University of Agder

FACULTY OF ENGINEERING AND SCIENCE

DEPARTMENT OF ICT

Master of Science Thesis

Combination of automatic and manual testing for web accessibility

Author: *Justyna Magdalena Mucha*

Degree programme: *Information and Communication Technology*

Supervisors: *Jaziar Radianti Dr (UiA) & Mikael Snaprud Dr techn. (Tingtun AS)*

Grimstad, 2018

Hereby I declare that this paper is my original authorial work, which I have worked out by my own. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Acknowledgements to my supervisors for their valuable remarks, support and excellent guidance through the research process.

Contents

List of Figures

List of Tables

List of Abbreviations

APC	Accessibility Pie Chart
AEM	Accessibility Evaluation Method
AT	Automated Testing
ATT	Automated Testing Tool
BW	Barrier Walkthrough
CR	Conformance Review
DOM	Document Object Model
EARL	Evaluation and Report Language
EIII	European Internet Inclusion Initiative
IR	Information Retrieval
POUR	Perceivable, Operable, Understandable and Robust
QA	Quality Assurance
SC	Success Criterion
SPA	Single-Page Application
UN	United Nations
UT	User Testing
UTT	User Testing Tool
UWEM	Unified Web Evaluation Methodology
W3C	World Wide Web Consortium
WAI	Web Accessibility Initiative
WAD	Web Accessibility Directive
WAI	Web Accessibility Initiative
WCAG	Web Content Accessibility Guidelines
W3C	World Wide Web Consortium

Abstract

Web accessibility is an indispensable medium for online communication and digital inclusion nowadays. With the recent adoption of the Web Accessibility Directive making the Internet resources accessible has become a legal obligation and strikes a need for more detailed and reliable ways of web accessibility evaluation of the websites.

Throughout the years, many tools have been developed for testing web accessibility as well as a plethora of metrics that are expected to convey the results. Unfortunately, in most cases the findings appear to be incomplete since the studies rely only on one testing method, i.e., automatic or manual. The study has set itself a goal to contribute with knowledge to solving three research questions. First, how to combine results from automated and manual evaluation of web accessibility? Second, how to express the integration results in a quantitative manner? Finally, what is the impact of the dynamic content on the integration results when the content of the website is frequently updated and personalized?

This thesis proposes a novel approach to integration of manual and automated accessibility testing, where the results of the evaluations are combined on the basis of accessibility guidelines. Additionally, a quantitative metric – Union Score, together with a graphical visualization called Accessibility Pie Chart, are propounded, as the means for expressing the outcomes of the accessibility evaluation with use of the combined approach.

The research has been grounded on the mixed-method approach and embedded the findings of the conducted interviews into a quantitative study. In order to find emphirically the most suitable method for combining manual and automated testing, fifteen web pages selected from two websites were chosen for evaluation with two testing tools: WTKollen Checker and WTKollen User-Testing Tool.

The findings of the analysis show that WCAG 2.0 may serve as a bridge between manual and automated evaluation outcomes and result in an increased coverage of the Success Criteria. The proposed metric has been preliminarily validated with regard to its application for benchmarking purposes and supplemented with a graphical way of presenting accessibility testing results. Furthermore, it is concluded that the suggested integration approach can be deployed. Yet, the challenge of dynamic content evaluation requires more research attention.

The study has contributed to the current state of knowledge about web accessibility evaluation and the results are expected to be used for implementation of the novel approach. For the future paths, a more extended study on the proposed metric's properties is advised. Also, the importance of further research in the area of dynamic content evaluation is highlighted.

Preface

This thesis is original, unpublished, independent work by the author, J.M. Mucha. This master project is connected to the WTKollen project, financially supported by the Swedish Post and Telecom Authority (PTS). The research focus of the thesis has been put on combining results from automated accessibility testing with those from manual accessibility evaluation.

1. Introduction

> The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect.
>
> Tim Berners-Lee

Statistics show that 81% of the population in the European Union (aged 16 to 74) is regularly using the Internet [1]. According to the World Health Organization (WHO) Report on Disability, over a billion people, which is about 15% of the world's population, have some form of disability [2]. Accessibility acts as the medium and facilitator for the full integration of all persons in society, regardless of disability. The goal of inclusion signifies that there is an obligation to create environments that provide access to all aspects for all people on an equal basis.

Following the World Wide Web Consortium (W3C) definition [3], *web accessibility* means that websites, tools, and technologies are designed and developed so that people with disabilities can use them. More specifically, so that people can perceive, understand, navigate, and interact with the Web as well as contribute to it.

A few years ago web accessibility was associated only with the possibility for the disabled to take advantage of the Internet. It is still the case, however web accessibility has broadened its meaning. Over time web accessibility has become a crucial consideration for every user of the Internet. The biggest change that happened in the recent years is that web accessibility has also become a concern for an average user. With all the technological advance that can be witnessed today: digitalization, Artificial Intelligence (AI) boom, accessibility has moved from the backs to the foreground. There can be observed a growing awareness of why web accessibility is vital to so many users. What is the benefit of technological progress, advanced applications if they cannot be fully used by the people?

Accessibility issues may hinder users from everyday activities like browsing the website or using a digital contact form. More and more communication between public sector agencies and users are conducted online. Moreover, thanks to the growth of accessibility consciousness, organizations are increasingly being expected to clearly state their accessibility statement, as

required by the WAD. Web accessibility is no longer an encouragement but it has become a legal obligation to allow that all people can participate in the digital world.

1.1. Background

The overall goal of web accessibility is to enable equal participation of people in the digital world. Various initiatives that are undertaken to make the Web more accessible set the directions for enhancements. In order to improve web accessibility of digital resources – an accessibility evaluation needs to be made. The issues causing accessibility barriers have to be first identified, so that they can be repaired.

Web accessibility evaluation is a broad field that encompasses both technical and non-technical aspects [4]. From the technical angle, assessment of conformance to accessibility regulations and guidelines can be named. On the other hand, non-technical aspects would involve users to the checking process. Regarding that, diverse approaches to evaluation can be adopted. Three main kinds of accessibility evaluation can be distinguished: manual, automated, and semi-automated testing. Manual evaluation assumes that tests are carried out by human evaluators, supported by given instructions or guidance. Automated testing approaches the evaluation in a fully-automated way, conducting the accessibility tests with aid of software tools that do not require human input. Finally, the third way of testing web accessibility – semi-automated checking, encompasses evaluation, partially done by the software tools, that is combined with human judgment, e.g. to decide on the unsure test results or conduct tests that were not implemented in the tools, like the tests that are currently hard to automate. For instance, verification whether a title is well describing the subsequent content may only be done by a human.

1.1.1. Need for Integration

Various reasons for combination of manual and automated testing can be identified. Recent advance in technology of web development creates new challenges for accessibility evaluation. More accurate and reliable methods, being at the same time efficient and affordable, are in high demand.

One cannot rely only on one tool.

It has been observed that neither automated nor manual evaluation can solely deliver a complete report on web accessibility of a website. There still remains a gap between the efforts of these testing approaches. No one evaluation method alone can identify all problems [5]. Vigo *et al.* [6] has conducted an extensive study that measures the sole reliance on automated testing. The effectiveness of six tools were investigated with a focus on their coverage, completeness and correctness with regard to Web Accessibility Guidelines (WCAG) 2.0 conformance. The authors claim that entrusting the accessibility evaluation only to the automated tools entails that one

out of two Success Criteria (henceforth SC) will not even be analyzed and among those involved, exclusively four out of ten will be found at the further risk of generating false positives. The research shows also that the coverage of SC is at most equal to 50%. Coverage, completeness and correctness of the tools were measured in comparison to the outcomes of an in-depth experts analysis. Another study reports that only approximately 20% of all accessibility tests can be automated [7]. The rest can be carried out manually.

Manual and automated methods can complement each other.

Both manual and automated testing have their unique advantages that can be utilized for the evolution of the assessment methodology. While automated checking is cost effective, robust and replicable, manual evaluation can be seen time consuming but necessary to cover the tests that cannot be automated. The strengths of one approach to accessibility evaluation may fill in the gap of the other approach. Manual and automated methods can complement each other. There is a need for integration so that there is no more dilemma over the Automated Testing Tools (ATT) and User Testing Tools (UTT) about which one is better. Instead – a mutual cooperation aimed at making the Web more accessible to people can be achieved. Moreover, manual testing can provide a more detailed report of accessibility issues. When combined, the approaches can work together towards a more accessible Internet.

Accessibility integration can support the efforts of Web Accessibility Directive (WAD).

Thanks to the combination of manual and automated testing, the accessibility statements imposed by enactment of WAD [8], can convey a more complete picture of the real accessibility the website provides, not only a statement of conformance to the accessibility guidelines. A more detailed report is possible to be shared. It is important to engage people in the evaluation process. Human input becomes in this case especially necessary.

1.1.2. Benefits of Accessibility Metrics.

Providing a quantified information about the web accessibility state of a website has several benefits. In order to gauge the level of accessibility, some measures are needed to be established. Following [9], a metric is a procedure for measuring a property of a web page or a website. In case of the web accessibility metrics, the failure-rate metric calculates the ratio between the number of accessibility barriers of a particular set of criteria over the number of failure points for the same criteria.

First, the metrics may aid developers in the Quality Assurance (QA) process to control the level of accessibility throughout the development cycle. Secondly, monitoring accessibility trends on websites can be helpful to capture how the page updates impact its accessibility as well as implementation of the adopted accessibility laws and standards. In addition to that, web accessibility metrics allow for benchmarking of web resources, whether the evaluation is

done within the same region or is more spread geographically. An example of such analysis is an assessment of the level of accessibility of e-governmental websites. Goodwin *et al.* [10] has run a global analysis of web accessibility of national government portals and ministry websites, which presents the results using the benchmarking methodology. Moreover, accessibility metrics support conformance testing of websites allowing for accessibility conformance claims issuing (A, AA, or AAA in case of WCAG 2.0). In addition to that, the quantitative accessibility scores may be used for generating benchmarking lists and supports indirectly a competition between the websites' owners in terms of accessibility. Last but not least, a quantified information about the web accessibility may serve for Information Retrieval (IR) purposes. Web resources can be retrieved not only accordingly to the information relevance but also to the accessibility level they present.

1.1.3. EIII Roots

The idea of manual and automated testing for web accessibility was pursued in the European Internet Inclusion Initiative (EIII) project. The project itself highlighted the need for harmonization of web accessibility evaluations and pointed out a handful of benefits and challenges of combining automated evaluation and user-testing [11]. While automated testing is robust, can encompass many pages and produce repeatable results, manual evaluation can deliver detailed and more complete reports for a smaller number of web pages. Designing automated tests in a way that they could produce three different results: *pass, fail, to be verified*, was the first attempt of involving humans into the process of web accessibility evaluation. The results labelled by *to be verified* tag became an input data for the UTT tool. A simple case of checking whether an image has implemented an *alt* attribute can reflect the idea of results verification. Even though the ATT tests if the attribute is present, it cannot verify if the text conveys the intended information – crucial for the user. At that point, human evaluation becomes necessary. Notwithstanding the efforts, no physical deliverables has been provided in the EIII project on integration methodology apart from an extensive report on that topic [12] and a separate UTT tool. Yet, the substantial experience from the EIII project can be taken further to the WTKollen project through this research.

Early automated methodologies were developed in the European Internet Accessibility Observatory (EIAO) project [13], aimed at providing the first large scale website evaluation. The EIAO was then followed by the eGovMon project [14] that implemented large scale evaluation of governmental websites with a conformance to WCAG 2.0 . Both projects brought a significant impact to the works of the EIII initiative at that time. For the expert evaluation, EIII benefited from Unified Web Evaluation Methodology (UWEM) [15], which was an early attempt to unify accessibility results. Noteworthy is to stress that the UWEM 1.0 sought to comply with the

WCAG 1.0. Later on, some efforts on updating the UWEM methodology to be compatible with the WCAG 2.0 standards were done. Moreover, the WCAG-Evaluation Methodology (WCAG-EM) [16] had a vital input to the idea of results combination, providing useful information on sampling and reporting.

1.2. Literature Review

Immense sources on web accessibility are available in the literature. Web accessibility evaluation is strongly determined by the reason of investigation. Recent literature offers a broad overview of the methods for performing web accessibility evaluation. The methods depend greatly on the purpose of the testing. This research focuses its efforts in particular on combination of diverse methods of web accessibility testing as a contribution to a more barrier-free Internet.

First, recent advances in web accessibility guidelines and legislations are presented. Next follows a description of the state-of-the-art of web accessibility evaluation with a separate section for semi-automated testing. Then, the literature review discusses the prevailing work on web accessibility metrics, to arrive at a proposed revision of the previous integration approaches.

1.2.1. Web Accessibility Documents

A desk research on the current accessibility policies and legislations across the world has been conducted. Some general guidance on accessibility in software development are covered by the ISO Standard $9241 - 171 : 2008$ (Guidance on software accessibility) [17]. The document provides ergonomics guidance and specifications for the design of accessible software for use at in various life contexts. It covers issues connected with designing products for people with the widest range of physical, sensory and cognitive abilities, including those who are temporarily disabled, and the elderly.

In October 2012, WCAG 2.0 got adopted as an ISO/IEC International Standard under the number ISO/IEC $40500 : 2012$ [18]. WCAG 2.0 has been also referenced in legislations of many countries across the globe, for instance Autralia, Canada, Japan, and New Zealand. The guidelines developed by the World Wide Web Consortium (W3C) along with the Web Accessibility Initiative (WAI) have become a base for many national standards [19]. For instance, the German *BITV* [20], the French *RGAA* [21], the Norwegian *Forskrift om universell utforming av IKT-løsninger* [22], the Spanish *UNE* $139803 : 2012$ [23], the *Swedish Discrimination Act* [24] or *Section* 508 in the United States [25] together with more recent *ADA* [26]. The *British Equality Act* [27] is also worth mentioning as an example of national initiative of ensuring an accessibility society for all. For the European Union (EU), Mandate 376 embraces WCAG 2.0 as an official accessibility standard [28].

Recent advances in technology have shown a need for an update of the WCAG 2.0. Works done under the umbrella of W3C have resulted in development of the next generation of accessi-

bility guidelines: WCAG 2.1. For the time being, the new version of the guidelines is a Candidate Release. The complete WCAG 2.1 is planned to be published by June 2018. WCAG 2.1. addresses more accessibility requirements for mobile accessibility. Also, the needs of certain disability groups have been put into focus. WCAG 2.1 devotes its efforts on specifications concerning people with cognitive and learning disabilities, as well as people with low vision. Seventeen new SC have been added to the WCAG 2.0.

On 26 October 2016 the European Parliament approved WAD on making the websites and mobile apps of public sector bodies more accessible [8]. The enactment of the WAD appears to be a significant milestone on the path to ensuring accessibility for all. The Directive aims to make public sector websites and mobile applications more accessible, and to harmonize varying standards within the EU. According to the WAD, public sector bodies must provide on regular basis a detailed, comprehensive and clear statement on how their websites and mobile applications comply with the Directive (Art.1, Sec.44). Moreover, WAD imposes that there should be provided a feedback mechanism to enable a user to notify the public sector body of any failures of the website or mobile applications (Art.1 Sec.46). Besides, all twenty-eight EU countries must monitor compliance and report to the European Commission on the results of monitoring every three years, starting from 23 December 2021. The Member States are obliged to bring into force the legislation necessary to comply with WAD by 23 September 2018.

1.2.2. Web Accessibility Evaluation

The task of accessibility evaluation involves a question about the method that should be used, as an unsuitable method may become the culprit [29]. A number of methods are available for an auditor to choose from: automated evaluation, user testing, expert review or a semi-automated method. Each of them has its strengths and weaknesses, thus is preferred to be applied in certain cases. A sound reasoning on the evaluation method choice has been made by Brajnik in [30] : taking into account that there exist several definitions of accessibility, various methods have to be used to evaluate the website. Assessing a website manually does not make sense if one is interested in benchmarking purposes. The evaluation method should suit the evaluation purpose. A comparative test of web accessibility evaluation methods has been conducted [31]. The study involves a comparison between conformance testing and the developed Barrier Walkthrough (BW) method.

Another study which has investigated effects of different computational approaches to web accessibility metrics has been done by Freire *et al.* [32]. The research has been based on an expert evaluation, i.e., no automated testing involved. It has been observed that the ranges and spread of the normalized values differ a lot. The results from checklist review inspections of accessibility are said to have a significant impact on the quantitative results. See also [33].

However, it seems to be difficult to compare evaluation techniques as they measure different variables, which is the case for manual and automated accessibility testing. Both are the servants

of the idea of contributing to a more accessible Web. The synergy can result in a situation where the combined efforts are greater than the total contribution achieved by each method working separately.

A methodology for web accessibility evaluation has been developed by W3C as a supplementary document for WCAG 2.0 [16]. It describes the steps that are common to processes for a comprehensive assessment of the website's conformance to WCAG 2.0. The methodology highlights considerations for evaluators to apply these steps in a context of the tested website. The methodology encompasses five iterative stages, namely Scope defining, Target website exploration, Sampling a representative probe, Audit of the sample, and Reporting. The WCAG-EM tool gives a possibility to generate a machine-readable reports that facilitate processing of the evaluation results. The reports are created in Evaluation and Report Language (EARL).

An attempt on investigating the advancements in web accessibility evaluation methods has been made by Baazeem and Al-Khalifa [34]. The study reveals a lack of substantial evolution of these methods for the years $2011 - 2015$. The most recent information regarding evaluation methodologies that has been found was done as a part of the Digital Single Market initiative [35]. The study maps various methodologies used for monitoring accessibility of the websites in the EU and supports the implementation of the WAD. The study suggests that the European web accessibility monitoring methodology should combine manual and automatic web accessibility monitoring methods.

1.2.2.1. Semi-automatic Testing

Lang in her study of website accessibility evaluation methods concludes that a fully integrated approach is the most appropriate method for accessibility evaluation. Ideally, the approach would combine semi-automatic, manual, and user testing of accessibility features [5]. Moreover, it is stated that for organizations, that have imposed time and cost constraints, combining automated and manual evaluation (called there a 'discount accessibility ') is the best approach for accessibility evaluation. Similar conclusion has been made by Harper and Yesilada [36], stating that optimal results are achieved with combination of various approaches of web accessibility evaluation. Owing that, the strengths of the specific methods can be appreciated.

A semi-automatic tool is proposed by Rowan *et al.* [37] as a remedy for the gap between user testing and automated evaluation. The authors emphasize a need for a meta-method that takes advantages of current methods, but which also bridges their shortcomings. According to the authors, such a meta-method would provide a standard for detecting all accessibility barriers present on the website. The approach evaluates all pages automatically and a sample of representative and/or frequently visited pages for all accessibility problems. In case of manual evaluation, testing is based on the W3C's WCAG Checklist [38]. In the total accessibility audit, usability evaluations are included.

Another tool for semi-automatic web accessibility evaluation has been suggested by

Fuertes *et al.* [39, 40]. Hera-FFX, as the add-on is called, carries out an automatic preliminary evaluation and then enables the user to view the results and continue with manual testing of the WCAG checkpoints. Even though the the tool succeedes in linking manual testing with automatic, it is stated to be focused on the manual evaluation.

An attempt to integrating manual and automated accessibility metrics has been undertaken by Naftali and Clúa [41]. The study addresses the integration of three web accessibility metrics into a semi-automatic testing process. The research outlines the beneficial side of incorporating quantitative measures into the assessment process as well as emphasizes its limitations.

1.2.3. Web Accessibility Metrics

Over the last few years, an increased interest within web accessibility research has resulted in many attempts of quantifying the level of accessibility of web resources. One reason for a numbered score is the fact that monitoring of the web accessibility demands quantitative metrics. In terms of the QA process, web accessibility metrics have become an indispensable aid. Even though it is not advisable to compare the scores produced by different tools, due to the fact that different tools use distinct methods of accessibility measurements, the idea of accessibility level quantification for benchmarking purposes has become desirable enough to investigate more on this topic. Nietzio *et al.* [42] addresses the topic with creation of an accessibility score function for benchmarking that complies with WCAG 2.0. The following properties are suggested as the most relevant for the score function: low sensitivity towards menial changes in the web page, adequacy of scale and range of the score values. The proposed score function is based on aggregation of tests on SC level. The advantage of this solution is explained with the possibility of conformance level prioritization and for the sake of further processing of the results.

Vigo *et al.* [43] has discussed the validity of the proposed accessibility metric in terms of its applicability in the fields of QA, benchmarking or IR. The metric is automatically generated from reports provided by automatic evaluation tools. The investigation on metric's *reliability* has led to the conclusion that the metric is tool dependent, yet can be used for ranking scenarios and accessibility monitoring.

An extensive overview of the web metrics developed by 2009 has been presented by Vigo *et al.* [44]. The authors have analyzed a set of metrics created with a distinction for the kind of the disability impairment, e.g. for visually impaired users, the blind. The paper makes a reference to the UWEM 1.2, which proposes the calculation of the mean value of every single page from the sample set as a metric for a website [15]. Later on, the same author together with Brajnik [45] revisits the state of metrics for automatic accessibility evaluation.

A more recent update on the web accessibility metrics has been delivered by Vigo and Brajnik [45]. The study provides a complete overview of the automatic web accessibility metrics. The authors address the quality issue of automatic accessibility metrics and present an ancillary framework as an aid for analysis of the metrics. Then, the framework is applied to seven

selected metrics, which allows to showcase their strengths and weeknesses for defined scenarios within QA, benchmarking, search engines, and user adapted interaction. The findings show that Web Accessibility Quantitative Metric (WAQM), Page Measure, and Web Accessibility Barrier (WAB) have scored highest in terms of quality among the evaluated metrics.

Song *et al.* [46] has lately contributed to the current state of knowledge on web accessibility metrics. The newly developed metric combines the idea of automatic evaluation with user experience. The Web Accessibility Evaluation Metric (WAEM) metric assumes pairwise comparisons between different websites performed by the users. The aim of the comparisons is to develop checkpoint weights necessary for later score calculation. The process is boosted by application of Support Vector Machine (SVM) to derive the optimal checkpoint weights for the evaluation. The effectiveness of the method is stated to be validated through experiments on real-world websites.

The Web Accessibility Metrics Symposium, organized by WAI R&D group, has fructified in a presentation of a number of recent studies done on web accessibility metrics [9]. Among introduced, a nascent idea of a template-aware web accessibility metric has been raised by Fernandes *et al.* [47]. The authors stipulate that in the light of the estimates that $40 - 50\%$ of the Web content built on templates, an accessibility metric should take it into account. It is indicated that an issue of a relatively small amount of accessibility barriers can contribute largely to the lower accessibility score. It inherits from the problem of numerous reports of the same accessibility errors in the tool, connected to a particular template. As a remedy, it is proposed to alter the accessibility metric by calculating a separate parameter for accessibility of the templates and adding it to the results of the assessment of the rest of web pages. The idea of an accessible Content Management Style (CMS), contributing significantly to the overall websites's accessibility, is pointed out by Bailey in [48]. The role of the CMS templates in the assessment of web accessibility has also been highlighted by Mucha *et al.* [51].

The most actual guideline on web accessibility metrics development and validation, that has been accessed, is provided by W3C. The Research Report on Web Accessibility Metrics [9] delivers a framework for quality assessment of the accessibility metrics. The framework seeks to contribute to improvement of web accessibility metrics quality. The report focuses on five most crucial attributes that an accessibility metric should have to be considered for application. *Validity, reliability, sensitivity, adequacy* and *complexity* attributes constitutes the framework. Even though there are a plethora of metrics out there, the validity and reliability of most of these metrics are unknown and those making use of them risk arriving at misleading outcomes.

1.2.4. Integration Approaches

As of 2014, there was reported no tool that combined automated and user testing results based on WCAG 2.0 [11].

Brewer in early sources from 2004 expresses a need for combination of automated and manual testing [49]. The author suggests running a number of accessibility evaluation tools on a website, where the tools indicate conformance problems that should be checked by an expert familiar with the WCAG 1.0. The proposed approach involves a conformance evaluation including testing by users with disabilities.

A Semi-Automatic Method for Measuring Barriers of Accessibility (SAMBA) has been proposed by Brajnik and Lomuscio [50] as a new methodology for measuring accessibility. SAMBA combines expert reviews with automatic evaluation of web pages. In addition, an associated metric has been created. SAMBA uses the *Barrier Walkthrough* (BW) method for evaluating web accessibility [30]. It combines output produced by a testing tool with human judgment on a sample of the output to yield an overall index of accessibility. The idea is to run an accessibility testing tool against a website to identify a set of potential barriers. In the next step, potential barriers are sampled (with use of a non-proportional stratified sampling method with no replacement) and a panel of judges are asked to analyze the sample, associating the severity with each potential barrier. In the third phase, accessibility indexes are computed, e.g. barrier density of a website or confidence interval severity matrix to arrive at weighted/unweighted accessibility indexes. The suggested approach puts more focus on understanding the accessibility of a website with regard to the particular disability groups than only on conformance to the guidelines. Therefore, severity of the barriers is estimated together with error rate of the tool. The method is reported to be more effective than conformance testing in detecting the more severe barriers and minimizing *false positives*. Nevertheless, it is not as efficient as conformance testing when it comes to the coverage of all possible accessibility barriers present on a website [29].

1.3. Research Challenges

Some challenges of integrating manual and automated accessibility evaluation can be found. Due to their distinct properties, integration of manual and automatic evaluation may be problematic. The following concerns are taken into account in this study:

- Timing – Ideally, tests should be performed at the same time to avoid any changes to the content.
- Content – The manual and automatic evaluation are performed on different tools with different rendering (headless browser, user agent). It becomes arduous to satisfy that the page elements will be presented the same way unless both ATT and UTT can be run on the same rendered page.
- Different coverage – It may be challenging to assure that the same subset of web pages will be served for manual and automatic evaluations. Checking all web pages of a website is not feasible in case of large-scale evaluations.

