I.J. Information Technology and Computer Science, **2017, 9, 18-44**
Published Online September 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2017.09.03

Towards Automated Web Accessibility Evaluation: A Comparative Study

Siddikjon Gaibullojonovich Abduganiev [1,2]
[1] Lecturer, Khujand Polytechnic Institute of Tajik Technical University named after academician M.S. Osimi, 735700 Khujand, Tajikistan
[2] Ph.D. student, Johannes Kepler University, 4040 Linz, Austria
E-mail: abduganievsg@gmail.com

Received: 20 March 2017; Accepted: 08 June 2017; Published: 08 September 2017

Abstract—With each passing day, the Web is becoming increasingly important in our lives. Hence, the need of making it more accessible to everyone, especially for the disabled and elderly spurred a great interest in automated tools, the total registered number of which has been continuously increasing and reached from forty-five software bids in 2014 to ninety-three in 2017. The purpose of this empirical research is to assess and compare eight popular and free online automated Web accessibility evaluation tools (AWAETs) such as AChecker, Cynthia Says, EIII Checker, MAUVE, SortSite, TAW, Tenon and WAVE with regard to the WCAG 2.0 conformance. As a result, significant differences were observed in terms of tool's coverage (a maximum of 32.4%), completeness (ranges between 10% and 59%), correctness (an average of 70.7%), specificity (reaches 32%), inter-reliability (lies between 1.56% and 18.32%) and intra-reliability (the acceptable score), validity, efficiency and capacity. These eight criteria can help to determine a new role played by modern AWAETs as dependent methods in Web accessibility evaluation. Moreover, consequences of relying on AWAETs alone are quantified and concluded that applying such approaches is a great mistake since subjective and less frequent objective success criteria (SC) failed to be automated. However, using a good combination of AWAETs is highly recommended as overall results in all the mentioned quality criteria are maximized and tools could definitely validate and complete each other. Ultimately, integrating automated methods with the others is ideal and preferably at an early stage of the website development life cycle. The study also provides potential accessibility barriers that make websites inaccessible, challenges AWAETs are currently facing, nineteen pros and fourteen cons and fifteen improvement recommendations for the existing and next generation of AWAETs. Fundamentally, achieving the objectives of this study was possible due to the elaboration and implementation of a new five-phased methodology named as "5PhM-for-AWAEMs" for successful selection, evaluation and/or comparison of AWAEMs. In addition to providing detailed descriptions of the estimation process, this methodology represents eleven key criteria for effective selection of suitable AWAEMs and necessary numbers of web pages and expert evaluators for acceptable, normal or ideal assessment.

Index Terms—Web accessibility; guidelines and standards; WCAG 1.0 and 2.0; automatic Web accessibility evaluation methods and tools; TP; FP; FN; human-expert review; tool's coverage, completeness, correctness, specificity, inter- and intra-reliability, validity, efficiency and capacity.

I. Introduction

An ever-growing range of Web-based services such as online communications, internet banking, ordering, government services, consultations, job searching and others that do not require leaving home are becoming increasingly important for people with disabilities whose number has been rapidly growing because of the demographic trends and reached to one billion people or 15% of the world's population [1]. However, inaccessible websites exclude this significant segment of the population from their fundamental rights to fully use, benefit and contribute to the Web. Also, websites with poor accessibility lead to decreased credibility [2]. Unfortunately, making the Web accessible for disabled and senior people still remains an urgent human-computer interaction problem despite the existence of numerous accessibility guidelines, the wide availability of conducted studies and free software programs as well as inexpensive solutions. In this regard, incorporating automated Web accessibility evaluation methods (AWAEMs) with the other testing methods to evaluate and ideally improve the accessibility of websites has a great potential and is the best way to address this problem.

The fast-paced growth of the Web imposes new challenges for AWAEMs. On the other hand, AWAEMs are being developed since many years due to the expansion of their track records of mistakes. New AWAEMs are emerging and non-competitive ones leave their place to stronger ones. With this connection, we hypothesize that a holistic picture of the accessibility assessment of dynamic websites by modern AWAEMs

has now changed too. It is, therefore, reasonable to conduct a new study to characterize these changes.

Accordingly, the goals of the paper are:

1. To define what role do modern AWAETs play in verifying accessibility issues and are they more effective than their predecessors.
2. To discover in what degree each opted tool is capable to independently reveal accessibility errors and suggest the best way(s) to improve the efficiency of using AWAETs.
3. What specific problems make websites inaccessible and to what extent they are automated.
4. To represent challenges AWAETs are currently facing, provide a new insight into their pros and cons and propose new theoretical and practical recommendations for improving their functionalities.

To address the above-mentioned goals, eight free AWAETs are benchmarked in the context of eight quality characteristics such as coverage, completeness, correctness, specificity, inter- and intra-reliability, validity, efficiency and capacity. This research has selected WCAG 2.0 for the compliance of the website since it is internationally accepted and implemented by the majority of AWAEMs as the basic standard. In addition, manual expert evaluations were conducted to confirm true and false positives found by each tool as well as identify missed true positives.

The remaining parts of this paper is organized into eight sections: S.II) comprises background information, S.III) gives an overview of obtained results, S.IV) discusses the findings and their implications, S.V) considers pros and cons of AWAETs, S.VI) provides recommendations for improving the quality and functionality of AWAEMs, S.VII) presents limitations and indications for future work and S.VIII) is about summary and concluding remarks.

II. Research Background

A. Accessibility: Terms and Definitions

Accessibility refers to technical, design and organizational barriers that make the use of the Web difficult or impossible for users, mostly the elderly and disabled. Subsequently, making the website accessible involves the removal of all existing barriers that are caused because of inattentive construction of sites [5]. In a broad sense, the term accessibility means that any user regardless of economic, geographic or physical circumstances can easily visit any place of a website [6], understand its content and have full interaction with it by having a well-known browser [7], operating system and device. A website is considered accessible if it can be reached, navigated, controlled [8-11], perceived, operated and understood [11, 13] by every user, regardless of his/her limited opportunities. As specified in Fig. 1, web accessibility is often considered as a crucial subset of usability.

Fig.1. Interrelations between usability, accessibility, manual and automatic tests

B. Web Accessibility Guidelines

Since organizations at various levels such as international levels (W3C/WAI [8], [14], ISO IS 9241-171 [15] and TS 16071 [16]), state or national levels (e.g. the state ICT development program (Tajikistan, [17,18]) and National Action Plan on Disability (Austria, [19]) and individual organizations (e.g. IBM, Microsoft, SUN or SAP) have undertaken to develop standards/guidelines for accessibility estimations, the concept of multiple Web accessibility guidelines has appeared. Basically, these guidelines share the same idea, but with small distinctions that need to be addressed in specific conditions when building websites. Further, over the last decades, accessibility became a legal requirement for all. In 2006, the UN Assembly passed a Treaty on Rights of Disabled that prohibits all kinds of discriminations and guarantees an equal access to the ICT for the disabled [20]. Thus, besides adopting international Web accessibility standards, most countries and communities across the world have enacted or are in the process of adopting their own accessibility legislations and policies for online content and sites. These legislations are mainly based on WCAG 1.0 or 2.0.

C. Web Accessibility Evaluation Methods (WAEMs)

Web accessibility evaluation is a process of measuring how well the website is accessible to people with various degrees of limitations [21] and if the disabled can also use the site with the same efficiency as people without disabilities [22]. Many attempts exist to perform estimations through several methods, including the standard's review, user testing, subjective assessment, screening technique and barrier/cognitive walkthrough [22-26]. However, there is no agreement regarding the best evaluation method that would guarantee the detection of all possible problems. Besides, their effectiveness is not yet proven and yet different techniques reveal different accessibility problems. Next, conformance review, which is also known as guideline, standard or expert review or manual inspections [25, 26], is the most widespread method [27]. It inspects web pages with predetermined checklists of guidelines and thus, depends on chosen guidelines. In essence, owing to the development of software, the conformance review approach can be automatized, which is considered as an important phenomenon in the field of accessibility evaluation.

D. Automated Web Accessibility Evaluation Methods (AWAEMs)

The automated accessibility testing is a fairly new method, which has a new phase of development since the publication of WCAG 1.0 in 1999. The AWAEM aims to automate the process of evaluation and keep websites compliant with Web accessibility regulations. They inspect the accessibility of source code written in (X)HTML, CSS and JavaScript as well as XML and hyperlinks in HTML/XML documents, determine a compliance level of the website with specific accessibility guidelines, examine web server productivity and logs of sites and optimize the site for search engines.

AWAEMs can be categorized in a number of different ways on the basis of their location (on local computers or servers), analysis of concrete sets of standards, cost of evaluation (free or commercial), platform (within a browser or authoring software, online services or offline), repair functionality (evaluation only or evaluation and repair), scope of evaluation (inspection of one page, many or all pages of a site at one time or specific items from the perspective of a group of disabled individuals) and report styles (text and/or graphic based to highlight accessibility barriers or machine-readable formats).

E. Related Works

Our extensive review of the early and recent published works carried out in the field of evaluative and comparative analysis of AWAEMs revealed that they are few in numbers and there is very little evidence so far about AWAEM's features such as efficiency, coverage, completeness, correctness, inter- and intra-reliability, validity, efficiency, capacity and others. Usually, studies have utilized different methodologies to measure the accessibility level of various websites. From a small number of the early studies, Ivory and her colleagues [28] explored the effectiveness of such tools as WatchFire Bobby, W3C HTML Validator and UsableNet LIFT from both designers and users' perspectives. The outcome of their study established that the three selected AWAETs were not as effective as expected in helping web developers to improve the accessibility and usability of the site. Nevertheless, these three tools turned out to be good assistants. Then, Brajnik (2004) [29] introduced comparison way for tools through three dimensions of software quality such as correctness, completeness and specificity. He found that LIFT Machine and Bobby were already capable of providing accurate and reliable results despite having more room for improvement. Again, Brajnik (2008) [30] explored various existing methods for evaluating Web accessibility and concluded that all reviewed methods treated context differently. AWAETs are quite popular, but they have to incorporate with manual means to define other non-examined checkpoints. Notably, in a recent paper, an empirical analysis of the state-of-the-art of six AWAETs was conducted by Markel et al. (2013) [31]. This evaluation was aimed at determining the effectiveness of tools by analyzing their coverage, completeness and correctness with regard to the WCAG 2.0 conformance across nine web pages by following ad-hoc sampling techniques. Particularly, damages with the reliance on automated tools alone were calculated and as a result, leaving out human-expert evaluation was not recommended. Other authors Kaur and Dani (2016) [32] conducted a study focused on using four automated tools and the manual way for assessing the adequacy of the mobile Web. The comparison of AWAETs was based on three quality factors i.e. correctness, completeness and coverage with respect to the conformance of Mobile Web Best Practices. Research findings claim that many mobile accessibility guidelines needed improvements because of the rapid growth of mobile technology and design enhancements. Also, device and platform features should be kept in mind when designing mobile testing tools.

Further, studies that compared one AWAET with others will be mentioned. So, Pivetta et al. (2014) [33] evaluated the usability of the ASES tool as compared with WAVE through a heuristic evaluation carried out by the three authors as experts. They showed that ASES was not stable enough and its interface should be redesigned. Al-Khalifa, (2012) [34] presented the first Arabic Web accessibility testing system to examine the accessibility of Arabic websites based on WCAG 2.0 (level A). The new tool was compared with varies tools such as TAW, Worldspace FireEyes, Total Validator, WaaT and AChecker. Consequently, it had distinct evaluation results from the others. Also, the faced difficulties were proper Arabic translations for many technical terms and readability of generated reports. However, it was an initial positive impact in the area of building AWAETs in the Arabic language. Similarly, Kaur (2012) [35] discussed his developed tool named as SITE CHECKER. As a result, this tool was able to validate CSS code, define code to text ratio on sites and find java script errors.

Some similar researches on AWAETs were also presented in the literature. Ashli et al. (2006) [36] measured the reliability of three tools (Lift, Bobby and Ramp) for the compliance with U.S. Section 508. Accordingly, there were substantial discrepancies in the inter- and intra-reliability between the tools, but a better level of inter-reliability could still be achieved. Centeno et al. (2006) [37] focused on assessing Web accessibility by using a mixture of automated (Bobby, Tawdis and WebXACT), manual and semi-automated methods. The researchers pointed out that Bobby and WebXACT could not provide a good automated coverage of WCAG 1.0. Also, U.S. Section 508 and WCAG 1.0 guidelines had a lot of common rules. Next, Xiong et al. (2007) [38] provided insights in using AWAETs for taking care of Web accessibility at the different stages of the development life cycle and demonstrated that tools are capable of evaluating some guidelines in early phases. Al-Ahmad et.al (2010) [39] compared five AWAETs and concluded that they did not cover all accessibility issues and provided results that may mislead developers. Last but not least, Akgül and Vatansever (2016) [40] evaluated the accessibility of thirty Turkish metropolitan municipal sites by the

disabled and employed the TAW tool. In conclusion, their investigations showed that most of the sample sites failed to follow the WCAG 2.0 guidelines. The most commonly detected barriers were of two types such as lack of suitable alternative text for non-text items and use of tags to create visual presentations.

F. A new Methodology and its Application

The author proposes a new methodology named as "The five-phased methodology for successful selection, evaluation and/or comparison of automated Web accessibility evaluation methods (AWAEMs) when analyzing Web accessibility" or shortly "5PhM-for-AWAEMs" that could be employed in the life cycle of an automated Web accessibility testing process, as given in Fig. 2.

The aims of the elaborated methodology specified in Fig. 2 are twofold. First, it enables successful selection of various AWAEMs in association with manual and other testing techniques to effectively estimate the accessibility of the website. Second, 5PhM-for-AWAEMs helps to gain improved assessment and/or comparison of AWAEMs in order to define how well they are comprehensive, complete, correct, specific, inter- and intra-reliable, valid, effective, capable and etc. in identifying Web accessibility issues. Moreover, within the framework of the new methodology, Section III explains structures and strategies of detailed analyses of each of the mentioned criterion. 5PhM-for-AWAEMs assess AWAEMs in terms of whether a single or multiple case study where a large number of the same sites should be analyzed by one or many different AWAEMs.

After analyzing quite a large number of the mentioned quality dimensions, the novel method gives assessors an ability to classify AWAEM's pros and cons and emphasize requirements for the improvement of AWAEMs themselves and their next generations. However, the application of 5PhM-for-AWAEMs will be quite time and effort consuming if to increase the number of sample sites and AWAETs for analysis. The outcome of assessment can be negatively affected by certain factors such as small numbers of tools and experts, improper choice of AWAETs, poor appraiser's knowledge of accessibility standards/guidelines, a low level of experience in using AWAEMs and lack of appraiser's skill in the field of web accessibility.

Fig.2. Description of 5PhM-for-AWAEMs

The described method 5PhM-for-AWAEMs in Fig. 2 includes the following five integrated phases:

The 1st phase: Planning and Estimation. The statements describing rationales, overall purposes, objectives and expectations of research should be ensured. Hence, the aims and objectives of the paper are completely considered in this phase.

The 2nd phase: Design

- *Selection of accessibility standards and/or guidelines and their interpretation, comparison and limitations*

In this phase, reasons for choosing standards or guidelines in accordance with research topics, questions and scope as well as comparison with other related standards should be outlined. Also, the researcher briefly describes the history of standard's development or other critical moments; characteristics, principles, categories and management practices; scope of application in the fields of science; usage examples in the past studies and frequency of updates, including what's new in the latest versions. The second phase should also contain an acknowledgment of limitations, disadvantages and other specific shortcomings of selected standard/guidelines.

The Web Content Accessibility Guidelines 1.0 (WCAG 1.0) [8] was issued by the world's leading organization – The World Wide Web Consortium (W3C) on May 5, 1999. Later on, it was accepted around the world as the beneficial recommendations for Web accessibility assessment [41] and was used as a landmark in the achievement of the Web accessibility in the many EU Member States [42].

WCAG 1.0 was replaced by a new improved version

called as WCAG 2.0 in December of 2008. WCAG 2.0 is also an ISO standard [43] since September 2012 and the best practice for Website Semantics [44]. It includes the twelve guidelines that are organized into the four specific principles at the top levels. A comparison summary of both standards is given in Table 1 in order to better understand how WCAG 2.0 is structured.

