web document collection provided by the Portuguese Web Archive (PWA) initiative.

### 1.1 Summary of contributions

Through the devised study we have discovered a set of properties that characterise Web accessibility, as well as those who create Web pages, including:

- (1) the drastic difference in interpretation of *warnings* in Web accessibility evaluation results;
- (2) the correlation between the accessibility quality of a Web page and its complexity (in terms of number of HTML elements);
- (3) the usage of templates and content management systems to improve the accessibility quality of a Web page; and
- (4) the lack of compliance with the simpler and most well-known accessibility guidelines (i.e. alternative texts).

### 1.2 Structure of the article

This article is structured as follows: in Section 2 we present the background for Web accessibility evaluation (particularly focused on *automated evaluation* techniques) and for the PWA—the core initiative upon which we built our experimental study—as well as discuss the research question that framed our work; Section 3 details the three main steps of the conducted experiment: *acquiring the document collection*, *accessibility evaluation* and *data analysis*; afterwards, in Section 4, we present four results from our study: *distribution of rates*, *rates and page complexity*, *rates and hyperlinking* and *alternative texts*; Section 5 delves into the discussion of the results, especially on the impact of our findings, as well as the limitations we faced in this study; we end the article with conclusions and a set of ongoing efforts of studying more aspects of accessibility at the large.

## 2. Background

### 2.1 Web accessibility evaluation

Web accessibility is an umbrella term for the study of the adequacy of Web technologies to users with special needs such as people with blindness, cognitive disabilities, etc. This adequacy can be viewed from two perspectives: (1) stricter, where accessibility means *the ability to access* (e.g. a person with blindness cannot grasp information conveyed in images); and (2) broader, where the term represents how easily these users can interact with a Web page. This second perspective has been the main direction taken by the Web Accessibility Initiative (WAI)<sup>1</sup> of the World Wide Web Consortium.<sup>2</sup> Through WAI, several recommendations have been proposed on how Web technologies should be used without posing barriers to people with disabilities. One of these recommendations is the Web Content Accessibility Guidelines (WCAG) (Caldwell *et al.* 2008). WCAG defines a set of guidelines that should be followed by *creators* (e.g. developers, designers, etc.) when constructing Web pages, to ensure a good level of accessibility for all users. Each guideline is composed by a set of verifiable checkpoints (e.g. Checkpoint 1.1 from WCAG 1.0 states “Provide a text equivalent for every non-text element”) that creators must follow accordingly on Web technologies (e.g. HTML, CSS, etc.). While WCAG can be used during the Web page creation process, it is also the basis for analysis by usability experts, developers, etc. Thus, it effectively serves as a way to measure the level of accessibility an existing Web page has, e.g. in a qualitative A/AA/AAA conformance level defined by WCAG.

On top of WCAG, there are several proposals for measuring how accessible a Web page is. Other than qualitative levels, work has been conducted on quantifying accessibility (e.g. a percentage), such as failure rate (Sullivan and Matson 2000), WAB (Zeng 2004), UWEM (Velleman *et al.* 2007), A3 (Biühler *et al.* 2006) and WAQM (Vigo *et al.* 2007). All of these metrics depend on the single results obtained from the verification of checkpoints conformance, which can be either manually performed by experts or automatically evaluated through software. While expert evaluation provides an in-depth answer on the accessibility quality of a Web page, being a manually performed task poses problems on the scalability of the process (e.g. analysing a dozen pages already requires a significant effort from seasoned practitioners), and brings potential bias when comparing the analysis of two pages. As detailed by Sloan *et al.* (2006), even the interpretation of existing accessibility guidelines leads to ambiguous answers on whether specific features of a Web page are accessible or not. Nevertheless, conformance checking is one of the keystone starting points for accessibility evaluation.

On the other hand, software-based evaluation, by being automated in its nature, has the benefit of scalability and objectivity. However, since some checkpoints cannot be fully machine-verified (e.g. for WCAG Checkpoint 1.1., verifying if a text correctly explains the information conveyed on a picture), the evaluation is less detailed when compared with expert analysis. Nonetheless, as pointed out by Vigo *et al.* (2007), the rate of non-verifiable

evaluation failures is typically proportional to machine-verifiable ones. This provides additional support on bridging the precision gap between expert and automated evaluation.

But having an automated way of evaluating the accessibility of Web pages opens the way to perform large-scale analysis. To our knowledge, no large-scale accessibility evaluations of the Web (and its evolution) have been performed before, which limits the kind of knowledge that could be grasped, e.g. how developers and designers are deploying accessibility concerns throughout the Web. We assume that this is due to the dependency of computational resources for large-scale analysis. To mitigate this problem, work is typically conducted on evaluating smaller scale collections of Web documents (Vigo *et al.* 2007, Mirri *et al.* 2009). These processes are based on sampling methods, such as those defined in UWEM (Bühler *et al.* 2006). However, as argued by Brajnik *et al.* (2007), there is always a significant sampling bias induced by these methods. Consequently, we claim that, for a large-scale analysis of Web accessibility and its macroscopic properties, and for a proper characterisation, sampling bias has to be reduced to a minimum, i.e. by evaluating collections of Web pages that are more representative of the entire Web.

## 2.2 Portuguese Web Archive (PWA)

A large amount of information is published exclusively on the Web everyday. However, after a few months, most of the published contents become unavailable and are lost forever (Ntoulas *et al.* 2004, Gomes and Silva 2006). The information published on the Web reflects our times and it is a valuable historical resource for future generations. For this reason, the Internet Archive (Kahle 2002) has been collecting and archiving Web pages since 1996. Several countries have also founded archival initiatives for their own national Webs (Philips 2003, Christensen 2005).

The PWA, a project of the Foundation for National Scientific Computing,<sup>3</sup> periodically crawls Portuguese language contents, mainly from the .pt domain (Portugal), and stores them into a repository for archival purposes. Besides supporting research, the PWA has a direct interest in contributing to monitor and enhance the quality of the contents published on the Web. The accessibility quality of the archived contents is, therefore, an important measure that is being taken into account by the PWA.

## 2.3 Research question

From what we could tell, there is barely any research work on investigating how accessibility is shaped on the Web, as well as how the Web, at a macroscopic scale, influences how people with disabilities use it. The experimental work detailed in this article aims at framing Web accessibility into Web Science, thus opening the way to understand what core issues might arise. Through the methodologies established by Web accessibility evaluation

and Web archiving disciplines, we formulated the following overarching research question:

What macroscopic properties emerge from Web accessibility?

This question aims at understanding not just what is the quality of accessibility of the Web, but also about Web accessibility evaluation procedures themselves. Furthermore, with this knowledge, we also intend a meta-analysis of the experiment, by making sense of the benefits and pitfalls of large-scale studies of Web accessibility. In order to find answers, we devised an experimental analysis of a Web document collection provided by the PWA initiative, which is detailed in the forthcoming Sections.