1.4. Problem Statement

An exponential growth in size of the Web and the advances in making web accessibility part of the legislation, create a need for an evaluation approach that would provide a more complete picture of website's accessibility. Integration of the traditional approaches, namely automated and manual testing, would be beneficial for assuring a decent quality of checking Internet resources, as neither of them alone can satisfy the needs for a thorough accessibility report of the website. Outcomes obtained with one evaluation method are simply not reliable enough to state about the accessibility .

It is important to provide a measurable determinant of web accessibility to produce a comparable and repeatable results. What is more, the approach has to be credible so that both users and policy makers can trust it. More benefits of methods combination have already been mentioned in Section 1.1. Integration of automated and user testing is an interesting idea because of the fact that it unites two different approaches to accessibility evaluation, which at the same time complement each other. Quality of the assessment is expected to be significantly improved while not affecting drastically the efficacy of the testing.

The challenges that have been pointed out above make the task of integration hard to accomplish. It leads to a situation when one has to either balance the expectations or choose between barriers coverage and evaluation efficiency.

A plethora of evaluation methods and metrics have already been developed. Various methods on how to quantify the results of web accessibility assessment can be found in the literature. Most of the studies focus their attention on automated testing and processing the results only from checkers. That is only one part of the effort that is needed.

There has already been produced enough scientific metrics that touch the topic and attempt to translate web accessibility evaluation results into numbers. In spite of the fact that they provide some valuable information about the accessibility, in most cases the findings are incomplete since they rely on testing results from one tool.

However, there can be recalled some studies, that have used a semi-automated approach to accessibility evaluation. Table 1.1 gives an overview of the most related approaches that can be encountered in the literature, which use a semi-automated approach to accessibility evaluation.

All of the selected approaches combine at least two Accessibility Evaluation Methods (AEM). In most cases, there is observed a trend to support automated checking with manual evaluation. In some studies, usability testing has also been noticed. It can be seen that the BW method has been quite popular among the researchers. Both *SAMBA* and *OceanAcc* incorporate it into their methods. The advantage of BW method is that it takes into consideration context of a website. Moreover, it proves to be effective in finding more severe barriers and in reducing false positives. Nevertheless is is less effective in detecting all possible barriers. The approach of the three-fold AEM proposed by Lang [5] encompasses automated testing, manual checking

and additionally usability testing. Although, the author proposes a broad and detailed way of evaluation, it is not practical enough to use it on a daily basis, let alone for benchmarking purposes, where the number of websites to check is considerable. In case of the browser extension *Hera-FFX*, the culprit of this method is that it is not able to evaluate the rendered version of a web page including locally displayed content. In the light of the current web development technologies, dynamic content testing for web accessibility becomes an indispensable task in order to produce sound results. Recent WAEM metric involves AI concepts to the evaluation process [46]. Song *et al.* postulates that the method responds to the need for user involvement in the accessibility testing by having the users perform pairwise comparisons between websites to indicate better browsing experience. It is necessary for finding the optimal weighting scheme of the WCAG checkpoints. One advantage of this approach is that it produces a quantified result of the evaluation. Withal, despite the novelty of the method, the idea hits the wall with the fact that it demands from the users to be acquainted with the WCAG principles, which is not always feasible to achieve.

A practical way of enacting more exhaustive evaluation is needed. An approach that would also be simple to apply and possible to use on a daily basis. Ergo, to address the requirements of reality, the study sets itself some standards that a desired evaluation method should fulfill:

- mature to convey a complete picture of accessibility, not only one side
- producing measurable, accurate results
- able to detect accessibility problems
- simple enough to use and put into practice
- clear to comprehend
- effective in terms of time and cost

Taking into consideration all of the revised work that has been done so far, and the pointed limitations, an alternative approach to web accessibility evaluation is proposed in this thesis. Considering the fact that it has been proven that combination of manual and automated testing yields the most complete results as can be achieved for the time being, the study embarks on an investigation on how to combine them to deliver a thorough report of the accessibility issues.

The idea is to support the automated evaluation with the results of independent manual assessment conducted next to the automated testing. The novel approach differentiates itself from the previously proposed approaches in the sequence of manual and automated evaluation. Both are meant to be carried out concurrently or nearly simultaneously. The approach is to start the checker as soon as the UTT bookmarklet is launched by the user. Thanks to that, the challenge of dynamic content evaluation can be addressed. The core integration is planned to be achieved by connecting automated and manual checking results through the WCAG 2.0 SC,

that works as a bridge between these two AEMs. Both perform the assessment by checking the compliance with the established accessibility principles. Besides, owing to the evaluation run both in the checker and in the browser via UTT bookmarklet, more SC can be applied, which can result in more accessibility violations captured. Moreover, the UTT utilized for user testing can be used by non-accessibility experts. Yet, some basic knowledge about the concepts of web accessibility is needed. What is more, to fulfill the requirements of the proposed integration method, a tailor-made web accessibility metric is propounded: a Unified Accessibility Score (UAS).

The following hypothesis can be posed:

Outcomes from manual and automated web accessibility evaluation can be combined on the grounds of implemented guideline, resulting in a single, quantitative accessibility score expressing to what extend a particular website is accessible for the users.

The aim of this research is to create a method that would help to uncover existing barriers on the websites. The metric is supposed to comply with the quality framework developed by W3C [9]. Additionally, different ways of graphical presentation of the integrated results of the accessibility testing are to be explored. Moreover, impact of the dynamic content on the integration results of the evaluation is going to be observed along the experiments.

With the proposed approach to accessibility evaluation, the process of accessibility assessment is expected to be taken a step further and contribute in the long run to a more accessible Web for all.

In order to confirm or reject the aforementioned hypothesis, following research questions have been formulated:

RQ 1 : How to combine automated and manual evaluation of web accessibility?

RQ 2 : How to express the integration results in a quantitative way and present them visually?

RQ 3 : How does the dynamic content influence the results of the integration?

1.5. Assumptions and Limitations

Due to the natural constraints of the study and the novelty of the approach a few assumptions are needed to be made. Sampling has been left beyond the scope of the study. The fact that most of the websites are built on CMS templates and the effect it has on accessibility assessment influences the approach to evaluation [47]. Embracing the recently proposed sampling approach by Mucha *et al.* [51], it is assumed that the pages selected for the evaluation comprise the set of website templates. Moreover, as a consequence of the previous assumption to sampling, it is presumed that all of the pages selected for evaluation are checked, both manually and by the automatic tool. Thus, the challenge of different coverage may seem to be resolved with this simplification. What is more, an assumption is made that the same accessibility guideline is utilized for manual and automated testing.

The study has also limitations that can be identified. One limitation of the applied method is that quantitative results obtained with use of the distinctive metrics cannot be directly compared against each other. It is caused by the fact that different calculation formulas for accessibility scores are implemented.

1.6. Contribution

The main contribution of this research is to pave the way for the more thorough evaluation method, that would deliver a more complete report of the accessibility state of the website. The technical outcome of the project can then contribute to adoption of a novel quality metric and producing more reliable accessibility audits. In addition to that, the integration methodology combined with the above-mentioned clustering approach to sampling may result in a substantially more complete and efficient web accessibility assessment.

1.7. Structure of Report

The report is organized in six chapters. Chapter 2 outlines key concepts needed to follow the research methodology and conducted experiments. A brief overview of web accessibility principles is provided at the beginning and supplemented with the information about the web accessibility metrics. Later on, the next part explains the way of accessibility score calculation. Then, diverse accessibility evaluation methods are explained. Finally, the idea of integration methodology is described.

Chapter 3 consists of the description of the proposed solution to combination of manual and automated accessibility testing. The chosen research strategy and design are presented. Secondly, multiple research methods, both qualitative and quantitative, that are used in the study, are presented. Next section elaborates on the details of the integration task: how to calculate the accessibility score, which approach should be used and how to assure the quality of the proposed metric. The following section provides information about the design of the experiment conducted in order to find the answers for the research questions. Dataset is also described. Then, further information about the quality assurance is outlined. A summary of methods utilized in the study closes the chapter.

In chapter 4 an extensive study of the results of the experiment has been delivered. Outcomes of the qualitative and quantitative methods are presented, together with information about the triangulation. The last section of the Results Chapter demonstrates the results connected to the developed score function. An attempt to metric's quality validation is made and a suggestion for visual presentation of the evaluation results depicted.

Chapter 5 discusses the outcomes of the research. Major findings are summarized and the research challenges addressed. Additionally, advantages and shortcomings of the propounded

solution are highlighted. At the end of the chapter, some limitations that have been identified are acknowledged.

Finally, the last chapter 6 summarizes the efforts of the study and makes a suggestion for future paths to follow.

Table 1.1

Pros and cons of related evaluation approaches.

Approach	Characteristics	Advantages	Disadvantages
BW [31]	• adapts a user-centered accessibility definition • based on context of use of the website • a number of predefined barriers are identified • uses AT, UT on a sample with severity assessment	• context of website usage is considered • considers severity of barriers • minimizes false positives	• not-experts may struggle with evaluation • 10% less reliable than CR [31]
SAMBA [50]	• AT results combined with experts input • builds on BW method	• effective in detecting severe barriers • two AEMs used	• a panel of judges needed • not checked if tool independent
OceanAcc [41]	• focused more on overall accessibility • uses BW approach • computes 3 metrics(WAB, FR, UWEM)	• UT of a sample • combines several AEMs	• not efficient time-wisely • users filter all AT results • extra parameters required • no guidance for testers
Lang [5]	• a fully integrated AEM • combines AT, UT and usability testing • website browsing as initial step	• multiple AEM applied • detects most of the barriers	• resource-demanding • costly • not suitable for large-scale evaluations
Hera-FFX [39]	• a mix of AT and UT • focused on manual evaluation • an add-on	• lightweight	• no evaluation of rendered content • no accessibility score produced
WAEM [46]	• user-experience influences weighting of checkpoints • involves pairwise comparisons between websites • a machine-learning model employed	• outputs a qualitative metric • modern • effectiveness verified through an experiment	• complex weighting scheme • require knowledge of WCAG

2. Theory

This chapter explains the key concepts needed to understand the investigation. First, the meaning of web accessibility is defined accordingly to the adopted definition. Next section provides information about various ways of accessibility evaluation with a special attention to the semi-automated testing. Then, accessibility metrics and criteria that determine a valid metric are explained. Finally, a connection between web accessibility evaluation and dynamic content is presented. Further details can be found in the referred sources.

2.1. Web Accessibility

Nowadays the Web is present in all of the fields of life and many people cannot imagine their lives without the possibility to use the Internet on a daily basis. However, a great deal of the users may encounter problems when the websites are not satisfying the basic level of accessibility. *Web accessibility* is a property that states to what extent the web resources are accessible for different groups of users. Over fifty different definitions of web accessibility are available in the literature [52]. Some of them explain the web more precisely whereas others define web accessibility as a vague term, e.g. the one provided by Thatcher *et al.* [53]: "(...)it is effective, efficient and satisfactory for more people in more situations.". Yesilada *et al.* explored perceptions of web accessibility using a survey approach [54]. It has showed that the respondents strongly agree that accessibility must be grounded on user-centred practices and that its evaluation should be more than just inspecting the source code.

The main distinction is made within the definitions for two groups:

– Web accessibility related to the conformance to the accessibility guidelines

– Web accessibility, which assumes that websites can be used by all people on equal basis

According to the W3C, web accessibility is defined as a situation when people with disabilities can perceive, use and interact with the web. Also, when the users can contribute to it and fully benefit from it [55]. Web accessibility encompasses all disabilities that affect access to the Web, including:

– auditory issues, e.g. deafness, as well as partial deafness;

- visual impairments;
- cognitive;
- neurological;
- speech;
- physical, mobility problems, (i.e., difficulties caused by a limited mobility that can be a result of certain diseases, accidents or the aging process);

Moreover, web accessibility addresses people without disabilities that can also benefit from accessible web resources. That includes the older people, people with short-term disabilities such as a broken arm or experiencing a "situational limitation", for instance due to a bright sunlight. Apart from that, a slow Internet connection is also regarded as an accessibility barrier. Also, web accessibility takes into consideration people using devices with small screens etc. such as mobile phones, smart watches etc.

Various examples of the problematic issues on the web pages can be called together with suggestions how to counteract them. A blind person may struggle when no textual equivalents are provided for the images, so that assisting tools like text-to-speech devices can process them. Often encountered flashing effects can lead to photo epileptic seizures. Cognitive needs can be fulfilled with developing a more simply-structured content. For the people with mobility issues, problems can be alleviated with providing a way of navigating through a page using keyboard, ensuring an adequate size of the buttons or even enabling usage of a simple voice switcher, in a situation when a person is affected by a muscle weakness.

Depending on the type of disability, different issues may prevent users from taking full advantage from website. What may cause an accessibility issue for one person, may not disturb the other one. For instance, missing captions would probably not affect as much a person with a broken arm as a deaf person, for which provided captions are the only way to understand the content of the video. G. Brajnik proposes a user-centered accessibility view and suggests a three-step accessibility model, that is meant to help to plan and perform accessibility assessment [56]. The proposed Properties-Context-Methods (PCM) accessibility model takes into consideration the type of user disability among others. With this in mind, accessibility considers as well the effect and severity that a badly-structured or programmed website may have on the various groups of users.

What is more, built upon which definition of web accessibility is chosen for the study, the way of testing it may slightly vary. The definition provided by W3C has been chosen for the purpose of this study.

2.2. Web Accessibility Evaluation

According to the W3C, accessibility evaluation is also called “assessment”, “audit”, and “testing”. W3C serves as a Consortium that develops international standards for the Web, including HTML, CSS. W3C is also responsible for creation of the WCAG 2.0 accessibility guidelines. Web accessibility evaluation can be described as a process of checking weather a particular website is accessible for people and determining to what extent it is accessible.

To assure and certify the fullfilment of the accessibility guidelines, miscellaneous accessibility evaluation methods have been designed. Following Lujan-Mora, S. and Masri, F., web accessibility evaluation methods can be classified into two types: qualitative methods (analytical and empirical) and quantitative menthods (metric-based methods) [57]. Evaluation methodology need to include different techniques and maintain flexibility and adaptability toward diverse situations [4]. Embracing the size of the web resources, it should be also robust and effective. The abundance of evaluation techniques makes it difficult to choose the most suitable one. Following Accessibility Evaluation Methods (AEMs) can be distinguished: conformance review (CR), subjective assessment, screening techniques, barrier walkthrough, and user testing [56].

Evaluation may be performed automatically, manually or semi-automatically. The difference between these methods is connected to the degree to which humans are involved in the process of testing. In case of the automated evaluation, all accessibility tests are conducted automatically by a testing software. No involvement from the evaluator is needed. For manual evaluation, called also user-testing, it is a human that tests the website manually with use of given tasks/questions about its accessibility. The tests may be performed by different users, e.g. accessibility experts, website users, professional testers. When it comes to the semi-automated evaluation, the website is analyzed both by an accessibility checker and a human. As a result of the evaluation a summary of the findings may be provided in form of a description or a quantitative metric – an accessibility score. All AEM follow a common procedure for evaluation of a website. Figure 2.1 presents the process of evaluation of a website. Evaluation of a website starts with sampling of a needed number of web pages for checking. Sampling of the pages is needed due to the fact that evaluation of all web pages may be infeasible because of the large number of pages. In the next step, an appropriate investigation method is applied to the sampled web pages, e.g. user-testing, automated check, or both. Optionally, a quantitative metric is applied and an accessibility score calculated for the website. Finally, an evaluation report about the discovered barriers is created.

2.2.1. Conformance Review

Conformance evaluation determines how well web pages or applications meet adopted accessibility standards. W3C’s WCAG-EM is an approach for determining conformance to WCAG. The W3C/WAI model of accessibility aims at universal accessibility and assumes that the website conformance to WCAG is necessary.

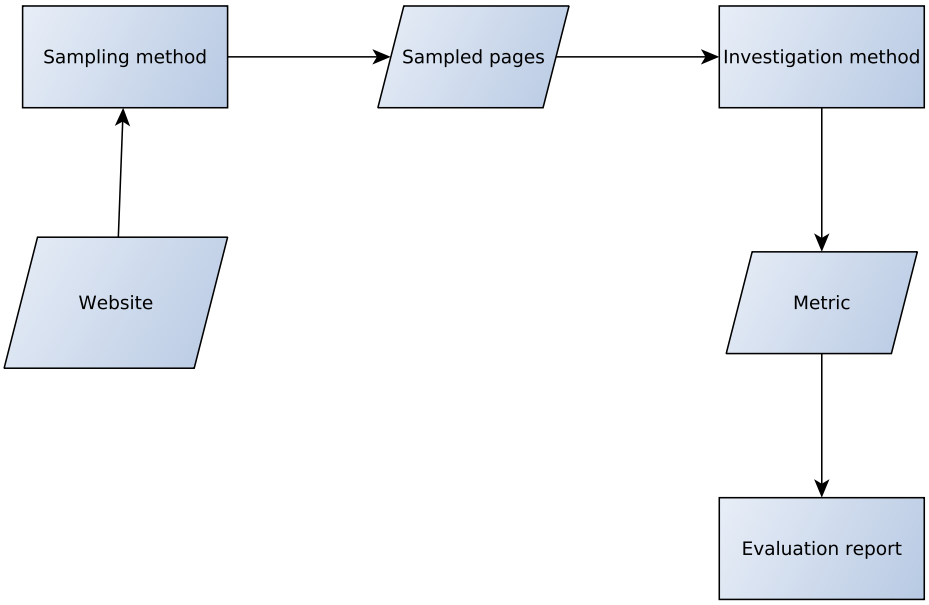


Figure 2.1
Process of web accessibility evaluation of a website.

The W3C develops Web accessibility guidelines addressing the three main components:

- Authoring Tool Accessibility Guidelines (ATAG), regarding authoring tools;
- Web Content Accessibility Guidelines (WCAG), that addresses Web content;
- User Agent Accessibility Guidelines (UAAG), which encompasses Web browsers and media players;

Focusing on Web content evaluation, the current version of WCAG 2.0 is the main guideline widely accepted that treats on creating accessible content and its testing [58]. WCAG 2.0 defines how to make Web content accessible to people with disabilities. Several layers of guidance have been defined, including overall principles, general guidelines, testable success criteria (SC) together with a collection of sufficient and advisory techniques, supplemented with documented common failures and examples. Current version of WCAG 2.0 consists of four principles of accessibility : *perceivable, operable, understandable, and robust.* Often the four accessibility principles are referred as POUR. The principles lay the foundation necessary for anyone to access and use the Web resources:

Perceivable indicates that all the information and user interface componets must be presented to the users in a way that they can perceive.

Operable means that the users must be able to fully operate the interface and navigate the website.

Understandable defines that information and user interface must be presented in such a way that allows the users to understand it.

Robust defines that the users must be able to access the content no matter how the technology evolves.

To help to address the aforementioned principles for people with disabilities, twelve guidelines refer to particular problems of people with disabilities. Among many existing guidelines, the ones that affect people more severly have been included. Examples of guidelines implemented under the first principle – *Perceivable* are: Text Alternatives, Time-Based Media, Adaptable, Distinguishable. More information can be found in [59].

Under each guideline, several SC have been defined that outline what must be provided in order to satisfy the particular standard. SC included in the WCAG 2.0 are designed as testable criteria, meant to be technology objective. It is worth to emphasize that even though some of the checking can be performed automatically with help of testing tools, others require human input to complete the testing.

Figure 2.2 presents a part of the structure for the first principle from WCAG 2.0 – *Perceivable*.

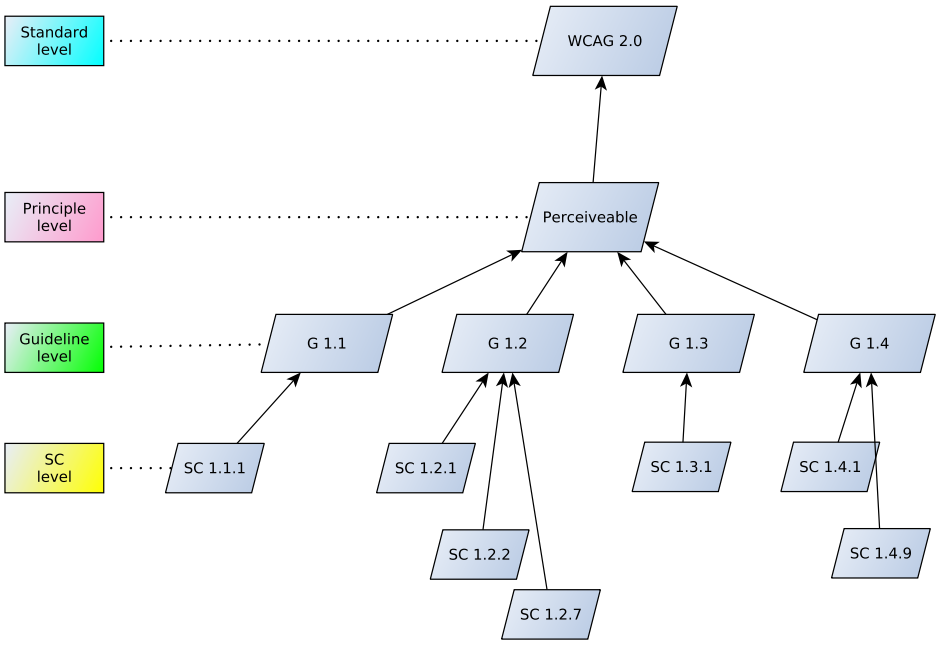


Figure 2.2
Structure of WCAG 2.0 on the example of the principle *Perceivable*.

WCAG 2.0 SC are categorized into three levels of conformance to meet the needs of different groups and various situations: A (the minimum level of conformance), AA (mid range), and AAA (the highest level). Level AA assumes that the web page satisfies all the Level A and Level AA SC. For the Level AAA conformance, the web page must satisfy all the Level A, Level AA,

and Level AAA SC. Otherewise, an alternate version is provided. The requirement states that no conformance is possible without at least satisfying all of the Level A SC.

2.2.2. Combined Evaluation Method

Since no method alone is sufficient to provide full-accessibility evaluation, many studies combine both quantitative and qualitative methods to guarantee more detailed and accurate results. By using the term *combined evaluation method* it is understood that more than one evaluation method has been used for evaluation of web accessiblity. It can be both combination of quantitative and qualitative methods, as well as a combination of several methods either qualitatives or quantitatives. Various advantages and shortcomings of each group have been presented in [57]. Quantitative methods outperform testing because of their usability and quick results while the manual evaluation provides a hand-on experience from the users. However, it has been apparent that no one tool alone, can determine if a website meets the requirements given in Web accessibility guidelines [57]. According to the same authors, a combined accessibility evaluation method must consist of a user test method and a qualitative metric. User test method has been selected because of its reliability and effectiveness, while the quantitative metric is supposed to control and monitor accessibility results.

2.2.3. Quantitative Metrics

In the web engineering field, a metric can be defined as a procedure for measuring a property of Internet resources such as a web page or a website [9].

When it comes to the web accessibility domain, a metric can express in a quantitative way among others the subsequent qualities:

- The number of images missing an alt attribute.
- The number of violations of SC on different levels, i.e., Level A, AA, AAA.
- The estimate of possible failure points where accessibility barriers may occur.
- The severity of an accessibility issue.
- The amount of time needed to perform a task.

As an outcome of the calculation of the accessibility metrics, various types of data can be produced as outcomes. Two main types can be named:

1. Ordinal values – expressing WCAG 2.0 conformance levels (A, AA, AAA), or "accessible/not accessible" scores.
2. Quantitative ratio values, e.g. 0, 45, 0.85.

Due to the fact that web accessibility can be defined in different ways, the metrics provide a feedback based on the assumed definition of web accessibility. Metrics, that have been based on SC and accompanied techniques are called *conformance-based metrics*. The conformace-based metrics evaluate the resources on the grounds of whether SC of given accessibility guidelines are met. In a situation where the web accessibility is viewed as a property that differs from conformance, other metrics can be defined. For instance, in case of the Section 508, where accessibility is defined as "accessible technology (...) can be used as effectively by people with disabilities as by those without it" [25], accessibility metrics are described as *accessibility-in-use metrics*.

Checklists and guidelines have been created to allow for assessment of the quality and reliability of various Internet resources. The purpose of using a quantitative metric is to synthesize a value that is assumed to convey the level of accessibility. Accessibility metrics help to understand the accessibility level of websites. Web accessibility metrics can be applied in several areas of web engineering [9, 45]:

- Quality assurance within web engineering – a fine-grade metric helps to keep track of accessibility during the iterative development cycle and contribute to a better implementation of the accessibility principles.

- Benchmarking – an accurate measurement is needed that will allow for monitoring of the accessibility websites or their conglomerates. Also, it becomes an indisposable aid in the domain of eGovernment.

- Information retrieval systems and search enginees – apart from providing the users with relevant content, users can be able to retrieve resources that are also accessible. Incorporation of the accessibility metrics into the searching algorithms may bring a possibility to sort the websites accordingly to their level of accessibility.

- Adaptive hypermedia techniques – metrics may help to deliver guidance or criteria to carry out interface adaptations such as adaptive navigation support.

2.2.3.1. A Framework for Quality of Accessibility Metrics

A Framework for Quality of Accessibility Metrics has been proposed by W3C [9] on the basis of the work contributed by Vigo, M. and Brajnik, G. [45]. For web accessibility metrics, following quality factors can be established:

Validity

A property defining the extent to which the results obtained by the metric express the accessibility of the website that has been evaluated. Two types of validity can be distinguished, namely validity with respect to accessibility in use and validity with respect to conformance to certain

guidelines. While the validity referred to the accessibility-in-use indicates how the interaction is perceived, the validity with respect to the conformance refers to the specific guidelines and principles. An estimate of error rate of tools is used for determining the metric's validity. What is more, validity is perceived as the most important quality for an accessibility metric according to the Report on Web Accessibility Metrics [9].