Table 1. Comparison between the two versions of WCAG

WCAG 1.0	WCAG 2.0
---	4 Principles: P-O-U-R
14 Guidelines	12 Guidelines
67 Checkpoints	61 Success Criteria
3 Priority Levels per Checkpoints: Priority 1,2,3	3 Levels per Success Criterion: Level A, AA,AAA
3 Levels of Conformance	5 Requirements for Conformance
Support Techniques	Support Techniques
---	Understanding

Each guideline contains certain testable SC as a basis for determining the compliance of sites with WCAG 2.0. All SC of WCAG 2.0 are classified into the three conformance levels: A (beginner), AA (intermediate) and AAA (advanced). E.g. achieving the level "AA" means that all SC grouped in the conformance levels A and AA are satisfied by a web page (See Fig. 3). Although all levels of SC are equally important, the A and AA levels are deemed fundamental to ensure equal access.

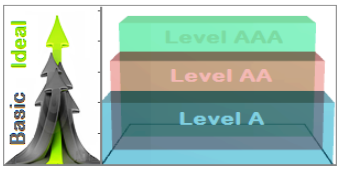


Fig.3. The WCAG 2.0 conformance levels and their dependence on each other

WCAG 2.0 [45] innovatively differs from WCAG 1.0 because it was built based on four general accessibility and universal design principles such as perceivable, operable, understandable and robust, which address the past, present and future of Web technologies.

Yet, WCAG 1.0 and 2.0 have several limitations regarding their validity and testability [46, 47] as well as poorly addressing the needs of people with cognitive impairments [48]. Besides this, only 50.4% of the problems found by the disabled were covered by WCAG 2.0 [49, 50], which means even those websites that meet standards may not be accessible [51, 52]. Basically, guidelines do not address some problems at the high levels like the structural complexity of pages or color combinations for website components. Further shortcomings of accessibility standards and conformance reviews can be found from here [53-56]. However, despite the mentioned problems, WCAG 1.0 and 2.0 aim at being universal accessibility standards [57]. Nowadays, this aim is achieved as they were laid as the foundations of Web accessibility legislations in many countries. Moreover, they remain the commonly utilized guidelines by AWAEMs [58, 59] and considered as a benchmark for analyzing web accessibility. Especially, WCAG 2.0 with its advent has been stimulating the emergence and improvement of new AWAEMs, i.e. the number of tools has increased from forty-five registered tools on December 18th, 2014 to ninety-three on 12th June 2017 [60]. This is because WCAG 2.0 became more testable than its predecessor and considers more advanced web technologies such as SMIL, client/server-side scripting and ARIA. In addition, if a website is constructed in accordance with WCAG 2.0, then it is guaranteed with a good level of accessibility [21, 61], as well as consistency and accuracy.

- *Exploration and collection of a representative sample*

In this type of the multiple case study, any sampling technique from the category of probability, random or specific harms the result of analysis because the possibility of the appearance of homogeneous elements in a sample is very high. Besides, using only home pages is truly incorrect since many other special issues can be found from the other types of web pages. A study by Hackett et al. [62] confirmed that home pages are not a true representation of the entire site or a good indicator of Web accessibility. Therefore, we highly recommend using a purposeful sampling approach to aid the process of multiple case selections. Generally, more detail on this sampling technique can be found in the Patton's (1990, p. 182-183) [63] sixteen purposeful sampling strategies.

Table 2. A minimum number of twenty-two various web pages and their types required for the acceptable Web accessibility assessment

Numbers of pages required	Types of websites
1 web page	Location (across continents or countries)
3 web pages	Accessibility (high, middle, low)
2 web pages	Content (dynamic, static)
8 home pages	Topic (governmental, news, educational, NGO, e-commerce, etc.)
8 web pages	Types of pages (sitemap, contact forms, photo gallery, search results, virtual tour, login, etc.) - can be chosen from the same Topic

Now, we propose a way of the analytical selection of sample web pages for 5PhM-for-AWAEMs. So, the choice of cases or sample units needs to be driven by two important factors such as appropriateness (a fit to the purpose of a study) and adequacy (how many cases). The purpose of a study of this type is conformance review, which could be performed by whether automatic, manual or integrating both of the methods. With regard to the adequacy, a well-chosen data sample for both single and multiple case studies should represent the whole set of issues of selected guidelines. Thus, as described in Table 2, only a large number of sites in terms of location (across continents), accessibility (high, middle and low), content (dynamic and static), topic (governmental, news, e-commerce, etc.) and type of web pages (sitemap,

contact forms, photo gallery, etc.) must be undertaken. In sum, a minimum of twenty-two different web pages is enough to conduct the acceptable evaluation. For the normal assessment, this same set of twenty-two web page types from Table 2 should be selected from two continents or countries, which totally would be forty-four web pages. For the ideal evaluation, a set of twenty-two various web page types, highlighted in Table 2 should be chosen from three, four and etc. continents or countries.

Table 3. A list of fifty-two web pages used in this multiple case study

№	Selected Tajik and Austrian websites	Description
1.	president.tj	Press and Information service of the President of the Republic of Tajikistan
2.	http://president.tj/en/taxonomy/term/5/13	Press and Information service of the President of the Republic of Tajikistan:Decrees
3.	khujand.tj	The city of Khujand
4.	http://www.khujand.tj/feedback/	Public reception: The contact form
5.	tajikembassy.at	The Embassy of Tajikistan in Vienna, Austria
6.	http://tajikembassy.at/index/registration_form/0-37	The Embassy of Tajikistan in Vienna, Austria: Registration Form
7.	somonair.com	Somon Air- The Tajik Air Company
8.	http://booking.somonair.com/oxygen/	Somon Air: Buying Tickets Online
9.	http://www.tajikairlines.com/en/content/passengers/flight-schedules.php#	Tajikistan National Air Carrier: A timetable of the flights
10.	tajikngo.tj	Information Portal of the Tajik NGOs - Internet Community
11.	http://tajikngo.tj/index.php?option=com_k2&view=itemlist&layout=category&task=category&id=5	The list of Tajikistan NGOs
12.	kbtut.tj	Khujand Polytechnic Institute of Tajik Technical University (KPITTU)
13.	http://kbtut.tj/index.php?pg=gallery	The photo-gallery of the Institute
14.	http://nbt.tj	The National Bank of Tajikistan
15.	http://nbt.tj/en/search/map.php	The National Bank of Tajikistan: Sitemap
16.	http://eskhata.com	Bank Eskhata
17.	http://eskhata.com/about/bank/branches.php	Bank Eskhata: The location of all branches and contacts
18.	http://marka.tj	The auto market in Tajikistan
19.	http://marka.tj/default.aspx?&god_ot=1970&god_do=2015&marka=Volkswagen	Search result: “Volkswagen”
20.	http://polyglotclub.com/language/tajik/forum	Tajik language forum
21.	http://polyglotclub.com/language/german/forum	German language forum
22.	http://www.toptj.com/login/	Tajik Information Portal. Ratings of Tajik sites. Tajikistan News: Login page
23.	http://www.ibnisino.tj/tg/darmongoh.html	International Clinic Ibn Sina: Clinic
24.	http://www.ibnisino.tj/tg/dorukhona.html	International Clinic Ibn Sina: Pharmacy photos
25.	http://ict4d.tj/category/vakancii/	ICT4D JOURNAL-News, jobs and events of ICT in Tajikistan and Central Asia: Vacancies
26.	https://www.facebook.com/Ict4dTJ	ICT4D JOURNAL on Facebook
27.	*bundespraesident.at*	*Press and Information Service of the Federal President of the Republic of Austria*
28.	*http://www.bundespraesident.at/historisches/geschichte-der-hofburg/*	*History of Hofburg*
29.	*http://www.linz.at*	*City of Linz*
30.	*http://www.linz.at/tourismus/tourismus.asp*	*Interesting and tips for your visit to Linz*
31.	*http://www.linz.at/zahlen/115_Archiv/*	*Statistical Yearbooks of Linz*
32.	*http://www.apa.at*	*The Austrian Press Agency*
33.	*http://www.apa.at/Site/index.de.html*	*Top theme: The Austrian Press Agency*
34.	*http://www.apa.at/Site/Presse/Pressefotos/APA-Management.de.html*	*APA-Management*
35.	*http://www.austrian.com*	*Austrian Airlines*
36.	*http://www.miles-and-more.com/online/portal/mam/at/homepage?l=de&cid=18001*	*Miles and More - Europe's largest frequent flyer and loyalty program*
37.	*https://book.austrian.com/app/fb.fly?pos=AT&l=de*	*Buying Tickets Online*
38.	*ngo.at*	*The World of NGOs in Austria*
39.	*http://ngo.at/ngos/literatur-und-links*	*Literature and Links*
40.	*http://ngo.at/component/search/?searchword=sitemap&searchphrase=all&Itemid=435*	*Search result: “sitemap”*
41.	*asyl.at*	*The Austrian asylkoordination*
42.	*http://www.asyl.at/links/links.htm*	*Member organizations asylkoordination Austria*
43.	*http://univie.ac.at*	*The University of Vienna*
44.	*http://blog.univie.ac.at/?d=0*	*UNIVIENNA BLOGS*
45.	*www.jku.at*	*Johannes Kepler University Linz (JKU)*
46.	*http://www.jku.at/content/e213/e152*	*JKU: Organization & Structure*
47.	*http://www.jku.at/content/e213/e161/e6998*	*JKU: Campus Map*
48.	*http://www.idv.uni-linz.ac.at/lehre/w16.ssi*	*Department of Data Processing in Social Sciences, Economics and Business: Preview*
49.	*http://www.bankaustria.at*	*Bank Austria*
50.	*http://www.bankaustria.at/mediathek-newsletter.jsp*	*Bank Austria: Newsletter*
51.	*https://mobile.bankaustria.at/IBOA/login.htm?language=*	*Bank Austria: Mobile Banking Login*
52.	*http://www.careesma.at*	*Jobs, Careers, Job Market, Job Search and Student Jobs*

There are two approaches for organizing sampling data: developing own test pages and using already existing real ones. Creating special test pages is time-consuming and labor-intensive due to the high-level requirements of the standards and web technologies as well as considering wider instances of violations from the real world [29]. The advantages of using real sites are based on the facts that they are built in many different ways and consequently, contain numerous real SC; they may have user-generated content, live and on-demand multimedia elements and frequent updates of content, design, services and features.

This study involves selecting a purposeful sample of fifty-two real pages according to the suggested ideal number of web pages described in this 2nd phase of 5PhM-for-AWAEMs. So, a sample of small, medium and large websites - twenty-six Tajik and twenty-six Austrian sites that comprise the best practices in their Web technology platforms were chosen (See Table 3).

- *A process of choosing the best AWAEMs*

The selection of automated methods should be based on the following eleven key criteria:

1. The ability to evaluate sites against opted standards or guidelines. In particular, the amount and types of checkpoints or SC that an AWAEM can adequately address.
2. A high frequency of the use of the AWAEM, being often used in top rankings and scientific studies and having extensive positive feedbacks by a great number of common users, web developers, accessibility testers and experts. A quantitative choice is defined by this scheme: 4 common users = 2 web developers = 2 testers = 1 accessibility expert. That is, if we have two feedbacks from accessibility testers, then it is equivalent to the feedbacks of four common users.
3. A good adaptation level of the AWAEM to the requirements of a study and its integration into Web development environment.
4. Free or commercial versions, including trial or demo periods and the openness of source code. Luckily, the majority of AWAEMs are freely and widely available today.
5. The coverage, completeness, correctness, specificity and reliability of delivered results by the AWAEM. It is possible to choose simultaneously all or some of these criteria for analysis based on research objectives.
6. Effective and efficient evaluation and repair suggestions for inaccessible websites.
7. Incorporating with relevant Web technologies: (x)HTML, CSS, JavaScript, SQL, Java, ASP.NET, PHP, SVG and etc.
8. The amount of pages that are automatically examined at a single-click, including single pages, a group of pages or a whole site; restricted or password protected pages and etc.
9. The AWAEM's type such as the authoring or browser plugin, command line mode, desktop application, mobile application or online software product.
10. Generating a convenient report with highlighting relevant guidelines, violated source code and numbers and percentages of passed and failed accessibility issues.
11. The accessibility of the AWAEM itself.

It should be noted that the eleven crucial criteria are listed by their levels of importance, i.e. the especially important criteria are listed first. Thus, the analyst may stop the selection process if he/she considers that an AWAEM meets some of the firstly listed requirements.

In the mid-1990s, the first online HTML Validator was offered to users at a website called "web techs" [64]. At the present time, a lot of Web accessibility evaluative and reparative tools have been elaborated and their numbers have been growing rapidly over the recent years: 45 tools on December 18th, 2014; 88 tools on March 2016 and 93 tools on June 12th, 2017 [60]. These mostly free AWAETs existed for a long time and have a great historical significance. Unfortunately, there is a study [65] showing that nearly 65% of the tools registered in 2014 seemed to be no longer available and about 50% of the available ones were not using the latest version of WCAG 2.0.

This study utilizes eight most frequently used AWAETs that are often used in the top rankings [66-69] and scientific studies [31-34,37-40,65], have extensive recalls by a great number of users, are able to evaluate sites against W3C WCAG 2.0 and could be used freely. They are AChecker [34, 70, 71], Cynthia Says [72, 73], MAUVE [65, 74], SortSite [75, 76], TAW [37, 77], Tenon [78, 79], EIII Checker [80, 81] and WAVE [82, 83]. On the other hand, the vast majority of the available tools [65] follow WCAG 1.0 that became outdated. The other famous tools, including Accessibility Valet, EvalAccess and FAE were not chosen because of their inability to check websites against the latest guideline version of WCAG 2.0.

The 3rd phase: Automated Inspection and Reporting. This phase involves data collection with the help of AWAEMs from sample sites and saving them to the PC in right file formats. It is advisable to select the file formats that are demanded by opted statistical software packages that can be found from the list made in the next fourth phase. Further, before starting our tests, tools were taught to consider the compliance level AAA despite the fact that the level AA of WCAG 2.0 is the best practically achievable level of the conformance for sites. In principle, the sample pages were tested by the AWAETs in the same way and time to get correct data, avoid changes in websites and reduce the implementation bias. Finally, generated HTML and PDF reports were converted into MS Word and MS Excel formats and saved to PCs for the further analysis.

The 4th phase: Manual Validation and Re-evaluation. In this penultimate phase, human judgment is necessary

to validate the objective data produced by AWAEMs in the third phase. Consequently, TP, TN and FP will be ultimately categorized manually. Besides this, expert re-evaluations of each sample web page must be performed to supplement the automatic analyses by AWAEMs. This is because expert reviews can uncover genuine problems that are unable to be discovered by employing AWAEMs alone. Accordingly, in order to sort automatically generated data into the TP, TN and FP and perform statistical calculations with them so that new discoveries could be made, we suggest some of the best statistical analysis software bids including MS Excel, IBM SPSS, GNU PSPP, AcaStat, Analytica, Develve, EasyFit, Forecast Pro, GAUSS, LIMDEP, MaxStat, NCSS, StatPlus.

It is worth noting that the WCAG guidelines clearly indicate to the accessibility problems that require human-expert judgments, describe the features of auxiliary technologies and ensure the approaches that can help experts to simulate situations. Since expert assessors play a key role in all kinds of WAEMs, including AWAEMs, they need to understand Web technologies and their development tendencies, technical skills and accessibility guidelines, evaluation methods and software, assistive devices and the spectrum of subjective barriers that the disabled or elderly face. Lastly, regarding the number of testers, at least two or three experts are enough in conducting an effective, reliable and valid test, as more experts may also mean more judgment-based disagreements. Moreover, each SC is explained properly with technical examples in WCAG 2.0.

For the implementation of this part of the evaluation, the author of this study as an expert and another expert appraiser have reviewed generated reports by the tools independently to ensure that all produced issues and warnings can be classified as TP, FP and FN as well as validated and counted with high reliability. Also, all sample web pages were re-evaluated manually so that misrepresented violations can be accurately detected. Those very small numbers of specific barriers, the recognition of their belonging to particular groups of SC caused problems, were discussed with other experts. Subsequently, affiliations of all kinds of such TP have been established as a result of finding the consensus.