### 3. Experiment

Since we predicted a certain level of complexity in this study, particularly due to the amount of data that would be analysed—a large collection of Web pages provided by the PWA—we segmented the experiment into three main steps, as follows:

- (1) *Acquiring the document collection.* This first step centres on acquiring the collection of Web documents, by following specific crawling and storage processes.
- (2) *Accessibility evaluation.* After obtaining the document collection, we will focus on computing data about the accessibility quality of each Web page, through a WCAG-based automated evaluation process.
- (3) *Data analysis.* The second step focuses on computing accessibility metrics and other analysis processes on top of the evaluation results.

The main reason for this rationale lays at the expected computational effort required for the evaluation of a large collection of Web documents, and the analysis of the evaluation results. The following Sections further detail each step.

#### 3.1 Acquiring the document collection

Collecting a representative sample of the Web for characterisation purposes is not a straightforward process (Henzinger *et al.* 2000), due to both its size and its linking structure. However, national Webs, due to their smaller nature, are more treatable. Chung *et al.* (2009) show that national Webs tend to share the same macroscopic characteristics, independently from their internal structure differences. For these reasons, national Webs can be crawled and analysed in a practical and relevant way.

A crawler is a software that harvests content referenced by URLs and extracts its links for further processing, iteratively. While in theory this would allow for the collection of the entire Web, spider traps (Heydon and Najork 1999) lead to corner cases on crawling. Thus, not every single Web page can

be crawled and stored. For this reason, we established the following criteria as halting conditions within the crawling process:

- maximum of 10,000 URLs crawled per Web site;
- maximum content of 10 MB;
- maximum number of five hops from the seed;
- respect for the robots exclusion protocol (Koster 1994); and
- a courtesy pause of two seconds between requests to the same Web site (considering a *Web site* by its full qualified domain name).

To bootstrap the crawling process, we used as URL seeds nearly 200,000 site addresses hosted under the *.pt* Top Level Domain (TLD). The document collecting process was undertaken between March and May 2008. Most of the Web pages were collected during the first month, but the process was left running for two extra months reaching convergence. In order to create this Web document collection, the PWA developed a crawling system based on the following components, as detailed in Miranda and Gomes (2009):

- the Heritrix Web crawler (Mohr *et al.* 2004) to collect Web pages; and
- a software component to store and manage all crawled documents in the ARC file format, ideal for the archival of large collections of Web documents (Burner and Kahle 1996).

### 3.2 Accessibility evaluation

After the creation of the collection of Web pages, we developed software components to process it according to automatable Web accessibility evaluation practices. Since the evaluation of large Web document collections can be a resource intensive process (i.e. CPU, memory, disk), we implemented the evaluation software in order to minimise its impact resource-wise.

While there are already freely available Web accessibility evaluators, most of them are either service-based (i.e. interactive, thus targeted at aiding experts and developers) or limited with respect to performance and configurability. Consequently, we opted to implement evaluation techniques from WCAG 1.0 (Chisholm *et al.* 1999). We chose version 1.0 of these guidelines instead of 2.0 for two main reasons: first, the crawling process and subsequent document collection creation were performed before WCAG 2.0 achieved a *W3C Recommendation* status; and second, the implementation of WCAG 1.0 checkpoints can be easily supported by existing Web accessibility evaluation practices, in order to ensure their correctness.

We used the hardware and software infrastructure created by the PWA for processing large document collections. On the hardware side, 10 blade servers were made available for this experiment, each with  $2 \times$  Quad-core 2.3GHz CPUs and 8-GB of RAM. Regarding software components, we developed an evaluator for WCAG 1.0 that sits on top of Hadoop (Apache Foundation 2010), an open-source implementation of Map/Reduce (Dean and Ghemawat 2004), leveraging the hardware cluster's scalability properties.

We implemented the techniques for 39 checkpoints from WCAG 1.0, focusing on HTML structure analysis of the collected Web documents, using UWEM (Velleman *et al.* 2007) as the benchmark implementation reference. Consequently, all checkpoints are either from priority 1 or 2, thus leaving out the most advanced—albeit less critical—checkpoints from priority 3. Following UWEM, we categorised the expected results of evaluating a checkpoint as follows:

- *PASS*: it is applicable to an HTML document *and* its compliance is verified automatically;
- *FAIL*: it is applicable to an HTML document *and* its compliance unachieved; and
- *WARN*: it is applicable to an HTML document *but* it is impossible to verify its compliance.

UWEM defines its suggested implementation of checkpoints with a mix of XPath 1.0 (Clark and DeRose 1999) expressions and verbose text. However, using XML-based technologies such as XPath put difficulties on parsing malformed HTML documents, which are significant in any Web document collection, as well as performance penalties when scaling up the evaluation process. Consequently, we have opted to implement our accessibility evaluation component in the Groovy programming language,<sup>4</sup> aided by the NekoHTML parser.<sup>5</sup> Groovy's native capabilities of writing terse source code allowed us to migrate the XPath expressions defined in UWEM into equivalent native code expressions, thus maintaining the semantics defined by UWEM evaluation, as well as build upon the optimisation techniques of the Java Virtual Machine.

The main evaluator component implements all checkpoints and applies them to all HTML elements of each Web page in the collection. In order to minimise the impact on CPU and memory and, consequently, comply with the desired scalability for this study, we took advantage of streaming and caching algorithms in the implementation of the Web accessibility evaluator (the discussion of these optimisation aspects is out of the scope of this article, though). All evaluation results were collected and stored for data analysis tasks, as detailed next.

### 3.3 Data analysis

Having the raw results of the accessibility evaluation, we processed data in order to have answers for our research question. For this task, we used the R Statistical Computing framework<sup>6</sup> and ancillary shell scripts. To assess a quantification value of the accessibility quality of Web page, we defined three evaluation rates based on the *failure rate* metric by Sullivan and Matson (2000).

We opted to adjust this metric by taking into account the amount of times a checkpoint fails in the same Web page. This option was based on the assumption that a Web page that fails once must yield a smaller penalty than



if it would fail several times. Each rate is normalised into a percentage, with the semantics from *not accessible* (0%) to *fully accessible* (100%). Based on these conditions, we defined our automated evaluation process and corresponding metric rates with the distinction of whether a checkpoint evaluation of a given HTML element results in PASS, FAIL or WARN, as follows:

- **Conservative rate:** *WARN* results are interpreted as *failures*. The semantics of this rate conveys the worst-case scenario on accessibility evaluation:

$$\text{rate}_{\text{conservative}} = \frac{\text{passed}}{\text{applicable}} \quad (1)$$

- **Optimistic rate:** *WARN* results are interpreted as *passed*. This rate is related to a best-case scenario where developers and experts dismiss warnings—often incorrectly, as explained in Sloan *et al.* (2006) as accessibility issues that were taken into account:

$$\text{rate}_{\text{optimistic}} = \frac{\text{passed} + \text{warned}}{\text{applicable}} \quad (2)$$

- **Strict rate:** *WARN* results are *dismissed*, thus accounting only the actual *FAIL* results:

$$\text{rate}_{\text{strict}} = \frac{\text{passed}}{\text{applicable} - \text{warned}} \quad (3)$$

Each rate was computed for each checkpoint, as well as their aggregation into a final score stating the accessibility quality of a Web page. Next, we present the results of this experiment.