Reliability

Reliability of a metric points to the extent to which independent evaluations yield the same results. It is widely known that accessibility checking tools produce different results when applied to the same website. It is due to different coverage of the implemented guidelines as well as its interpretation [45]. The *reliability* is related to the ability of the metric to produce reproducible and consistent scores. It is measured as the extent to which the metric delivers the same results in changing context e.g. different tools, testers, diverse goals and different time of the evaluation.

Sensitivity

Sensitivity of a metric is a property that estimates how changes in metric results are reflected in the real changes to the evaluated resources. It is desired for the metric to have a low sensitivity in order to be robust and applicable to websites with frequently changing content.

Adequacy

Provided that the metric's validity and reliability are covered and sound, various aspects of metric's use are considered. First and foremost, the overall goal of the metric is to provide the users with a meaningful information about the website's accessibility. Most importantly the metrics's suitability and usefullness for particular users in different contexts. Again, contexts can be defined twofold: as a purpose of use, and with a distinction to miscellaneous disabilities to show to what extent the website is accessibility for a particular group e.g. for the blind or people with motor disabilities. Adequacy assumes as well analysis of the suitability and usefullness of the metric's values for users in different scenarios, as well as metric's visualization and presentation issues. The type of data used to present the scores, the precision and normalization of the metric's values should be taken into consideration.

Complexity

Complexity of a metric expresses how computationally demanding is the metric when it comes to particular aspects of the resources such as time, memory, computational power. Above that, the complexity takes into consideration as well human resources that are needed for the evaluations, especially for user-testing. When performing automated checking, crawling of large websites or insufficient storage capacity may become process bottlenecks.

2.3. Web Accessibility and Dynamic Content

Web pages can be either *static* or *dynamic*. Static content means that the web page remains constant. It does not change each time the page is loaded. Whereas a dynamic web page contains elements that are changing lively. The difference is that a dynamic web page can be generated on-the-fly while a static page stays unchanged. Standard HTML pages are considered static. It is also the case for Cascading Style Sheets (CSS), and Images. Also, JavaScript (JS) files residing on the server are considered to be static. They can be served equally well by an application server such as Tomcat, or a web server such as Apache etc. For instance, a navigation menu, or a logo of the website, would not require any input from users. On the other hand, PHP, ASP.NET programming, and JSP web pages belong to the group of dynamic web pages. The content presented to the user is created each time uniqely by the server when the page is accessed. This type of content is usually dependent on inputs from visitors and their individual accounts [60].

Similar behaviour can be observed when it comes to single-page applications (SPA). They interact with the user by dynamically rewriting the currently viewed page without loading complete pages from an external server. An SPA can appear to work as a desktop application, however interaction with it often combines in dynamic communication with the web server.

Static pages are simpler and more secure than dynamic pages in a sense that it does not require any code to execute, nor any external database to be accessed. Thus, it is more secure. Another advantage of static pages is that they are compatible with every type of webserver technology. Static pages have also their disadvantage – they cannot adjust the content to the users. In case of dynamic pages, they are capable of adjusting the content to various users from the same code. That is to say, it can be accustomed to the viewer. The code is generated at the time they request the page. It exists only at that moment. Another attempt on hitting the same web page may result in a slightly different content.

For web accessibility testing, this distinction becomes a crucial factor when approaching the checking task. As it is pointed out in [61], some screen readers may struggle with dynamic content by not being able to detect changes that were done through modification of the content. Oftentimes such Javascript libraries as jQuery utilize both DOM and *innerHTML* methods to manipulate the content of the web page. The interoperability problem with assistive technologies may be triggered.

Dynamic content can be seen as a challenge for user testing. It becomes difficult to assess something that is constantly changing. Different appearance and often slightly different functionality generate possibilities for ambiguities. For the blind and low vision users, unless there is an audible cue to changes, the updated information is virtually invisible. Furthermore, an unexpected change of focus may be disorienting when no previous warning is provided. When performing a web accessibility audit one cannot guarantee that the dynamic page can be awarded status accessible since the same code may be rendered differently on other devices or browsers

and cause accessibility barriers. In addition to that, evaluation of dynamic content is tightly connected with the *sensitivity* attribute of an accessibility metric. Low sensitivity of a metric is particularly important when evaluating highly dynamic websites.

Most implementations of the ATT focus on page content that is transferred through the first HTTP request, which may significantly vary after application of dynamic content techniques. Fernandes *et al.* [62] has conducted an experimental study which revealed that there are deep differences in the accessibility evaluation carried out in command line and web browser (via bookmarklet) environments. The numbers of detected HTML elements by accessibility evaluation procedures varied between checking performed in the command line and the browser. Therefore, regarding web pages with dynamic content, developers and designers may be faced with different HTML DOM structures. Additionally, nearly 67% of the analysed cases for command line environment has yielded *false negatives*, i.e., those SC that were unable to be applied in the command line, compared to 13% of *false positives*, i.e., SC that could be applied in the command line but not in the browser. It shows that automated web accessibility analysis in the command line can yield incorrect results, principally on the applicability of SC. That can be related to the application of Javascript/CSS to the page before the check is carried out.

3. Methodology

> Research that produces nothing but books will not suffice.
>
> Kurt Lewin

3.1. Research Strategy

A complete research strategy needs a few components to be complete [63]:

– Research paradigm – how to approach the research

– Research design – a structured plan of action

– Research problem – a specific goal

3.1.1. Mixed Methods Research approach

Pragmatism became a philosophical foundation for the research. It underpins the mixed methods approach and separates the study from the concepts leaning solely towards positivism or interpretivism [63]. A solution to the problems can be pluralistic, yet there is no single method that can pave the way to arriving at the valid results. In pragmatism, knowledge is based on practical results and therefore assessed in terms of its usefulness and applicability in solving the problem [64]. Another reason for choosing mixed methods approach is that empirical enquiry of an exploratory nature is supposed to test what works best for the approached project. Additionally, the study done at this point, with a constant technological advance, may become obsolete one day. Its provisional side is also acknowledged. More information on mixed-method approach are discussed in [65].

The research is based on a hypothesis from the beginning. Yet, being a novel study, it is still focused on discovering things through means of the exploratory investigations. Rather than theory-driven, the mixed methods approach selected for this study is practical, problem-driven.

The benefit of the mixed methods approach is that the data obtained as a result of application of diverse methods can be complementary. Combined, they provide a more hollistic view of the

case. Moreover, by taking advantage of contrasting methods, things can be seen from alternative perspectives, which then can contribute to getting a more complete picture of the subject [65–67].

On the other hand, some of the shortcomings of embracing this research strategy may include additional resources needed for the project completion or development of extra skills from qualitative and/or quantitative approaches. Also, there exists a risk that the findings from different methods may not corroborate one another and thus a valid explanation for that might be necessary.

3.2. Research Design

For the research design, seen as a constructed plan of action, Johnson [67] lays out a mixed methods research process comprised of eight steps, which follows:

1. Determining the research questions
2. Verifying the appropriateness of using a mixed design
3. Selecting the mixed method
4. Data collection
5. Analyzing the data
6. Data interpretation
7. Validating the findings
8. Drawing conclusions together with research documenting

Selected research design involved a concurrent implementation of the qualitative and quantitative methods. When it comes to the paradigm emphasis, a higher priority and weight was given to the outcomes of the quantitative study. The decision to embed qualitative data within a quantitative design was made to better understand the challenges of the research problem [64]. Qualitative results provided insights on mechanisms that related variables. In this case, the different data sets (qualitative and quantitative) are not intended to converge. The Embedded Design, applied in the study has been explained with aid of the Figure 3.1.

The investigation started with collecting and analysing a relatively small amount of qualitative data as an exploratory phase of the research. That helped to explore how the problem was perceived in the research environment and provided some external thoughts on the possible solution. In the meantime, an introductory investigation with use of the quantitative methods would be embarked on.

To link the quantitative data obtained from the WTKollen Checker and the User-testing tool, a concept of data triangulation was utilized.

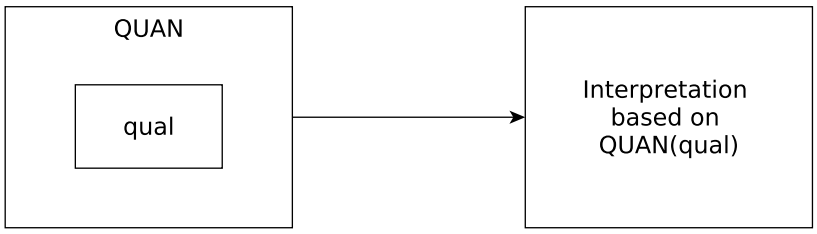


Figure 3.1
Embedded Design Procedures. Qualitative methods were embedded in quantitative study.

In case of this research, qualitative data is treated as subsidiary to the quantitative one. The qualitative part was supposed to complement the current state-of-the-art and provide with hints on possible directions for the study. The rationale for using the contrasting methods is the novelty of the study one one hand, and the practical case of the WTKollen checking tools, linked to this research. Both methods are interlinked in such a way that the output from the qualitative part became an input for the further process design and quantitative approach.

3.3. Research Methods

3.3.1. Triangulation

Triangulation can be defined as a combination of methodologies in the research of the same phenomenon [68]. Coined by Denzin [69], the term describes a process of studying the problem utilizing miscellaneous methods to get a broader perspective [70].

According to Flick [71], triangulation can have four distinctive forms: triangulation of data, investigator triangulation, triangulation of theories and methodological triangulation. For this study, triangulation of data and methodological traingulation were used.

Triangulation of data– combining data from different sources and at different time;

Methodological triangulation– applying divergent methods to data for the sake of increased validity and broadened overview.

Applying the triangulation can result in an increased knowledge about the studied subject in form of the improved accuracy, as a mean of validation. Another outcome of triangulation is a better picture, that enhances the completeness of the findings. In this case, a triangulation of data generated by the ATT (the checker) and the UTT (bookmarklet) was employed to complement the results from one evaluation methods with results from the other one. The purpose of applying triangulation was to obtain different but complementary data on the same topic to better understand the level of accessibility. User testing captured different aspects of accessibility that the automated check could do. Data provided by the ATT was in a quantitative form,

while the results from the UTT were in qualitative form. Figure 3.2 presents the idea of data triangulation applied in this research. The results from manual and automated testing were collected during the same time frame and with equal weight. The data transformational model [64] was chosen to combine the qualitative measures from UTT with quantitative generated by ATT. After the initial analysis, the qualitative findings were quantified to allow the data to be mixed during the analysis stage and facilitate the interrelation.

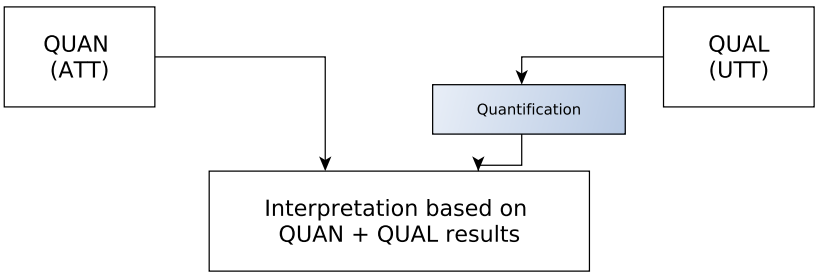


Figure 3.2
Triangulation of data generated with use of UTT and ATT. Interpretation of the results is based on quantitative outcomes from the checker (ATT) and qualitative results from the bookmarklet (UTT), which have been quantified beforehand.

3.3.2. Qualitative Methods

In the first phase of the study, qualitative reseach was carried out by conducting interviews with a group of researchers and accessibility experts acquainted with accessibility concepts. The aim of the interviews was to understand more the challenges of the accessibility results integration and to gather the ideas on how to accomplish the task. The interviewees involved in the process were the researchers working either directly in the web accessibility field or in related fields, e.g. universal design, usability, and acquainted with topic of this research. A series of the semi-structured interviews was done with use of the following questions prepared beforehand:

Question 1.

What do you think about the idea of combining different evaluation methods?

Question 2.

What kind of possible challenges can you see in the integration of manual and automated accessibility testing?

Question 3.

On which level should the integration be performed in your opinion? (websites/web pages/individual tests/page object elements or maybe the guidelines)

Question 4.
What impact the dynamic content can have on combined accessibility evaluation?

Question 5.
Which presentation form of combined evaluation would be more preferable for users? Graphs/numerical scores?

3.3.3. Quantitative Methods

The second stage encompassed testing various approaches for combination of the results using real data gathered from the WTKollen checking tools: the checker and the UTT bookmarklet. Quantitative methods utilized for the study emphasized objective measurements and a numerical analysis of data collected from the ATT and UTT.

Data triangulation were applied to the results coming from the checker and the bookmarklet. Combining the qualitative data, in form of manual accessibility evaluation results, with the quatitative data, given as an output of automated testing tools, called for a structured method of integration.

3.4. Integration Approach

Different approaches to combination of manual and automated evaluation results were tested in the experiment. To be able to integrate two sets of data, there has to exist a common ground between them. It can be imagined as a bridge linking one bank of the river with another. In case of the accessibility evaluation, an accessibility guideline can be seen as a connecting bridge since both automated and manual test sets have been created on the ground of particular guidelines. Whether it is the WCAG, Section 508, or any other international/national standard, the principle remains the same. A crucial assumption to this approach is that both of the tools support the same guideline, as it is the case for the WTKollen tools.

Various levels of integration have been considered. Aggregation can be seen from two perspectives: considering the components that a website is built of i.e., its structure, or choose to look at the evaluation scheme comprised of the tests associated with SC. Figures 3.3 and 3.4 illustrate the discussed concepts of the aggregation levels.

Concievable levels of integration with regard to *Document Object Model* (DOM) structure :

- Website
- Web page
- Element on page

Possible levels of integration with regard to evaluation scheme from the guideline:

– SC from WCAG 2.0 for the above

– Individual test result for the above aggregations

Grounding the integration of the evaluation results on the object level would be beneficial to creating a more exact report about page elements that cause accessibility barriers. The current implementation of the ATT does have the functionality of showing code snippets. However, the UTT bookmarklet lacks this component for the time being.

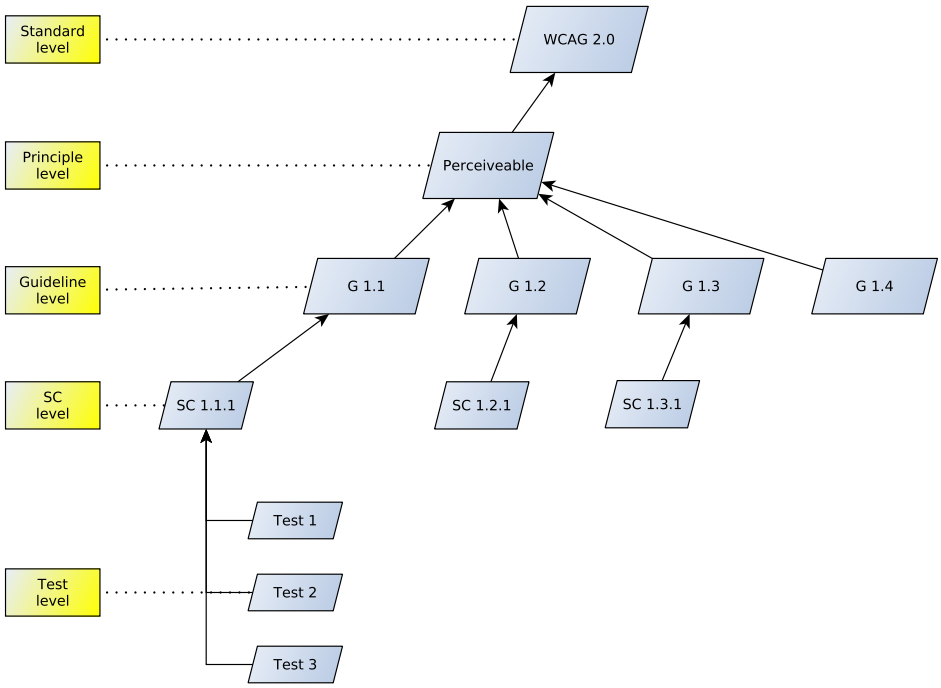


Figure 3.3
Overview of the possible levels of aggregation considering accessibility guidelines. Starting from the most general one – the standard level, it can be boiled down to the test level, through principle, guideline and SC levels.

For this research project, WCAG 2.0 was decided to become a bridge connecting manual and automated evaluation, owing to the common practice of its use both in ATT and UTT. Subsequently, further integration of the results was conducted on a more detailed level of the guideline, that is SC. Both ATT and UTT tests have been built on the *Perceivable, Operable, Understandable and Robust* (POUR) principles from WCAG 2.0 and can be linked to the appropriate SC.

Integration of the accessibility assessment can be expressed in many ways. The study propounds taking a closer look on the numerical aspects of integration by developing a score calculation function and further on the graphical possibilities for accessibility results illustration.

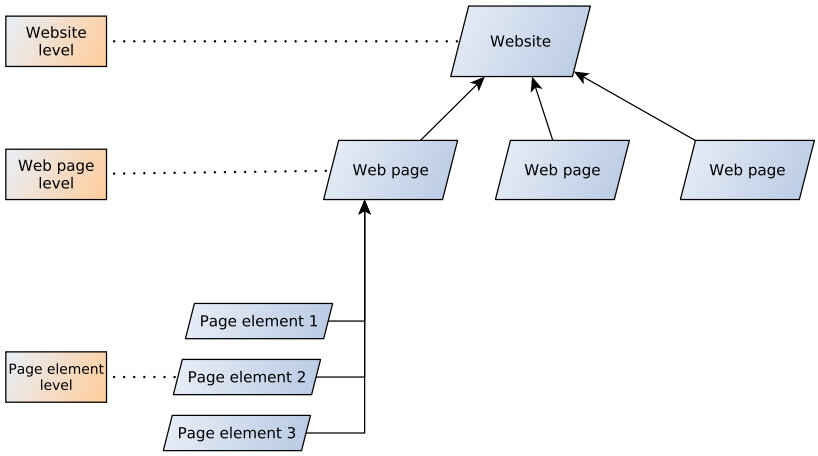


Figure 3.4
Overview of the levels of aggregation regarding the DOM structure. Results can be integrated on the page element level, as well as on web page or website level.

3.4.1. Unified Accessibility Score

During the study, two viable approaches for the integration were identified. Figures 3.5, 3.6 present a graphical illustration of the unified score calculation approaches. The first difference between those approaches lies in the stage, when the integration takes place. One approach would be to combine the separately calculated scores on the web page level i.e., after the calculations have been done for the ATT and UTT. In this case, two independent accessibility scores are calculated, and the integration is performed on the accessibility scores. Figure 3.5 shows the process of the scores integration. However, there arises a question at this point: How should the scores be treated? Should they be simply combined using an aritmetic mean or maybe a better solution would be to apply SC as weights, similarly as it is proposed in [12] for the page score calculation?

The other possibility would be to follow the concept of results integration and score calculation on the SC level. Figure 3.6 indicates the notion of the idea. For this solution, there emerged two probable paths. The first one would propose score calculation based on the results from mutual SC i.e., only those that were applied during manual and automated evaluation. Whereas the second option would be to use a union of SC. The advantage of the first solution is that the score would then mirror the common barriers detected by the two evaluation approaches. What is more, it may support the QA of the evaluation process. This can be seen as a mutual validation of the tools to some degree. The score would be more grounded, unified, inasmuch as two independent tools have yielded corroborating results. Nonetheless, the disadvantage of this path can be that the score would reflect a smaller spectrum of the accessibility barriers present on the website. That can be perceived as a downside when the assumed priority of the testing is to discover as many distinct barriers as possible, as well as the amount of those present on

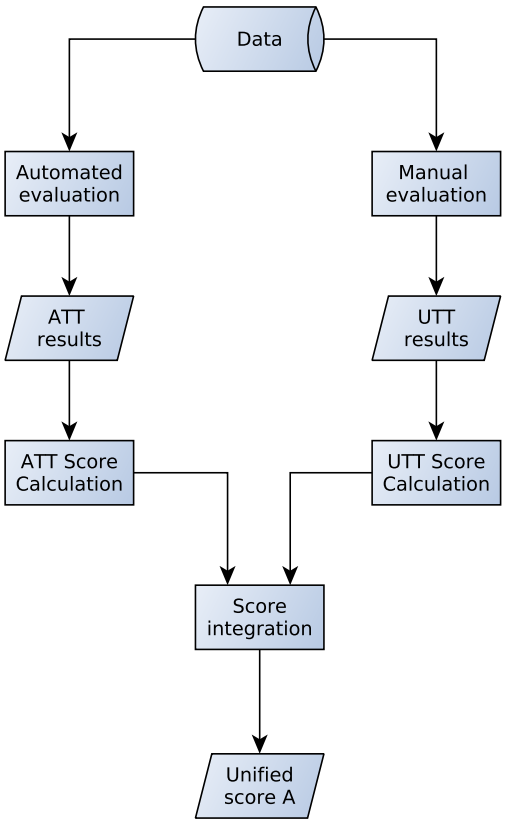


Figure 3.5
Scores integration

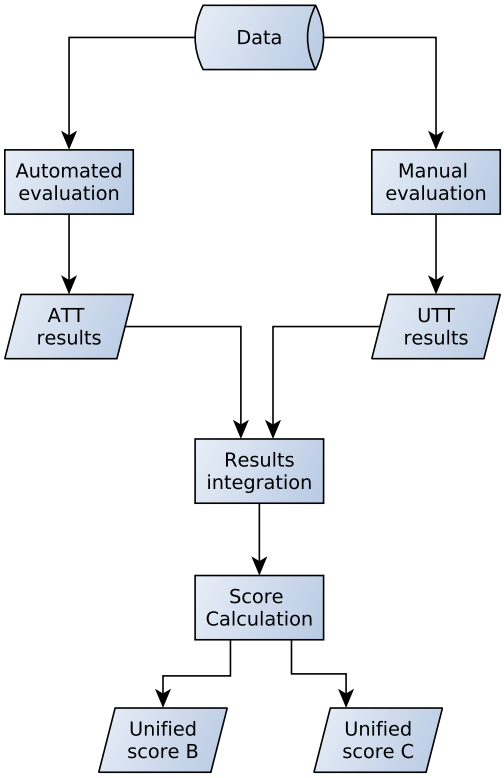


Figure 3.6
Results integration.

the web page. Computing the score using all the data provided by both tools can lead to a more complete coverage of the accessibility issues, where one tool supplements the findings of the other one.

Taking into consideration the identified pros and cons, it is proposed to empirically test the possible approaches to the accessibility testing integration.

3.4.2. Accessibility Metric for Unified Score

Current accessibility score function for the ATT has been described in details in [42] and can be further refered to the *Integration Methodology* proposed by Nietzio and Berker [12]. Accessibility score for a single page is calculated as an average of the SC-level page results. Then, the score for a website is computed as a weighted mean of accessibility scores of all web pages belonging to that website, where numbers of test instances serve as mean's weights.

With the novel approach to web accessibility evaluation, that involve combination of manual and automated results, it is proposed to adjust the score fuction. Depending on the adopted integration approach, three approaches can be distinguished for further investigation:

1. : Unified Score by ATT and UTT scores integration – scores integration approach

2. : Unified Score via results integration – intersection approach

3. : Unified Score via results integration – union approach

Let p denote a web page, c a SC, and C the set of SC for which there have been carried out tests at the checking stage. Additionally, $S(p)_{ATT}$ represents the page score from the ATT for the web page p, $S(p)_{UTT}$ the page score from the UTT, and $U(p)$ an unified score for the page p.

Calculation of the score for a single web page (Approach nr 1: Scores Integration)

In the first case: scores integration, it is suggested to be calculate the score as a weighted mean of the ATT and UTT scores. Equation 3.1 presents the proposed formula.

$$U(p) = \frac{S(p)_{ATT} * |C_{ATT}| \ + S(p)_{UTT} * |C_{UTT}|}{|C_{ATT}| + |C_{UTT}|} \tag{3.1}$$

Calculation of the score for a single web page (Approach nr 2: Intersection Approach)

For the case number 2, where the results are integrated on the intersection of SC applied for UTT and ATT, formula given by the equation 3.2 is applied. The remaining SC are not considered for the score calculation.

$$U(p) = \frac{1}{|C_m|} \sum_{c \in C_m} S_c(p), \tag{3.2}$$

where $S_c(p)$ denotes the intermediate result per c on page p, and $C_m = C_{ATT} \cap C_{UTT}$.

Calculation of the score for a single web page (Approach nr 3: Union Approach)

The last, third case, assumes results integration on the basis of the union of SC from the ATT and UTT. Assigning $S(p)_m$ to the page score calculated on the basis of mutual SC that were applied in testing phase both by UTT and ATT, following formula 3.3 yields:

$$U(p) = \frac{S(p)_{ATT} * |C_{ATT}| \ + S(p)_{UTT} * |C_{UTT}| - (S(p)_m * |C_m|)}{|C_{ATT}| + |C_{UTT}| - |C_m|} \tag{3.3}$$

Calculation of the Unified Score for a single website

Eventually, an accessibility score for a website s, where $s = \{p_1, p_2, \ldots\}$ is computed as a weighted average of page scores, given by the formula 3.4. A number of all instances within page p, where tests for c were applied is denoted by $n_c(p)$. Subsequently, $N_p = \sum_{c \in C} n_c(p)$ refers to the number of test instances for the page p, and $N_s = \sum_{p \in s} N(p)$ within the website s.

$$S(s) = \sum_{p \in s} \frac{N(p)}{N(s)} U(p) \tag{3.4}$$

3.4.3. Quality Assurance for the proposed metric

In order to confirm that the proposed novel metric can be deployed, the quality of the metric needs to be assessed. For the quality assessment, a framework proposed by WAI group [9] was selected. As mentioned in the previous chapter, the framework lays out the guidelines on how to approach the five metric's quality characteristics, i.e., *validity, reliability, sensitivity, adequacy and complexity*. It indicates the facets of the metric which should be deeply investigated.