5th phase: Maintenance and Improvements. Finally, this kind of studies must be re-conducted with optimal organizational conditions and expert capabilities as well as updated guidelines and software features in order to maintain a high quality of Web accessibility. More details about this phase are summarized in Fig. 2 (5th phase).

III. Findings from the Utilization, Evaluation and Comparison of Awaets

AWAEMs can be estimated or compared on the basis of a general quality scope that usually contains a set of quality criteria such as coverage, completeness, correctness, specificity, effectiveness, validity, efficiency, usefulness [84-88], inter-reliability and intra-reliability, capacity and others. These are the necessary indicators of software quality, which are also considered in this article to give an in-depth overview and actually highlight strengths and weaknesses of the opted tools. Further, some of the quality criteria can be divided into sub-groups to be more specific. AWAEMs provide three kinds of information: false positives, true positives and false negatives as a result of their activities. These data are widely used to describe characteristics of the opted AWAETs in terms of eight various quality criteria:

- True positives (TP) - commonly referred to as "precision" and are real accessibility barriers correctly determined or predicted by AWAEMs. It covers the numbers of the correct classifications of barriers and indicates how well an AWAEM can identify only true barriers.
- False positives (FP) – which are also called "false alarms", mean that reported accessibility issues are inaccurate, irrelevant or mistakenly reported. The more imprecise an automated test is, the more likely that FP will be uncovered. FP create noise when expert inspection does not capture them and as a result, they will be accepted as the actual errors (TP). Again, only the human accessibility expert will establish the inaccurateness or irrelevance of detected problems. For instance, a SC called "H30: Providing link text that describes the purpose of a link for anchor elements" was considered as FP, where the text of a link was too long.
- False negatives (FN) – are existing true accessibility problems in a site that an AWAEM cannot reveal. In other words, FN are certain missed TP. E.g. not catching the fault type H42 as TP every time when the first headline in web pages is not H1. Nevertheless, the absence of TP does not mean they absolutely do not exist at all rather they emerge seldom or not at all. Unfortunately, FN can occur in any automation test. When all used AWAEMs have missed TP, they can even go unnoticed, which may cause erroneous consequences. So, additional and different well-established testing technique, e.g. a human expert's visual inspection or user testing is required to improve the outcome of AWAEMs.

To sum up, the Sharipo – Wilk's test (originally restricted to a small sample size, which is less than 50) and the visual inspection of histograms confirmed that all types of collected data that has impacted the results of this study such as TP, FP and FN were not normally distributed, as given in Table 4.

Table 4. Results of the normality statistics for the three variables

The Shapiro-Wilk test										
AWAETs		Statistics						df		Sig.
		TP		FP		FN		TP,FP,FN		TP,FP,FN
		.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj .at
1.	AChecker	,396	,260	,233	,206	,706	,682	45		,000
2.	Cynthia Says	,542	,555	,463	,442	,641	,623	45		,000
3.	Mauve	,451	,467	,458	,486	,614	,567	45		,000
4.	SortSite	,516	,394	,373	,373	,682	,707	45		,000
5.	TAW	,550	,399	,280	,271	,646	,621	45		,000
6.	Tenon	,516	,336	,464	,375	,659	,649	45		,000
7.	EIII Checker	,422	,245	,452	,473	,681	,684	45		,000
8.	WAVE	,378	,266	,301	,185	,660	,683	45		,000

Note: .tj – Tajik sites and .at – Austrian sites.

A. Coverage

Coverage is the extent to which AWAEMs detect numbers of different SC that contain at least one TP. In principle, any accessibility testing activity is based on at least one coverage strategy. In a broad sense, coverage measures scope and readiness of AWAEMs to ensure that preferably all SC are tested. Hence, it is the best indicator for testing completeness and correctness. The following Table 5 explains how thoroughly the numbers and types of SC were caught from our sample websites.

Table 5. Unique WCAG 2.0's SC that are covered by all the tools and Web accessibility experts

№	AWAETs	The violated success criteria	N	%
1.	AChecker	F65,H37,H44,H65,H42,G18,G17,H64,G91,F89,H57,H93	12	16.9
2.	Cynthia Says	F65,H37,F30,H44,H65,H2,H39,H42,H43,H63,H73,G18,F24,C17,C12,G17,H64, H57,H32,H83,G71,H93,G91	23	32.4
3.	Mauve	F65,H37,H44,H65,H42,H63,H73,G140,C12,C14,C20,C21,G91,H57,H32,H93	16	22.5
4.	SortSite	F39,F65,F30,H44,H2,F91,H42,F49,F89,H57,F70,H93,F68	13	18.3
5.	TAW	F65,H37,H44,H2,H42,G140,H71,G90,F89,H57,H32,G91,G134	13	18.3
6.	Tenon	F65,H37,H44,H65,F91,H42,H43,F88,G91,H33,H57,H93,F59	13	18.3
7.	EIII Checker	F65,H37,H44,H65,H71,H64,G91,F89,H57,G134,H93,F59	12	16.9
8.	WAVE	F39,F65,H44,H2,F91,H42,G18,F88,H57	9	12.7
9.	Experts	F3,F38,H45,H85,F81,C30,PDF13,SCR31,G107,H30,H37,F54,G131,H67,H25,H42, H51,G63,G14,G145,G148,G174,G21,G159,G158,G69,G78,G8,G87,SCR,G141,F59	32	45.1

Note: Column names are denoted as "N" - total numbers of SC. "%"- percentages of total numbers of SC that are calculated by the following equation: Coverage=Automation or manual coverage/Total coverage by automation and manual.

Actually, three measurable SC are only partially verified by the AWAETs, i.e. H37 – was counted when the alt attribute of images had no text, H42 – when no H1 heading is used on pages and G91 – when such text fragments were met: "<a href="#"", "#inhalt" or "/373.asp". Table 5 shows that there are large varieties in coverage results from the usage of multiple tools and experts. Out of a total of 71 unique WCAG 2.0's SC violated, a maximum of 39 (54.9%) were found to be covered by using the combination of all AWAETs and 32 (45.1%) additional missed SC were captured by expert efforts (See Table 5). The majority of tools have similar coverage rates of 16.9 -18.3%, while WAVE has the lowest of 12.7% and Cynthia Says reaches the highest of 32.4% (23) SC.

Next, significant differences are observed in the coverage of SC across the principles. According to data of Fig. 4, the scope of coverage varies among tools from a maximum of 27 (28.1%) SC for Perceivable to a minimum of 6 (6.3%) SC for Operable, while around 7 (7.3%) SC belong to each of the other two remaining principles. Furthermore, 5, 1, 2 and 2 SC of the Perceivable, Operable, Understandable and Robust principles, respectively were considered to be covered by at least one TP. In this case, however, some tools could substitute others by adequately addressing the poorly covered principles.

The most frequently observed SC that have led the twenty-six Tajik and twenty-six Austrian websites into troubles are shown in Table 6.

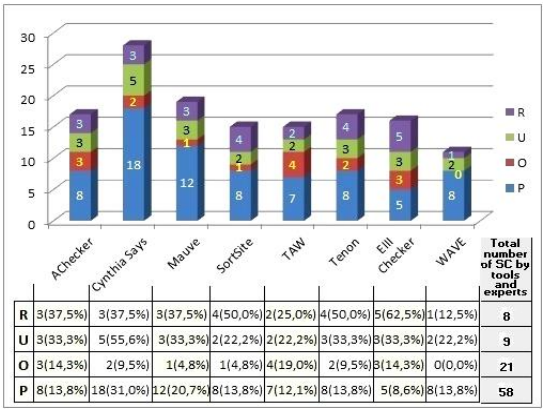


Fig.4. Data on numbers and percentages of SC that were captured by AWAETs at least once

Note: "P" stands for Perceivable, "O" for Operable, "U" for Understandable and "R" for Robust principle. Some SC covered by the tools (e.g. H44, G18, G90, etc.) belong to different principles and levels of WCAG 2.0 and accordingly, represent different conditions. Because of this, their total number was increased from (71 to 96) in Fig. 4.

Table 6. Often violated SC of the fifty-two sites

№	Error categories (Success criteria)	Total issues across types of sites		Total issues
		.tj	.at	
1.	1.4.4, 1.4.5 - C12	483	304	787
2.	1.4.6 - G18	378	380	758
3.	1.4.6, 1.4.8 - F24	291	386	677
4.	1.4.6 - G17	375	223	598
5.	1.1.1 - F65	179	267	446
6.	1.4.4, 1.4.5 - C14	236	173	409
7.	1.4.3 - G18	203	205	408
8.	1.4.8 - C21	195	174	369
9.	4.1.1 - G134	183	143	326
10.	1.4.3 - F24	125	166	291
11.	1.3.1, 1.4.5, 1.4.9 - G140	97	131	228

AWAETs performed well for the SC such as [C12] that has 787 TP, [G18] – 758 TP, [F24] – 677 TP, [G17] – 598 TP, [F65] – 446 TP, [C14] – 409 TP and so forth (See Table 6). On the other hand, frequently encountered accessibility problems determined by the experts and that could not be found with the eight selected tools are: 1.1.1 - [H37] – 179 TP; 1.1.1 - [F38] – 152 TP, 1 1.1.1, 1.2.1 - [H67] – 121 TP, 1.1.1 - [H45] – 81 TP, 1.1.1- [F3] – 68 and 1.4.3 - [G145] - 67 TP. Note that, none of the AWAETs could target the SC like "Guideline 1.2 Time-based Media: Provide alternatives for time-based media" and "Guideline 1.3 Adaptable: Create content that can be presented in different ways without losing information or structure", which significantly benefit not only the disabled but also, the majority of users if not everyone.

B. Completeness

Since AWAEMs are not able to find all potential problems, to consider "completeness" as an important criterion is necessary to evaluate them. Completeness refers to identifying maximum possible and only "true" accessibility faults that exist in a website. Here, the AWAEM should increase numbers of caught TP to get closer to real numbers of existing ones and correctly display them while further reducing FN. If coverage indicates to the length, then completeness points out to the depth of problems captured. Most scholars consider completeness as the most difficult quality feature to measure.

Usually, a SC is recognized to be violated if at least one TP is reported by an AWAEM. A maximum possible number of violations – 7356 from 52 selected websites were caught with the combination of all eight tools: for the Tajik sites, a total of 3962 TP across 39 SC and for the Austrian sites - 3394 TP across the same number of SC were found. Similarly, Fig. 5 presents a description of research findings on the proportion of identified TP and FN in general and Fig. 6 does it across the WCAG 2.0's principles.

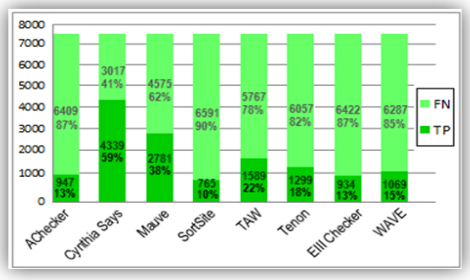


Fig.5. A relationship between the overall numbers and percentages of the generated and missed TP

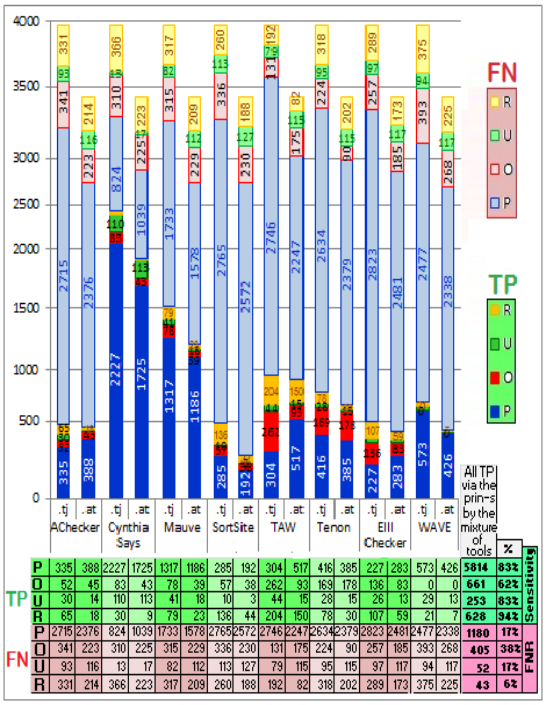


Fig.6. Summary of the differences on completeness across the tools, WCAG 2.0's principles and sites

Note: "P" stands for Perceivable, "O" for Operable, "U" for Understandable and "R" for Robust; ".tj" – Tajik sites and ".at" - Austrian sites; FNR - the false negative rate

The statistical comparison summary provided in Fig. 5 and 6 confirm that not all problems can be automatically discovered. It can be seen that AChecker, SortSite, EIII Checker and WAVE could catch nearly similar, but the lowest percentages of TP, which range from 10% (SortSite) to 15% (WAVE). Conversely, only Cynthia Says got an exceptionally higher number of classified TP – 59% (4339) than the rest tools. The main reason for obtaining the high score for Cynthia Says was argued in the section Specificity.

Sensitivity measures the probability of determining TP. It can be stated as the proportion of TP over a total value of the reported and unreported TP: TP/(TP+FN). The estimated sensitivity grade of the mixture of all the tools in all the WCAG 2.0 principles (the last column of Fig. 6) is around 80% on the average. Furthermore, the FNR for AWAETs can be expressed by subtracting sensitivity from 100. Thus, it is inversely proportional to sensitivity. Consequently, using the combination of all the tools has reached lower FNRs (ranged from 6 to 38% for all principles) than using the tools alone.

Based on Fig. 6, more TP were identified in the Perceivable principle than the others. Yet, even more TP are missed to complete this principle along with the others. Nonetheless, almost all TP - 94% (626) belonging to the Robust principle were completely captured with all the tools together. Likewise, both tools Cynthia Says - with 3952 flagged TP for Perceivable and 223 for Understandable and TAW - with 355 for Operable and 354 for the Robust principle outperform the remaining tools. However, one exception occurs with Tenon, which flagged the highest amount of TP – 178 in the Operable principle for the Austrian sites, instead of the expected TAW tool. On the other hand, Cynthia Says and TAW have the minimum numbers of FN in the same mentioned principles than the rest of tools.

AWAETs report no errors for the principles and conformance levels such as Operable (AA) and Robust (AA, AA), as shown in Table 7. In the same spirit, SortSite and EIII Checker tools could not conduct any automated tests for the Level AA and AAA. In turn, AChecker, Tenon and WAVE were not able to find TP across Operable, Understandable and Robust (Level A and AA). Contrariwise, tools have much higher completeness values in those SC that are considered in the level A of the conformance rather than the levels of AA and AAA in general. Here, Cynthia Says, Mauve and WAVE are exceptions, which demonstrate high completeness in the levels (AA-AAA) too. Another interesting finding is that looking ahead to Table 7 (e. g. the last three columns), the large numbers of SC with the small numbers of TP were produced by the tools for the level A and vice versa for the levels AA and AAA. More detailed assessments in this regard can be found in Table 6.

Table 7. Absolute numbers of TP through the AWAETs, WCAG 2.0 principles and levels of conformance

	AChecker			Cynthia Says			Mauve			SortSite			TAW			Tenon			EIII Checker			WAVE			The combination of tools		
	A	AA	AAA	A	AA	AAA	A	AA	AAA	A	AA	AAA	A	AA	AAA	A	AA	AAA	A	AA	AAA	A	AA	AAA	A(SC)	AA(SC)	AAA(SC)
P	519	16	188	591	1475	1885	903	1231	369	477	0	0	821	0	0	664	0	137	510	0	0	505	408	86	1276(16)	1999(6)	2539(5)
O	97	-	0	126	-	0	117	-	0	95	-	0	285	-	70	347	-	0	219	-	0	0	-	0	558(5)	-	103(1)
U	44	0	0	38	18	167	50	9	0	13	0	0	42	17	0	43		0	39		0	42		0	53(3)	21(2)	179(2)
R	83	-	-	39	-	-	102	-	-	180	-	-	354	-	-	108	-	-	166	-	-	28	-	-	628(7)	-	-
Sum by levels	743	16	188	794	1493	2052	1172	1240	369	765	0	0	1502	17	70	1162	0	137	934	0	0	575	408	86	2515	2020	2821
Total	947			4339			2781			765			1589			1299			934			1069			7356		
Median	90	8	0	83	747	167	110	620	0	138	0	0	320	9	0	228	0	0	193	0	0	35	408	0	593	1010	179

Note: "P" stands for Perceivable, "O" for Operable, "U" for Understandable, "R" for Robust and SC for Success criteria. Also, ".tj" – Tajik sites and ".at"- Austrian sites. The frequency distribution of our data is skewed, as these are values of TP for some SC that is found to be too small or large. In this situation, the median is a better measure of central tendency, which is not as strongly influenced by skewed values as the mean.