#### 4. Results

The evaluation of the Web document collection yielded results for 28,135,102 HTML documents, out of 48,718,404 Web content documents (Web pages, images, PDFs, etc.) that have been crawled in total (nearly 58%). A total of 40,831,728,499 HTML elements were analysed, an average 1451 HTML elements per Web page.

Of these, 1,589,702,401 HTML elements successfully met all applicable Web accessibility criteria, an average of 56 HTML elements per Web page (around 3.89%). Regarding failures, 2,918,802,078 HTML elements failed to comply, corresponding to an average of more than 103 errors per Web page (approximately 7.15%). Finally, 36,323,224,020 HTML elements were detected for belonging to a *warning* status, accounting for an average 1291 warnings per Web page (nearly 89%).

While these numbers provide a crude characterisation of the document collection and the evaluation process, they do not answer our initial research question. Thus, we analysed all the computed data by focusing on distributions of the different aspects of accessibility that were evaluated. The next sections present a more detailed set of results in the following fronts: *distribution of rates, rates and page complexity, rates and hyperlinking and alternative texts*.

#### 4.1 Distribution of rates

We aggregated the accessibility quality rate by permillage values, and generated a frequency distribution of how many Web pages belong to each aggregation. Figure 1 presents the distribution of *conservative rate* metric versus page count. Since all *warnings* are interpreted as errors, and no Web page was missing the HTML elements detectable in the checkpoints that yield warnings, no Web page was able to reach the maximum value of accessibility quality. The depicted exponential decay starts around 5% of compliance, where the number of pages with good quality is minimal.

Figure 2 presents the distribution of *optimistic rate* metric versus page count. Since this metric takes into account all *warnings* as positively complied, all checkpoints that cannot be univocally evaluated have a significant positive effect on its page count distribution. Here, we observed that there is a rapid progression of the number of pages for each aggregated rate, with a lower bound of accessibility quality around 50% and a 90% mean value.

When analysing from the perspective of the 100% detectable problems (i.e. just *errors*), we found that there is a near constant distribution of Web pages

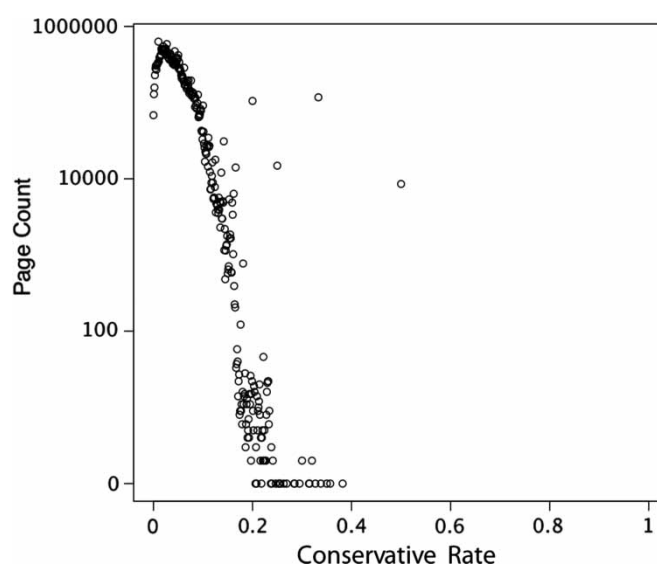


Figure 1. Accessibility distribution for *conservative rate*.

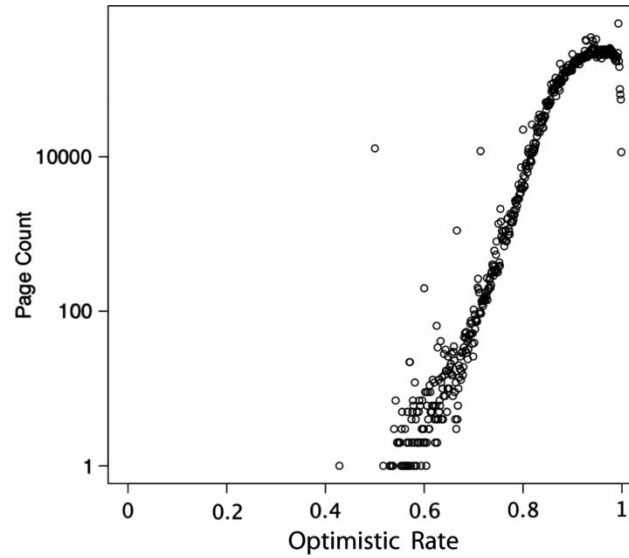


Figure 2. Accessibility distribution for *optimistic rate*.

according to their accessibility quality, as depicted in Figure 3. The only exceptions to this are at the edges of the distribution, especially when approaching full compliance with the detectable errors. Here, the decay on the page count is still measurable, despite the fact that it is less steep comparing to *conservative rate*.

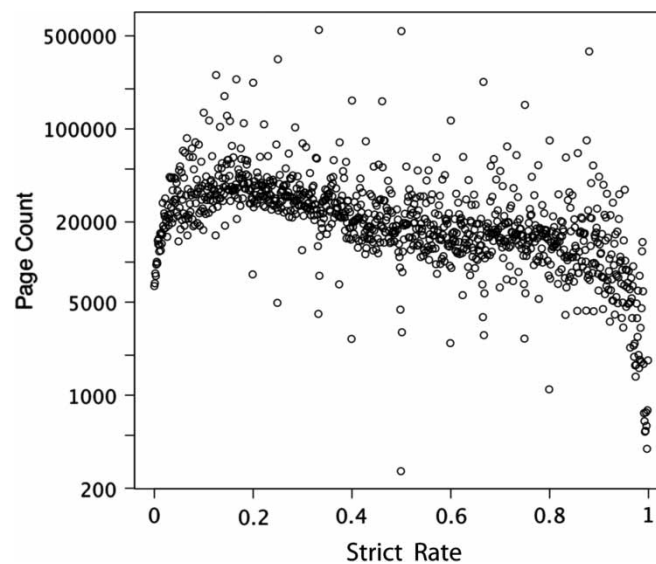


Figure 3. Accessibility distribution for *strict rate*.

#### 4.2 Rates and page complexity

Our second incursion on this study relates to the verification of a correlation between the rate and complexity of each Web page, i.e. the number of HTML elements present in a Web page, encompassing both the breadth and depth of the Web pages' HTML element tree.

Regarding *conservative rate*, with the exponential decay of HTML elements count, the accessibility rate approaches 10% quality, as presented on Figure 4. When taking into account the *optimistic rate* metric, there is no obvious correlation between elements count and accessibility quality, as depicted on Figure 5. Nevertheless, there is homogeneity on the distribution of *optimistic rate* regarding elements count. Lastly, Figure 6 depicts the distribution for the *strict rate* metric. Like in the *conservative rate* metric, we discovered the same kind of exponential decay between elements count and the metric. However, in this case, the rate approaches 100%, since *warnings* were not taken into account.