To address the metric's *validity*, following questions were askedin the QA process:

- Does validity of the metric change when the underlying accessibility guidelines are changed?
- Does changes in number of the SC influence validity of the metric?
- Is validity dependent on the genre of the website?
- Does the type of the data being provided by the testing tool affect validity?
- Does validity depend on the tool utilized to collect data? What in case of providing merged data, coming from different testing tool?
- Are there any quick ways in which validity of the metric can be estimated?

When it comes to the second characteristic – *reliability*, relevant questions were:

- How results of the accessibility assessment of a particular website vary when produced by different tools?
- How applying another set of guidelines change the metric scores when used for evaluation of the same website and with the same tool?
- What impact has page sampling on the metric scores?
- How does *reliability* change when fed with data delivered by two or more different tools?
- What kind of correlation between *reliability* and *validity* can be found, if at all?

Other qualities, such as *sensitivity, adequacy and complexity*, were also taken into consideration when evaluating the metric's quality. Subsequent questions were pondered:

- For *sensitivity* – Which accessibility barriers could have a more or less strong impact on conformance?
- For *adequacy* – Are the values produced by the metric suitable and useful for the users, considering different scenarios? How should the results be presented or visualized?
- Finally for *complexity*, a few research questions can be named:

- Does complexity on a metric guarantee more valid and reliable accessibility evaluation results?
- Are there any trade-offs that can be made in order to lower the complexity of the metric?
- Is the metric difficult to adopt and deploy?

3.5. Experiment Design

The purpose of the experiment was to answer the established research questions RQ1-RQ3.

3.5.1. Experiment Environment

The accessibility results were generated with use of the tools developed in the WTKollen project. For automatic evaluation, the ATT checker [72] was deployed. The checker assumes testing and score calculation based on the SC from the WCAG 2.0, level AA. 44 HTML accessibility tests, mapped to the 27 SC were available for testing against accessibility violations.

For manual testing, the UTT bookmarklet tool [73] was put to work. The bookmarklet encompasses a set of 13 applicable questions presented to the user with regard to the content on the page. The questions are grouped into 10 tests. The manual tests can be matched against 15 SC, 11 of which corresponding to the WCAG 2.0 level AA. Apart from that, the tool contains two additional tests satisfying the requirements of the Swedish National Guidelines for Web Accessibility – *Webbriktlinjer* [74], which WCAG 2.0 does not entail. An explanation for using this particular tool for evaluation comes with the fact that at the time of evaluation no other UTT tool with an API is known.

3.5.2. Dataset

Two sources of data were used for the data collection. Appropriately, from interviews and from the checking tools.

3.5.2.1. Interviews

The target group to reach out were the researchers conducting studies in the accessibility field. The interviews were carried out via Internet media such as web teleconferences, email exchange and Skype. The investigation abided by the research ethics rules such as data protection laws and confidentiality of information. In accord with these principles, the informants were guaranteed anonymity and given free will to respond the questions. The task of information collecting were approached with a self-put rule, which stated that the interest of participants should be protected. All of the research subjects were informed about the purpose of the interviews and gave a clear consent to participate in the study. All of the quotations were included in the

thesis after the interviewees' review. The interviews were approached with awareness that there is confidentiality involved and privacy policy, which protects the solutions. It was acknowledged as a part of the reality and the research limitations the researcher faces.

3.5.2.2. Accessibility Evaluation Results

A few general requirements were imposed for a valid dataset:

1. The same set of web pages used for manual and automated testing
2. Web pages coming from the same website
3. Manual and automated accessibility evaluation was conducted with no more than one week inbetween

The data used for the research consisted of two sets of web pages selected from two independent websites. The sets contained five and ten web pages appropriately. The criteria set for selection of data, imposed that the website had to be in English due to the manual evaluation purposes, as English was commonly-understood. Set number one, labelled by S encompassed five static web pages selected from the `http://eksempelsamling.medialt.no/` website. The website aims at providing simple examplary pages for web accessibility testing. Table 3.1 lists out the elected static web pages. Page S_1 contains a Norwegian anthem text with some images, page S_2 plain text, page S_3 lists out a few facts about Norway decorated with an image, page S_4 hits *page not found* landing page, while page S_5 contains a newsletter subscription form.

Table 3.1
Static web pages used for the experiments.

ID	Web page url
S_1	`http://eksempelsamling.medialt.no/wcag-errors/p001.html`
S_2	`http://eksempelsamling.medialt.no/wcag-errors/p003.html`
S_3	`http://eksempelsamling.medialt.no/wcag-errors/p006.html`
S_4	`http://eksempelsamling.medialt.no/wcag-errors/p008.html`
S_5	`http://eksempelsamling.medialt.no/wcag-errors/p011.html`

Let us denote D as a set of web pages with a dynamic content, chosen for the experiments from the United Nations' (UN) website `http://www.un.org/en/`. Table 3.2 outlines web pages belonging to the set D. The set D was required to contain the main page, contact page and 'about' page, due to the fact that most of the websites utilize CMS templates for those standard pages. The rest of the web pages were selected randomly and contained daily news, articles and blog posts.

3.5.3. Data Preparation

Before approaching the analysis of the data, some preparatory actions were taken:

Table 3.2
Dynamic web pages used for the experiments.

ID	Web page url
D_1	http://www.un.org/en/index.html
D_2	http://www.un.org/en/contact-us/
D_3	http://www.un.org/en/about-un/
D_4	http://www.un.org/en/events/waterday/
D_5	http://www.un.org/sustainabledevelopment/blog/2017/11/cop23-liveblog/
D_6	http://www.un.org/en/sections/general/meetings-and-events/index.html
D_7	https://www.un.org/press/en
D_8	http://www.un.org/en/sections/issues-depth/climate-change/
D_9	http://www.un.org/en/sections/issues-depth/women/index.html
D_{10}	http://www.un.org/en/women/endviolence/

– Both static and dynamic web pages were verified on the code level in order to confirm that they fulfilled the criteria of static/dynamic pages. More information on dynamic content can be found in Chapter 2, Section 2.3.

– Datasets S and D were automatically tested by the researcher with use of the checker. As a result of that, two new data sets were identified, namely SA and DA.

– Two additional datasets, SM and DM were created by manual testing with help of the UTT bookmarklet. The manual assessment was performed by the researcher with an accessibility background. All sensitive information collected by the tool during evaluation, like ip address or user-agent, were left out due to the data protection policy.

3.5.4. Experiments Procedure

Figure 3.7 depics the steps that were undertaken in the experiment. Usually, with a larger set of web pages, the process of website accessibility evaluation starts with web crawling in order to download a set of URLs. The URLs are then simultaneously stored in the database and sent to the sampling module, where uniform random sampling without replacement is applied. For this study, the process of sampling was beyond the scope. It was assumed that the pages elected for the experiment were the sample pages chosen for evaluation at the sampling stage. In the next stage, the sample of the selected data was sent to the following testing component. The data were then evaluated both automatically with use of the ATT and manually by being tested by a human.The results of the assessment were stored in the database and some QA employed. In the end, integration actions were applied and a new unified accessibility score computed.

The experiment involved an empirical study of the unified score calculation approaches, resulting in score calculation by utilization of outlined in 3.4.2 accessibility metrics. For each dataset created for the experiment, the accessibility scores were calculated three times using

examined approaches to integration: Score Integration approach, Intersection approach, and Union approach.

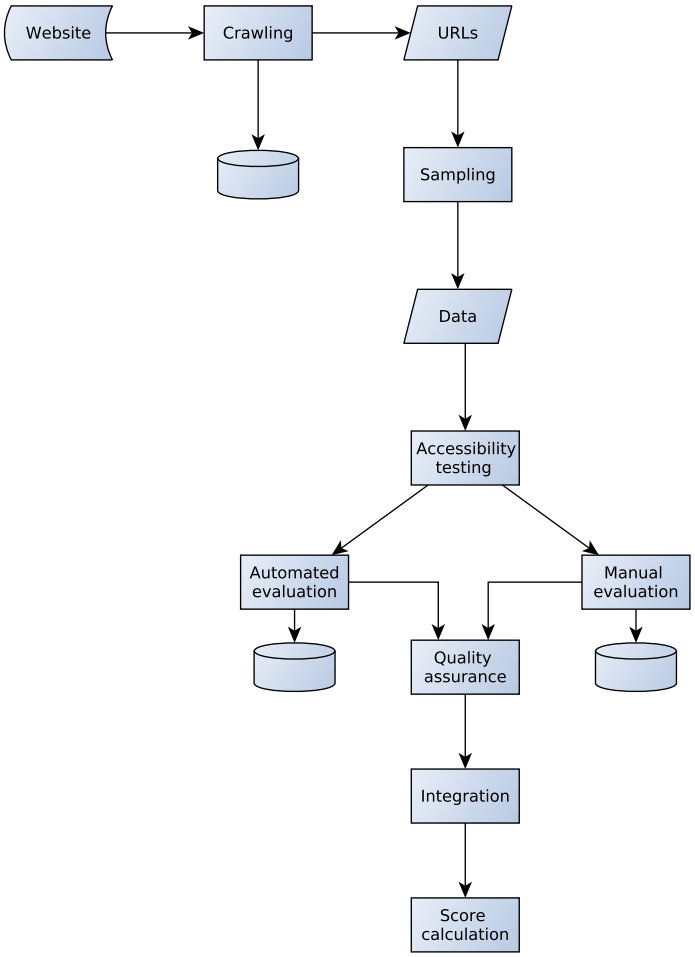


Figure 3.7
Workflow of manual and automated accessibility evaluation integration process.

3.5.4.1. Accessibility score for manual testing

In order to approach the task of score calculation of the UTT results, the manual tests implemented in the tool had to be assigned to the relevant SC from WCAG 2.0. The tests have been implemented in such a way that there exists a *m:m* – multiple-to-multiple relationship between a single test t and a SC c. Figure 3.8 presents the type of the relationship.

A framework for score calculation of the manual accessibility test results was proposed with the following table as a basis for calculation. Table 3.3 presents a matching between the implemented UTT tests, WCAG 2.0 SC and the Webbriklinjer.

Having established the mapping relations between SC and the implemented tools, manual evaluation was conducted on the sets S and D to create the subsets SM and DM. Following the proposed integration approaches *Approach 1 – Scores Integration*, *Approach 2 – Intersection*

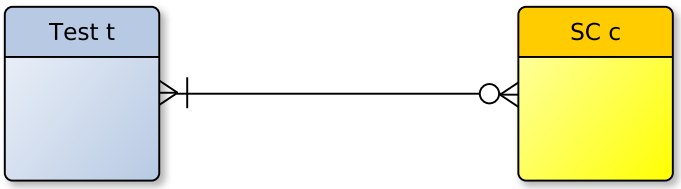


Figure 3.8
A multiple-to-multiple relationship between a test t and a SC c. One test can be linked to none or many SC, while one SC has to be related to at least one test.

Table 3.3
UTT test id indicates a test identifier. *Question id* points to the atomic question presented to the tester. *Success Criteria* informs about the test reference to the WCAG 2.0 SC. *Webbriktlinjer* shows the connection between a paticular test t_i to the Webbriktlinjer.

UTT Test id	Question id	Success Criteria	Webbriktlinjer
t_1	q_1	3.1.3, 3.1.4, 3.1.5	10
	q_2	3.1.3, 3.1.4, 3.1.5	10
t_2	q_3	-	24
t_3	q_4	-	51
t_4	q_5	2.4.6	61
t_5	q_6	1.4.4	91, 111
t_6	q_7	1.4.4, 2.1.1, 2.1.2, 2.4.3, 2.4.7, 4.1.2	34
	q_8	1.4.4, 2.1.1, 2.1.2, 2.4.3, 2.4.7, 4.1.2	34
t_7	q_9	1.4.8	91, 111
t_8	q_{10}	1.2.2, 1.2.4	-
	q_{11}	1.2.2, 1.2.4	-
t_9	q_{12}	1.1.1	115
t_{10}	q_{13}	2.4.2	115

approach, and *Approach 3 – Union approach*, outlined in 3.4.2, the accessibility scores were calculated for each subset.

3.5.4.2. Accessibility score for automated evaluation

Automated assessment of a page was immediately commenced after the manual check of that web page to simulate as much as possible concurrent testing. For the calculation of the accessibility score for automated evaluation, the established score function from the EIII project was used. The function is currently utilized in the ATT. As a result of the checker testing of the sets S and D, subsets SA and DA were obtained and the scores computed.

3.6. Quality Assurance

QA of the accessibility checking is an intricate part of the web accessibility assessment. In fact, there is no guarantee at all that the evaluation results are valid 100%. Since no ready-made implementation that would be flawless is available to the checker or the tester, it is hard to return a verdict. How can one state that the results are sound if they do not have the solution to compare with? That remains a situation when trying to do QA of the accessibility test results.

Assumptions are the foundation of automated testing. ATT searches for symptoms of accessibility violations. Every automated test works on assumptions in one way or another. What is considered inaccessible by one tool, may be at the same time regarded as allowed by another accessibility checking tool. Accessibility checkers mostly cannot unanimously determine whether a SC was met or not. The absence of issues highlighted by a tool does not state that the website has no accessibility problems. It means that that tool has not encountered any. The aim of the tool is to provide its user with a notion of the potential obstacles. However, there exist actions that can be taken in order to verify whether the produced results are viable or not.

First step of the QA was undertaken after the testing had been finished. In case of the automated testing, the following QA criteria have been applied:

- Verify that all evaluated web pages have gotten the check results.

- Verify a sample code excerpts from a few individual pages to see if the results provided by the checker reflect the actual accessibility barriers present on the page.

When it comes to manual testing, the task of QA was also conducted. Special attention was put on verifying the quality and completeness of the provided input.

For the QA of the propounded integration approaches and the score calculation metrics some aspects were considered. Due to the novelty of study, hence not many reference points, the experimental nature of the research limits possibilities for validation. In fact, the only reference that can be made is the score from automated checking, which can be considered stable and reliable enough. However, the score obtained as a result of integration expounds it evaluation territory to the greater coverage thanks to incorporating the manual testing outcomes. This contribution to the score computation makes the comparison less deciding in terms of validation, yet still interesting to explore the relationship between the automated and the integration scores. The automated score cannot be tantamount to the decision of the integration score correctness, but still valuable to look at.

Similar is the case for the validation of manual score calculation. That is to say, automated score provides a taste of the accessibility state of the page, yet it is not a determinant of its accuracy [5].

3.7. Summary

Following methods were used in this thesis as aid in order to answer the research questions posed at the begining of the study. Table 3.4 lists the methods together with the purpose of their application and associated research questions (RQ).

Table 3.4
Methods used in the study and their relationship to the research questions.

Method	Purpose	RQ(s)
Interviews	To understand the challenges and gather ideas about integration	RQ1, RQ2, RQ3
Descriptive statistics	To explore the central tendency and understand datasets	RQ1, RQ3
Boxplot	To understand the spread and variation of the accessibility results	RQ1, RQ3
Scatter plot	To understand the assosciacions between the pairs of the scores	RQ1
Correlation (Pearson)	To measure the strength of the relationships between the scores	RQ1
Euclidean distance	To investigate the difference between the evaluation results computed with use of different AEMs	RQ1
Manual page inspection	To check whether the same page elements have been evaluated both by ATT and UTT	RQ1, RQ3
Metric's quality validation	To examine the properties of the proposed web accessibility metric	RQ2
Accessibility Pie Chart	To present the evaluation results visually	RQ2

4. Results

4.1. Interviews

Interviews have been conducted as a qualitative part of the study. The aim of the interviews was to provide the study with remarks on possible directions, understand the challenges of the manual and automated evaluation methods integration. Also, the purpose was to get to know the opinion of the accessibility experts about the undertaken endeavour.

In total, seven interviews have been carried out. The interviews were standardized, open-ended, which means that the users were asked the same questions and given freedom to answer the questions. Two of the interviews were conducted via Skype, three via email exchange and one over the phone. All interviewees were the researchers actively working in the area of web accessibility, universal design and usability.

The interviews have served as a method to answer the research questions RQ1, RQ2, and RQ3. The responses were gathered and analyzed with regard to the research questions they refered to.

RQ1: How to combine automated and manual evaluation of web accessibility?

Analysis:

The interviewees were all in agreement that despite being a difficult task, combining automated and manual evaluation of web accessibility is important and necessary since different methods are able to expose different types of issues. What is more, it was emphasized that automated tools cannot find all accessibility barriers. Moreover, not only combining manual and automated evaluation, but also combining various methods for manual evaluation as well as different tools for automatic evaluation, would be beneficial. Following things have been identified by the responders as possible challenges of manual and automated methods combination:

- context/situation and current state of the manual tester
- order of checking manually may influence the evaluation results
- format of data and test results may be problematic to merge
- in case of manual testing, different evaluators produce different accessibility results

- a great amount of pages to check may impede manual evaluation and further integration
- going beyond the paradigm where there is performed an assessment by the tool and a report is generated
- going beyond the static evaluation of the page, i.e., also testing scripts, watching the behaviour of the page
- it may be challenging to supplement automatic testing with manual check 1:1 due to the very time-consuming nature of manual testing
- understanding what type of accessibility issues each of the methods can identify

As the responses showed, the interviewees referred mostly to the challenges related to dynamic content, subjectivity of the manual evaluation and feasibility of manual evaluation of a great number of pages. Additionally, the attention was paid as well to extending the usual evaluation, which focuses on the checking the conformance manually and provide a broader perspective on the real accessibility of the website.

The interviewees were also asked about the level of possible integration, whether the results should be integrated on the level of websites, web pages, page objects or maybe accessibility guidelines. Two third of the responders pointed out that the level depends on the intended use. It was suggested by one of the researchers that the tools should provide a sample of pages and ask questions to the users about the cases in which the statement could not be made by the tool. Another interesting approach was proposed by an accessibility expert Giorgio Brajnik:

"One approach would be to extract widgets from the pages and base the evaluation on behaviour testing of these widgets. Combining this novel way of thinking, perhaps with neural networks could better address the challenge of dynamic content evaluation."

Knowing the potential of neural networks, this idea could be perhaps put as a future work.

RQ2: How to express the integration results in a quantitative way and present them visually?

Analysis:

Most of the users answered that the way of presenting the results depends very much on the intended audience. It became clear from the interview with the Technical Director of Accessibility Foundation in the Netherlands and known accessibility expert Eric Velleman, that generally users prefer graphical visualizations to textual information. Yet, as the author underlines, all depends on the users:

"Generally, all depends on the target group. The users or policy makers do not need the statistical details to know that the page is accessible enough. However, the developers would probably like to delve into the data and see what exactly causes barriers and in that case numerical results would be handy."

The intended use of the data were indicated together with the purpose what do the users need the data for.

RQ3: What is the impact of the dynamic content on the integration results?

Analysis:

The interviewees accentuated that dynamic content has a significant impact on the evaluation results as it is where the tools differ in responses. It is a challenge to ensure that the same content is evaluated. Facebook may serve as an example. The content feed is changing with each page refresh. Having that in mind, it was pointed out that dynamic content is a good argument for mixed evaluation, where the automated testing stands back quantity and manual takes care of quality of the evaluation.

Summary:

The answers of the responders were taken into consideration for the further design of the quantitative experiments. More focus in the study was put on the key aspects of integration underlined by the interviewees. The way of presenting the results was designed in a way to best suit the needs of the target group. The challenge linked to the evaluation of the dynamic content was acknowledged. The interviews have contributed to the better understanding of the endeavour. Moreover, through the interviews, the experts provided inspiration for the ideas of future research avenues to take.

4.2. Data Triangulation

Accessibility evaluation results from the ATT are quantitative in form of a number of applied accessibility tests and their outcomes with the distinction for *passed, verify, failed*. On the other hand, UTT bookmarklet produces qualitative data of a categorical character: *pass/fail/incomplete*. In order to link the outcomes of the ATT and the UTT accessibility evaluation, the concept of data triangulation was utilized for combining results from ATT and UTT. The first step that was taken in that direction was to quantify the qualitative UTT results. The *pass* and *fail* possible user responses were assigned binary values, based on the context of the accessibility question. The responses with status *incomplete* were excluded from the score calculation. Table B.1 in the Appendix B presents the questions from the UTT and possible user answers that got assigned numerical values.

4.3. Data description

An introductory QA of the data was conducted. All of the evaluated pages were tested by the checker. Additionally a few individual pages were inspected for the soundness of the results.

In case of manual testing, the completeness of the answers was verified. It turned out that there were missing two responses about the *alt text* of the pictures for pages D_9 and D_{10}. It was decided not to remove those pages for the sake of the observation how the missing results influence the accessibility score after integration. When it comes to the quality of the assessment, the provided input was of high quality.

The produced final data consists of two data frames containing calculated accessibility scores of two sets of web pages (S and D) with use of five evaluation methods as indicated in 3.4. The methods namely ATT, UTT, Scores Integration approach, Intersection approach and Union approach are referred hereafter as features while the single web pages are referred as observations. The main focus was put on understanding the relationships between the data features more than investigating the relationships among the observations. Yet a simple descriptive statistics is provided to give the notion of the central tendency measures. In order to better understand the dataset, Exploratory Data Analysis (EDA) has been applied. More information on EDA can be found in [75]. Figure 4.1 presents a brief summary of the accessibility scores among web pages from the set D. All of the attributes are quantitative variables of a discrete character.

Figure 4.1
Table shows a summary of evaluation results of the dataset D. Considered methods of score calculations were used for testing.

	att	utt	integration	intersection	union
Min. :	82.73	70.76	77.37	81.85	80.28
1st Qu.:	85.10	80.38	84.90	94.14	85.74
Median :	87.29	97.89	92.04	96.83	91.80
Mean :	88.83	90.30	89.49	94.54	90.16
3rd Qu.:	92.37	99.58	93.58	98.51	94.33
Max. :	97.18	100.00	98.50	99.46	98.27

Similarly for the dataset S, the central tendency measures are presented in Figure 4.2.

Boxplots analysis

In the studied case, the values of the ATT scores are not spread a lot, while variation of the UTT scores proved to be high for the set D. Considering these two variables, Union approach seems to be the least prone to the changes of UTT Score values among the three investigated AEMs. An interesting behaviour of the Intersection score can be observed in both cases. The interquartile range (IQR) for the Intersection score calculated for dataset S is significantly bigger than for others boxplots. There arises a question how does the spread of the UTT values influence the Intersection score? According to the statistics presented in Figure 4.3 the bigger IQR for UTT score, the less spread are the values for the Intersection score. Yet, there is not so much

Figure 4.2
A summary of evaluation results of the dataset S. Accessibility scores were calculated with use of the investigated in this study AEMs.

	att	utt	integration	intersection	union
Min. :	55.04	38.89	45.81	47.29	46.05
1st Qu.:	84.44	93.45	89.36	74.75	87.11
Median :	86.89	95.83	94.04	88.89	94.39
Mean :	84.64	85.40	84.90	81.13	84.33
3rd Qu.:	96.83	98.81	97.37	96.97	96.21
Max. :	100.00	100.00	97.91	97.73	97.88

difference of the ATT between the datasets. Taking these two observations and recalling that the Intersection score is calculated on the basis of common SC from ATT and UTT evaluation, it is noticed that there is no rule stating that the more agreement between the ATT and UTT tools, the higher the Intersection score. Besides, medians for Integration and Union scores are nearly the same in both datasets.

Scores for single pages

The calculated scores for the single pages were separately analyzed one by one. Figure 4.4 presents the results graphically for the dataset S. Apart from the results for the page S_1 the ATT and UTT tools yielded similar notes for the accessibility of the pages. The Union score and the Integration score resulted in close values. The Intersection score varied most compared to the two remaining methods of the combined score calculation.

Similarly, the scores for pages from the dataset D were inspected and the results depicted in Figures 4.5 and 4.6. Evaluation results for the dataset D were less spread within than the scores in the dataset S. For the pages which received similar evaluation by ATT and UTT, all three approaches to combined score function produced nearly equal results, for instance for pages S_1, D_5 or D_4.

Scatter plot matrix

In order to understand the associations between the calculated scores, a scatter plot matrix was created. Figure 4.7 and 4.8 present graphically relationships between pairs of scores for sets D and S.