In this manner, after grouping TP identified by the eight tools in accordance with the four principles, it became clear that the Perceivable principle shows extremely higher completeness score – 5814 (64.3%) than the rest, as illustrated in Table 8. The Operable and Robust are next, with 1067 (11.8%) and 671 (7.4%) TP respectively. On the contrary, AWAETs exhibit the lowest completeness in the Understandable principle with 253 (2.8%) TP. Subsequently, if to consider completeness per principle, then Perceivable and Operable need more attention for improvement.

If home pages are not accessible, then users will not be able to access other pages. Hence, it is necessary to give a special attention in making them accessible. Our comparative statistical analysis states that the best accessible Tajik home pages were – tnu.tj (23 TP), kbtut.tj (41 TP) and tajikembassy.at (47 TP) and Austrian ones include linz.at (14 TP), ngo.at (28 TP) and jku.at (33 TP). However, Tajik home pages such as tajikngo.tj (397 TP) and toptj.com (161 TP) as well as Austrian ones - asyl.at (496 TP) and careesma.at (153 TP) were champions on the quantity of having barriers.

In principle, the scope of completeness in this research is not limited to the SC that were automatically found by the eight tools. Expert reviews were also accomplished to reveal maximum numbers of existing problems. In this way, experts have additionally discovered the rest of 1681 violations from 49 SC: for Tajik sites a total of 811 TP across 49 SC and for Austrian sites a total of 870 TP across exactly the same 49 SC were found (See Table 6).

As expected, the total amount of TP and SC decreases from the Perceivable (1180 (31SC)) to Robust (43(SC)) principle. No TP were accounted for Understandable and Robust in the levels AA-AAA.

Table 8. Statistics on TP covered by the four principles

№	The four WCAG 2.0 accessibility principles	The sum of TP found by all the tools		Total sum & percentage	The sum of all issues found by experts		Total sum & percentage	Total sum of TP found by the tools and experts
		.tj	.at		.tj	.at		
1.	Perceivable	3050	2764	5814(64.3%)	562	618	1180(13.1%)	6994(77.4%)
2.	Operable	393	268	661(7.3%)	194	212	406(4.5%)	1067(11.8%)
3.	Understandable	123	130	253(2.8%)	31	21	52(0.6%)	305(3.4%)
4.	Robust	396	232	628(6.9%)	24	19	43(0.5%)	671(7.4%)

".tj" – Tajik sites and ".at" - Austrian sites

Table 9. Resulting numbers of TP discovered by experts

	A		AA		AAA		Sum	
	Σ	SC	Σ	SC	Σ	SC	Σ	SC
Perceivable	928	18	149	7	103	6	1180	31
Operable	182	6	73	3	151	6	406	15
Understandable	52	2	0	0	0	0	52	2
Robust	43	1	0	0	0	0	43	1
SUM	**1205**	**27**	**222**	**10**	**254**	**12**	**1681**	**49**
Median	**117**	**4**	**37**	**2**	**52**	**3**	**229**	**9**

Note: "Σ"- total numbers and "SC"- success criteria

The common conclusion of this study about coverage and completeness is that 47 SC and about 7356 (81.4%) TP are checked for definitively with the combination of the eight automated means (See Table 7), while expert analyses revealed the remaining 1681 (18.6%) TP out of 49 SC, as displayed in Table 9. In this way, the transition from the automatic-only to automatic-manual scenario increases both coverage and completeness in numbers due to the fact that applying more heterogeneous test cases means that fewer properties remain unverified.

The calculation of total numbers of TP in order to find out whether Tajikistan or Austrian sites are more inaccessible was difficult because some numbers of TP are close to each other: the sum of the additional barriers explored by human experts is 811 for the Tajik sites and for the Austrian sites is equal to 870. However, if to analyze the completeness of found barriers using all AWAETs, then the numbers of all TP will be changed significantly: Tajik sites will have 3962 TP and Austrian ones will obtain 3394 TP. In both cases, Tajik sites were found to be more inaccessible than Austrian sites.

Now, the relationship between two closely related factors such as quantities of TP and their types (or SC) will be revealed. As Fig. 4-5 suggest, TAW has twice more automated test runs than SortSite, despite the fact that they have the same number of SC, which is 15. Alternatively, WAVE, having the smaller amount of SC than SortSite was able to produce 1.5 times more TP. Similarly, all manual checks with 49 SC are 4.4 times less than the combination of all automated checks with 47 SC (See the last columns of Tables 7 and 9). It can, therefore, be concluded that the number of SC does not affect the number of TP.

C. Correctness

Correctness, also mentioned as precision or accuracy, is the percentage of problems detected by the AWAEM which are indeed true problems. Hence, it refers to how often FP are reported by an AWAEM and thereby, how well it minimizes their quantity. This follows that the accurate AWAEM is the one that finds only true accessibility mistakes. Until today, the only way to characterize the correctness of AWAEMs is to conduct manual tests over automated ones. The following Fig. 7 illustrates evaluation results for correctness.

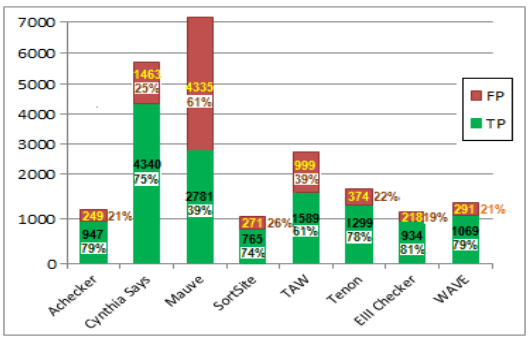
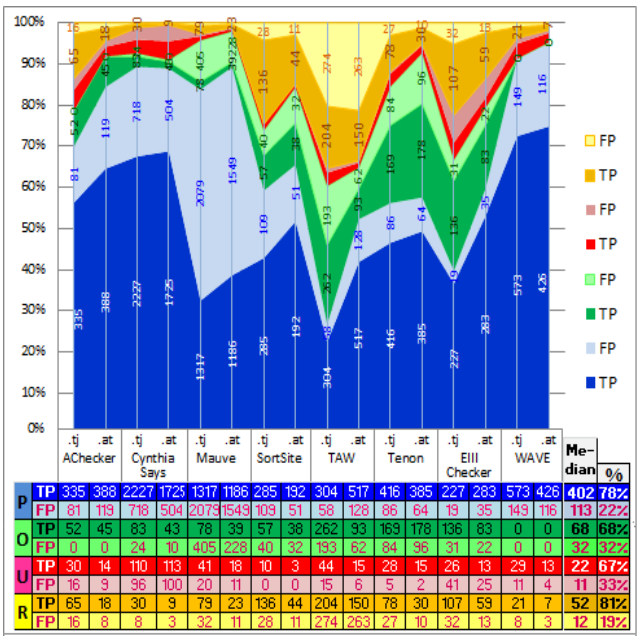


Fig.7. Ratios between the numbers of both correctly and incorrectly identified problems by the eight tools

Note: The percentages of FP in stacked columns represent the incorrectness grades computed as a ratio of FP to a sum of FP and TP reported (incorrectness=FP/(FP + TP)) and in turn, the percentages in columns of TP introduce the precision of tools calculated by subtracting incorrectness from 100.

Generally, all eight tools show 70.7% average level of correctness, while the incorrectness level is 29%. Further, the highest percentage of the correctness for half of the tools such as EIII Checker, AChecker, WAVE and Tenon appears to be quite similar, around 79.5%. The reason why Mauve has a very poor indicator of correctness (39%) is because it categorizes even probable barriers as real. The most serious source of FP in AChecker is due to its failure to check F65; Cynthia Says – failed to check H37, C12; Mauve - H42, C12, C14, C21; TAW - F65; Tenon - H33 and EIII Checker - F65 and H65. For a more expanded review of the correctness aspects, this study yielded the data reflected in Fig. 8.



		.tj AChecker	.at AChecker	.tj Cynthia Says	.at Cynthia Says	.tj Mauve	.at Mauve	.tj SortSite	.at SortSite	.tj TAW	.at TAW	.tj Tenon	.at Tenon	.tj EIII Checker	.at EIII Checker	.tj WAVE	.at WAVE	Median	%
P	TP	335	388	2227	1725	1317	1186	285	192	304	517	416	385	227	283	573	426	402	78%
	FP	81	119	718	504	2079	1549	109	51	58	128	86	64	19	35	149	116	113	22%
O	TP	52	45	83	43	78	39	57	38	262	93	169	178	136	83	0	0	68	68%
	FP	0	0	24	10	405	228	40	32	193	62	84	96	31	22	0	0	32	32%
U	TP	30	14	110	113	41	18	10	3	44	15	28	15	26	13	29	13	22	67%
	FP	16	9	96	100	20	11	0	0	15	6	5	2	41	25	11	4	11	33%
R	TP	65	18	30	9	79	23	136	44	204	150	78	30	107	59	21	7	52	81%
	FP	16	8	8	3	32	11	28	11	274	263	27	10	32	13	8	3	12	19%

Fig.8. Classification results for the correctness across the sites and WCAG 2.0 principles.

Note: "P" stands for Perceivable and etc. ".tj" – Tajik sites and ".at"- Austrian sites. The last column "%" represents the incorrectness and correctness grades respectively.

The median column in Fig. 8 indicates that tools don't produce the same amount of FP across the principles, but from a small percentage of the incorrectness value - 19% for Robust to a high and almost the same percentage of the incorrectness values – 32% and 33% for Operable and Understandable, respectively. Next, the medium of TP values found per principle ranged significantly from 22 (67% of correctness values) for Understandable to 52 (81% of correctness values) for Robust.

It is interesting to note that even if tools uncover more TP, yet they get more FP. In our example, for the Understandable principle - Cynthia Says reports the greatest amount of TP (100 for Tajik and 113 for Austrian sites), but also relatively more FP (96 for Tajik and 100 for Austrian sites), for the Robust principle - TAW reports the greatest amount of TP (204 for Tajik and 150 for Austrian sites), but also more FP (274 for Tajik and 263 for Austrian sites) and Mauve generated the greatest amount of FP in the remaining principles (See Fig. 8). On the contrary, if tools uncover a few TP, then they also receive a few FP. Likewise, our eight selected AWAETs behave this way in general. EIII Checker for Perceivable, WAVE for Operable and SortSite for Understandable define the minimum values of TP and FP too. A graph in Fig. 8 reflects a clear picture of such interrelations.

D. Specificity

Specificity measures the certain ability of the AWAEM expressed in reporting additional or distinctive numbers of those existing true accessibility errors that others cannot. It also considers more details for characterizing problems and providing specific warnings, comments and suggestions. Ensuring maximum specificity is an extremely important aspect of the AWAEM's capability, as it enriches coverage and completeness. Hence, the best AWAEM is the one that has improved the quality of its specificity in addition to determining common errors. Further, the analytical investigations from specificity assessments of the eight AWAETs can be found in the data of Table 10.

Table 10 points out to the big differences in the specificity among the tools. Cynthia Says is relatively the best tool that could reach a good specificity level with around 16.7% (1241) across 6 SC, while SortSite, TAW and Tenon are worse tools with 2.2-3.1% confirmed true positives. Perhaps even worse, AChecker, EIII Checker and WAVE could define no TP within the confines of specificity.

Table 10. An overview of the specifically captured TP by AWAETs

№	Prin-ciples	Levels	Error categories	AChecker		Cynthia Says		Mauve		SortSite		TAW		Tenon		EIII Checker		WAVE		The sum of TP	Percentage of the sum of TP
				.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at		
1.	P	A	1.3.1 - H39			25	13													38	0.5%
2.	P	A	1.3.2 - F49							32	11									43	0.6%
3.	P	AA	1.4.3 - F24			125	166													291	3.9%
4.	P	AA	1.4.4 - C17			28	41													69	0.9%
5.	P	AA	1.4.4.1.4.5-C14					121	47											168	2.3%
6.	P	AA	1.4.4 - C20					16	19											35	0.5%
7.	P	AAA	1.4.6.1.4.8-F24			291	386													677	9.1%
8.	P	AAA	1.4.8 - C21					195	174											369	5.0%
9.	O	A	2.1.1 - G90									73	21							94	1.3%
10.	O	A	2.4.4.2.4.9-H33											91	139					230	3.1%
11.	O	AAA	2.1.3 - G90									54	16							70	0.9%
12.	U	AAA	3.2.5 - H83			56	90													146	2.0%
13.	U	AAA	3.3.5 - G71			10	11													21	0.3%
14.	R	A	4.1.1 - F70							67	29									96	1.3%
15.	R	A	4.1.2 - F68							34	9									43	0.6%
Sum by types of sites				0	0	535	707	332	240	133	49	127	37	91	139	0	0	0	0	2389	32%
Sum by the tools				0		1241		572		182		164		230		0		0			
Percentage of sum by the tools				0		16.7%		7.7%		2.5%		2.2%		3.1%		0		0			

Note: "P" stands for Perceivable and etc. ".tj" – Tajik sites and ".at"- Austrian sites.

E. Reliability

Since one of the purposes of this research is to evaluate a degree of agreement and consistency among AWAEMs, inter- and intra-reliability types are considered. The inter–reliability (between testers) is the degree to which two or more different and independent AWAEMs produce the same results when examining the same phenomenon. Inter–reliability values are calculated by considering those cases where all eight AWAETs judge and evaluate barriers with a severity level greater than 0. Here, it is important to note that the reliability between tools is questionable [89] and hardly reaches 100%.

Surprisingly, the obtained absolute inter-reliability values for all eight tools is only 116 (1.56%) out of 7423 total issues (See Table 11). Here, all AWAETs could define identical errors in only three unique SC: F65 - 83, H44 - 28 and H57 – 5 errors. Furthermore, the inter-reliability weakens when numbers of tools are reduced by the piece from eight to two, but the values of the simultaneous discovery of TP naturally increase. Even in such a scenario, pairs of different tools have the few numbers of identical accessibility vulnerabilities. Their total coincidence value is 1360 (18.32%).

On the other hand, intra–reliability (within testers) is the degree to which measurement results taken by the same AWAEMs are stable and consistent for the same phenomenon under the same conditions. The internal consistency or similarity of AWAEMs in relation to numbers of verified TP can be measured using Cronbach's alpha, which is the most common numerical coefficient to reflect internal reliability. The resulting α coefficient for the twenty-six Tajiks sites is 0.568 and for the twenty-six Austrian sites is 0.793, which means our eight AWAETs have poor and acceptable internal consistencies, respectively.

Another option for exploring "similarity" between AWAETs is the Euclidean distance, which can be employed for both inter and intra-reliability tests, but under different conditions. That is, SC are considered to have the same importance for inter-reliability and the other case for intra-reliability. Actually, the Euclidean distance method measures the length of a segment connecting two points in any number of dimensions. If to consider all SC as equally important, then Euclidean distance information also allows ideally defining the best AWAET in terms of associating both coverage and completeness. Ordinary distances between TP and their SC across each pair of tools describe how far apart these tools are located (See Fig. 9). Basically, the farther away are tools from each other, the less similar they are.