#### 4.3 Rates and hyperlinking

As discussed earlier, the typical accessibility evaluation of Web pages is performed over a single Web page or a Website. While this type of measurement is important by itself, it does not take into account one of the core aspects of the Web—*hyperlinks*. Based on this assumption, we analysed the accessibility quality of Web pages from the perspective of their *outdegree*—how many outbound links does a Web page contain, as presented below.

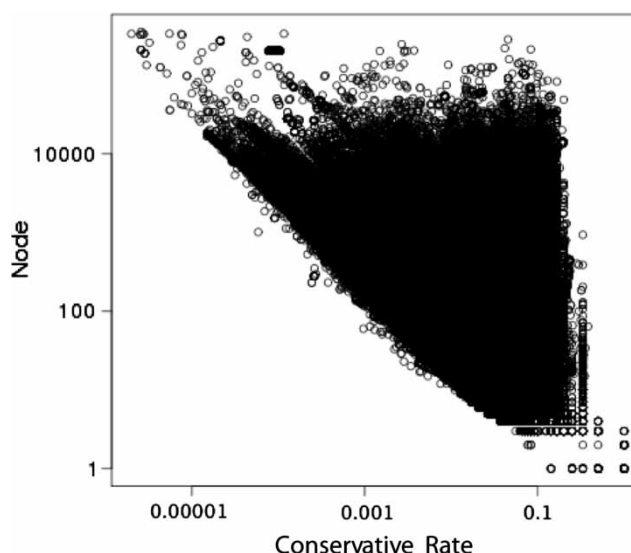


Figure 4. Accessibility *conservative rate* versus page complexity.

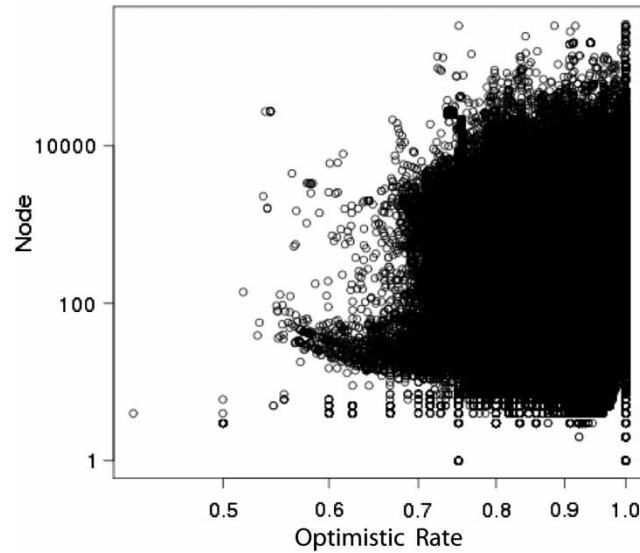


Figure 5. Accessibility *optimistic rate* versus page complexity.

This resulted in the discovery that Web pages having few links tend to have a smaller accessibility quality, a property that manifests similarly in the three rates. When accessibility quality improves, the variability of the outdegree increases proportionally. In other words, Web pages with a lot of links have a high variability on their accessibility quality, but those with few links tend to have their accessibility quality worsened.

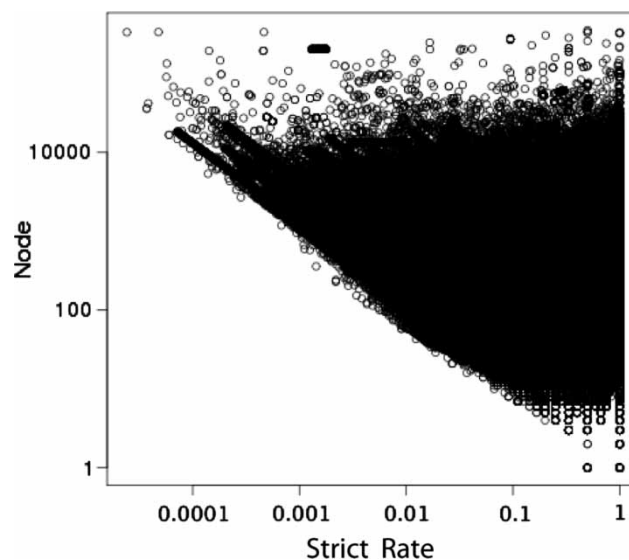


Figure 6. Accessibility *strict rate* versus page complexity.

#### 4.4 Alternative texts

While it is important to characterise Web accessibility at the large based on consolidating the results from all checkpoints, it is also relevant to analyse the shape of single checkpoints at a large scale. In this Section we discuss one particularly important checkpoint from WCAG 1.0, Checkpoint 1.1, which states: “Provide a text equivalent for every non-text element (e.g. via alt, longdesc or in element content)”.

By principle, all checkpoints are of uttermost importance, for accessibility guideline compliance. However, Checkpoint 1.1 deals with both *the ability to access information*—accessibility on its stricter sense (e.g. people with blindness are unable perceive the information depicted on an image)—and affords easily detectable success criteria (*alt* or *longdesc* attributes, or in element text).

We applied the same three rate metrics to the results from WCAG 1.0 Checkpoint 1.1. Here, we also wanted to understand the distribution of the rates, and how it maps into Web page count for each rate aggregation, as presented in the Figures 7, 8 and 9 (log-log distribution).

In the case of the *conservative rate* distribution, as depicted in Figure 10, we verified that page count decreases exponentially with the increase on Checkpoint 1.1 compliance, closely resembling a power-law distribution. When going into the *optimistic rate* distribution, the opposite situation occurs. There is an exponential increase between page count and the rate distribution due to *warnings* being accounted as positive assertions of Checkpoint 1.1, as depicted in Figure 11. Lastly, regarding the *strict rate*

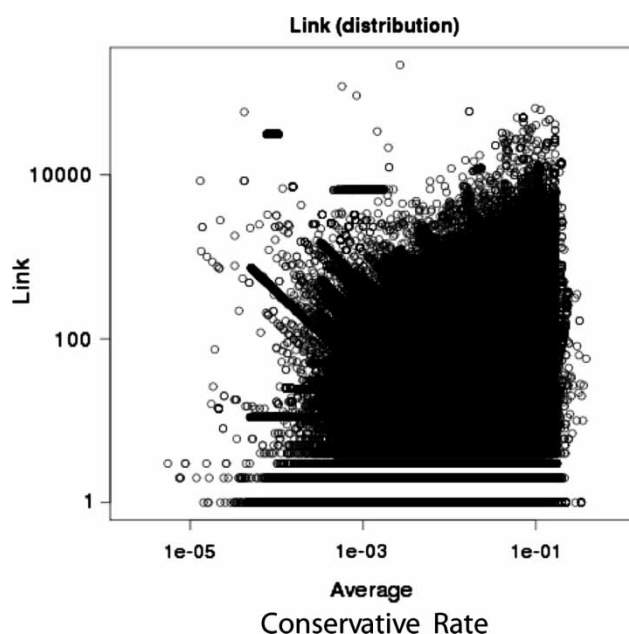


Figure 7. Accesibility *conservative rate* versus outdegree.