For the set D there can be observed a positive trend for the relations (ATT-Integration, ATT-Union, UTT-Integration, UTT-Union). It is a moderate relationship and it follows a linear pattern. The strongest association between the variables is noticed between Union and Integration scores. For the Integration and Union approaches, the trend curves for ATT and UTT are

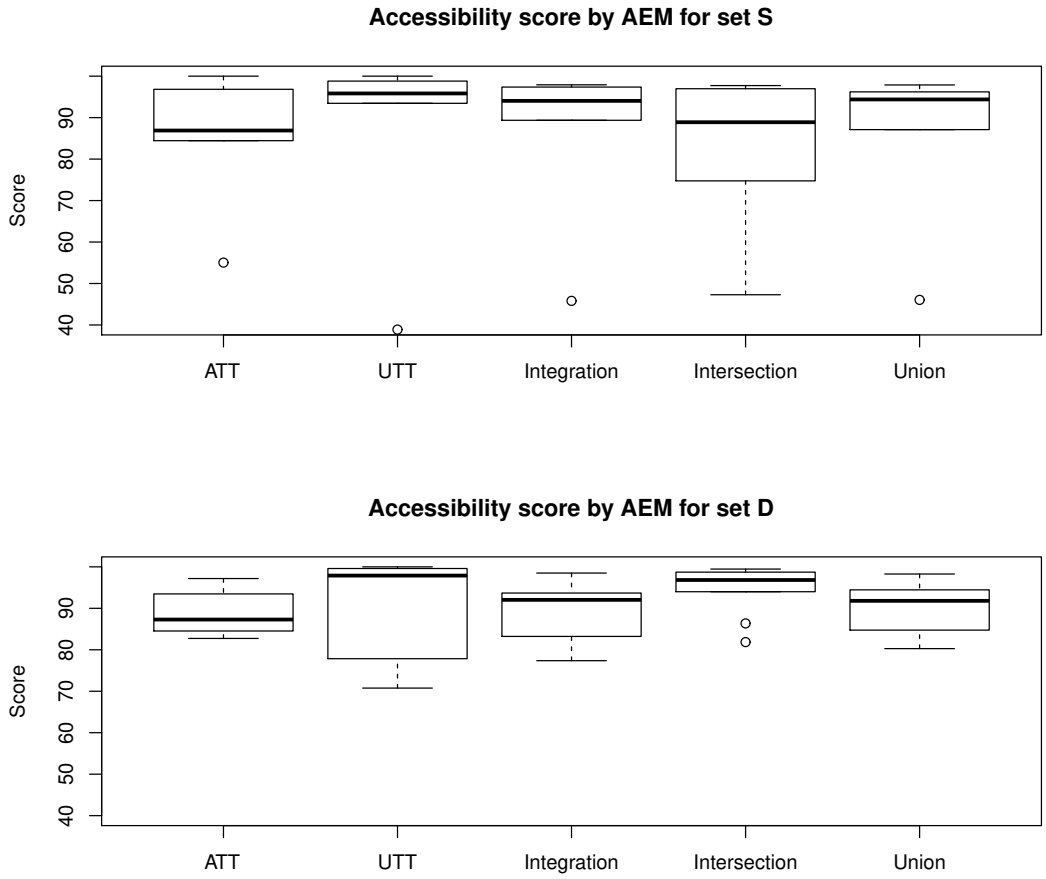


Figure 4.3
Graphical illustration of the spread of the score values by evaluated AEMs for the dataset evaluated datasets S and D.

very similar in a sense that the relationships UTT-Integration and UTT-Union resemble each other. Likewise ATT-Integration and ATT-Union. Quite many outliers can be observed in the relationship between ATT and UTT, which suggests that ATT results were not always aligned with the UTT results.

When it comes to set S, all of the relations have positive direction of a linear shape. A stronger relationship between the variables can be observed. Compared to the dataset D, relationship between ATT score and UTT score is stronger, with no outliers. That could probably be explained with differences of automated and manual evaluations of the dynamic content present on the pages in the dataset D. Another possible explanation could be the quality of the manual testing, where the tester could have missed out some barriers. Integration Score behaves itself similarly when being the dependent variable against the ATT and UTT scores. The same pattern can be noticed for the Union Score dependent on ATT and UTT scores.

Looking at the data holistically, in case of the set D, much weaker associations are present

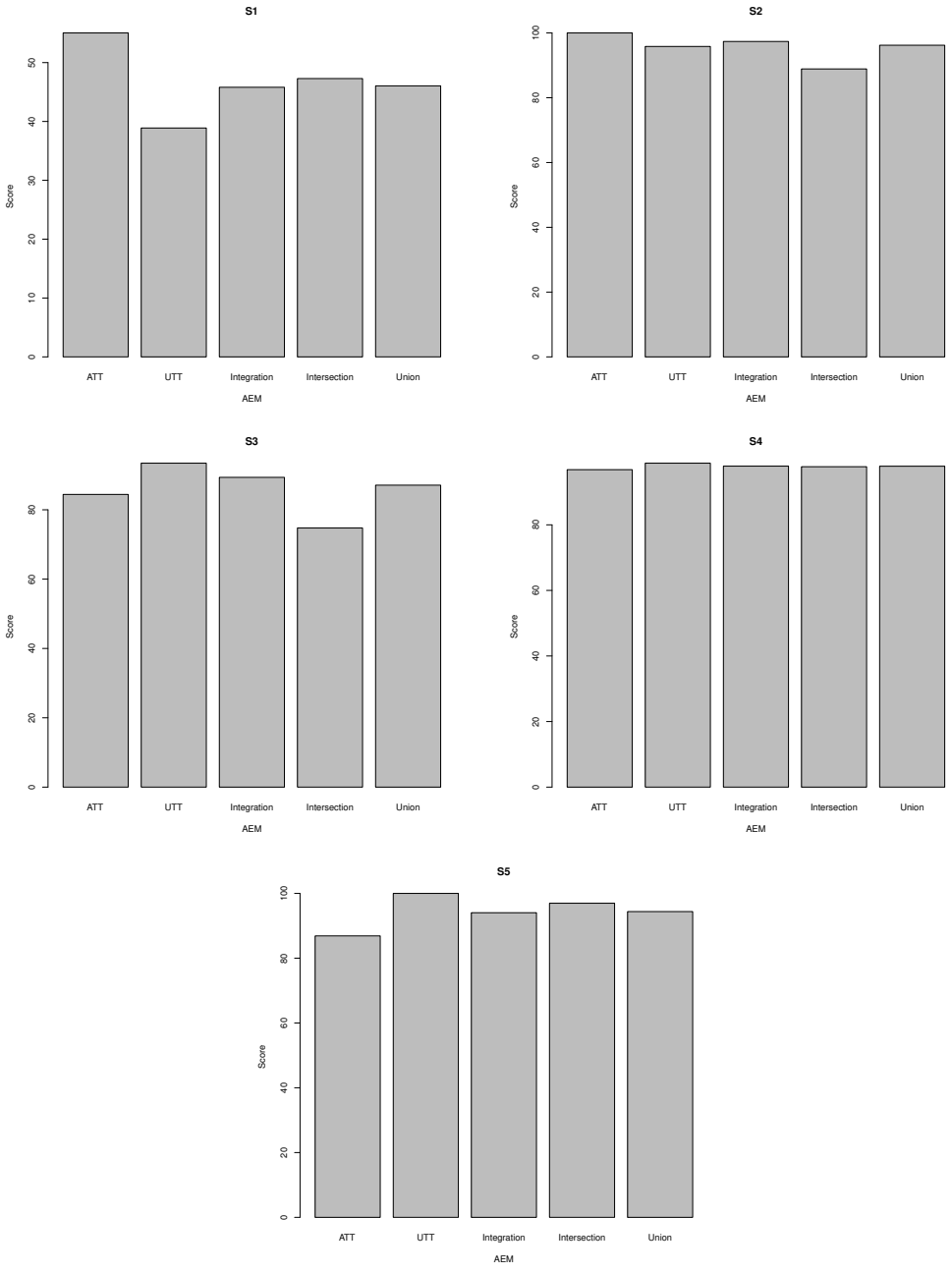


Figure 4.4
Individual presentation of calculated scores with use of five AEMs: ATT, UTT, Score Integration, Intersection, and Union for the dataset S.

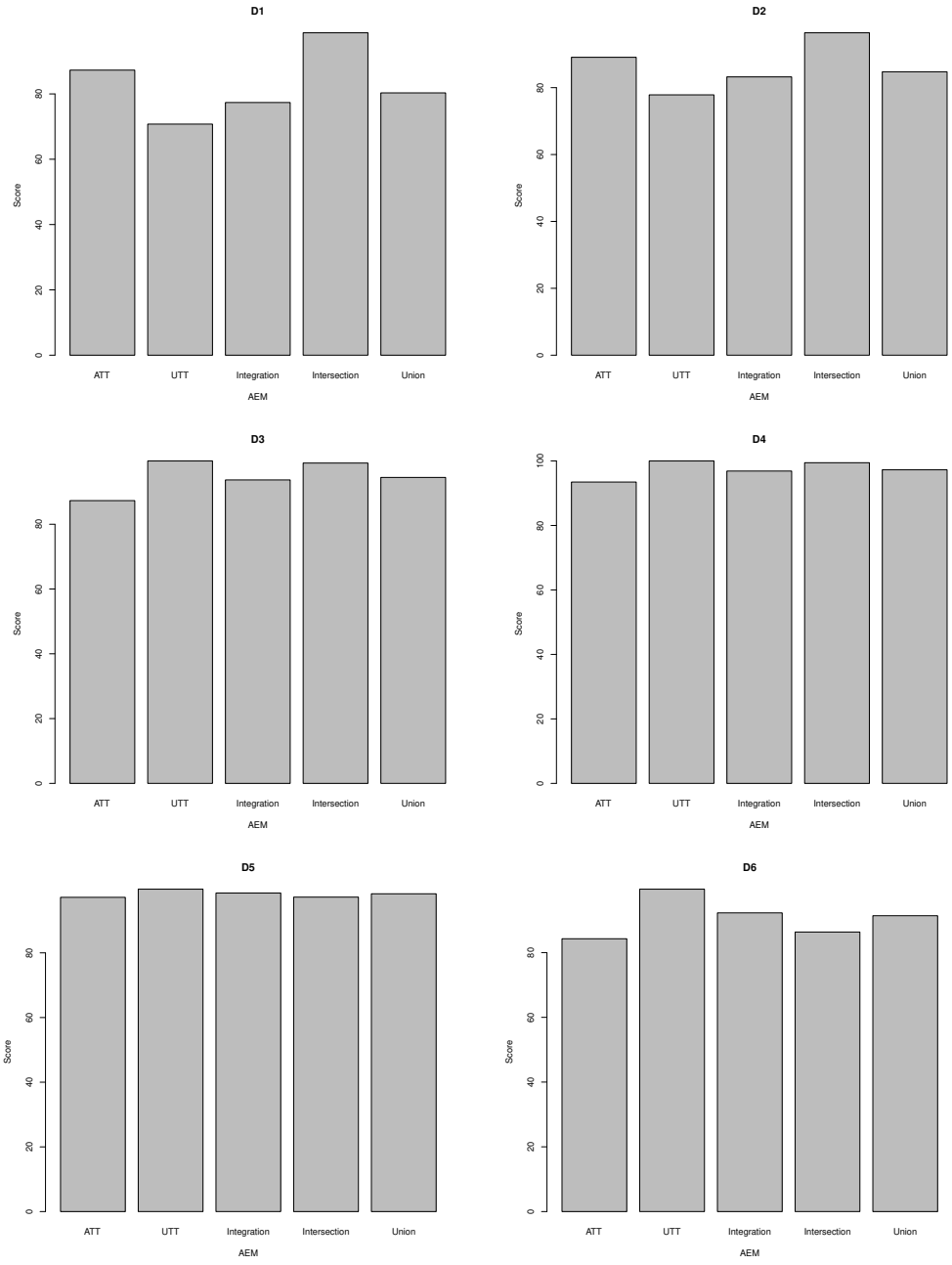


Figure 4.5

Individual presentation of calculated scores with use of five AEMs: ATT, UTT, Score Integration, Intersection, and Union for the pages $D_1 - D_6$ from the dataset D.

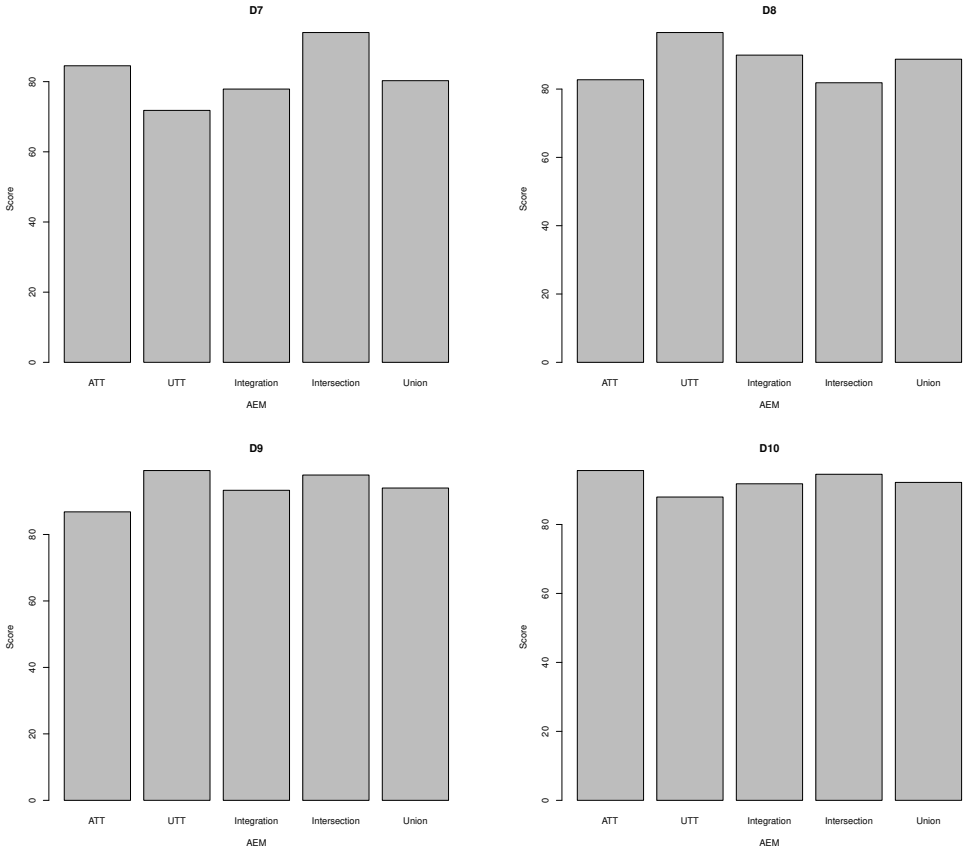


Figure 4.6

Individual presentation of calculated scores with use of five AEMs: ATT, UTT, Score Integration, Intersection, and Union for the pages $D_7 - D_{10}$ from the dataset D.

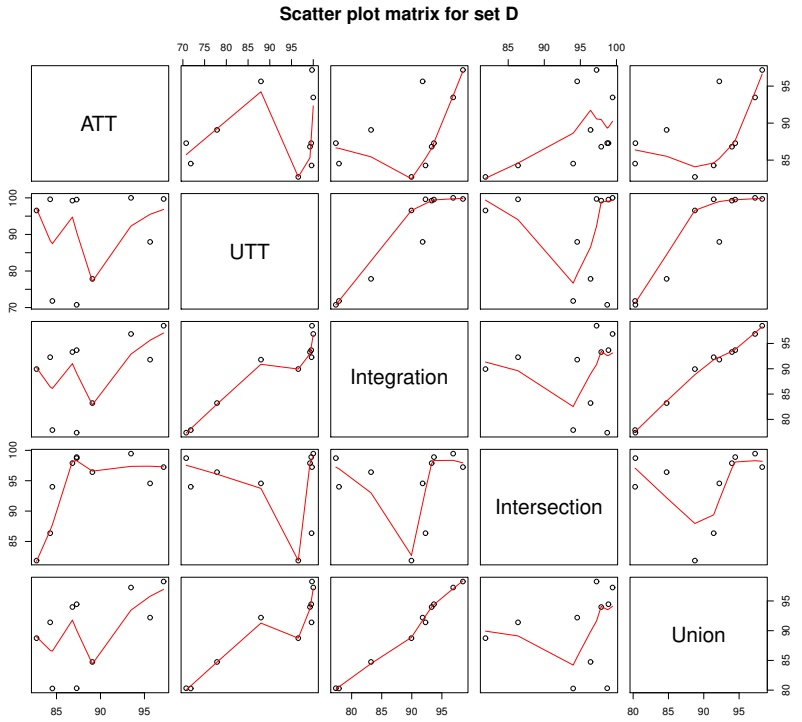


Figure 4.7
The matrix presents a pairwise study of relations between calculated scores using examined AEMs-(ATT, UTT, Score integration, Intersection, and Union) for the dataset D.

than for the set S. Moreover, the relationships between computed scores for the dataset S show more linearity than for the dataset D.

Correlation between accessibility scores

In order to measure the strength of the relationships between particular accessibility scores, Pearson's correlation coefficients for the computed accessibility scores were calculated. Figures 4.9 and 4.10 show the correlation panels for the datasets S and D, respectively.

From the point of the study, the most interesting relationships to investigate were the relationships between the ATT/UTT and the combined scores (Scores Integration score, Intersection score, and Union score). By this mean, it can be understood how changes in the ATT/UTT scores are reflected in the combined scores. Little knowledge for this research is gained by analyzing the correlations between the combined scores. In case of the dataset S, a very strong positive relationship was noted – in all cases $r\ value = .90$. For the dataset D, strong positive relationship ($+.40 - +.69$) was observed between the ATT and Scores Integration score ($+.49$), ATT and Intersection score ($+.54$), and ATT and Union score ($+.55$). It suggests that the ATT results determine significantly the integration scores. Analysis of the correlation coefficients between the UTT and the integration scores revealed very strong positive relationships between the UTT and the Scores Integration score ($+.95$) as well as the UTT and the Union score ($+.92$), which

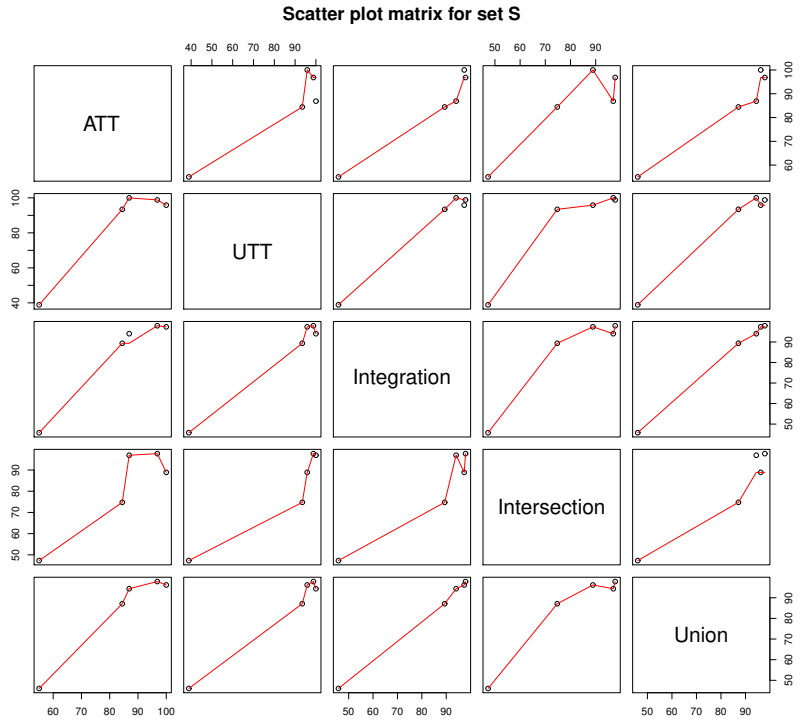


Figure 4.8
The matrix gives an overview of a pairwise study of relations between calculated scores using examined AEMs for the dataset S.

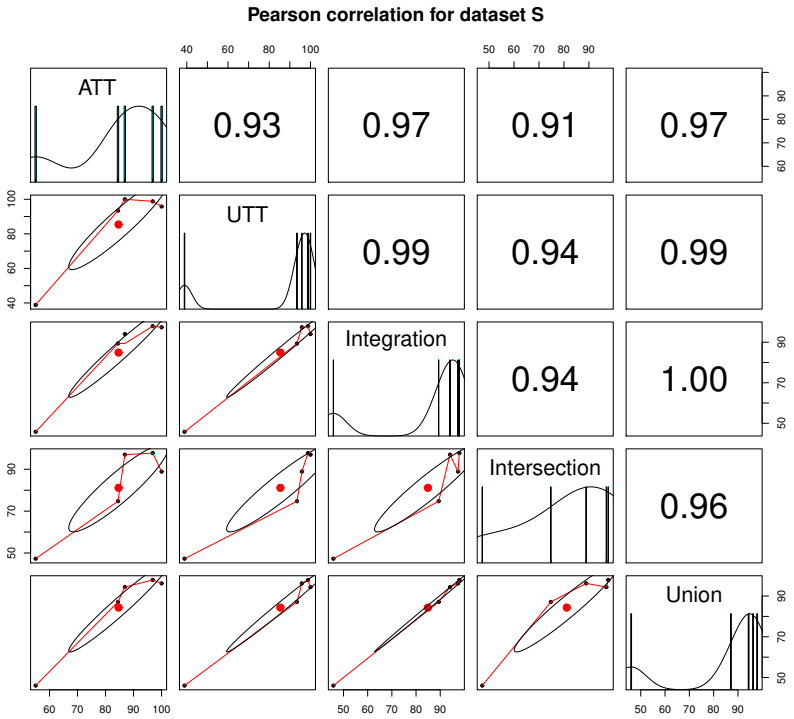
proves that the UTT outcomes influence more the Integration and the Union scores than the ATT results. A negligible negative relationship was found ($-.17$) between the UTT and the Intersection scores. At the same time, there was observed a negligible positive relationship between the ATT and the UTT scores ($+.18$).

Euclidean distance between accessibility scores

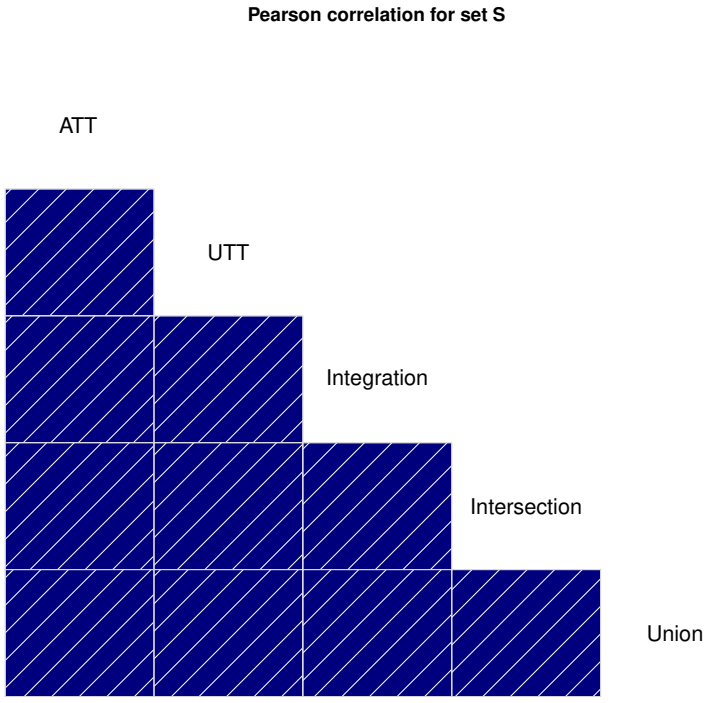
Euclidean distances between the accessibility scores were measured. The purpose of this information was to investigate the difference between the evaluation results computed with use of various AEMs. Figures 4.11a and 4.11b present graphical heatmaps representing the Euclidean distance between pairs of the accessibility scores: ATT, UTT, Scores Integration, Intersection, and Union scores for the dataset S and D.

The results for the the dataset S showed that the biggest distance was noted between the ATT and the UTT (23.13), while the shortest one between the Scores Integration and the Union scores (2.57). It was observed as well that the ATT score leaned more towards the Union score. Conversely, the UTT was closer to the Integration score – approximately one unit of a difference between them. In case of the dataset D, the shortest distance was found between the Scores Integration and the Union scores (4.51). On the other hand, the biggest distance was found between the UTT and the Intersection scores (45.41). The distance between the ATT and the

Figure 4.9
Analysis of the accessibility scores and relationships between them for the dataset S.

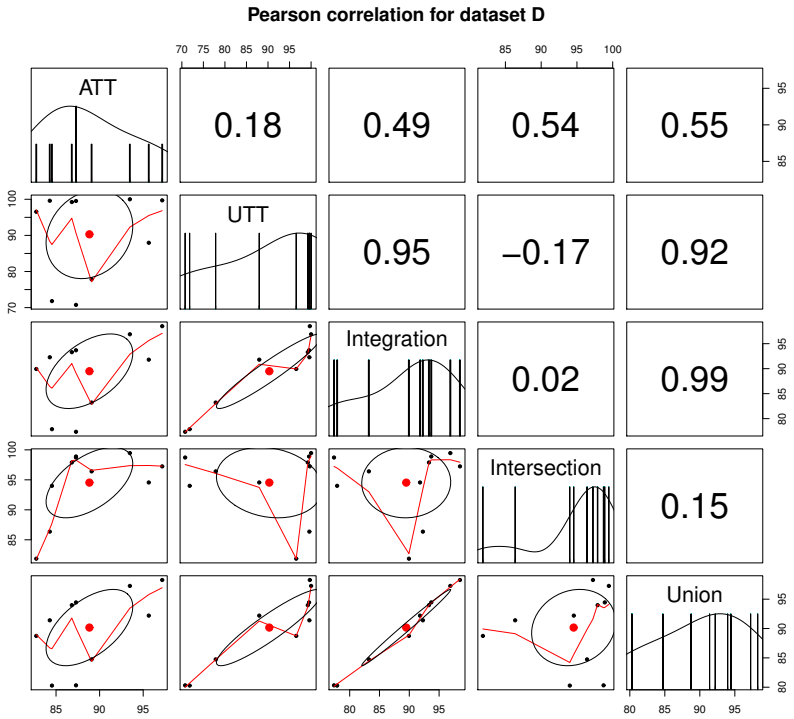


(a) The panel presents calculated Pearson's correlation coefficients for the accessibility scores computed for dataset S using examined AEMs.