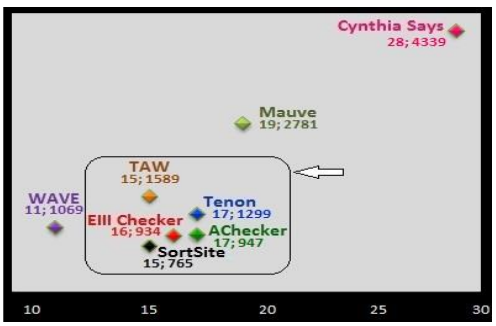


Fig.9. Two-dimensional Euclidean distances between the AWAETs

According to the calculations, given in Fig. 9, the average distance between all the tools is 1300. In the plane, the lowest located tool - SortSite and the highest - Cynthia Says have the longest distance of 3574. Lastly, the average distance of the five AWAETs that are grouped and pointed in Fig. 9 is 403, which means that they are relatively close to each other. Hence, it can be

concluded that most of the tools are quite similar in the way they deal with TP but still, the tool's performance is not optimally reliable in all cases.

Table 11. Dataset statistics of the inter-reliability of the AWAETs that operated according to WCAG 2.0 (levels A-AAA)

№	Principle	Levels	Error categories	AChe-cker	Cynthia Says	Mauve	Sort - Site	Taw	Tenon	EIII Checker	WAVE	Identical TP		Sums of similar TP	Percentages of the sums	Numbers of tools
												.tj	.at			
1.	P	A	1.1.1 - F65	☑	☑	☑	☑	☑	☑	☑	☑	66	17			
2.			1.1.1,1.3.1 - H44	☑	☑	☑	☑	☑	☑	☑	☑	6	4			
3.	U	A	3.1.1 - H57	☑	☑	☑	☑	☑	☑	☑	☑	2	3	**116**	**1.56**	**All tools**
4.			3.3.2 - H44	☑	☑	☑	☑	☑	☑	☑	☑	6	3			
5.	R	A	4.1.2 - H44	☑	☑	☑	☑	☑	☑	☑	☑	6	3			
6.	P	A	1.3.1 - H42	☑	☑	☑	☑	☑	☑	☒	☑	**3**	**3**	**6**	**0.08**	**7 tools**
7.	P	A	1.1.1 - H37	☑	☑	☑	☒	☑	☑	☑	☑	7	8			
8.	O	A	2.4.4,2.4.9 - G91	☑	☑	☑	☒	☑	☑	☑	☑	44	18	**83**	**1.12**	**6 tools**
9.	R	A	4.1.1 - H93	☑	☑	☑	☑	☒	☑	☑	☑	3	3			
10.	P	A	1.1.1,1.3.1 - H65	☑	☑	☑	☒	☒	☑	☑	☑	3	2			
11.	U	A	3.3.2 - H65	☑	☑	☑	☒	☒	☑	☑	☑	1	2	**11**	**0.15**	**5 tools**
12.	R	A	4.1.2 - H65	☑	☑	☑	☒	☒	☑	☑	☑	2	1			
13.	P	A	1.1.1 - H2	☒	☑	☒	☑	☑	☒	☒	☑	36	8	**56**	**0.75**	**4 tools**
14.	O	A	2.4.4,2.4.9 - F89	☑	☒	☒	☑	☑	☒	☑	☒	2	10			
15.	P	A	1.3.1 - F91	☒	☒	☒	☑	☒	☑	☒	☑	20	2			
16.	P	AA	1.4.3 - G18	☑	☑	☒	☒	☒	☒	☒	☑	9	7	**56**	**0.75**	**3 tools**
17.	O	A	2.4.1,4.1.2 - H64	☑	☑	☒	☒	☒	☒	☑	☒	5	4			
18.	U	AA	3.2.2 - H32	☒	☑	☑	☒	☑	☒	☒	☒	7	2			
19.	P	A	1.1.1 - F39	☒	☒	☒	☑	☒	☒	☒	☑	4	18			
20.	P	A	1.1.1,1.2.1 - F30	☒	☑	☒	☑	☒	☒	☒	☒	3	2			
21.	P	A	1.3.1 - H43	☒	☑	☒	☒	☒	☑	☒	☒	16	11			
22.	P	A	1.3.1 - H73	☒	☑	☑	☒	☒	☒	☒	☒	23	14			
23.	P	A	1.3.1,1.4.5, 1.4.9 - G140	☒	☒	☑	☒	☑	☒	☒	☒	94	131			
24.	P	A	1.3.1,3.3.2 - H71	☒	☒	☒	☒	☑	☒	☒	☑	1	1			
25.	P	A	1.3.1 - H63	☒	☑	☑	☒	☒	☒	☒	☒	16	15	**1360**	**18.32**	**2 tools**
26.	P	AA	1.4.4,1.4.5 - C12	☒	☑	☑	☒	☒	☒	☒	☒	483	291			
27.	P	AAA	1.4.8 - F88	☒	☒	☒	☒	☒	☑	☒	☑	0	0			
28.	P	AAA	1.4.6 - G18	☑	☑	☒	☒	☒	☒	☒	☒	8	2			
29.	P	AAA	1.4.6 - G17	☑	☑	☒	☒	☒	☒	☒	☒	94	84			
30.	R	A	4.1.1 - G134	☒	☒	☒	☒	☑	☒	☑	☒	11	16			
31.	R	A	4.1.2 - F59	☒	☒	☒	☒	☒	☑	☑	☒	5	17			

Note: "P" stands for Perceivable and etc. ".tj" – Tajik sites and ".at"- Austrian sites. The symbol "☑" points out that a given AWAET could detect TP for corresponding SC and the symbol "☒" is used in the opposite case.

Tools' behaviors when detecting violations were further investigated with the Kruskal–Wallis test. Results show that all the tools do not behave the same when detecting TP, FN and FP, except for FN missed from Austrian sites, where significance is greater than .05 (See Table 12).

Table 12. A summary of the Kruskal-Wallis tests on various data distributions for assessing accessibility among the tools

	TP		FN		FP	
	.tj	.at	.tj	.at	.tj	.at
Chi-square	19,613	22,916	15,054	12,441	20,290	17,195
df	7	7	7	7	7	7
Asymp. sig.	,006	,002	,035	,087	,005	,016

Note: ".tj" – Tajik sites and ".at"- Austrian sites

F. Validity

Validity is also called as expert or guideline review as well as manual inspection [90, 91] and considers accepted decisions of the beginner which correspond to the expert ones [50,54-56, 87]. In other words, it means defined problems by AWAEMs are not different than those found by expert assessors. Thereby, any difference

in knowledge of experts and ways AWAEMs interpret guidelines leads to less reliable and valid results. Next, as generated issues by AWAEMs must be confirmed with the help of experts, so all found TP are already valid. In turn, those errors that are not valid are included as false positives. Based on this, we consider the validity of AWAEMs as a dividing red line between correctness and completeness together with coverage. Lastly, validity can be further divided into correctness and reliability.

In order to understand how close the outcomes of the AWAETs with the total real errors across all SC are and how strong the tools associated among themselves, correlation coefficients between them are studied. Table 13 includes the results of these measurements.

Table 13. A correlation matrix for TP defined by the different tools and experts

		AChecker		Cynthia Says		Mauve		SortSite		TAW		Tenon		EIII Checker		WAVE		All TP	
		.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at	.tj	.at		
AChecker	Correlation Coefficient	1,0	1,0																
	Sig. (2-tailed)																		
Cynthia Says	Correlation Coefficient	,367*	,326*	1,0	1,0														
	Sig. (2-tailed)	,013	,029																
Mauve	Correlation Coefficient	,367*	,335*	,123	,050	1,0	1,0												
	Sig. (2-tailed)	,013	,025	,420	,745														
SortSite	Correlation Coefficient	,261	,268	-,196	-,251	,004	-,032	1,0	1,0										
	Sig. (2-tailed)	,083	,075	,197	,097	,981	,836												
TAW	Correlation Coefficient	,346*	,468**	-,020	-,071	,309*	,339*	,252	,251	1,0	1,0								
	Sig. (2-tailed)	,020	,001	,898	,644	,039	,023	,094	,097										
Tenon	Correlation Coefficient	,518**	,421**	-,008	-,028	,322*	,329*	,224	,096	,240	,263	1,0	1,0						
	Sig. (2-tailed)	,000	,004	,960	,855	,031	,027	,139	,531	,112	,081								
EIII Checker	Correlation Coefficient	,649**	,556**	,003	-,033	,320*	,301*	,211	,126	,498**	,499**	,563**	,544**	1,0	1,0				
	Sig. (2-tailed)	,000	,000	,985	,829	,032	,045	,164	,408	,000	,000	,000	,000						
WAVE	Correlation Coefficient	,371*	,488**	,088	,162	,077	,207	,516**	,474**	,285	,422**	,430**	,359*	,136	,198	1,0	1,0		
	Sig. (2-tailed)	,012	,001	,566	,287	,616	,172	,000	,001	,057	,004	,003	,016	,372	,192				
All TP	Correlation Coefficient	-,017	,106	,252	,304*	,064	,042	-,103	-,163	,071	,038	-,102	-,152	-,184	-,230	-,021	-,050	1,0	1,0
	Sig. (2-tailed)	,913	,489	,094	,043	,676	,786	,499	,286	,642	,805	,506	,319	,226	,129	,889	,742		

Note: * Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).
".tj" – Tajik sites and ".at" - Austrian sites.

Table 13 shows correlation values applied to each possible pair of tools. Subsequently, only AChecker could become barely close to all TP found and even worse, the other tools had almost no linear relationship with all TP. Furthermore, AChecker with 6 tools, EIII Checker with 4 tools, WAVE with 3.5 tools (extra 0.5 point for having similarity with TAW for Austrian sites), Tenon with 3 tools, TAW with 2.5 tools (extra 0.5 point for having similarity with WAVE for Austrian sites), Mauve with 2 tools, Cynthia Says with 1 tool and SortSite with 1 tool have moderate uphill positive relationships or moderately vary together.

G. Efficiency

Efficiency is a quality factor that relates to the amount of resources such as efforts, time, skills, money and facilities needed to conduct the evaluation that brings to a specified degree of effectiveness and usefulness [21, 23, 92]. Here, the accessibility and usability of AWAEMs themselves are fundamental aspects. The below summary Table 14 introduces an assessment overview of the efficiency parameters for the AWAETs. Next, each detailed scan by all the tools took an average of 1.8 minutes for pages with few accessibility barriers and 2.6 minutes for pages with lots of barriers. Usually, AWAETs that ask for commercial licenses often provide compelling desktop environment versions, while free tools are generally available online and are in a great number, including all our selected tools. Besides being free and available online, there are downloadable multi-platform desktop versions for SortSite and TAW. WAVE and Tenon have Chrome extensions with automatic updates. As for EIII Checker, it ensures a bookmarklet, designed for the Web accessibility evaluation. Lastly, Tenon offers full access to all of its services within thirty days after registration (See Table 14).

The final result of each AWAEM on a tested site ends with the creation of HTML, XML and/or PDF reports of all caught accessibility faults by the frequencies or numbers of occurrences. Typically, reports are structured in accordance with chosen standards. The quality of these reports has a significant impact on the efficiency of the AWAEM. Generally, this experimental study with the eight tools indicates that forms and structures of provided reports have become much better as well as easier to understand and interpret. However, there are problems with providing multiple links to violated source codes. The top levels in the reports of Cynthia Says and Mauve are the entirely copied text of relevant WCAG 2.0's SC without any interpretation, which may not be intuitive to users who are not already WCAG scholars. Even worse than imagined, AChecker, WAVE and

Tenon provide their own interpretation of WCAG with human-readable words, but no SC's numbers or names. In these cases, it is very difficult to recognize affiliation of those SC. Optimally, TAW and SortSite label errors according to their own interpretations of WCAG and plus indicate to SC's numbers or names, which increases the value of reports.

Table 14. Efficiency statistics afforded by our sample AWAETs

№	Tools	Report	Languages	Browser plugins	License	Report formats
1.	AChecker	Reporting the results	English, German and Italian	-	Free Software and Open Source	HTML, PDF,XML and EARL
2.	Cynthia Says	Reporting the results	English	-	Trial or Demo, Commercial and Enterprise Online checker	HTML
3.	MAUVE	Reporting the results	English	Chrome, Firefox	Free Software, Online checker, Browser plugins: Chrome and Firefox	HTML
4.	SortSite	Reporting the results	English	-	Trial or Demo, Commercial, Online checker, Hosted service and Server installation	HTML and CSV
5.	TAW	Reporting the results	Galician, Catalan; English and Spanish	Firefox, Chrome	Desktop software and online tool	HTML
6.	Tenon	Reporting the results	English	Chrome	Trial or Demo, Commercial, Enterprise. Online checker, Hosted service and Server installation	HTML, XML, CSV and JSON
7.	EIII Checker	Reporting the results	English	-	Open Source, Online checker, Hosted service and Server installation	HTML
8.	WAVE	Reporting the results within web pages	English	Chrome, Firefox	Free Software, Trial or Demo, Commercial, Online checker, Hosted service and Server installation	HTML and XML

Besides, yet an intensive work needs to be done to improve the presentation of guideline's violations. Cynthia Says and SortSite do provide neither total numbers of TP nor total numbers of TP calculated by groups. AChecker, MAUVE and Tenon ensure no total numbers of TP, but total numbers of TP by groups. TAW, Tingtun and WAVE report both numbers of detected and not detected faults in details. Another problem is that the same accessibility errors are often determined repeatedly and appear several times with unnecessarily increased numbers in the reports created by Cynthia Says, SortSite and Tenon.

Table 15. Capacity comparison of AWAETs

№	Open source tools	Website submission mode			Accessibility standards				Number of pages in one check	Supported formats
		URL	File	Paste*	WCAG1	WCAG2	Sec508	Others		
1.	AChecker	☺	☺	☺	☺	☺	☺	*W,B,S*	Single and password protected web pages	CSS, HTML and XHTML
2.	Cynthia Says	☺	☹	☹	☹	☺	☺	☹	Single and password protected web pages	CSS, HTML and XHTML
3.	Mauve	☺	☺	☺	☹	☺	☹	*S*	Single and password protected web pages	CSS, HTML and XHTML
4.	SortSite	☺	☺	☹	☺	☺	☺	*W*	4-10 Web pages	CSS, HTML, XHTML and PDF
5.	TAW	☺	☹	☹	☺	☺	☹	*M*	Single and password protected web pages	CSS, HTML and JavaScript
6.	Tenon	☺	☹	☹	☹	☺	☺	☹	Single and password protected web pages	CSS, HTML and XHTML
7.	EIII Checker	☺	☹	☹	☹	☺	☹	☹	Single and password protected web pages	HTML and XHTML
8.	WAVE	☺	☹	☹	☺	☹	☺	☹	Single and password protected web pages	CSS, HTML and XHTML

Note: W = W3C HTML Validator, B = BITV, S = Stanca Act, *Paste (HTML Markup) - pasting complete HTML source code from the clipboard. BITV- accessibility criteria test of the German BIK project, RGAA- French Web accessibility Law. M = Own Mobile Heuristics

H. Capacity

We claim that one more quality characteristic is necessary to widely estimate the AWAEM, which is "capacity". Capacity is the ability of an AWAEM to simultaneously examine lots of web pages of various formats and complexities with many code monitoring modes against different kinds of guidelines and finally, ensuring multi-format reports. The scope of capacity must be extended with the emergence of new features of the AWAEM. Table 15 depicts capacity comparisons of our tools. The symbol "☺" points out that a given AWAET satisfies a corresponding evaluation criterion and the symbol "☹" is used in the opposite case. According to Table 15, each tool has different dimensions of capacity for estimating even the same websites.

The capacity comparison has emphasized effective and

advantageous AWAETs based on the four aspects, as shown in Table 15. Among the eight well-known tools, AChecker and SortSite could test the accessibility of sites based on the wide set of international guidelines. None of the tools could scan an entire site at one time. In particular, only SortSite checks the first 4-10 pages of a website with one click. Regarding web page code monitoring modes, only AChecker and Mauve can operate in the three different ways: ensuring URLs, uploading files or pasting HTML codes. Finally, only Cynthia Says and Mauve could analyze web pages against CSS (See Table 5), while the producers of all the tools, except EIII Checker stated the opposite case (See Table 15).