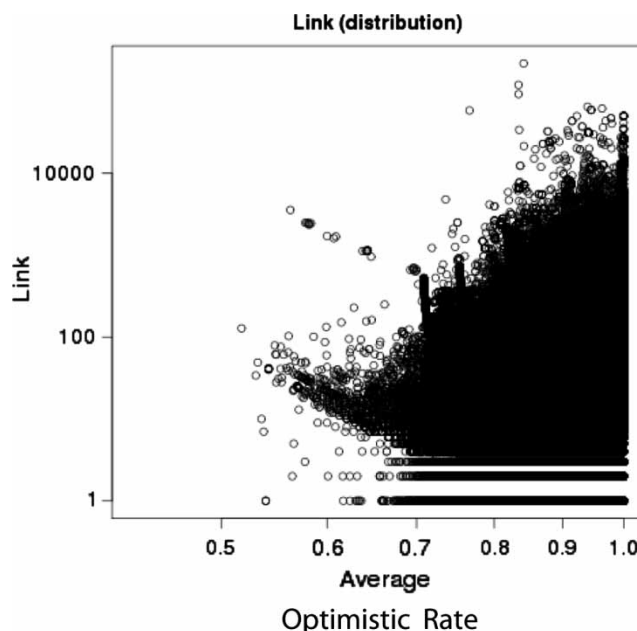


Figure 8. Accessibility *optimistic rate* versus outdegree.

distribution, *errors* are distributed in a less steep path in comparison with the *conservative rate*, as presented in Figure 12. The page count variability is stable until quality rate reaches approximately 5%, where page count variability increases proportionally to the quality rate.

## 5. Discussion

One of the interesting aspects arising from this experiment is the distribution of *conservative*, *optimistic* and *strict* rates. When looking at *errors* distribution (i.e. *strict rate* metric), its linearity implies that critical accessibility problems are likely to be encountered with the same probability by end users who depend on proper accessibility. However, when taking into account *warnings*, the picture of accessibility on the Web is not clear. When *warnings* are perceived as positive, accessibility quality quickly reaches high levels. But, as discussed by Vigo *et al.* (2007), the rate of warnings goes hand in hand with error rate. This result has the direct consequence that indeed such small-scale studies are verified at the large scale.

Another important result from this experiment concerns the relationship between the number of HTML elements in a Web page and its accessibility quality. While applying the *optimistic rate* is insufficient to reach a significant conclusion, the *conservative rate* and *strict rate* both provide further insights of Web accessibility. In both cases, we discovered that a high number of HTML elements on a Web page have a strong correlation with its accessibility

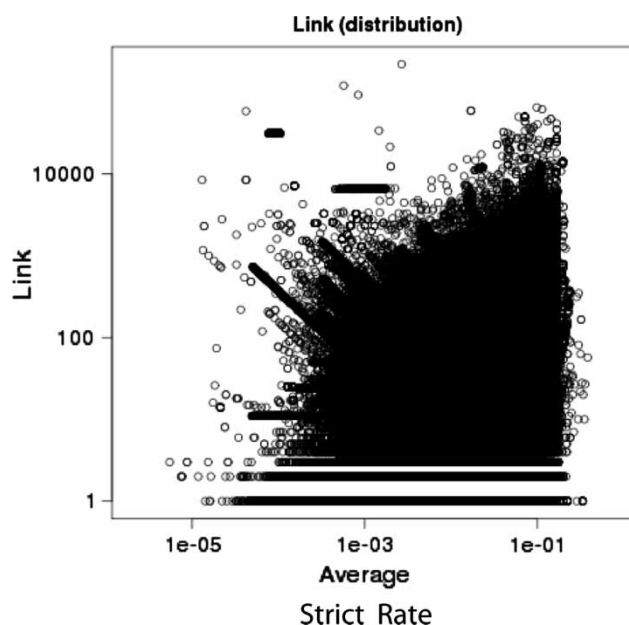


Figure 9. Accessibility *strict rate* versus outdegree.

quality. There was no single Web page in the evaluated document collection that had both a small HTML element count and a poor accessibility quality rate. We hypothesise that this happens due to the complexity of Web pages: simplicity leaves out several HTML structural compositions that hinder accessibility, and also that smaller Web pages are more manageable while designing/implementing them.

Finally, when going deeper into this evaluation, Checkpoint 1.1 also provides some clues about the state of accessibility on the Web. This checkpoint is constantly being used as an example poster child for accessibility issues, since it is easy to conceptualise the problem behind the checkpoint. Even so, the distribution of its compliance rates implies that the message is not being disseminated enough. In this checkpoint, *warnings* signal those cases where it is not possible to automatically infer if there is a textual equivalent for a given non-textual element.

But for those cases where it is possible to algorithmically decide that a media element does not have a textual equivalent (e.g. the *alt* attribute on an *img* element), the distribution for the *strict rate* also denotes a strong decay in this checkpoint's compliance. Based on this distribution, we argue that Web page creators still have a difficulty with providing accessible contents, even if there are direct mechanisms within the HTML specification to encompass this aspect. Consequently, we believe that the distribution for *conservative rate* on Checkpoint 1.1 is close to reality, since complying with this checkpoint's *warnings* is less obvious due to insufficient support from the HTML specification. Recent advancements in textual analysis of alternative texts, as described by Olsen *et al.* (2010), with the aid of machine-learning



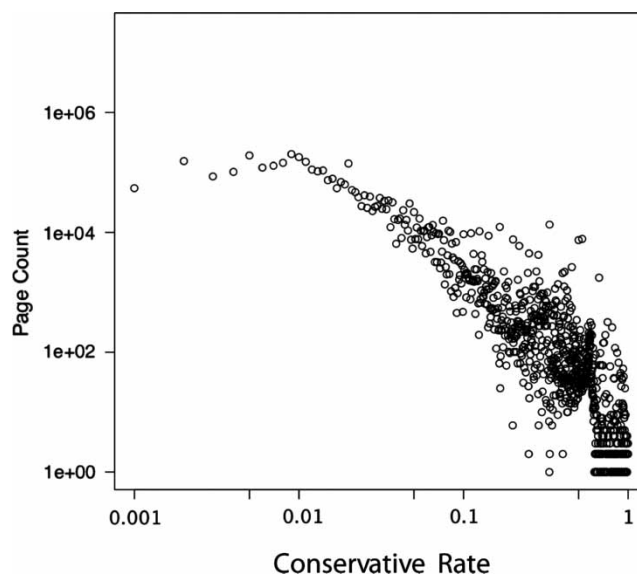


Figure 10. Compliance with WCAG Checkpoint 1.1 based on *conservative rate* evaluation.

techniques, is starting to open the way to better understand this type of issue in evaluating Web accessibility compliance.