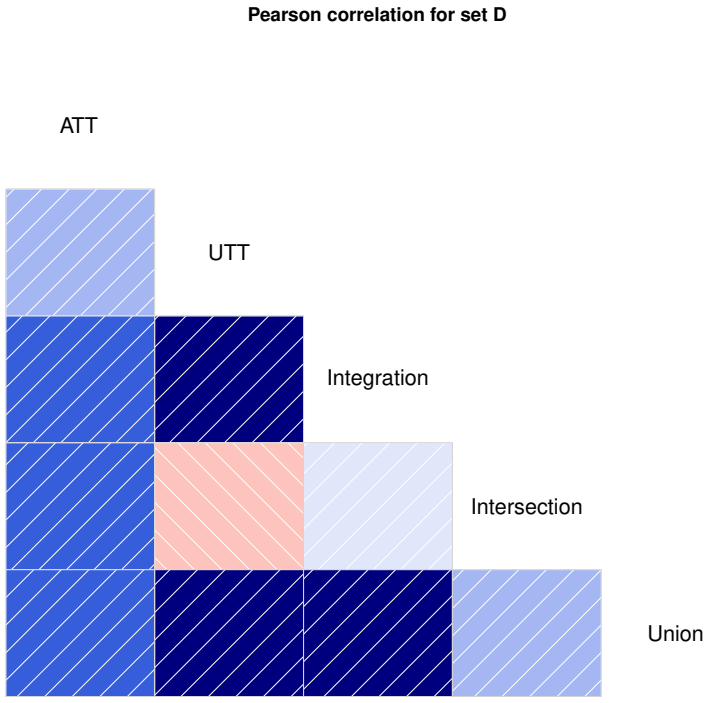


(b) Pearson correlation for dataset S supplied with a heatmap presenting the strength of the relationships.

Figure 4.10
Analysis of the accessibility scores and relationships between them for the dataset D.



(a) Pearson's correlation coefficients for the accessibility scores computed for dataset D accompanied by scatter plots.



(b) Heatmap presenting the strength of the relationships between accessibility scores for dataset D.

UTT scores was 37.41. Similarly as for the dataset S, the ATT score vector lay closer the Union score vector and the UTT score vector inclined more to the Scores Integration vector. The difference of about three units was noted.

4.4. Quantitative Study Results

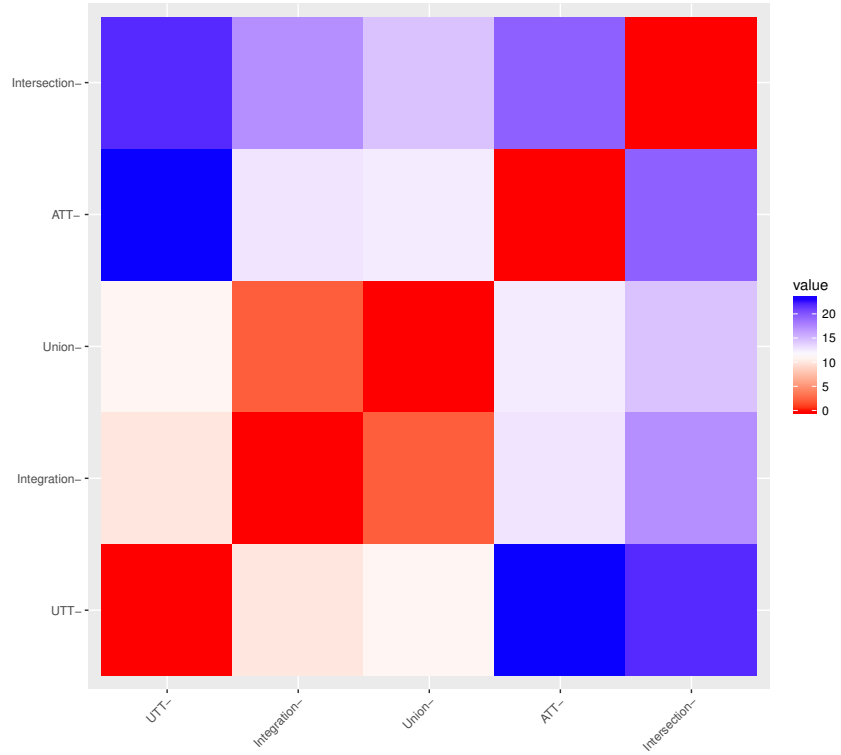
The main part of the research involved quantitative study of the possibilities for combining manual and automated accessibility testing, as well as an investigation on the most appropriate ways to conduct it. In this section, first, results of the integration on SC level are discussed. They are focused principally on conformance testing and addressing the coverage of the POUR principles from WCAG 2.0. Later, integration on page object element is examined by a thorough study of the accessibility of a selected as an example web page D_7.

4.4.1. Integration on Success Criteria Level

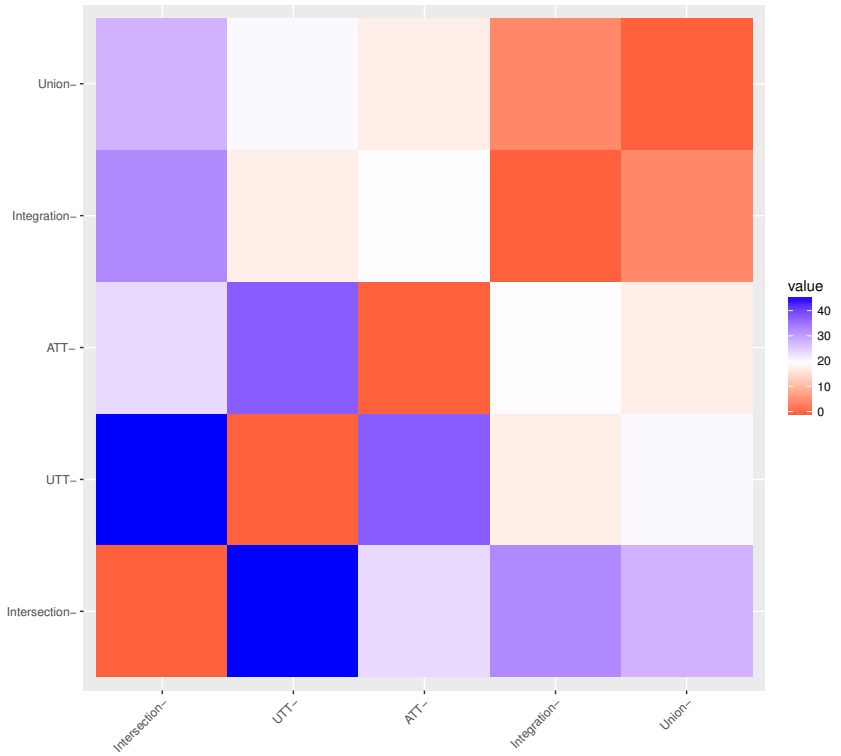
It was decided to conduct the combination of the accessibility test results on the level of SC from WCAG 2.0, due to the fact that the tests both in ATT and UTT can be aggregated on SC. WCAG 2.0 acts at this point as a linking bridge between the tools. In order to be able to link the test results, the tests from ATT and UTT were grouped under the corresponding SC. Table 4.1 illustrates the mapping of the implemented tests to the appropriate SC from WCAG 2.0. A numerical comparison of SC covered by distinct evaluation tools, i.e., ATT, UTT and by means of the integration is delivered in Table 4.2. In total, the WTKollen testing tools cover 27 SC of all levels, where ATT implements 17 and UTT 15 SC. Among them, five SC are covered both by ATT and UTT. On level AA, combination of the ATT and UTT covers 22 SC, with a distinction to 16 provided by the ATT and 11 by the UTT. Figure 4.12 presents a coverage of implemented SC in relation to the WCAG 2.0 guidelines level AA. It can be seen that the combination of the ATT and UTT may be advantageous for conformance review purposes. Owing to the integration of the ATT and the UTT assessment, 57.9% of the WCAG 2.0 SC level AA can be covered, compared to 42.1% and 28.9% when the testing is conducted using only one method, either automated or manual, respectively.

The coverage was also studied with regard to the POUR Principles on which WCAG 2.0 is based. In this case, again level AA was taken into consideration. Figures 4.13 and 4.14 present the coverage of the individual principles by the studied accessibility evaluation approaches. By combining ATT and UTT 42.9% of the SC from the principle *Perceivable* was covered. When it comes to the second principle – *Operable*, integration of approaches resulted in 83.3% coverage of the guidelines classified under this principle. Results showed that for principle *Understandable*, ATT and UTT combined together can cover up to 40.0% of the SC from WCAG 2.0, all provided by ATT testing. UTT did not make any contribution for this particular principle. Finally, the last principle *Robust* was covered in 100.0%, as the analysis indicates.

Figure 4.11
Analysis of the distance between computed accessibility scores for the dataset S and D.



(a) Euclidean distances for the dataset S.



(b) Euclidean distances for the dataset D.

Table 4.1

Tests from the ATT and the UTT were mapped against appropriate SC from WCAG 2.0.

SC	ATT test id	UTT question id
*SC*1.1.1	H45, H35, H37, H46, H36, H53, H2	Q12:image
*SC*1.2.2		Q10:captions, Q11:desc
*SC*1.2.4		Q10:captions, Q11:desc
*SC*1.3.1	H39, H48	
*SC*1.4.1	SC1-4-1-a	
*SC*1.4.4		Q6:small:screen, Q7:nav, Q8:focus
*SC*1.4.8		Q9:scroll
*SC*2.1.1	F54, F55	Q7:nav, Q8:focus
*SC*2.1.2		Q7:nav, Q8:focus
*SC*2.2.1	F41, F40	
*SC*2.2.2	X'HasMarquee, X'HasBlink	
*SC*2.2.4	F41, F40	
*SC*2.4.2	H25, F25	Q13:title
*SC*2.4.3		Q7:nav, Q8:focus
*SC*2.4.4	H30, F63, H33, H24	
*SC*2.4.5	G125	
*SC*2.4.6	G130	Q5:heading
*SC*2.4.7		Q7:nav, Q8:focus
*SC*3.1.1	SC111text, SC3-1-1-xml-lang, SC3-1-1-html	
*SC*3.1.2	SC3-1-2-xml-lang, SC3-1-2-lang	
*SC*3.1.3		Q1:text, Q2:words
*SC*3.1.4		Q1:text, Q2:words
*SC*3.1.5		Q1:text, Q2:words
*SC*3.2.2	G13, H32	
*SC*3.3.2	H71, G167, G131	
*SC*4.1.1	SC4-1-1-id, SC4-1-1-idref, SC4-1-1-accesskey	
*SC*4.1.2	F59, H65, H44, H63, H91, F89, F68	Q7:nav, Q8:focus

Summary results of the comparison study of the WCAG 2.0 Level AA are presented in Figure 4.15. More detailed data about the SC coverage, divided by guidelines, are available in A.1. The diagram shows that the Robust principle was covered completely. The smallest coverage of SC was observed for the principle *Understandable*. N.B. it should be mentioned that the UTT bookmarklet tool has implemented additional three tests that are aimed to verify conformance with the guideline 3.1 from WCAG 2.0. Those tests were not taken into calculations since they are on Level AAA. Similarly, one extra automatic test can be found in the checker for the guideline 2.2, Level AAA. Yet, their presence needs to be highlighted. The reason for choosing Level AA for the analysis is the compability with the WAD regulations, which adopts conformance to SC Level AA.

Table 4.2

Coverage of SC from WCAG 2.0 according to various levels.

SC Level	ATT coverage	UTT coverage	Common SC	Integration coverage
A	13	7	4	15
AA	3	4	1	6
AAA	1	4	0	5

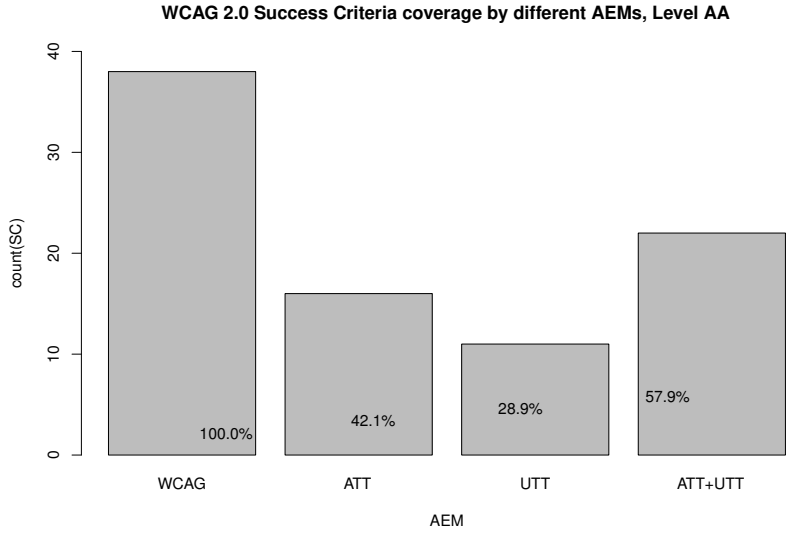


Figure 4.12

WCAG 2.0 SC coverage by ATT, UTT and combined approach ATT+UTT indicated as A+U.

Thanks to the combination of the ATT and UTT, the percentage of the accessibility guidelines can be significantly increased, compared to the situation when a website is evaluated only with use of the ATT. Integrated approach brought a considerable increment in CR of the respective accessibility principles from WCAG 2.0. Graphical illustration of the coverage gain is depicted in Figure 4.16. Two principles (Perceivable, Operable) have gained on coverage of the WCAG 2.0 principles. When it comes to the remaining two principles, integration did not alter the coverage of the *Understandable* principle since the UTT had no contribution for this particular principle. In case of the *Robust* principle, no increase in coverage could be obtained by implemented in the UTT tests since the ATT had already covered the assigned to this principle guidelines.

On the other hand, benefits of combining manual and automated evaluation have their cost. The price in this case is the additional work that needs to be done to perform the integration on the SC level – finding mutual SC in the ATT and UTT evaluations and extra calculations of the unified score.

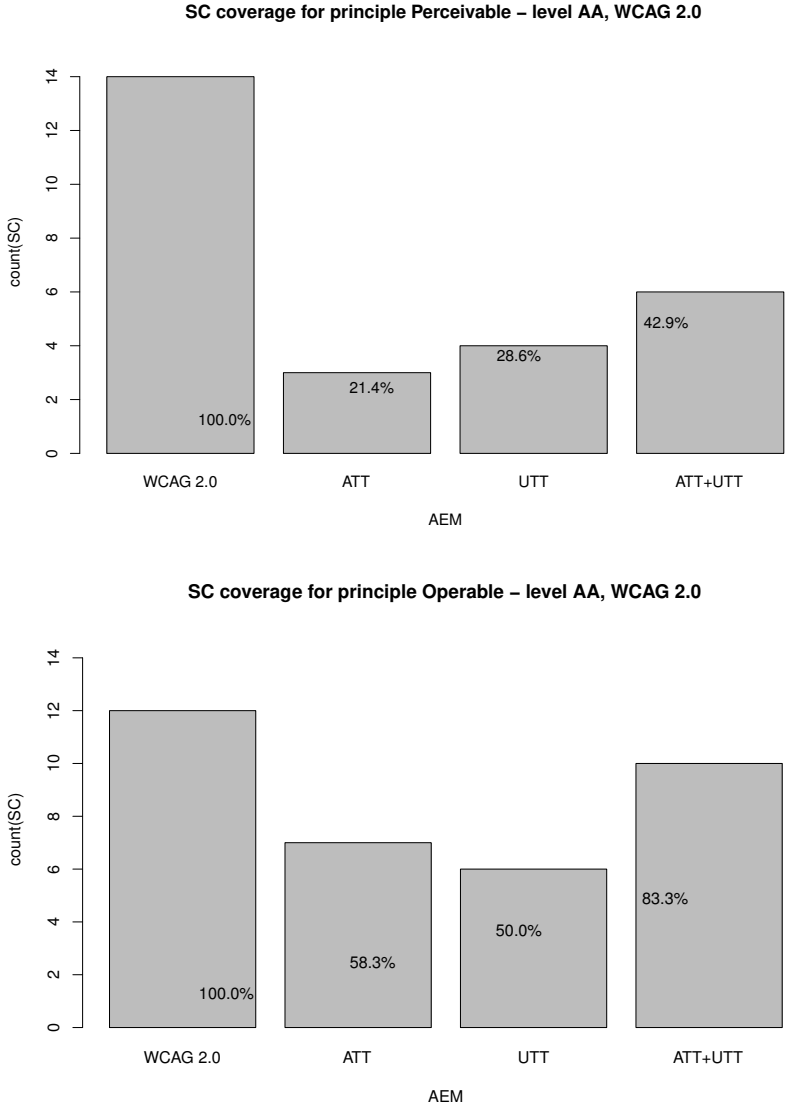


Figure 4.13
Coverage of WCAG 2.0 Perceivable and Operable principles with regard to investigated techniques: ATT, UTT, ATT+UTT using Union approach.

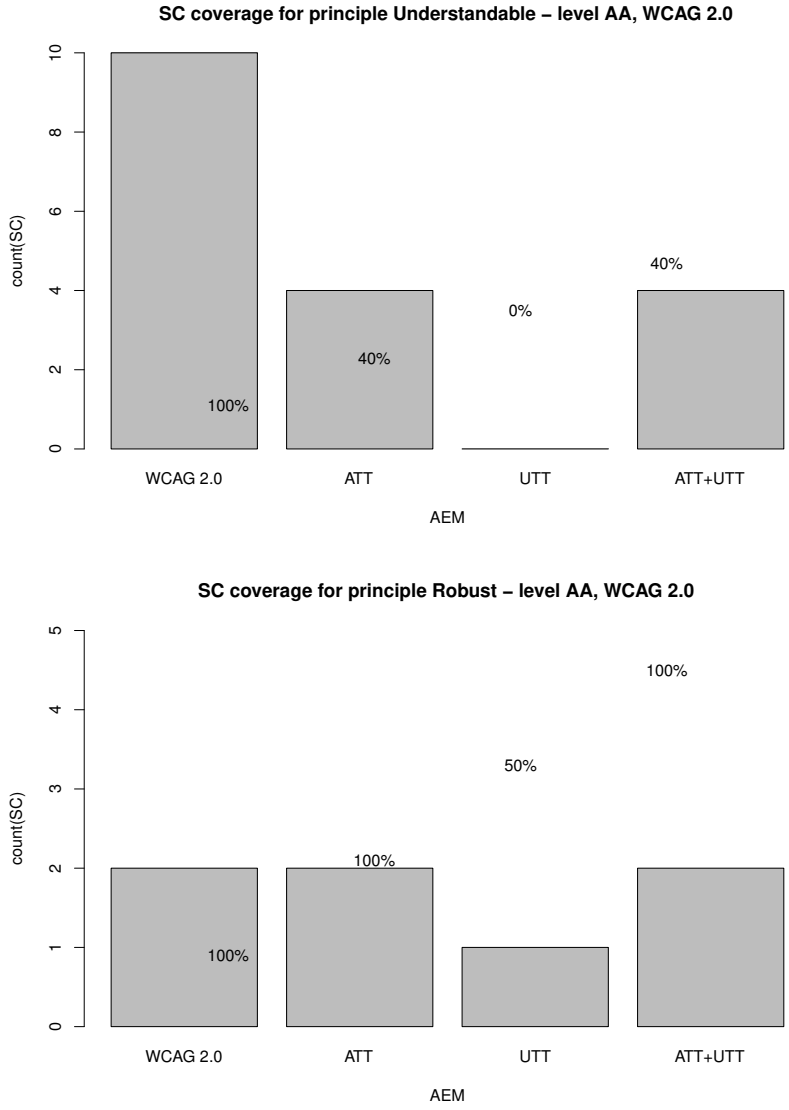


Figure 4.14
Coverage of WCAG 2.0 Understandable and Robust principles with regard to investigated techniques: ATT, UTT, ATT+UTT using Union approach.

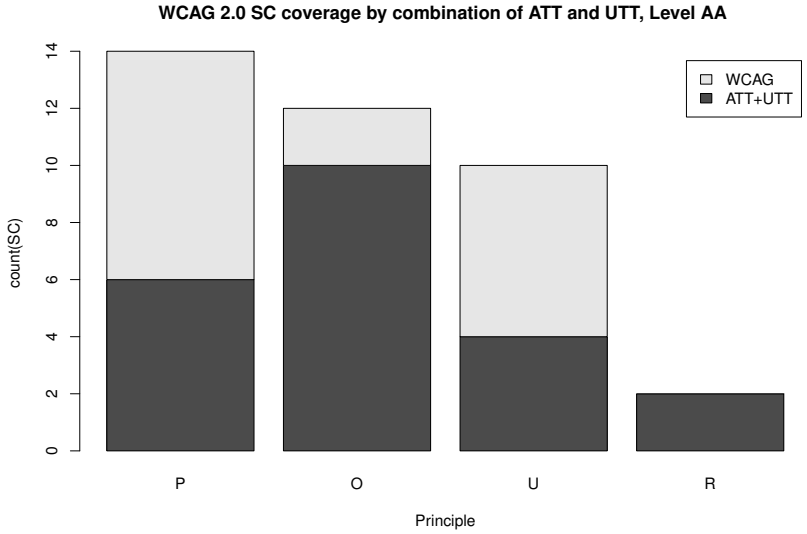


Figure 4.15
Summary of the coverage that can be attained by combining ATT and UTT in relation to the distinct principles from WCAG 2.0 SC level AA.

4.4.2. Integration on Page Element Level

To investigate the possibility of integration of manual and automated testing on the level of page elements, one page from the evaluated set was selected for a closer inspection. In order to confirm experimentally that the testing tools targeted the same DOM elements on the page, a manual analysis of accessibility test results was carried out. The results were then compared with a special focus to check whether the same page elements were assessed and what kind of relationship was present between them. UTT test set contains questions of different type, namely general ones regarding the perception of the page, e.g. whether the website is well displayed on a small screen, as well as more specific questions, e.g. checking if the given *alt text* describes good enough the particular image. To be sure that one evaluates the same object in UTT and ATT, those objects need to be identified in the code.

Page `https://www.un.org/press/en`, denoted as D_7, was both checked by the ATT and the UTT. The page contains meetings coverage and press releases of then UN. The news feed is updated daily. It includes a few images and a slider. Following elements were inspected:

– Images

  Analysis showed that the UTT had not detected four images present on the page: three JPEG images, 60 x 60 pixels, and a page logo in PNG format of 196 x 46 pixels dimensions. However, those images were evaluated by the ATT. The remaining five images on the web page were tested by both tools.

– Title

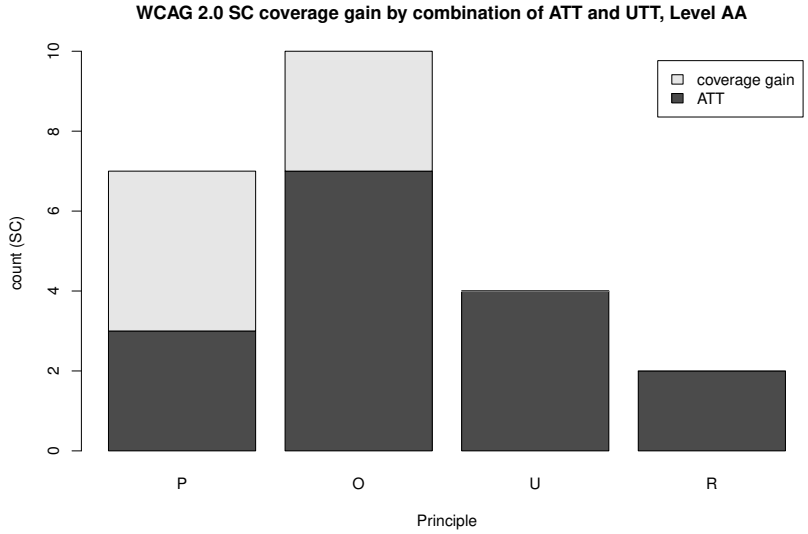


Figure 4.16
Illustration of the gain in coverage that can be attained by combining ATT and UTT in relation to the distinct principles from WCAG 2.0 SC level AA.

Manual inspection confirmed the consistency of the title evaluation. Both ATT and UTT evaluation resulted in "passed" outcome as depicted in Figure 4.17

– Headings

Results showed that manual testing was accordant with automated evaluation. Additionally, UTT detected three errors of fail ordering of the headings, which ATT did not cover.

– Buttons and links (Navigation and focus)

• Search box: submit button with accessibility barriers.

According to the UTT testing, there was reported a navigation problem with the search box. The box was assessed to be inaccessible. Similar accessibility problems were detected by the ATT under the violation of the H91 test ("Use HTML form controls and links"), which disclosed barriers with submit button. Further manual inspection of the code confirmed that the inaccessible button was located in the search box. Figures 4.18 and 4.19 depict the search box and a snippet of the extracted code that defines the search box. The snippet is supplied with the ATT outcome of the test performed on that particular page element.

• Highlights section – empty *alt* attribute.

The links lacked *alt text* attribute in clickable images from the "Highlights" section, which would be useful for assistive technologies to determine the destination link. The ATT did discover those shortcomings, but there was no feedback provided on that in the UTT results. Figure 4.20 enlights the mentioned section.

```
▸ test:                {…}
▾ 1:
  @type:               "Assertion"
▸ assertedBy:          "ATT:version/webpage-wam/…9e46ec899a7658b1ede3c85"
▾ result:
  ▾ 0:
    @type:             "TestResult"
    date:              "2018-04-02T09:58:14Z"
    description:       "http://checkers.eiii.eu/en/tests/"
  ▾ info:
      Warning:         "HumanInputRequired"
    outcome:           "passed"
  ▾ pointer:
      @type:           "charSnippet"
    ▾ char:            "<title>Meetings Coverage and Press Releases</title>"
    ▸ startPointer:    {…}
    reference:         "F25-pass1"
▾ test:
    @id:               "F25"
    @type:             "TestCriterion"
    isPartOf:          "wcag20:navigation-mechanisms-title"
    title:             null
▾ 2:
```


(a) ATT evaluation results in *pass* for the page title on page D_7.


```
▾ results:
  ▾ 0:
      module:          "title"
      type:            "all"
      status:          "pass"
      value:           "Meetings Coverage and Press Releases"
      info:            ""
      created_at:      "2018-04-02 11:59:16"
      info_id:         "utt::title:all"
  ▾ 1:
      module:          "title"
      type:            "feedback"
      status:          "pass"
      value:           "Meetings Coverage and Press Releases"
      info:            ""
      created_at:      "2018-04-02 11:59:16"
      info_id:         "utt::title:feedback"
  ▾ 2:
```


(b) UTT confirmed the accessibility of the page title on page D_7.