IV. Discussion

A. …about Coverage

We have learned that one AWAEM informs about the existence of specific numbers of SC and TP and another does the same job, but slightly wider than the first. One of the basic problems of AWAEMs is that they ensure poor coverage of standards, including WCAG 2.0 and each one uses different SC to assess even the same pages. Generally, all the tools individually could cover from the lowest of 12.7% (WAVE) to the highest of 32.4% (Cynthia Says) out of 71 unique SC, which is 20% on average, while in studies by [31] and [37] are around 23-50% and 7.1-35.7% respectively. However, there are similarities with [31, 40] in identifying SC by the principles, as Perceivable and Operable got greater numbers of violated SC 60.4% and 21.9% respectively. The research by [40] also states that the Robust principle is the second highest violated principle after Perceivable (52%) across all metropolitan municipalities with a total of 867 errors (25.20%). As for the combination of all eight tools, it improves the situation with 25% from the average value. In addition, expert analyses predicted slightly more SC (a total of 45.1%) than AWAETs (32.4% for the best tool) and wherein in those SC that the tools were unable to find. To this end, some SC that just need a verification of certain HTML tags such as 2.4.2 Page Titled [31] and "F70": Failure of Success Criterion 4.1.1 due to the incorrect use of start and end tags or attribute markup" are not even covered by the majority of the AWAETs.

B. …about Completeness

Websites with low accessibility levels possessed sets of common violations and vice versa for inaccessible sites. Out of the overall 9037 TP established, 7356 issues were considered TP derived from the combination of all the tools and 1681 issues were considered not-identified TP by the tools, and thus received by means of manual inspections. Yet, at this time, the AWAETs did not provide a good completeness as we would expect. The best tool - Cynthia Says generated 59% (4339) of all TP and this degree is drastically reduced for the remaining tools up to 10% (765). It means that AWAETs are able to determine only 4 out of 9 problems in the best case scenario. It follows that more than a half number of TP remains unverified if tools are applied separately. Our results for completeness match with [31], which is (14-59%) and are more likely improved when compared with those that ranges between 30% (21TP) and 49% (50TP) [29] as well as between 14% (91TP) and 38% (249TP) [31].

In addition, accessibility faults found as warnings can be transformed into TP after double-checking by manual means, and thus, interpreted as the improvement factor for completeness. Some types of "Visual, warnings or N/A" violations offered by tools such as Cynthia Says and TAW have no specific sources of code. In this case, it is even impossible to view indicated mistakes.

C. …about Correctness

The correctness of AWAEMs is imperative since they are used in practice. In fact, tools under study have possessed a higher level of correctness than any other characteristics, which is on average 70.7% (See Fig. 7). The best correctness score is 81% (934) for EIII Checker, whereas the worst is 39% (2781) for Mauve. These data findings slightly differ from previous investigations, e.g. 66% (202) - 96% (192) by Vigo et al. [31] and even insignificantly increased with those established by G. Brajnik [29], which is between 79% (118) and 93% (139). Yet, all studies indicate to the considerably positive correctness of AWAETs. Also, we observed that tools with the best accuracy have the lowest completeness level. Our estimation showed that all the tools report a smaller number of FP when checking sites with the WCAG 2.0 guidelines. However, multiple and unnecessary references to the same errors as well as repeated increase of errors in numbers were revealed in Cynthia Says, SortSite and Tenon.

Fundamentally, completeness and correctness have an important causal relationship and play a key roles in characterizing effectiveness. Furthermore, complete but incorrect AWAEMs can catch various faults from a site and thus, generate a lot of FP. Conversely, incomplete but correct AWAEMs could generate no FP, but a large number of FN, which means that many other TP remain hidden.

D. …about Specificity

Naturally, the good AWAEM focuses on more specific SC. More specific the AWAEM is, more special and important it becomes. Unfortunately, very few studies have investigated the specificity of AWAEMs. In our case, although the chosen tools looked for the same and common kinds of problems, we could easily observe the different specificity ratios among them. Notably, TAW as having the poorest specificity result with 164 TP is slightly superior to early tools such as Bobby and LIFT Machine that implemented 70 and 103 automated tests, respectively [29].

E. …about Reliability

When considering inter-reliability, a detailed analysis

of fifty-two sites discovered that the reliability of today's AWAETs is still low. There were only a few TP recorded simultaneously by all the eight tools – 116 (1.56%) through only 3 unique SC. Even, the views of pairs of tools are the same in only 1360 (18.32%) true positives. In 2003, Diaper and Worman [89] in their experimental analysis of two tools "A-prompt vs. Bobby" have emphasized that the inter-reliability of these tools was questionable and there was no guarantee for any of them to produce the same results. Subsequently, there was no agreement between A-Prompt and Bobby when looking at priority 2 of WCAG 1.0. Moreover, there is also a claim that our eight tools under study as well as Bobby, Lift and Ramp [36] and A-Prompt and Bobby [89] provided the different reliability results.

Despite this, the experiment results for the intra-reliability indicate that all our AWAETs do not behave the same when producing TP in both Tajik and Austrian sites. These facts confirm their similarity with [32]. In sum, imprecise results, not defining accessibility obstacles (FN) or displaying non-existent problems (FP) decrease the reliability level.

F. …about Validity

To investigate origins of TP, validity was computed as the percentage of problems caught by automated means that matched with those caught by manual checks. Out of 9037 (96 SC) possible TP violated, the eight given tools could fully automate an average of 23.3% of TP, which is 8.3% more than in a previous investigation in the year 1999 by [93]. Given the fact that the best tool Cynthia Says was able to find 59% of TP, we can state that about 40% of TP require manual checks, which is again 10% more than in the mentioned study [93]. Preferably, if to use a combination of AWAETs, then the percentage ratio of automated tests will increase significantly up to 81.4%. Other previous studies on automated accessibility and usability assessments [93, 94] that were conducted in the early 1990s, state that only around 44-55% of tests can be automated. Thus, another main conclusion is that tools have become more intelligent after two and a half decades.

G. …about Efficiency

AWAEMs have different evaluation techniques, strategies, characteristics, benefits and drawbacks. Making a right choice of them depends on those priorities set by the factors discussed in more details in Section V. E.g. one drawback should be mentioned, which is AWAEMs' failure to distinguish between important and non-important errors. This point was also emphasized by [30]. Additionally, using AWAEMs requires experience and knowledge about them, as well as a better understanding of reported TP and alerts. Finally, looking back in the year 2006, there were few AWAETs freely available [37]. Luckily, numerous excellent and free software applications with ninety-three registered and other non-registered ones are available for stakeholders to assess the accessibility of their sites and applications.

H. …about Capacity

All previous investigations and yet, this study claims that different AWAEMs propose different faults and none of them find all faults of a site. Some barrier identification processes are automated in one AWAEM and semi-automated in the other one. Also, AWAEMs are generally limited to particular technology platforms. File formats such as HTML, CSS, PDF, GIF and Flash; JavaScript, AJAX, Java and SMIL; various browsers as well as integrated design environments are not supported in all the eight tools simultaneously. As for the number of web pages to be checked, almost all the tools evaluate only a single page with a single click, which usually is the homepage. Moreover, our findings together with [97] indicate that still most of the AWAEMs statically analyze HTML source code and rarely the design itself. However, only WAVE could visually analyze the accessibility of website design. Lastly, all the selected AWAETs are exclusively focused on the Web accessibility assessment, whereas only SortSite provides the already working additional functionalities, including broken links and usability checks, browser compatibility, testing with search engine guidelines and more.

V. Pros And Cons of AWAEMs

Using AWAEMs to support testing activities leads to the following nineteen pros and fourteen cons, as described in Table 16.

If to consider quantities, the benefits of using AWAEMs turned out to be more than their limitations (See Table 16). This increases the motivation for using AWAEMs, which is possible during the construction, implementation and maintenance phase of the site's development. In turn, high productivity and efficiency of AWAEMs cannot be compared with the other methods as they can catch certain issues that might otherwise be time, cost and resource consuming. AWAETs could identify larger numbers of potential problems than web designers [28]. Moreover, the web designers have experienced difficulties and were not effective in interpreting and employing design constraints and accessibility guidelines without using AWAETs [28, 96].

Table 16. A summary of pros and cons of AWAEMs

Automated Web accessibility evaluation methods		
№	**PROS**	**CONS**
1.	- Many automated software applications and tools are widely available on the market today	- Often incompatible for many major technology platforms such as (x)HTML 5, CSS, PDF, JavaScript, AJAX, Java, .Net, PHP etc. that are in a wide use today
2.	- Cost effective in general, often free and easy to implement. The assessment gets less labor-intensive and because of it, all expenses will be decreased	- There could be an expensive purchase of commercial versions for single or multiple users and domains
3.	- Practical and highly effective with large evaluation projects and websites with lots of pages - Executed by software, thus quickly and conveniently runs tests with far more code	- Free software versions evaluate a limited number of pages, often one page at one run
4.	- Determining a large and diverse spectrum of objective non-compliance issues - Performing a wide range of certain tasks	- Unable to detect a set of subjective barriers. Hence, human cognitive skills are required
5.	- Less time-consuming - Saving a huge amount of time	- Enough time and effort should be spent on learning and getting experience to work with AWAEMs - Basic computer literacy is necessary
6.	- Revealing issues that might be missed by other methods. Therefore, helps to focus in the manual or other kinds of the checking techniques	- Unrealistic expectations such as missing actual problems and sometimes flagging non-existing accessibility barriers
7.	- Manufacturers have been improving the quality of AWAEMs with each passing year. E.g. this study claims that completeness, correctness, intra-reliability and specificity of AWAETs are enhanced than in the previous years	- Still unable to ensure a comprehensive validation of all accessibility problems. E.g. less than 50% of the WCAG 2.0's SC can be examined automatically - The scope of coverage and completeness vary among AWAEMs
8.	- Non-technical appraisers can run tests and cut down the necessity of expert knowledge among certain appraisers - AWAEMs train developers on accessibility best practices by presenting necessary information in their reports	- Standards/guidelines can be too difficult to read and interpret, too abstract and/or too much in size
9.	- Easy to conduct a set of tests repeatedly and at regular intervals	- Entirely dependent on the quality of chosen standard/guidelines and/or some of their SC or checklists
10.	- Allowing a huge number of evaluators to collaborate as a great team	- Strong expert knowledge and experience for understanding specific standards, interpreting reports and performing skilled evaluation are required
11.	- Easy to employ at any stage of site development and design. At early stages, the AWAEM helps to minimize an eventual cost of fixing violations related to guidelines/standards	– Conformance to a standard or guideline does not indicate to the real accessibility issues
12.	- Combinations of two or more AWAEMs is possible and ideal	- AWAEMs alone are not sufficient to explore the full compliance of sites with standards or guidelines
13.	- Can be performed remotely in space and time	- There is a need to have an active Internet connection. Hence, the connection might be unreliable at time of carrying out assessments
14.	- Automatic grading of found issues into groups - Statistical evidence of progress - Providing quality comparisons between different sites	– Assigning priorities and severity of accessibility issues are not supported
15.	– Appropriate for diagnostic as well as formative and summative types of assessments	
16.	- Helpful when code and interfaces are changed frequently	
17.	- Can be run on different machines, OS and browsers at the same time - Easy configuration of settings	
18.	- More reliable and robust as performed by software applications	
19.	- Able to program new worthy tests or expand an existing suite of tests to cover more features	

VI. Recommendations for Improving the Quality and Functionality of Awaems

Our review of the current state of the automatic accessibility evaluations highlights an important list of fifteen recommendations in different aspects to assist in the improvement of AWAEMs:

1. The user interfaces of AWAEMs themselves are supposed to be accessible and usable to be operated by all types of users, including the disabled and elderly. Otherwise, purchasing a difficult to use AWAEM will be money wasted.
2. The AWAEM should ensure a high quality and reliable results with well examining TP and removing FP since inaccurate, incomplete or misleading results are potentially harmful to organizers and websites and risky for project budgets. For example, concluding decorative images with empty alt text as inaccessible to assistive technologies e.g. screen readers, when this is not always the case.
3. In today's conditions, it is necessary to employ a variety of particular accessibility guidelines to develop websites and satisfy the needs of

numerous users. On this basis, existing and new AWAEMs that intend to be created need to support at least two or more prominent international standards, including a necessary number of specific guidelines, e.g. search engine guidelines and/or those that are adopted in a given country or company.

4. The AWAEM has to include functionalities that automate repairs across sites with further facilitation by communicating with testers or without them. For instance, if a set of the same images used repeatedly in a site, then a repair process can be included by asking the developer to enter alternate texts for those images. Even then, Lee and Hanson focused on producing the AWAEM that automatically allows deep changes in the structure of sites to make them more accessible [97]. Basically, repair-based AWAEMs have great influences in substantially reducing time, cost and effort required to make websites accessible. Moreover, I strongly believe that automatic accessibility repairs are the future priority direction for the improvement of AWAEMs.
5. Developing features to facilitate scheduled monitoring of web accessibility, where e.g. new content is being released daily or even several times a day. Further, when carrying out regular monitoring, AWAEMs should ignore re-checking of pages, where changes have not occurred.
6. The growing trend in modern Web technologies is the creation of dynamic and interactive websites through the use of CSS3, HTML5, AJAX, jQuery and various scripting and programming languages that are leading to new challenges in automatic accessibility evaluation. In this case, the existing AWAEMs must be able to validate web pages built with the association of these advanced techniques.
7. The majority of AWAEMs evaluate only one page (e.g. all the eight tools except SortSite) or a small number of pages with one click. Hence, comprehensive testing of an entire website with any amount of web pages is a desirable feature.
8. Testing uploaded files and/or source code entered directly to avoid accessibility errors in early stages of website's development. For this reason, content management systems should also have access to AWAEMs during testing of code in the site's development phase without uploading to a publicly accessible URL.
9. To be able to follow, process and assess the accessibility of non-standard links to the other non-HTML file formats such as Flash, PDF, Microsoft Office applications and etc.
10. To expand the list of other services within an AWAEM that will effectively increase the accessibility of analyzed sites, such as browsers' compatibility tests (with a variety of modern browsers such as Chrome, Firefox, Safari, Opera, IE, iOS, Android, BlackBerry, etc.); performing a spell check to facilitate proofreading since misspelled words are most likely to be mispronounced by screen readers; conducting page preview filters such as applets off, images off, scripts off, without color and so on with various types of color blindness; testing document object models and detecting broken links, Java, ASP.NET, PHP and etc. script errors and the like.
11. AWAEMs should clearly list all existing violations together with their source code and indicate to employed guidelines and numbers of accessibility errors. Therefore, it would be more efficient to report accessibility assessment results by: a) clearly indicating the names of violated issues and source codes; b) providing own interpretation of violated issues and ways to eliminate them c) grouping all accessibility failures in accordance with principles or sub-groups of a chosen standard; d) clearly showing the relationships between quantities and percentages of measured problems: total tested issues, past, failed and those issues that need to be verified by experts as well as the general scores of examined websites.
12. Providing functionalities to export and save data reports in numerous file formats and store raw evaluated data in XML documents with subsequent access.
13. The presentation of summary reports by the AWAEM should be customizable so that testers could select the levels of details they want to see otherwise the reports might be overwhelmed by full-detailed data.
14. AWAEMs may offer a full or partial integration with other existing accessibility checking systems, toolsets and processes to improve functionalities, modify test logics, integrate specific guidelines and/or add new rule sets for re-inspecting from scratches.
15. Nowadays, AWAEMs should be applied and integrated dynamically at all stages of the website development process. We support the view of the authors Xiong et al. [38] that most guidelines' checkpoints can be examined at very early stages of the website development rather than the post-implementation and integrated into AWAEMs to lower the cost of corrections over final realization.

VII. Limitations and Future Work

WCAG 2.0, being the updated version of WCAG 1.0 is the international standard that makes a great contribution to the digital inclusion and harmonization of accessibility rules globally. Strictly speaking, however, its success criteria (mostly in the Level AAA) may impose restrictions on web page design (choosing colors or styles), aesthetic (look and feel), components and functions; content presentations (e.g. more images and tables, but less text) and author's freedom of expression.