### 5.1 Impact on designing accessible Web pages

The results of our experiment can also be discussed with regard to more practical matters, i.e. how people who create Web pages (e.g. designers,

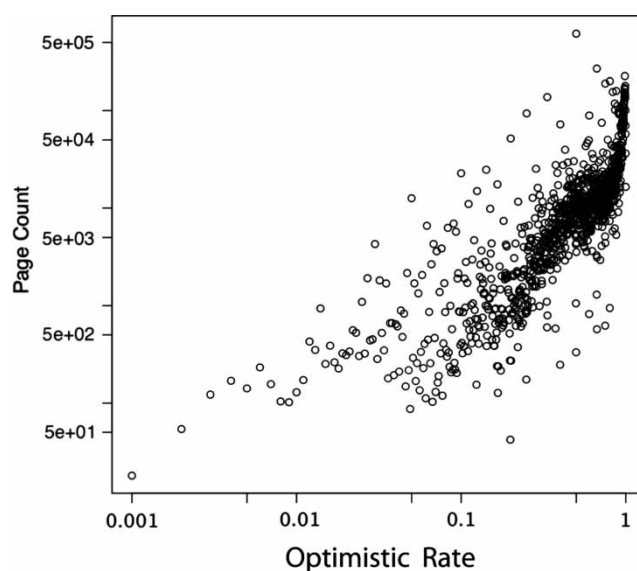


Figure 11. Compliance with WCAG Checkpoint 1.1 based on *optimistic rate* evaluation.

developers, etc.) can mitigate the recurring accessibility problems encountered on the Web.

Paying attention to detail in the structure, rhetoric and discourse of a Web page conveying information is critical for its accessibility success. The *warnings* raised by the evaluation process concern the lack of usage of HTML structural elements that help building the discourse of a Web page. Therefore, we believe that there is a strong need for a better education and dissemination of best practices for properly using the semantics of HTML elements.

Another issue concerns the aforementioned problem of the relationship between Web accessibility quality and the complexity of Web pages. Our position on this issue, with respect to designers and developers, is that Web accessibility is more manageable in *smaller chunks*. Our advice for Web page creators is to follow a *simplicity* approach to defining the structure of Web pages, which lowers the burden of verifying accessibility compliance during development. This goes hand in hand with the usage of templates for lowering the burden of maintaining the quality of a Web page.

## 5.2 Impact on the perception of accessibility

Our results show the profound difference between the opposite perspectives of accessibility given by the *conservative* and *optimistic* rates. Overall, the *conservative* results are in pair with the *strict* analyses performed. This discovery confirms the expectations and model followed by the WAQM accessibility metric (Vigo *et al.* 2007), in that errors and warnings tend to occur proportionally.

On the other hand, when comparing with the *optimistic* rate, it shows that developers and designers might interpret the accessibility quality of the Web

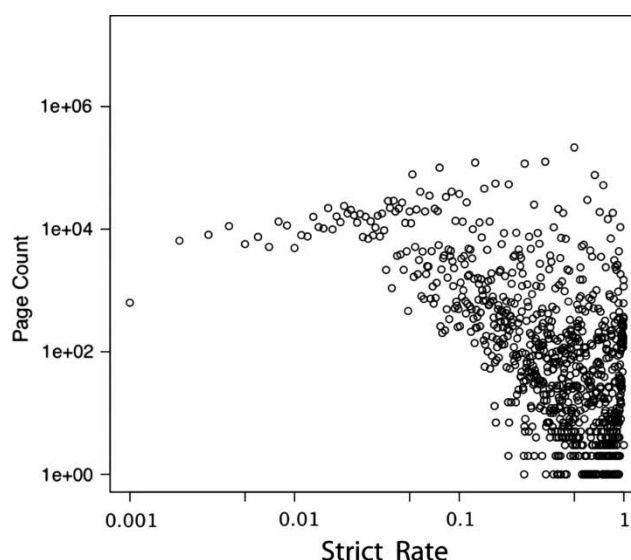


Figure 12. Compliance with WCAG Checkpoint 1.1 based on *strict rate* evaluation.

sites they create, i.e. having an *optimistic* view. Since most developers and designers are not accessibility experts, and since non-experts tend to incorrectly evaluate accessibility (Yesilada *et al.* 2009), we hypothesise that the *optimistic* rate might shed light on the real perception of accessibility by non-experts in the main. As presented in the Results Section, around 90% of the HTML elements that were evaluated yielded a *warning* result. Consequently, non-experts will misinterpret most of the results from evaluation.

This discrepancy further shows that guidelines and encompassing evaluation procedures are just starting points for proper accessibility adequacy. Consequently, we believe that improvements must be made to communicating guidelines and when presenting accessibility evaluation results to motivate developers and designers to investigate the nature of accessibility evaluation *warnings*.

### 5.3 Hyperlinking and accessibility

We believe that accessibility on the Web must be analysed not just in a single Web page (or Web site) perspective, but also by taking into account the relationship between Web pages. This distinction separates existing Web accessibility evaluation studies (microscopic scale) from a Web Science perspective on Web accessibility (macroscopic). In this article we explored one important factor of hyperlinking, the *outdegree* of each Web page, and analysed the relationship with its accessibility quality.

Dill *et al.* (2002) discovered that the Web follows a fractal structure in the way Web pages are linked between each other: strongly connected clusters of Web pages (such as national Web graphs) function as both inbound and outbound hyperlink hubs. The frequency distribution of these connections is scale-free—most Web pages have few links, and are linked by few Web pages—a feature greatly exploited by Web page ranking algorithms (Brin and Page 1998).

The *outdegree* of a Web page reflects if its content tends to be generated (high outdegree) or hand-made (low outdegree), as pointed out by Baeza-Yates *et al.* (2007). Therefore, this metric can be used as a compass to understand whether hand-made Web pages tend to have a higher accessibility quality, in comparison to generated content (e.g. with the aid of content management systems).

Based on these assumptions, our results show that in fact, the outdegree of a Web page provides some guidance on characterising accessibility at the large. For Web pages that have a high number of hyperlinks, the accessibility quality is higher. It is worth noticing that this correlation must not be interpreted as cause-effect. On the contrary, it does show that using mechanisms such as HTML template languages (e.g. used in content management systems) help improving the accessibility of Web sites. Our large-scale experiment confirms this pattern, as we show at a smaller scale (Lopes and Carriço 2008).

The other side of this analysis also provides some insights on the applicability of Web accessibility guidelines by designers and developers. For

Web pages that had a low outdegree, we found that they had a high variability on their accessibility quality. We hypothesise that this reflects the disparity on the expertise of designers and developers: those who hand code Web pages (i.e. detected through the tendency towards lower outdegrees) are part of a diverse group of people with different knowledge about accessibility.

#### 5.4 Limitations of the experiment

While the results of our experiment yield new insights over the state of accessibility on the Web, they must be interpreted in the light of its inherent technology limitations. Here we discuss the main limitations on *automated accessibility evaluation*, *Web collections*, and *scalability constraints*.