Figure 4.17

Evaluation results of the page title.

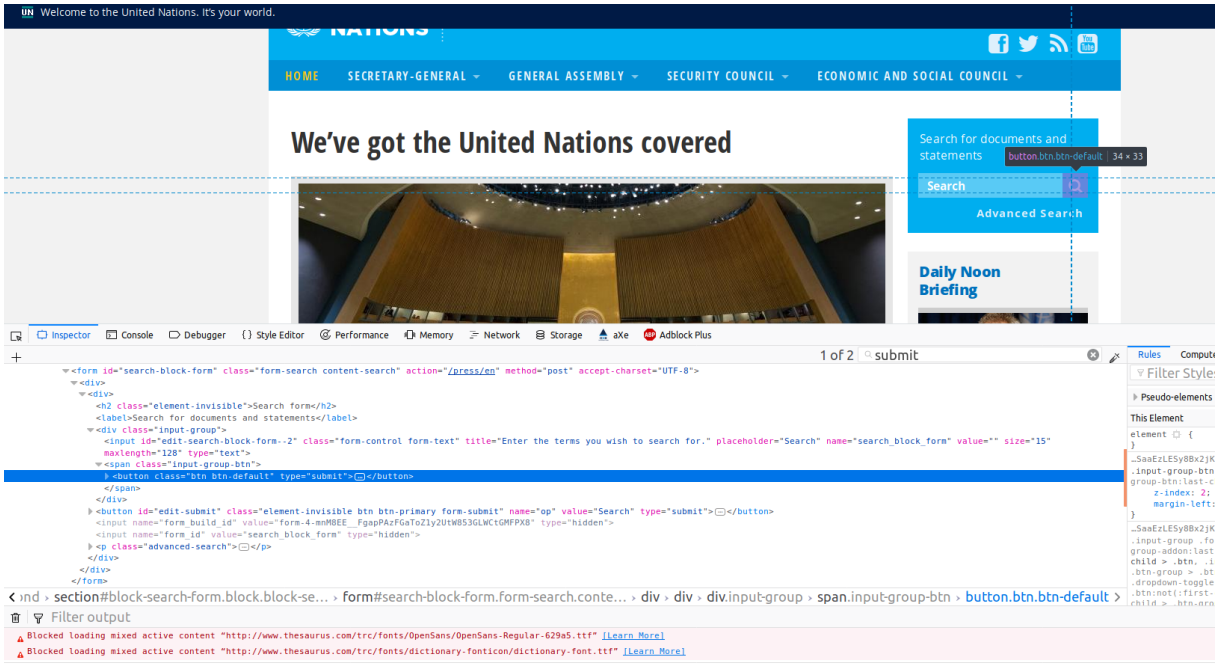


Figure 4.18

Inaccessible search box on page $D7$ identified in the HTML code.

Figure 4.19

ATT test outcome for the button on page D_7. ATT returned *fail* as the test outcome referencing the same page element as indicated in 4.18.

- "Follow Us" section – missing link names for images.

  Besides, both UTT and ATT were unanimous in evaluation of the "Follow Us" section, which contains four clickable icons, i.e. Facebook, Twitter, RSS and Youtube. The ATT results showed that the icons did not comply with the accessibility standards – the link names were empty for all images (failed test H91). For the UTT, the tester noted problems with focus. The icons were not designed properly to show that the pictures contained links. Figure 4.21 illustrates the lack of enough highlighting of the provided YouTube channel icon.

Figure 4.22 presents a comparison between the scores calculated for the page D_7:

Based on the obtained results, one can conclude that the Intersection score resembles the results from ATT and UTT for the mentioned elements. The Intersection score of 94.40 reflects

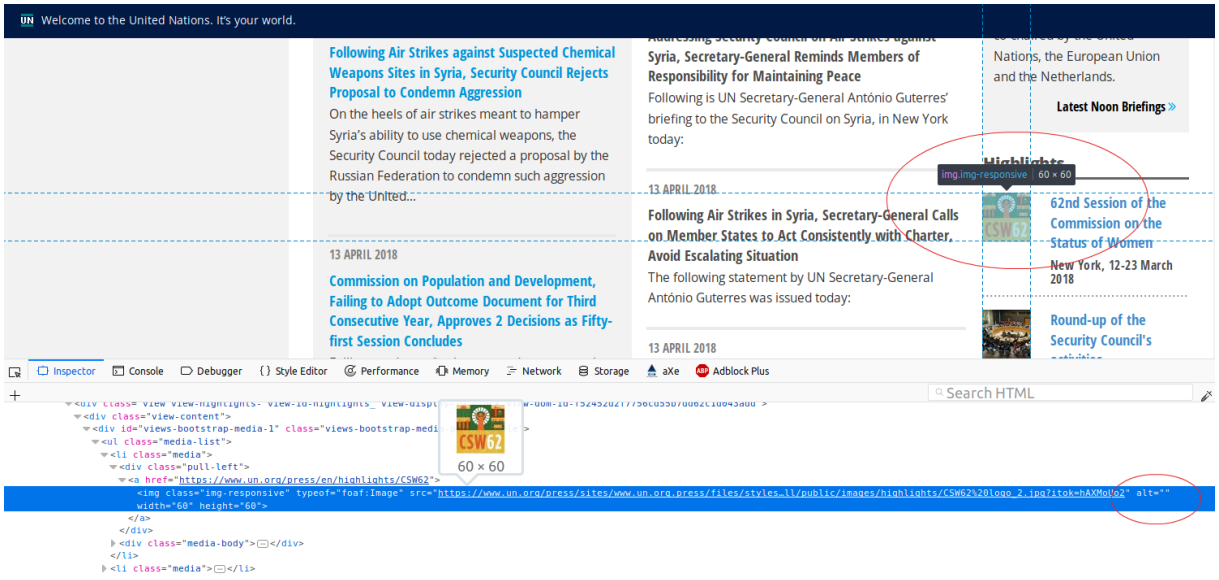


Figure 4.20
Missing alternative text on the clickable images in the "Highlights" section on page D_7.

the degree of consensus of the automated and manual evaluation for the common SC. The remaining six score points could be explained with the differences in image evaluation and icons without alternative text from the Highlight section. Those image icons could not be tested in the UTT since the icons had not been detected at all by the tool and served to the evaluator. The Union Score gives the most complete picture of the accessibility in numerical form, owing that it encompasses results from a broader spectrum of the accessibility guidelines. Equation 4.1 presents the formula used to calculate the number of reviewed SC for the Score Integration and the Union approaches. Score Integration embodies an average of manual and automated evaluation scores. Integration and Union scores were the closest to each other with a distance of roughly 2.5%.

Twelve SC were applied in the UTT evaluation, compared to eleven SC in the ATT checking. Four common criteria were found between the ATT and the UTT, which were used for the calculation of the Intersection score.

$$SC_{UTT \cup ATT} = SC_{UTT} + SC_{ATT} - SC_{UTT \cap ATT} \tag{4.1}$$

Applying formula 4.1 to the given example of the page D_7, following calculation can be done in terms of the SC coverage:

$$12 + 11 - 4 = 19$$

By integration of the ATT and the UTT evaluation, 19 SC were checked. Mutual effort resulted in an increased coverage of the tested SC, 58.33% in case of the UTT and 72% in case of the ATT, compared to the evaluation performed only with one tool.

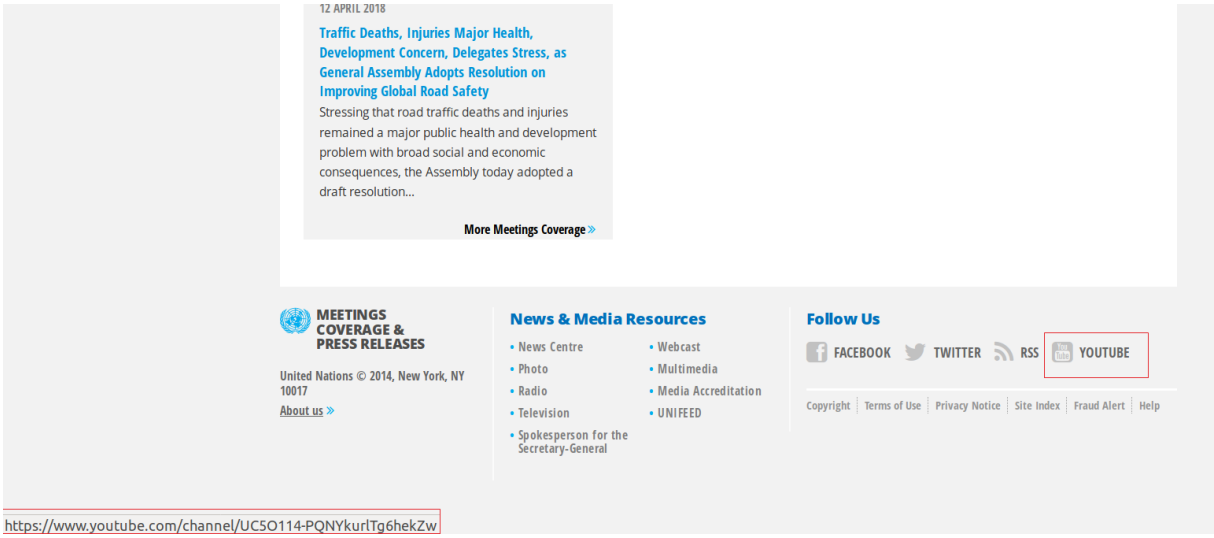


Figure 4.21
The link names were empty for all images presented in the "Follow Us" section on page D_7.

4.4.3. Impact of the dynamic content

As a side observation for the study, impact of the dynamic content on the integration approaches was analyzed. One of the reasons for creating two datasets – one with static pages and the second one with dynamic content, was to observe the behaviour of the scores. Recalling one of the challenges mentioned earlier, evaluation of the dynamic content is problematic due to the changing content depending on the user, the agent and the time. To combine the evaluation results from two separate AEMs, namely the UTT and the ATT, there has to be ensured that they had evaluated the same content. With the assumption that the number of checked pages by one tool equals the number of pages checked by the second tool and the pages were the same.

EDA revealed that quite many outliers were found in the relationship between the ATT and the UTT in case of the dataset D. That may suggest that the ATT results were not always aligned with the UTT results, which might be attributed to the differences in the evaluated content. Whereas for the dataset S, containing only static pages no outliers were found. What is more, there existed a stronger relationship between the ATT and the UTT scores.

The most detailed experiment, which touched the topic of the dynamic content, was the manual inspection of the web page and its evaluation results, described in the previous subsection 4.4.2. The meticulous examination showed that the same page elements were tested both by the ATT and the UTT. However, there were noted some differences in a sense that one tool detected some objects, which the other missed out. Yet, the accessibility assessment was performed for the same content as far as the conducted study managed to verify it.

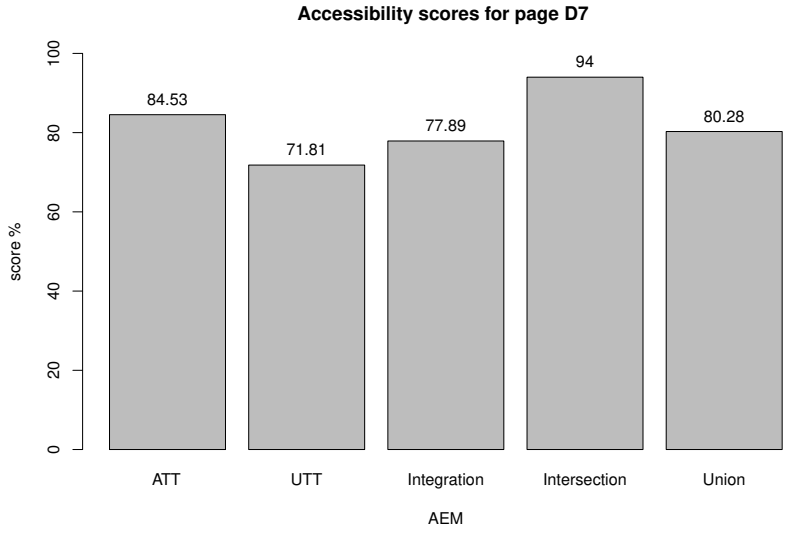


Figure 4.22
Accessibility scores for page D_7 computed with use of invesigated AEMs.

4.5. Score Function

Three score functions were created for the integration approaches experiments as described in 3.4.2. Calculations of the unified accessibility scores with use of the aforementioned approaches to integration, i.e., Scores Integration, Intersection and Union approaches, were performed for the given datasets of websites. The relationships between the outcomes from different approaches were studied. Figure 4.24 presents graphically outcomes of the investigation. For the sake of the reference, the plots contain also the ATT and UTT scores. The results showed that the Scores Integration score and the Union score were very close to each other, which was observed both for the dataset D and S. In 70% of the cases in the dataset D, slight differences were present as higher values of the Union scores than the Scores Integration scores. In the dataset S, 40% of the pages obtained higher Union score. As in case of the dataset D, Scores Integration and Union scores in the dataset S were very similar. This can probably be explained with the fact that, according to the formula 3.1 for calculation of the Scores Integration Score, the weights used for the mean are the numbers of SC tested for by the ATT and the UTT. No mutual SC are taken out from the calculation, which would mean that the results for the common SC are calculated twice. Taking into consideration the fact, that in general the number of mutual SC can vary between zero and five, the accessibility results of Scores Integration score may become higher when the page presents a high level of accessibility, and similarly it may be lower when the accessibility of the page is susceptible to improvement. Venn diagrams in Figure 4.23 illustrate the concept of mutual SC incorporation. From the set theory the mutual part, containing five SC, can be noticed.

When it comes to the values of the Intersection Score, they varied greatly among the pages

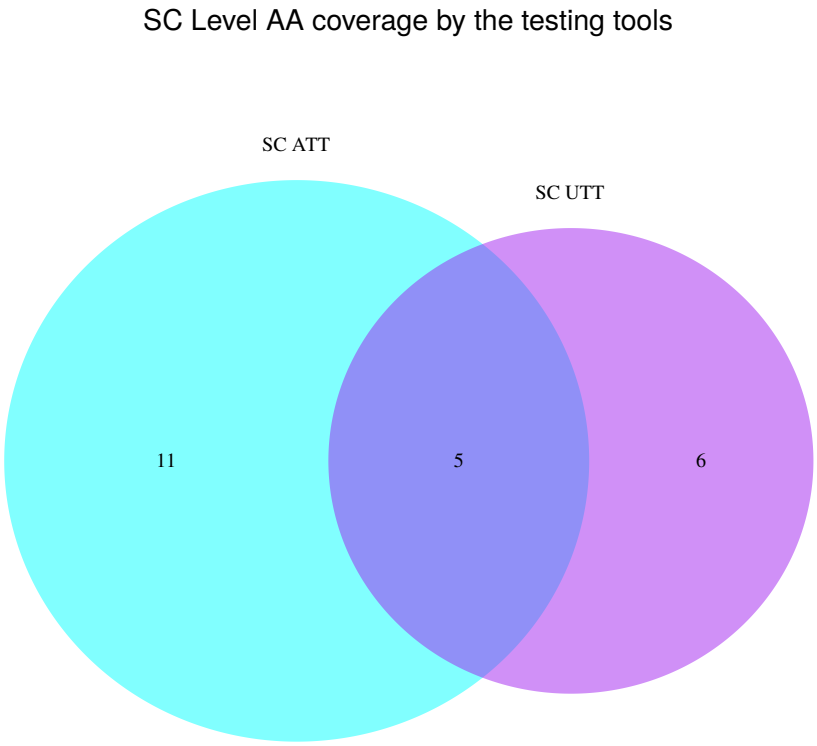


Figure 4.23
SC coverage by the testing tools used in the experiments. The common part of the diagrams indicates the mutual SC for the checker and the bookmarklet tool.

for both datasets, especially for the dataset D. Intersection Score demonstated itself not to reflect the accessibility of the evaluated page. Considering the example results of D_1 and D_7 page evaluation, there turns out to be a discrepancy between the high Integration Score and relatively low results from the ATT and the UTT. Even though the results for mutual SC were very high, oscilating around 100.00%, the applied tests for the remaining SC found many barriers, which resulted in lower scores for the remaining SC. Looking from the broader perspective and the aims defined for the study, the Scores Integration score did not add much value to the development of the combined score function for accessibility evaluation.

Taking into account the aforementioned double calculation of mutual SC in the Scores Integration score and the little information from Intersection score, the attention was turned to the the function for Union score for a closer investigation. Figure 4.25 demonstrates the relationship between the Union score and the scores produced by the UTT and the ATT for both datasets.

4.5.1. Metric's quality validation

In order to adopt a new metric, proposed by this paper, its validation needs to be performed first. The challenge with metric validation is that there is no ready-made solution which could be used for comparing the output produced by a candidate metric. It becomes difficult to produce a single value that characterizes the level of web accessibility of a page. Likewise, when it comes to the validation of manual testing results, subjectivity and diverse level of accessibility knowledge may influence the final output. For instance, certain users may not be aware of the accessibility problems they encounter during the testing.

For the validation of the Union Score, a research roadmap for web accessibility metrics developed by the WAI [9] was used. The framework for the assessment of the metric's quality became a base for the investigation. The Union Score was evaluated according to five metric's characteristics: *validity, reliability, sensitivity, adequacy and complexity*, where the first two are emphasized to be the most crucial to be satisfied. Following [45] with the distinction of the metric's quality assessment for various purposes of accessibility checking, the framework utilized in this study is oriented on delivery of a metric well-suited for benchmarking purposes.

Validity

For the benchmarking purposes, validity with respect to the conformance to the accessibility guidelines is desired. Validity is to respond the question of how well the score mirrors all and only the true barriers of the guideline's checkpoints. An estimation of false postitives and false negatives should be conducted additionaly.

As assessed, the Union Score is guideline dependent since the metric is based on the WCAG 2.0 and its SC. To apply the metric to the evaluation results obtained with use of another guideline set, e.g. Section 508, the score function would have to be adjusted. However, if limiting the guidelines set to the WCAG 2.0 conformance guidelines, the metric can be used both with a subset of the guidelines and the complete set of the SC. Also it could be used with the latest WCAG 2.1 that encompasses seventeen new SC in addition to WCAG 2.0. The genre of the website does not influence the validity of the metric. The study has showed that it can be applied both for static websites and websites with dynamic content. The metric performs an integration of the outcomes of more of one tool. It successfully implements the merging of the manual and automatic evaluation results.

The metric is tool-independent, in a sense that it can be used not only with the tools that it was created for, namely the WTKollen testing tools: the checker and the UTT bookmarklet. All of the checking tools that implement accessibility tests based on WCAG could be able to adopt this calculation method, provided that a link between the guidelines, ATT and UTT is established.

On the other hand, for the Union Score, validity with respect to the users is prone to the evaluator effect. It takes place in terms of manual evaluation as the outcomes are incorporated

in a qualitative form into the score function. What is more, the weight of the contribution of the UTT results are regulated in the same way as the contribution of the ATT results – by a number of SC for which the tests were applied. It was observed that the incomplete evaluation might have had an effect on the final accessibility score produced for the web page.

Reliability

Figure 4.25 presents graphically the relationship between the scores obtained from the ATT, UTT and the Union Score. A study of the differences between the accessibility scores produced by the ATT and the UTT revealed a tight relationship between them. Yet, it is not always desired since one tool may find the barriers that the other tool did not and thus the scores vary. However, as a whole, thanks to merging the outcomes from two tools, the Union Score is much more reliable than when applying data from a single tool only. The case with use another accessibility guideline, e.g. Section 508 has not been tackled.

Another aspect to consider in case of the Union Score's reliability is the soundness of the manual evaluation. The repeatability of the automatic evaluation results is taken for granted. The tool produces repeatable results each time when the same web pages/websites are evaluated. For the UTT results, the outcomes may slightly vary because of the human factor. The tool is designed with an idea that even a non-accessibility experts should be able to effectively use it and provide with a sound feedback. Still, what for one user causes problems may not be considered a barrier for another. Such situations may introduce a variance.

The question of sampling appears to be relevant for the study of the metric's reliability. For this research it has been assumed that the selected for evaluation web pages from the given websites were regarded as website templates. Following this idea, the evaluation of another set of web pages from that website is likely to return similar results. Nevertheless, considering the situation when page sampling is based on some stochastic methods – the final outcomes may be affected.

Sensitivity

Sensitivity of a metric is closely related to the topic of dynamic content where it can drastically change itself very quickly. From the benchmarking perspective, low-reliability is desired to assure that small changes in accessibility do not lead to big changes in the metrics. In other case, a huge variation may cause the rankings to run out of control.

An interesting research question to follow up becomes to examine closely the variability of the metric. In other words to compare the changes in the metric caused by small changes to the content. The changes in metric because of the changes in content should be smaller to state that it can be utilized to deliberately monitor the accessibility of a website.

Adequacy

It seems doable to use the metric for investigation of the accessibility for groups of disabilities.

A simple classification of SC according to the disability group it affects would make it possible to apply the metric for the Union Score calculation. However, this approach to web accessibility evaluation assumes that all of the barriers are treated equally. None of the barriers gets more impact on the score in the end.

Secondly, the Union Score metric is particularly designed for benchmarking purposes. Yet it can be used as well for QA within web engineering as it is sensitive enough when applied to the same web pages and at the same time precise to yield the accessibility barriers. Thanks to the code extracts provided by the ATT, the issues may be repaired more quickly.

Accessibility scores produced by the Union Score metric are given in a ratio. The values are normalized and range in $0.00 - 100.00$. Precision was taken into consideration to allow for comparisons between various websites and done over time. Metric's visualization and presentation issues are tackled in the next subsection 4.5.2.

Complexity

For the benchmarking purposes the metric should posses a low-internal complexity. Monitoring takes place on a regular basis and certain resource limitations should be taken into consideration. The Union Score metric is characterized by a low complexity since the only operations used are multiplication, addition and division. The complexity is greater than the complexity of the metric utilized in the ATT due to the additional part coming from manual evaluation and the common SC part. In spite of that, as the metric merges the results from more than one tool, the complexity on the metric ensure more valid and reliable results. Besides, lower complexity will allow more people to understand how it works.

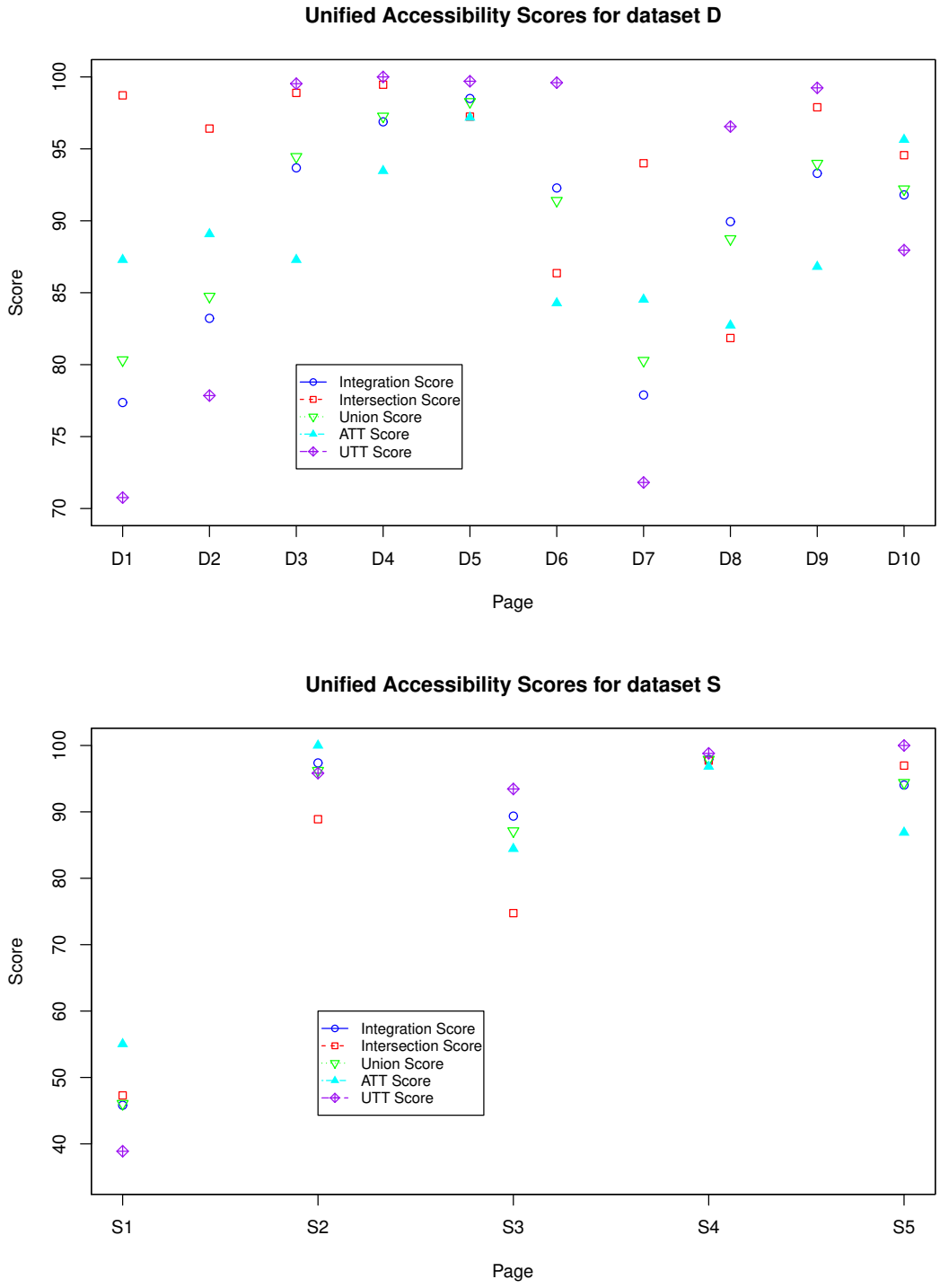
4.5.2. Score presentation

The way of presentating the score is crucial for the understanding of the accessibility level of a website. It is not uncommon to come across a website that scores 100% in accessibility level and find it not completely accessible. Testing tools usually present the outcomes of the evaluation using the scale $0 - 100\%$. However, that kind of information may be misleading for the users and convey a utopian message that the website is completely accessible, while in reality it has resulted to be perfectly accessible only for the subset of implemented tests and guidelines in the testing tool. Thus, the 100% accessibility means that it is accessible to a maximum grade that the tool can cover. Yet, it is not equal to 100% since the total coverage of the accessibility barriers may be attained only by a fully integrated approach, i.e., combining automated and manual evaluation and accessibility testing, as stated by [5]. There are also some tests that are hard to implement in the UTT. In case of the examined study, where the conformance to the

WCAG 2.0 is checked, assuming that the adopted evaluation approach covers 60% of the SC, should one present 60 or 100 if the tools have not found any barriers for a particular website?