Furthermore, some cross-site scripting (XSS) codes embedded in sites can manipulate or block some functions of AWAEMs when testing web pages. We encountered several cases with XSS codes, where our tools AChecker, SortSite, Tenon and WAVE were not able to check certain web pages and because of this, we had to select those web pages that all the eight tools could estimate.

Another limitation is that the current study is purely focused on the conformance with WCAG 2.0 without exploiting other features ensured by AWAETs, such as grammar and spell checking, implementing HTML5, ARIA, search engine and SEO best practice guidelines, browser compatibility testing, identifying website usability problems, broken links and unexpected HTTP error codes, missing images and script errors and etc. Further, one of the major limitations is that AWAETs do not appear to be robust enough to analyze content in different non-HTML formats, even though web content is considered the most important factor for high-quality sites [98]. It is, therefore, desirable that such aforementioned aspects will be also taken into account in future studies on web accessibility.

Obviously, increasing numbers of various websites, software tools and human experts affect testing results by extending implementation of WCAG 2.0 and as a result, brings to a more precise and valid view of web accessibility. Therefore, as a future work, more detailed analysis with hundreds of web pages could be carried out to find out if there is a correlation at the levels of websites. Also, only eight free of charge tools have been utilized for this research. Since eighty-five other tools have been developed that offer features and functionalities that are distinguishable than those mentioned in this study. That is why the study can be extended by further using other decent free and multifunctional tools. From this point of view, separate research can be carried out only with commercial AWAETs to compare summary outcomes, i.e. whether there are genuine differences between the two groups of free and commercial tools.

The results of this study may not cover the real-life issues that the disabled face since this research relies on the automated and expert assessments of Web accessibility based on the standard conformance. Consequently, testing sites with real disabled users can encompass a wide range of physical problems coming out from all complex and subtle interactions between web content and assistive technologies. Reasonably, conducting parallel studies with disabled users, older adults as well as with accessibility professionals to re-evaluate preferably the same sample web pages and comparing problems found in these studies would be necessary to generate more broad and accurate results. Finally, an increasing number of accessibility guidelines proposed for the Web makes the implementation of AWAEMs that work with these guidelines more complex. Hence, it is necessary to conduct more research to test guidelines introduced in AWAEMs and make guidelines and AWAEMs themselves more usable.

VIII. Summary and Concluding Remarks

The evaluation and comparison of AWAEMs are difficult because they are evolving fast and different methods measure different variables. Accordingly, various aspects of the quality or new quality criteria for measuring AWAEMs will arise, which must be taken into account. Apparently, this is causing a problem since AWAEMs are usually studied through the lens of a limited number of quality features and unfortunately, there is a lack of research on web accessibility. Furthermore, quality criteria for AWAEMs must be clearly defined in order to evaluate and benchmark them. As for this research paper, it focused on the relatively large amount of quality criteria such as coverage, completeness, correctness, specificity, inter- and intra-reliability, validity, efficiency and capacity to achieve its goals. Notably, considerable attention has been dedicating for these eight quality criteria in the software engineering discipline.

Next, in order to produce viable, effective and precise results in accordance with the targeted eight quality criteria, a new methodology is elaborated in the study. This new integral five-phase methodology called as “5PhM-for-AWAEMs” is applicable for an effective - a) selection and b) assessment and/or comparison of AWAEMs towards analyzing the accessibility of sites and web applications. Within the framework of this methodology, strategies and techniques of the analyses for each of the eight specified criteria and the relevant statistical results are disclosed in detail in Section III. More than this, the eleven key criteria for the selection of appropriate AWAEMs, the required numbers of web pages and experts for the acceptable, normal or ideal Web accessibility assessment have been proposed. Notably, for the first time, twenty-six Tajik and twenty-six Austrian sites have been analyzed based on the conformance with WCAG 2.0 as long as this latest standard is being accepted and replicated globally. To obtain final results, expert evaluations classified and confirmed all TP, FP and FN produced by AWAETs as well as uncovered the remaining ones. To sum up, our research findings are interpreted in the context of the prior knowledge and assumptions. To facilitate comparisons between studies, we strived for the comparability that covers methods, research design and investigations.

The ideal AWAEM is the one that can provide accurate predictions of all accessibility problems that may occur in a website. However, such an AWAEM still does not exist. Substantially, eleven relevant conclusions can be drawn from the employment and analyses of the eight AWAETs such as AChecker, Cynthia Says, EIII Checker, MAUVE, SortSite, TAW, Tenon and WAVE, as related to the goals of this study:

1. The scope of coverage varies among the tools from 12.7% to 32.4%, which is in fact far away from having a good automated coverage. This is not surprising that those SC that are less common or subtle are still hardly targeted or not

covered at all. Besides, even the most frequent and expected SC are not completely covered by the majority of AWAETs under study. Further, AWAETs should inform not only about TP violated, but also FN. Such an approach will help the appraiser to choose correct AWAETs for assessments and make plans to minimize numbers of FN.

2. In fact, the AWAETs exhibit low levels of completeness as all the tools determined on average less than 33.8% of TP (only 1 out 3 TP) in 8 types of SC. As for the best tool, Cynthia Says determines 59% of TP. The study has also revealed that the tool's completeness changes between various types of sites as well as certain guideline principles and levels of conformance. Indeed, SC of the Perceivable principle with the A level exhibit an excessively high level of completeness (5814 (64.3%) TP), as compared with rest. Sites such as tajikngo.tj (397 TP), toptj.com (161 TP), asyl.at (496 TP) and careesma.at (153 TP) were the most inaccessible ones. It can be observed that there is no tool that can perform best across all the principles and types of websites. Besides, almost all automation tools provide a great number of tests that are influenced by sheer numbers of "warnings" tests to verify. In such situations, AWAEMs should employ smarter algorithms to reduce those numbers of tests "to verify".
3. As a matter of fact, the correctness of AWAETs is the highest among all the considered quality criteria (70.3% on the average). However, there are cases where a tool that exhibits the highest coverage and completeness values, simultaneously has the largest number of incorrectly identified issues. This indicates that automating as much as possible TP in the world of software programming leads to the negative consequences of increasing numbers of unwanted FP, with the exception of a very few AWAETs like Cynthia Says. Preferably, the superior tool is the one that gives exact determination and forecasts of relatively high numbers of accessibility violations of a website with a minimal incorrectness score. That is why Cynthia says can be relatively considered to be such a tool since it shows the better coverage [23 SC – the maximum value], completeness [59% (4339TP) - the maximum value], correctness [more than average value-75%] and specificity [16.7% (1241TP) - the maximum value] as compared to the others.
4. The study adds further evidence that the inter-reliability of the eight assessment tools is the lowest among the other tested quality factors and quite variable between different combinations of tools. Findings indicate that the inter-reliability of the AWAETs lies between 1.56% (116 TP) - for all the eight tools and 18.32% (1360 TP) – for the pairs of various tools. Thus, another finding is that tools detect not identical violations. In turn, the intra-reliability evaluation suggests that the AWAETs were stable. E.g. according to the statistical analysis with the Kruskal-Wallis approach, AWAETs behave more similar in reporting TP, FP and FN, except in the case of FN, missed from Austrian sites. The Cronbach's alpha test revealed that the tools have poor (0.568) and acceptable (0.793) similarity results with respect to the number of TP flagged for Tajik and Austrian sites, respectively.
5. Even, if AWAETs are less consistent and do not have enough reliable results, it is because they perform well in some specific situations, while others are weak. Further, the amount of detected different and distinctive types of TP varies noticeably from zero to 16.7% (1241) of TP. Our three tools out of eight had no test that could be categorized as specificity. A total of 32% (2389) TP belongs to specificity, meaning that AWAETs focus on more frequent TP.
6. The research findings indicate that the AWAET have pros and cons, tend to produce false or misleading outcomes, inform correct code as incorrect, miss true positives or "flag" them as warnings. Therefore, overall damage caused by the use of tools alone was calculated – coverage lies in a range of 12.7-32.4% of the violated SC, completeness occurs in a wide range of 10-59%, correctness ranges between 39% and 81% and specificity lies in a range of 0-16.7%. Drawing on these facts, we do not recommend for the careful appraiser to solely rely on AWAETs in order to properly assess the accessibility of their sites. Nonetheless, today's AWAETs are capable of producing a bit more TP in terms of completeness and specificity than in the previous years.
7. This research paper investigated the importance of using combinations of AWAETs and strongly recommends such approaches. The mixture of tools can validate each other and find some other non-detected SC and TP. However, it should be based on existing strengths and weaknesses of AWAEMs, e.g. they might be employed on those SC they ensure high levels of coverage, completeness, correctness, and/or specificity. As demonstrated, using the multiple tools boost the overall coverage to 54.9% (39) SC, completeness up to 81.4% (7356) TP as well as helps to reach the best case scenarios for all the remaining measured quality indicators. However, expected results from combinations may be worse if not to pick up necessary AWAEMs. In fact, there is a broad list of ninety-three registered tools [60] that could be utilized in actual practice to maximize assessment results. As far as the combination of AWAEMs is concerned, it is the best way if not the only one to increase the efficiency of AWAEMs in all aspects. However, any suitable combinations cannot guarantee the detection of all

TP that exist in sites.

8. Based on all the above statements, issues identified with AWAETs should not be trusted as they are inadequate and insufficient in all respects. Instead, it is imperative to always revise and complete automated evaluations manually. Besides, in certain cases, AWAETs are technically oriented and represent negative or positive results against accessibility rulesets without considering contextualized interpretations. Hence, comprehensive knowledge of human experts about guidelines, accessibility requirements, general characteristics of AWAEMs and personal skillset are the final determining factors for ensuring a high-level of Web accessibility.
9. As claimed previously, it is also concluded that AWAEMs are not independent assessment methods, but rather useful additions to the other evaluation techniques, e.g. heuristic and user evaluation methods. As a matter of fact, making a website accessible requires considerable time and efforts in handling web page code, numerous accessibility as well as design and usability guidelines. In this perspective, the AWAEM as a primary toolkit of a careful assessor can significantly improve testing results by severely reducing time and effort. Also, due to their nature, AWAEMs can easily find certain SC and TP that are particularly hard to find with other approaches. As a result, the larger a site, the more it is needed to rely on AWAEMs.
10. Unfortunately, the current evaluation results showed that reaching and even more, maintaining the accessibility of sites still remain an actual problem as the vast majority of our sample of modern websites are far away from being accessible. Also, the researchers Pils et al. (2009) [99] came to the same unsatisfactory conclusion after analyzing thirty-two government websites from Upper Austria in compliance with WCAG 2.0. Of the evaluated websites, none of them fulfilled the required minimum criteria (the level A) with 100% and only four sites reached an overall score of "Good" [99]. In our case, around 4-6% of the surveyed websites passed both automatic and manual assessments with very fewer errors, which means they have met the minimum requirement for the Web accessibility presented by WCAG 2.0. Furthermore, the reported common accessibility barriers by automated tests that caused entire sites to become fully or partially inaccessible are (listed from the largest to smallest numbers of obtained TP): C12, G18 F24, G17, F65, C14, G18, C21, G134, F24 and G140. The rest or additional common issues that were found by the experts are: H37, F38, H67, H45, F3 and G145. It should be emphasized that whether web pages include equivalent information in different forms (e.g. simultaneously in the text, audio and video formats) is not addressed by the AWAETs.
11. The subjectivity of certain issues in standards, including WCAG 2.0 is problematic for the reliable functioning of AWAEMs. Many SC of WCAG 2.0 in the category of ease of use, look and feel, informative and accurate text of form controls, manners in which content is displayed, accessibility of error messages etc. are either too subjective or too complex to be accurately tested or caught at all with AWAEMs, and therefore fall outside the scope of fully automatic evaluations. Conversely, engaging human experts throughout the test process could additionally verify 1681 (18.6%) subjective TP. Hence, again the same conclusion with the above point 8 is that relying on AWAEMs alone to validate sites for compliance is a great mistake. Instead, our findings, taken together with previous research, explicitly suggest an association of automated, manual and user testing methodologies to achieve the best results in accessibility evaluation.

As the average accessibility level of sites is very low, it is necessary to emphasize that one should not forget that the disabled are also human beings and they are in large numbers. In addition, laws require all sites, especially governmental ones to be accessible to all. Therefore, from a good quality of Web accessibility benefits everyone and thus, its achievement is mandatory - at the request of the law, necessary - as a sign of respect to people with disabilities and elderly and finally, beneficial - to significantly increase revenue due to the growing number of users, clients and other stakeholders.

In the future, the presented new methodology "5PhM-for-AWAEMs" and outcomes of this paper improve the ease of evaluation and comparison of different AWAEMs, assist to an appropriate choice of AWAEMs and consequently contribute to rational use of AWAEMs in evaluating and enhancing Web accessibility. Hopefully, the reported nineteen pros and fourteen cons of AWAEMs, fifteen recommendations for the AWAEM's quality improvements and the comparison results of this study can lead to more intense competition among producers of AWAEMs in order to make AWAEMs themselves smarter and more efficient. In the end, I firmly believe that automatic Web accessibility repair will be the main research and development priority for AWAEMs in the future.

ACKNOWLEDGMENTS

The work reported in this paper has been funded by the TARGET II Project of the Erasmus Mundus Program and conducted at the Johannes Kepler University of Linz, Austria.

REFERENCES

[1] World Bank, Disability Overview, Apr 04, 2016. Available online: http://www.worldbank.org

[2] Fogg B. J., Swani P., Treinen M., Marshall J., Osipovich A., Varma C., Laraki O., Fang N., Paul J., Rangnekar A.

and Shon J., Elements that affect Web credibility: Early Results From a Self-Report Study, Proceedings of CHI'00, Extended Abstracts on Human Factors in Computing Systems, 2000, pp. 287-288.

[3] Killam B. and Holland B., Position Paper on The Suitability of Task Automated Utilities for Testing Web accessibility Compliance, Usability professionals' association conference, 2001. Available online: http://www.upassoc.org/conf2001/

[4] Lindenberg J. and Neerinex M.A., The Need for a "Universal Accessibility" Engineering Tool. Proceedings, Interact '99 workshop: Making designers aware of existing guidelines for accessibility, August 1999.

[5] Zeng X., Evaluation and Enhancement of Web Content Accessibility for Persons With Disabilities. Ph.D. Thesis, University of Pittsburgh, 2004.

[6] Good A., An Investigation of a Method for Improving Accessibility to Web-Based information for Users with Impairments, Doctoral Thesis, University of Portsmouth, Portsmouth, 2008.

[7] Letourneau C., Accessible Web Design – a Definitiion, 2001. Available online: www.starlingweb.com/webac.htm

[8] World Wide Web Consortium W3C, Web Content Accessibility Guidelines – WCAG 1.0, 5 May 1999.

[9] Paciello M., Web accessibility for People with Disabilities. CMP Books, 2000, ISBN: 1929629087.

[10] Jim Thatcher, Cynthia Waddell, Shawn Henry, Sarah Swierenga, Mark Urban, Michael Burks, Bob Regan and Paul Bohman., Constructing Accessible Web sites. Glasshouse 2002.

[11] Slatin J. and Rush. S., Maximum Accessibility: Making Your Web Site More Usable for Everyone. Addison-Wesley, 2003.

[12] Paciello M., Web accessibility for People with Disabilities. CMP Books, 2000, ISBN: 1929629087.

[13] Michigan State University, Web Accessibility, 2017. Available online: http://webaccess.msu.edu

[14] Web Content Accessibility Guidelines 2.0, W3C Working Draft, 11 February 2005.

[15] ISO IS 9241-171:2008 Ergonomics of Human-System Interaction - Guidance on Software Accessibility. (a restructured version of ISO TS 16071).

[16] ISO TS 16071: 2002 Guidance on accessibility.

[17] The order of the government of the Republic of Tajikistan, The State Program of Development and Implementation of Information and Communication Technologies in the Republic of Tajikistan, December 3, 2004, No. 468.

[18] The order of the government of the Republic of Tajikistan, The State Program of Development and Implementation of Information and Communication Technologies in the Republic of Tajikistan for 2014-2017, July 3, 2014, No. 428.