**5.4.1 Automated accessibility evaluation.** The depth of evaluating Web accessibility with algorithms is shallower than that of expert evaluation or usability evaluation with people with disabilities. Therefore, the results we have presented and discussed are based on the existence of a concrete upper limit of accessibility quality that can be detected, i.e. all implemented checkpoints yield a positive turnout.

Another aspect that limits the results of the experiment concerns the correct implementation of HTML analysis techniques. We based our implementation on existing techniques (i.e. following UWEM's XPath expressions) to ensure that they correctly express their corresponding accessibility guidelines. However, the issue of *correctness/truthfulness* of automated accessibility evaluation—both on existing evaluation software and expert manual verification modes—is still an open issue in the Web accessibility research field.

Furthermore, there might be a difference between the Web page being served—its HTML and associated resources—and its rendered layout on a Web browser, which is what the user is really going to perceive and interact with. The flexibility of CSS allows Web designers to create complex visual layouts for a given HTML file, which can result in a different rhetoric being conveyed on a Web page, such as wrongly styling some text as a header, while not having the HTML source reflect this. While the process of computing the styling of a Web is automatable (Lie 2005) (*Web browsers do it*), it has an inherent cost—computational complexity—that hinders the essential scalability properties of large-scale evaluations.

Along the same lines, the ever-increasing use of Javascript has also created limitation with respect to the *true* content of a Web page. Techniques such as AJAX allow changing the structure of a Web page and, consequently, altering its accessibility quality. Hence, to capture this, an automated evaluator would have to execute all scripts, in order to reach the same content result. This process in itself is also difficult to scale due to computational complexity issues, such as long-running scripts, the *halting problem*, etc. Nevertheless, recent proposals on making AJAX crawlable might help to address this issue (Probst *et al.* 2009).

**5.4.2 Web collections.** As explained earlier, the document collections were crawled from the Web using traditional techniques such as *spiders*. Consequently, the evaluated document collections are also limited by Web page crawling capabilities (named *spider traps*), including: difficulty of reaching the *deep Web* through HTML forms and AJAX, infinite generation of Web pages through server-side scripts, *robots.txt* exclusion protocol, etc.

Another issue regarding the document collection that has been evaluated relates to its national nature. It has been studied that national Web collections typically follow the trends, structure, etc. of the Web in general (Baeza-Yates *et al.* 2007, Gomes *et al.* 2008). Therefore, we believe that the characterisation bias of the conducted experiment is not statistically significant.

**5.4.3 Scalability constraints.** Due to the way document collections were created and maintained (especially due to its large size), the architectural aspects that allow evaluation processes to scale freely also pose hindrances onto the *cleverness* of the evaluation algorithms. One of such instances relates to associated resources—images, external scripts and applets—and its inclusion on the evaluation process. While more checkpoints could be implemented in an automated evaluation fashion, the dependency on other content poses severe limitations to the free scalability of the system. Therefore, we opted for discarding those checkpoints that rely on external resources.

## 6. Conclusions and future work

This article presented a large-scale study of accessibility on the Web conducted over a Web document collection provided by the PWA. The results of this study revealed a set of characterisations of Web accessibility, both on how it impacts users and how Web page creation hinders it. We have discovered the effects of Web page quality with respect to accessibility, and how it hinders the expected universality aspects of the Web. One of the aspects studied leveraged the confirmation that simpler, smaller Web pages tend to have a better accessibility quality. We hypothesise that this is due to providing less margin of error for Web designers and developers. Our results also show that accessibility communication must be further improved. This was also shown through the poor compliance levels of WCAG Checkpoint 1.1 (alternative texts for media elements), as well as on the disparity between *conservative* and *optimistic* perspectives over Web accessibility evaluation results. This study also uncovered how the fundamental hyperlinking properties of the Web differ when taking account accessibility.

Through the QualWeb research project,<sup>7</sup> we are devising a series of studies about the accessibility quality of the Web and, therefore, ongoing work is being conducted in studying different facets of evaluation of accessibility at large scale, including:

- compare document collections from different years to study the evolution of the Web with respect to accessibility (lack of) compliance, in order to

provide more clues for educating Web stakeholders on the issues of accessibility;

- study personalised Web accessibility, to discover the differences of accessibility compliance for different users (e.g. *is the Web more accessible to people with blindness than to those who are partially sighted?*);
- study vertical cross-cuts of document collections, such as site aggregation, government, online shopping, news sources, countries, etc.;
- expand our studies to larger Web *corpora*, including multi-country and other general purpose collections; and
- study different accessibility evaluation techniques (other guidelines, algorithms, scoring formulas, etc.) and compare their effect at the large scale.

Since, to our knowledge, this is one of the first large-scale experiments on automatically evaluating the state of accessibility on the Web, there is still a lot of room to improve on the conducted research. Hence, we will continue to work on the challenge of improving the accuracy of the automated evaluation process of our work, to mitigate the limitations of the experiments raised in the previous Section. Finally, we are also working in conjunction with the PWA in order to make available all data sets from the evaluation, as well as open sourcing the implemented evaluation framework.

### Acknowledgements

This work was funded by Fundagao para a Ciencia e Tecnologia (FCT) through scholarship SFRH/BD/29150/2006 and the QualWeb national research project PTDC/EIA-EIA/105079/2008, the Multiannual Funding Programme and POSC/EU.

### Notes

- [1] Web Accessibility Initiative: <http://www.w3.org/WAI/>
- [2] World Wide Web Consortium: <http://www.w3.org/>
- [3] Foundation for National Scientific Computing: <http://www.fccn.pt>
- [4] Groovy programming language: <http://www.groovy.codehaus.org/>
- [5] NekoHTML parser: <http://nekohtml.sourceforge.net/>
- [6] R Statistical Computing: <http://www.r-project.org/>
- [7] QualWeb research project: <http://hcim.di.fc.ul.pt/wiki/QualWeb>

### References

- Apache Foundation, 2010. Hadoop. Available online at: <http://www.hadoop.apache.org/> (accessed 9 November 2010).
- R. Baeza-Yates, C. Castillo and E.N. Efthimiadis, “Characterization of national Web domains”, *ACM Transactions on Internet Technology*, 7(2), pp. 1–33, 2007.
- T. Berners-Lee, W. Hall, J.A. Hendler, K. O’Hara, N. Shadbolt and D.J. Weitzner, “A framework for Web science”, *Foundations and Trends in Web Science*, 1(1), pp. 1–130, 2006.
- C. Böhler, H. Heck, O. Perlick, A. Nietzio and N. Ulltveit-Moe, “Interpreting results from large scale automatic evaluation of Web accessibility”, in: K. Miesenberger, J. Klaus, W. Zagler and A. Karshmer (Eds.) *Computers Helping People with Special Needs*, Berlin: Springer, pp. 184–191, 2006.

- G. Brajnik, A. Mulas and C. Pitton, "Effects of sampling methods on Web accessibility evaluations", in *Assets'07: Proceedings of the 9th International ACM SIGACCESS Incollection on Computers and Accessibility*, New York, NY: ACM, pp. 59–66, 2007.
- S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", in *WWW7: Proceedings of the Seventh International Incollection on World Wide Web 7*, Amsterdam, The Netherlands: Elsevier Science Publishers B. V., pp. 107–117, 1998.
- M. Burner and B. Kahle, 1996. Arc file format. Technical report, The Internet Archive. Available online at: <http://www.archive.org/web/researcher/ArcFileFormat.php> (accessed 9 November 2010).
- B. Caldwell, M. Cooper, W. Chisholm, L.G. Reid and G. Vanderheiden, 2008. Web content accessibility guidelines 2.0. W3C Recommendation, World Wide Web Consortium (W3C). Available online at: <http://www.w3.org/TR/WCAG20/> (accessed 9 November 2010).
- W. Chisholm, G. Vanderheiden and I. Jacobs, 1999. Web content accessibility guidelines 1.0. W3C Recommendation, World Wide Web Consortium (W3C). Available online at: <http://www.w3.org/TR/WCAG10/> (accessed 9 November 2010).
- N. Christensen, "Preserving the bits of the Danish Web", in 5th International Web Archiving Workshop (IWAW05), 2005.
- S. Chung, D. Shiowattana, P. Dmitriev and S. Chan, "The web of nations", in *WWW'09: Proceedings of the 18th International Incollection on World Wide Web*, New York, NY: ACM, pp. 1159–1160, 2009.
- J. Clark and S.J. DeRose, 1999. XML path language (XPath) Version 1.0. W3C Recommendation, World Wide Web Consortium (W3C). Available online at: <http://www.w3.org/TR/xpath> (accessed 9 November 2010).
- J. Dean and S. Ghemawat, 2004. MapReduce: Simplified data processing on large cluster. OSDI. Available online at: <http://www.labs.google.com/papers/mapreduce-osdi2004.pdf> (accessed 9 November 2010).
- S. Dill, R. Kumar, K.S. Mccurley, S. Rajagopalan, D. Sivakumar and A. Tomkins, "Self-similarity in the Web", *ACM Transactions on Internet Technology*, 2(3), pp. 205–223, 2002.
- D. Gomes, A. Nogueira, J. Miranda and M. Costa, *Introducing the Portuguese Web Archive Initiative*, Aarhus, Denmark, 2008.
- D. Gomes and M.J. Silva, *Modelling Information Persistence on the Web*, Palo Alto, CA and New York, NY: ACM Press, 2006.
- M.R. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork, "On near-uniform URL sampling", in *Proceedings of the 9th International World Wide Web Incollection on Computer Networks: The International Journal of Computer and Telecommunications Networking*. Amsterdam, The Netherlands: North-Holland Publishing, pp. 295–308, 2000.
- A. Heydon and M. Najork, "Mercator: A scalable, extensible Web crawler", *World Wide Web*, 2, pp. 219–229, 1999.
- I. Jacobs and N. Walsh, 2004. Architecture of the World Wide Web, Volume One. W3C Recommendation, World Wide Web Consortium (W3C). Available online at: <http://www.w3.org/TR/webarch/> (accessed 9 November 2010).
- B. Kahle, "The Internet archive", *RLG Diginews*, 6(3), 2002.
- M. Koster, 1994. A standard for robot exclusion. Technical report. Available online at: <http://www.robotstxt.org/wc/robots.html> (accessed 9 November 2010).
- H.W. Lie, 2005. CSS3 module: Cascading and inheritance. W3C Working Draft, World Wide Web Consortium (W3C). Available online at: <http://www.w3.org/TR/css3-cascade/> (accessed 9 November 2010).
- R. Lopes and L. Carriço, "The impact of accessibility assessment in macro scale universal usability studies of the Web", in *W4A '08: Proceedings of the 2008 International Cross-Disciplinary Incollection on Web Accessibility (W4A)*, New York, NY: ACM, pp. 5–14, 2008.
- J. Miranda and D. Gomes, *An Updated Portrait of the Portuguese Web*, Aveiro, Portugal, 2009.
- S. Mirri, L.A. Muratori, M. Rocchetti and P. Salomoni, "Metrics for accessibility on the Vamola project", in *W4A '09: Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibility (W4A)*, New York, NY: ACM, pp. 142–145, 2009.
- G. Mohr, M. Kimpton, M. Stack and I. Ranitovic, "Introduction to her-itrix, an archival quality Web crawler", in 4th International Web Archiving Workshop (IWAW04), Bath, UK. 2004.
- A. Ntoulas, J. Cho and C. Olston, *What's New on the Web? The Evolution of the Web from a Search Engine Perspective*, New York, NY: ACM Press, 2004.

- M. Olsen, M. Snaprud and A. Nietzio, "Automatic checking of alternative texts on Web pages", in K. Miesenberger, J. Klaus, W. Zagler and A. Karshmer (Eds.), *Computers Helping People with Special Needs*, Berlin: Springer, pp. 425–432, 2010.
- M. Philips, "PANDORA, Australia's Web archive, and the digital archiving system that supports it", *DigitCULT.info*, p. 24, 2003.
- K. Probst, B. Johnson, A. Mukherjee, E. van der Poel and L. Xiao, 2009. A proposal for making AJAX crawlable. Technical report, Google. Available online at: <http://www.googlewebmastercentral.blogspot.com/2009/10/proposal-for-making-ajax-crawlable.html> (accessed 9 November 2010).
- D. Sloan, A. Heath, F. Hamilton, B. Kelly, H. Petrie and L. Phipps, "Contextual Web accessibility – maximizing the benefit of accessibility guidelines", in *W4A '06: Proceedings of the 2006 International Cross-Disciplinary Workshop on Web Accessibility (W4A)*, New York, NY: ACM, pp. 121–131, 2006.
- T. Sullivan and R. Matson, "Barriers to use: Usability and content accessibility on the Web's most popular sites", in *CUU' 00: Proceedings on the 2000 Incollection on Universal Usability*, New York, NY: ACM, pp. 139–144, 2000.
- E. Velleman, C. Meerveld, C. Strobbe, J. Koch, C.A. Velasco, M. Snaprud and A. Nietzio, Unified Web Evaluation Methodology (UWEM 1.2), 2007.
- M. Vigo, M. Arrue, G. Brajnik, R. Lomuscio and J. Abascal, "Quantitative metrics for measuring Web accessibility", in *W4A '07: Proceedings of the 2007 International Cross-Disciplinary Incollection on Web Accessibility (W4A)*, New York, NY: ACM, pp. 99–107, 2007.
- Y. Yesilada, G. Brajnik and S. Harper, "How much does expertise matter? A barrier walkthrough study with experts and non-experts", in *Assets '09: Proceedings of the 11th International ACM SIGACCESS Incollection on Computers and Accessibility*, New York, NY: ACM, pp. 203–210, 2009.
- X. Zeng, *Evaluation and enhancement of Web content accessibility for persons with disabilities*, PhD Thesis, University of Pittsburgh, USA, 2004.