[19] National Action Plan on Disability 2012-2020, Strategy of the Austrian Federal Government for the Implementation of the UN Disability Rights Convention (Nationaler Aktionsplan Behinderung 2012-2020. Strategie der Österreichischen Bundesregierung zur Umsetzung der UN-Behindertenrechtskonvention), 2012.

[20] UN, Convention and Optional Protocol Signatories and Ratification. Available online: www.europarl.europa.eu

[21] Harper S. and Yesilada Y., Web Accessibility, Springer, London, United Kingdom, 2008.

[22] Coyne K. and Nielsen J., How to Conduct Usability Evaluations for Accessibility: Methodology Guidelines for Testing Websites and Intranets With Users Who Use Assistive Technology, Nielsen Norman Group, Oct. 2001.

[23] Brajnik G., Web accessibility testing with barriers walkthrough, March 2006.

[24] DRC, Formal investigation report: Web Accessibility. Disability Rights Commission, April 2004.

[25] Henry S.L. and Grossnickle M., Accessibility in the User-Centered Design Process. Georgia Tech Research Corporation, Atlanta, Georgia, USA, 2004.

[26] W3C/WAI, Conformance Evaluation of Web Sites for Accessibility, 2008. Available online: www.w3.org/

[27] Dey A., Accessibility Evaluation Practices – Survey Results, 2004. Available online: http://deyalexander.com

[28] Ivory M.Y. and Chevalier A., A Study of Automated Web Site Evaluation Tools. The Information School University of Washington, Aline Chevalier Department of Cognitive Psychology University of Provence, Technical Report UWCSE-02-10-01 October 8, 2002.

[29] Brajnik G., Comparing Accessibility Evaluation Tools: A Method for Tool Effectiveness. Universal Access in the Information Society, Springer Verlag, Oct. 2004, 3(3-4), pp. 252-263.

[30] Brajnik G., Beyond Conformance: the Role of Accessibility Evaluation Methods. In S. Hartmann et al., editor, WISE 2008: 9th Int. Conference on Web Information Systems Engineering – 2nd International Workshop on Web Usability and Accessibility IWWUA08, LNCS 5176, 2008, pp. 63–80, Auckland, New Zealand, Sept. 2008c. Springer-Verlag. Keynote speech.

[31] Vigo M., Justin B. and Vivienne C., Benchmarking Web accessibility Evaluation Tools: Measuring the Harm of Sole Reliance on Automated Tests, In: Proceedings of 10th International Cross-Disciplinary Conference on Web accessibility (W4A), ACM Press, May 13-15, 2013, Rio de Janeiro, Brazil.

[32] Kaur A. and Diksha D., Comparing and Evaluating the Effectiveness of Mobile Web Adequacy Evaluation Tools. Article in Universal Access in the Information Society, May 2016. DOI: 10.1007/s10209-016-0466-z.2.

[33] Pivetta E. M., Saito D. S., Flor C. S., Ulbricht V. R., Vanzin T., Automated Accessibility Evaluation Software for Authenticated Environments - A Heuristic Usability Evaluation. HCI (7) 2014, pp. 77-88

[34] Al-Khalifa H. S., WCAG 2.0 Semi-automatic Accessibility Evaluation System: Design and Implementation. Computer and Information Science, 2012, Vol. 5, No 6.

[35] Sukhpal K., An Automated Tool for Web Site Evaluation, International Journal of Computer Science and Information Technologies, 2012, 3 (3), pp. 4310 – 4313.

[36] Molinero A. M. and Kohun F. G., Reliability in Automated Evaluation Tools for Web accessibility Standards Compliance. In issues in Information Systems, Volume VII, No. 2, 2006.

[37] Centeno V. L., Kloos C. D., Fisteus J. A. and Alvarez L. A., Web accessibility Evaluation Tools: A Survey and Some Improvements. Journal Electronic Notes in Theoretical Computer Science(ENTCS) archive, 2006,157(2), pp. 87-100.

[38] Xiong J., Farenc C. and Winckler M., Analyzing Tool Support for Inspecting Accessibility Guidelines during the Development Process of Web Sites. In Proceedings of 1st IWWUA workshop inside WISE. Springer LNCS 4832, 2007, pp. 470-480.

[39] Ahmad Al-A., Ibraheem Y.Y. and Ahmaro M.M., Comparison Between Web accessibility Evaluation Tools. Almadinah Islamic Studies, 2010, 1(66): Malaysian Studies.

[40] Akgül Y. and Vatansever K., Web Content Accessibility of Municipal Web Sites in Turkey. February, 2016, 7(1), pp. 43-48. doi: 10.12720/jait.7.1.43-48
[41] Hackett S., Parmanto B. and Zeng X., A Retrospective Look at Website Accessibility Over Time. Behaviour and Information Technology, 2005, 24 (6), pp. 407-417.
[42] Chisholm W. Vanderheiden G. and Jacobs I. (Eds.), Web Content Accessibility Guidelines 1.0, May 5 1999.
[43] International Organization for Standardization: ISO/IEC 40500:2012.
[44] Caldwell B., Cooper M., Reid L.G. and Vanderheiden G. (eds.), Web Content Accessibility Guidelines 2.0, W3C Recommendation, 2008. Available online: www.w3.org
[45] World Wide Web Consortium (W3C). Web Content Accessibility Guidelines (WCAG) 2.0, 2008. Available online: https://www.w3.org/TR/WCAG20/
[46] Brajnik G., A Comparative Test of Web accessibility Evaluation Methods. In Proceedings of the 10th international ACM SIGACCESS Conference on Computers and Accessibility (Halifax, Nova Scotia, Canada, October 13 - 15, 2008), Assets 08, ACM, New York, NY, pp. 113-120.
[47] Brajnik G., Mulas A. and Pitton C., Effects of Sampling Methods on Web accessibility Evaluations. In Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility, 2007, Assets 07, pp. 59-66.
[48] Sevilla J., Herrera G., Martínez and Alcantud F. Web accessibility for Individuals with Cognitive Deficits: A Comparative Study Between an Existing Commercial Web and its Cognitively Accessible Equivalent. ACM Transactions on Computer Human Interaction, 2007, 14(3), p. 12.
[49] Power C., Freire A., Petrie H., and Swallow D., Guidelines are Only Half of the Story: Accessibility Problems Encountered by Blind Users on the Web. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, CHI '12, 2012, pp. 433-442.
[50] Gray W.D. and Salzman M.C., Damaged Merchandise: A Review of Experiments That Compare Usability Evaluation Methods, Human–Computer Interaction, 1998, 13(3), pp. 203– 261.
[51] Microsoft and HiSoftware 2009, Microsoft Web accessibility Handbook.
[52] Petrie H. and Kheir O., The Relationship Between Accessibility and Usability of Websites. In: Proc. CHI'07, ACM, CA, 2007, pp. 397–406.
[53] Mankoff J., Fait H. and Tran T., Is Your Web Page Accessible? A Comparative Study of Methods for Assessing Web Page Accessibility for the Blind. In Proc. SIGCHI Conference on Human Factors in Computing Systems, CHI'05, 2005, pp. 41-50.
[54] Hertzum M. and Jacobsen N.E., The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. Int. Journal of Human-Computer Interaction, 2001, 1(4), pp. 421–443.
[55] Lang T., Comparing Website Accessibility Evaluation Methods and Learnings from Usability Evaluation Methods, 2003.
[56] Sears A.: Heuristic Walkthroughs: Finding the Problems Without the Noise. Int. Journal of Human-Computer Interaction, 1997, 9(3), pp. 213–234.
[57] Kelly B., Sloan D., Brown S., Seale J., Petrie H., Lauke P. and Ball S., Accessibility 2.0: People, Policies and Processes. In: W4A 2007: Proc. of the 2007 international cross-disciplinary conference on Web accessibility (W4A), 2007, pp. 138–147. ACM, New York.
[58] Carter J. and Markel M., Web accessibility for People with Disabilities: An Introduction for Web Developers. IEEE Transactions on Professional Communication, 2001,44(4), 225- 233.
[59] Christopher B. and Elaine P., Development and Trial of an Educational Tool to Support the Accessibility Evaluation. Proc. In: Proceedings of the 2011 International Cross-Disciplinary Conference on Web accessibility W4A, 2011, pp. 2. Available online: http://dx.doi.org/10.1145/
[60] Web accessibility Evaluation Tools List, Updated June 2017 (first published March 2006). Available online: https://www.w3.org/WAI/ER/tools/
[61] Lopes R. and Carric L., Macroscopic Characterizations of Web Accessibility. New Review of Hypermedia and Multimedia, 2010, 16(3), pp. 221-243.
[62] Hackett S. and Parmanto B., Homepage Not Enough When Evaluating Web Site Accessibility, Internet Research, 2009, 19(1), pp. 78–87.
[63] Patton M.Q., Qualitative Evaluation and Research Methods, Sage Publications, Newbury Park, California, 1990, p. 532 .
[64] Web Tools for Quality, Accessibility, Standards Compliance. Available online: http://valet.webthing.com
[65] Schiavone A. G. and Paterno F., An Extensible Environment for Guideline-Based Accessibility Evaluation Of dynamic Web applications. Univ Access Inf Soc, 2015, Vol. 14, pp.111–132, DOI 10.1007/s10209-014-0399-3.
[66] Sean P. Aune, 12 Tools to Check Your Site's Accessibility, July 06, 2009. Available online: www.sitepoint.com
[67] Oleg Mokhov, 20 Наиболее Необходимых Инструментов для Проверки Отображения Сайта, 18 march 2011. Available online: habrahabr.ru/company/aiken/blog/
[68] Justin Mifsud, 10 Free Web-Based Web Site Accessibility Evaluation Tools, August 22, 2011. Available online: http://usabilitygeek.com/10-free-web-based-web-site-accessibility-evaluation-tools/
[69] Simon Heaton, 9 Tools for Website Accessibility Testing, Jun 29, 2016. Available online: shopify.com
[70] Achecker (Web accessibility Checker), Inclusive Design Research Centre, 2011. Available online: achecker.ca/
[71] Gay G. and Li C.Q., AChecker: Open, Interactive, Customizable, Web accessibility Checking. In: International Cross Disciplinary Conference on Web Accessibility-W4A, 2010, pp. 1–2.
[72] Cynthia Says, HiSoftware Inc, 2003. Available online: http://www.cynthiasays.com
[73] Koutsabasis P., Vlachogiannis E. and Darzentas J.S., Beyond Specifications: Towards a Practical Methodology for Evaluating Web Accessibility, Journal of Usability Studies, August 2010, 5(4), pp.157-171.
[74] MAUVE (Version: 1.3), Human Interfaces in Information Systems Laboratory - ISTI-CNR. Available online: http://hiis.isti.cnr.it:8080.
[75] SortSite, Power Mapper Software, 1996. Available online: http://www.powermapper.com.
[76] Declaring Conformance on Web Accessibility, ANEC Print version 21, May 2011, pp. 1-44. Project Reference: ANEC-R&T-2009-ICT-001final
[77] CEAPAT, Fundación CTIC, Spanish Ministry of Employment and Social Affairs (IMSERSO), Online Web accessibility test. Available online: www.tawdis.net
[78] Tenon (Version: 1.0), by Tenon. Available online:

tenon.io
[79] Quickly Check Your Website for Common Accessibility Problems with tenon.io. Available online: marcozehe.de
[80] European Internet Inclusion Initiative (EIII Checker). Available online: http://checkers.eiii.eu/
[81] Snaprud M., Benchmarking Results from 1000 European Websites. EIII supported by the European Union Seventh Framework Programme (Grant agreement no: 609667).
[82] WAVE (Web accessibility evaluation tools), WebAIM, 1999. Available online: http://wave.webaim.org
[83] Pivetta E.M., Flor C.F., Saito D. S. and Ulbricht V.R., Analysis of an Automatic Accessibility Evaluator to Validate a Virtual and Authenticated Environment. International Journal of Advanced Computer Science and Applications, 2013, Vol. 4, pp. 15-22.
[84] Sears A., Heuristic walkthroughs: finding the problems without the noise. Int. Journal of Human-Computer Interaction, 1997, 9(3), pp. 213–234.
[85] Gray W. and Salzman M. Damaged merchandise: a review of experiments that compare usability evaluation methods. Human–Computer Interaction, 1998, 13(3), pp. 203–261.
[86] Hertzum M. and Jacobsen N., The evaluator effect: a chilling fact about usability evaluation methods. Int. Journal of Human-Computer Interaction, 2001, 1(4), pp. 421–443.
[87] Lang T., Comparing website accessibility evaluation methods and learnings from usability evaluation methods, 2003. Available online: www.peakusability.com.au
[88] Hartson H. R., Andre T. S. and Williges R. C. Criteria for Evaluating Usability Evaluation Methods. Int. Journal of Human-Computer Interaction, 2003, 15(1), pp. 145–181.
[89] Diaper D and Worman L.: Two Falls Out of Three in the Automated Accessibility Assessment of World Wide Web Sites: A-prompt v. Bobby. In: People and Computers, Vol. 17, 2003, pp. 349-363, Springer.
[90] Henry S.L. and Grossnickle M.: Just Ask: Accessibility in the User-Centered Design Process. Georgia Tech Research Corporation, Atlanta, Georgia, USA; On-line book, 2004.
[91] W3C/WAI., Conformance Evaluation of Web Sites for Accessibility. Available online: www.w3.org/WAI/eval/
[92] Centeno V. L., Kloos C. D., Fernandez L. S. and Fernandez N. G., Device independence for Web Wrapper Agents Workshop on Device Independent Web Engineering (DIWE'04), 26 July 2004, Munich.
[93] Cooper M., Limbourg Q., Mariage C. and Vanderdonckt J., Integrating Universal Design Into a Global Approach for Managing Very Large Web Sites. In Proceedings of the 5th ERCIM Workshop on User Interfaces for All, 1999.
[94] Farenc C., Liberati V. and Barthet M.F., Automatic Ergonomic Evaluation: What Are The Limits? In Proceedings of the Second International Workshop on Computer-Aided Design of User Interfaces, CADUI '96, 1996, pp.159-170.
[95] Ivory M.Y. and Hearst M. A., State of the Art In Automating Usability Evaluation of User Interfaces. ACM Computing Surveys, December 2001, 33(4), pp. 470–516.
[96] Chevalier A. and Ivory M. Y., Web Site Designs: Influences of Designer's Experience and Design Constraints. Submitted for Publication, 2002.
[97] Lee A. and Hanson V., Enhancing Web Accessibility, Proceedings ACM Multimedia 2003, pp. 456-457, ACM Press.
[98] Abduganiev S.G., Pils M. and Roithmayr F., Elicitation of Criteria Weights for the Web Quality Evaluation Method Universal Star: By Using Different Ranking Methods. The Strategies of Modern Science Development: Proceedings of the X International scientific–practical conference. North Charleston, USA, 12-13 April 2016. - North Charleston: CreateSpace, 2016, pp. 11-24.
[99] Pils M., Ganglberger M. and Höller J., Barrierefreiheit von Behörden Websites - Anspruch Und Realität. in: H. Wandke /S. Kain/D. Struve (Hrsg.): Mensch und Computer 2009; Grenzenlos frei!? München Oldenburg Verlag, 2009, S.3-12.

Authors' Profiles

Siddikjon G. Abduganiev obtained his Master's Degree of Science in Systems Engineering with honors from the Khujand branch of the Technological University of Tajikistan (KBTUT) in 2007, Khujand city, Tajikistan. Until 2012, after university graduation, he has worked as a junior and senior lecturer in KBTUT and the Khujand Polytechnic Institute of the Tajik Technical University (KPITTU) in the departments of Computer Programming and Information Technologies and Higher Mathematics and Informatics, respectively. At the present time, he is pursuing a Ph.D. degree in Business Informatics at the Johannes Kepler University Linz, Austria. His core research areas are software developing and testing. He has published and presented around twenty research papers in national and international conferences and peer reviewed journals.

How to cite this paper: Siddikjon Gaibullojonovich Abduganiev, "Towards Automated Web Accessibility Evaluation: A Comparative Study", International Journal of Information Technology and Computer Science(IJITCS), Vol.9, No.9, pp.18-44, 2017. DOI: 10.5815/ijitcs.2017.09.03