One of the objectives is to propose a way that would convey in a clear way the degree of accessibility that a checked website shows. Therefore, it is suggested to present the results of the evaluation in form of a pie chart, with a distinction to the evaluation methods and their maximal coverage of the guidelines. Following Figure 4.26 indicates the idea of the results presentation in form of an Accessibility Pie Chart (APC). Figure 4.27 shows an application of the idea for the evaluated web page D_7. The pie chart shows the percentage of the checked SC that passed the tests next to the percentage of the checked SC that resulted in fail. The remaining part of the pie chart indicates the part of the WCAG 2.0 Level AA that was not tested.

There emereged an idea of visualizing the results with a distinction to the testing tools and their outcomes. Figure 4.28 presents the results for web page D_7 in an alternative way. The size of the 'not covered' slice has shrunk. It can be explained with the double calculation of the common SC for ATT and UTT. Because of that, this way of visualization is not flawless either.



(a) Accessibility scores for page D_7 computed with use of invesigated AEMs.

Figure 4.24

Accessibility scores for dataset D and S computed with use of the studied approaches to the integration of manual and automated accessibility evaluation.

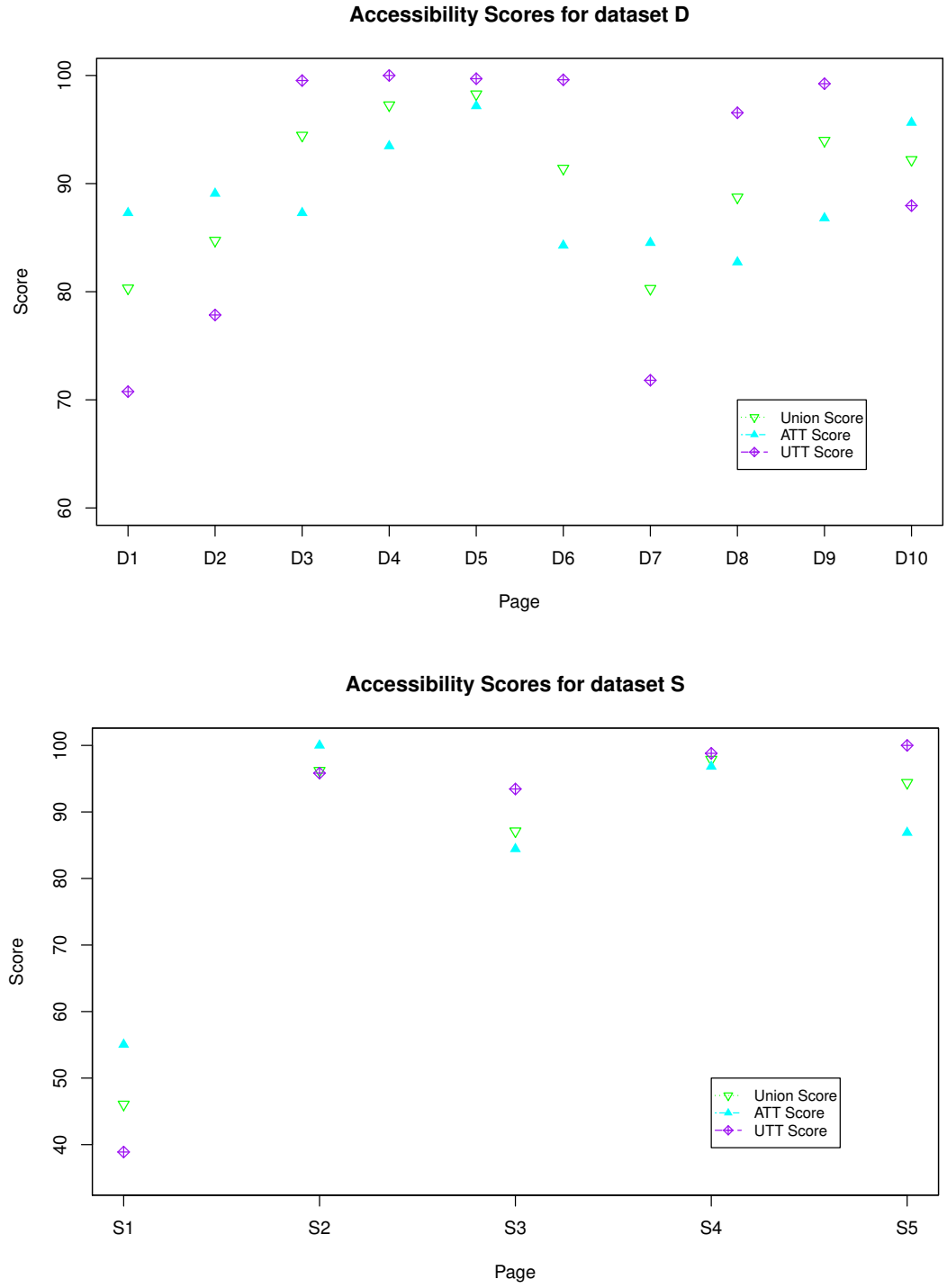


Figure 4.25

Relationship between Union score and ATT, UTT scores for dataset D and S.

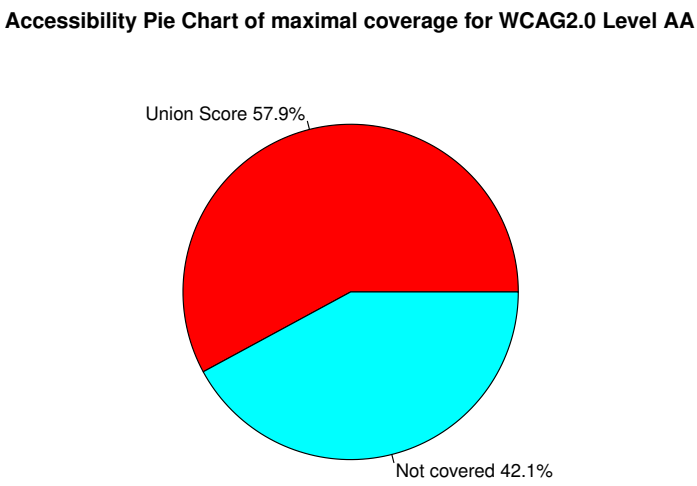


Figure 4.26
Accessibility Pie Chart as an idea for graphical presentation of the Union Score. At maximum, integrated tools can cover 57.9%, which means that 42.1% is left beyond evaluation. The percentages indicating coverage are valid for the studied WTKollen tools case.

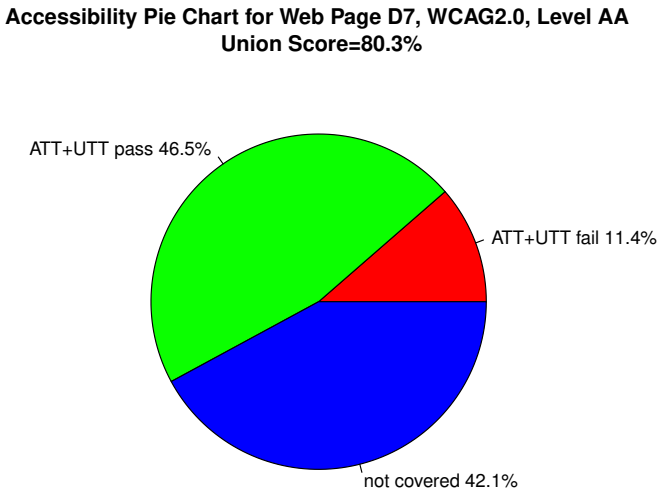


Figure 4.27
Accessibility Pie Chart for web page D_7, with Union Score equal to 80.3%.

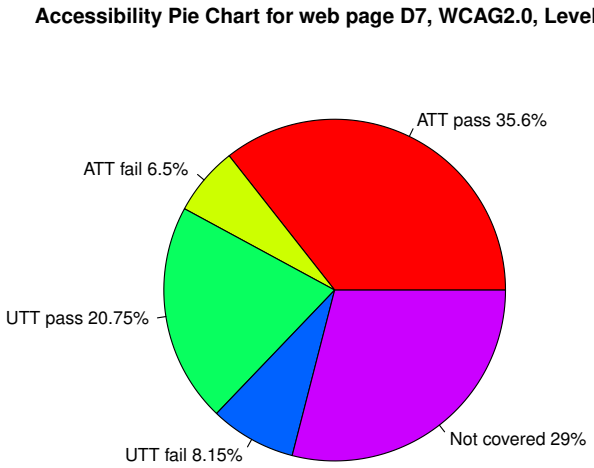


Figure 4.28
Accessibility Pie Chart alternative for web page D_7, with Union Score equal to 80.3%, ATT score 84.53% and UTT score 71.81%.

5. Discussion

> Today, I want to challenge us all to have greater ambitions for the web. I want the web to reflect our hopes and fulfill our dreams, rather than magnify our fears and deepen our divisions.
>
> Tim Berners-Lee, W3C Director and Founder in a speech for the 29th anniversary of World Wide Web

The research has set its course on a journey to explore possible ways of combining manual and automated evaluation of websites' accessibility. The goal has been set to attempt integration on the basis of implemented accessibility guideline to arrive at a single accessibility score, that would express to what extend a particular website is accessible for the users. Moreover, a graphical presentation of the novel metric was supposed to follow the quantitative results.

5.1. Major findings

The findings suggested that results of manual and automated web accessibility evaluation can be combined. The integration has been achieved through combining the test results on the basis of the SC from WCAG 2.0. Experiments have shown that the same page objects were extracted for accessibility evaluation both by the UTT and the ATT. Three approaches to integration have been discussed and one – the Union approach, experimentally selected as an appropriate method for the implementation. Thanks to the integration of manual and automated evaluation methods, the coverage of the SC from WCAG 2.0 Level AA has been increased to 57.9%, compared to 42.1% when using only ATT and 28.9% when basing the testing solely on the UTT. As an outcome of the research on integration approaches, a novel metric – the Union Score, has been proposed for the purpose of combined accessibility results quantification. The metric has been assessed with use of the framework for the metric's quality analysis [9]. The preliminary evaluation proved the Union Score to satisfy the requirements for benchmarking purposes. Additionally, a visualization method for presenting the outcomes of the integrated

results has been created. Accessibility Pie Chart has been introduced as a suggested way of conveying the results of the accessibility analysis graphically.

5.1.1. Accessibility Metric

Union Score Metric's preliminary validation has showed that the proposed metric for a combination of manual and automated accessibility evaluation results is well-suited for application in benchmarking context. Both static and dynamic web pages can be evaluated with use of the developed approach. However, the metric has turned out to be prone to the evaluator effect due to its manual part. Yet, with a properly conducted manual testing, done by a person knowledgeable in basic accessibility topics, it can deliver reliable and sound results. Moreover, the metric at the current state is guideline-dependent. It is compatible with WCAG 2.0 or its subset as well as it is supposed to work with WCAG 2.1. Yet, it is speculated that some adequate adjustments would probably make it possible to deploy the Union Score metric to other guidelines. When it comes to the sensitivity of the Union Score, which is closely related to the topic of dynamic content, more experiments are required to examine the variability.

Adequacy of the metric has been investigated. Research has shown that the metric can provide the users with a meaningful information about the website's accessibility, based on the two-step assessment combining automated testing and manual input. The calculated accessibility score is supplied with a graphical illustration of the results in form of the Accessibility Pie Chart, depicting the fractions of *passed, failed*, and *not covered* results. Due to the possible application for benchmarking purposes, is has been aimed to hold the complexity of the metric on a low-level to make it possible for large-scale monitoring on regular basis.

5.2. Importance of the study

Integration of manual and automated assessment is necessary for providing the users with a sound report about the web accessibility of the website. It has also been emphasized by the accessibility experts that contributed to the study through their involvement in the interviews. One tool is not enough to assess thoroughly the website for its conformance to the accessibility guidelines. The choice either/or between ATT and UTT will not suffice. It cannot provide a complete picture of the accessibility level of a website. It has been reported that approximately only 20% of all accessibility tests can be automated [7], whereas the time and cost of manual testing alone make it unsuitable for large scale evaluations. Both ATT and UTT have their weaknesses and strengths. Neither of them is perfect. Per contra, when combined, they complement each other in delivering an exhaustive accessibility audit. With the proposed way of accessibility score calculation, the results are comprised of the outcomes from two tools, thus become more reliable and able to deliver a more complete overview of the website's accessibility.

The new integration approach to web accessibility checking requires a new way of calculating the accessibility score. For the time being, no suitable score function has been found that could be adapted for the needs of the integration. Therefore, a novel score function called Union Score has been introduced. The presented metric builds on the approach developed for the EIII project and its efforts put into creating the integration methodology [12]. Another reason for a unified score is the fact that monitoring of the evolution of web accessibility demands quantitative metrics. Quantitative results make it possible to compare the results and observe the changes in the accessibility of the website. In this case, the proposed integration approach manages to deliver a metric that can be used for benchmarking purposes.

5.3. Research challenges

The study has tried to address the well-known challenges connected to combining manual and automated accessibility evaluation. Three main challenges have been identified at the beginning of the study, namely evaluation timing, dynamic content, and different testing coverage of the website.

Evaluation timing

In this research, it is proposed to mitigate the consequences of the different evaluation time by setting a time frame between the automated and manual testing. The exact amount of time needs more research due to the various time of the content's update. In the described experiment, manual evaluation of a web page done by the author was performed first and immediately followed by an automated check. Thanks to that, it has been possible to assure a nearly concurrent testing and limit the time gap between both evaluations. Other possible solution to consider in the future might be to run automatically the ATT check of the web page once the tester have started the manual testing with use of the UTT.

Dynamic content

The second challenge of integration – dynamic content, has been approached with a comparison study of the accessibility results done for the two sets of data: dataset D and dataset S. The first one consisted of static web pages and the second one of the dynamic web pages. Focus has been put on observation how the results of the accessibility evaluation relate to each other. It has been tried to capture the differences between the ATT and the UTT evaluation to relate them later to the Union Score. The analysis has revealed some slight variations between the results, however they have not been drastical. The aforementioned differences could possibly be explained with a diverse test suits in the ATT and the UTT. Moreover, in order to examine which elements had exactly been evaluated, a manual inspection of the targeted page elements has been performed for one web page. The analysis indicated that both tools had performed

the accessibility check on the same page objects. Notwithstanding the fact that the results have shown to be promising, more experiments should be performed on this topic as the sample has been too small to allow for generalization about the sensitivity.

Testing coverage

Integration of the results from the WTKollen checking tools on the level of one page seems to be straightforward. When it comes to the evaluation of the whole website, comprised of hundreds of web pages, the challenge of different coverage becomes relevant to address. This study has assumed that the web pages selected for analysis were the representants of the website. By representants it is meant the pages being identified as the website's templates. In reality, there is no choice other than to rely on automated tools when evaluating large-scale datasets for accessibility. Manual evaluation of the whole website turns out to be unworkable. For that reason, there is a need for an approach that would combine precise automated testing with a complementary user testing, resulting in a complete overview of the accessibility. The stage is open for involving the concepts of AI into the accessibility evaluation. Supposing that, only a relevant fraction of a large website could be evaluated in an efficient (and fast) way, and at the same time delivering an exhausive report of the accessibility level of the website without a need for trade-offs between robustness and completeness of the results. A recently conducted research by Mucha *et al.* [51] has showed that a cure for the problem with large-scale evaluation may lie in applying clustering techniques of machine learning to the checking process. By grouping similar web pages into clusters, a small sample of website templates can be obtained. The preliminary results have revealed that it is possible to reduce the number of web pages needed for evaluation by at least 70.6%. By virtue of significantly reducing the number of pages needed for evaluation, the set of identified templates, may become an input for combined website accessibility evaluation. In such a way, there disappears the problem of page coverage since the sampled pages may be thoroughly evaluated both by ATT and UTT.

Alternatively, a simpler way of selecting the pages to check with integration approach would be to first evaluate the website automatically and cluster the pages with regard to the automatic accessibility scores. Then, through sampling a representative page/pages from each cluster, a sample set could be obtained. Those selected web pages could be evaluated manually. One advantage with this approach is that both manual and automated checking takes place on the same web pages, assuring that the evaluation is conducted on the same dataset. However, in that case, there is created a time gap between the manual and automated checking, which for the frequently updated content may result in some noise. What is more, such a solution requires also carrying out the automated evaluation beforehand.

5.4. Similar studies

D4.1: Integration methodology by Nietzio, A., and Berker, F.

Motivation for further investigation of the possiblities for integration came from the Integration Methodology [12], delivered as one of the outcomes of the EIII project. The research has managed to depict the results on the same diagram, yet still with a distinction for manual and automated testing. The approach proposed in this study, summarizes the results of the accessibility analysis in form of an Accessibility Pie Chart (APC), which shows the results for passed, failed, and not tested. The main difference is that the APC presents the outcomes after the integration. For instance, the percentage of tests that resulted in *pass* relates to both ATT and UTT. Distinguishing the results on the evaluation source may have its pros and cons if one is interested in detailed information about the performance measured by different evaluation methods. When it comes to the quantitative way of expressing accessibility of a website, the proposed by Nietzio and Berker methodology misses a score function that could deliver a metric producing a single accessibility score for integration.

SAMBA approach by Brajnik and Lomuscio.

The study has also drawn inspiration for the research from the work done by Brajnik, G. and Lomuscio, R., which developed the SAMBA method for measuring barriers of accessibility [50]. Similar research question has been posed in both studies: How can a metric merge results produced by accessibility evaluation tools and by human reviews? As a solution, the authors proposed a methodology and associated metric for measuring accessibility that combines expert reviews with automatic evaluation of web pages. Compared to the Union Score, suggested in this project, the metric is focused more on accessibility that goes beyond the conformance testing. Whereas the Union Score is mostly aligned with satisfying the requirements for benchmarking purposes. What is more, SAMBA methodology takes into account different user groups and makes a distinction of the accessibility barriers when assessing their severity and impact on particular groups of users. Similarly, the unified approach requires input from the users in the UTT part and some entry level of knowledge of the web accessibility concepts is needed. However, the degree of involvement needed to complete the evaluation in case of the SAMBA methodology, may be an obstacle for an average tester. On the other hand, the Union Score, even though still dependent on human judgment, appears to be more manageable to use as it does not encompass severity considerations. Nonetheless, it is also acknowledged that additional information about the barriers' severity may be beneficial to some users. Yet, a trade-off between the efficiency of evaluation, together with costs involved in hiring the accessibility experts, and the extent of the details has to be made. When it comes to the graphical presentation of the results, it would be interesting to see how the results obtained with use of the SAMBA methodology could be presented visually.

5.5. Alternative explanations of the findings

Different scores UTT-ATT-Union score

The study has shown that combining manual and automated accessibility evaluation methods and produced by them results, is doable. For some pages assessed with the unified approach, manual and automated accessibility scores varied, which might result in a different Union Scores. The fact that the Union Score does not mirror the score obtained as a result of automated evaluation cannot state that it is not valid. Since a different thing is measured and with use of different methods. Certainly, there is some substantiated correlation between manual and automated evaluation results, as showed for example by Martínez *et al.* [76]. However, that should not forejudge the new approach. Comparing the scores from different tools would not bring much advantage since they have a different coverage of guidelines in terms of accuracy and quantity.

A considerate spread of the scores in the dataset S

Another thing to consider is the spread in the accessibility scores observed in the dataset S. The scores ranged from 55.04 to 100.00 from the automated evaluation and 38.89 to 100.00 in the manual evaluation. It is not expected to receive the same outcomes, however a less spread in the evaluation results was awaited since the pages belong to the same website. Such deviations in the dataset S may be explained with the fact that the web pages that comprised the dataset were artificially created for the accessibility testing purposes. Thus, they do not represent a real-life website. Many accessibility barriers were put on the web pages on purpose. That may explain such a variation in the results.

Integration approaches

In the course of the study, three approaches to combining the manual and automated web accessibility evaluation have been investigated, i.e., Scores Integration approach, Intersection approach, and Union approach. The study aimed at delivering a method that would be suitable and reliable to utilize for assessment. The results have revealed that the Intersection approach does not provide enough meaningful information since it calculates the accessibility score using only the evaluation results linked to the common for the UTT and the ATT SC. The maximum number of common SC is five, compared to twenty-two possible to cover by the both tools. The rest of the results is discarded in this approach if the results do not belong to the common SC. There are implemented too few mutual SC, that could be use as an independent source of information about the accessibility of a web page. Therefore, due to such a loss of data, this approach has been excluded from further analysis.

Further on, the remaining two aprroaches were investigated: the Scores Integration approach, which calculates the average of the manual and automated accessibility scores, and the Union approach, which bases the score calculation on the sum of the evaluation results from both tools.

The main difference is that the integration is done after the score calculation in the first case and in case of the Union approach the integration is performed on the raw results from the tools and followed by score calculation. The quantitative study has showed that the experimental results of the score calculations with use of these two approaches were very close to each other. Moreover, the calculated correlation between the Scores Integration and the Union Score values, indicated a very strong, positive relationship between those variables. For the dataset D the Pearson correlation p-value= 0.99 and for the dataset S the p-value= 1.00. The approaches have produced nearly the same output. With this information, there has arised a question whether it would be possible to allow for a simplification of the way in which the accessibility score is calculated. The Union approach to integration requires identifying the mutual results in the output from the ATT and the UTT, which is an extra work to be done. Whereas the Scores Integration approach proposes calculating the average mean of two numbers – the score produced by the ATT and the score produced by the UTT, both computed with the same score formula. However, a closer investigation in a search of the clarification of the similar results and the reasons unveiled that the difference between these two approaches lies in the way the methods approach the topic of the mutual for the ATT and the UTT SC. While the Union approach applies the logical formula for calculation of the sum of two sets, and thus counts the common part of the sets only once, the Scores Integration approach counts the common SC twice. First time as a part of the ATT results and second time as a part of the UTT results. So small differences in the values of the scores obtained with use of those two approaches to integration can be explained with the very small number of common SC, that had not so much impact on the final accessibility score. Yet, from the methodological point of view, implementation of the Scores Integration approach would be incorrect. Also, it has been assumed that all barriers are equal and none of them will be prioritized. Therefore, the whole attention has been finally turned to the Union approach to integration and the experiments proceeded with the results obtained using only this particular approach.

5.6. Limitations to the study

The research has shown an ample potential in combination of web accessibility evaluation results from manual and automated checking. However, due to the exploratory nature of the study, there can be found certain limitations to the proposed solution.

Sampling

First of all, in order to simplify the case, there has been made an assumption that the evaluated set of web pages consisted of already selected web pages from the website. In that way the effort connected to the sampling has been diminished. The sampling has been left beyond the scope of the project. In case of the evaluation 1:1 – each page evaluated both manually and automatically,

the integration is straightforward. However, when the checker evaluates let us say 600 pages and $10 - 20$ pages are further checked manually, the method of sampling may be deciding on the accessibility score. Further research on the topic of relationships between the sampling and the integration approach would be beneficial to the knowledge.

Dataset size

Secondly, a small number of the evaluated web pages may be considered as a limitation of the findings. Nonetheless even such a scale experiment provides encouraging results for a more wide spread investigation on the topic and can possibly contribute to enhanced accessibility testing practices. Nevertheless, testing the approach with more websites and the automatization of the score calculations are crucial for the approach to be largely implemented. What is more, due to the small size of the dataset, the impact of the dynamic content on the integration results cound not been deeply studied. The manual inspection of the evaluation results, performed for the web page D_7 has showed that both tools targeted the same page elements. Yet, examination of two websites is not sufficient to state clearly to what extent the differences in evaluation done with help of the UTT and the ATT are caused by the presence of the dynamic content or the subjectivity of the tester.

Metric

The novel metric – Union Score has appeared to be a suitable candidate for a quatitative metric for integration of the web accessibility results. The conducted preliminary validation of the metric pointed out both the advantages and the disadvantages of the metric. The solution has been tailor-made to the needs for the particular accessibility evaluation purpose, namely benchmarking. Also, the evaluation has been done with the given tools: WTKollen checker and the UTT bokmarklet. However, for the time being, it can be stated that the metric is tool-independent, but guideline-dependent. The concepts of the metric can be used with any tool that implements WCAG guidelines and accompanied with an appropriate mapping test-SC. What is more, the current investigation does not provide the user with the information about the false positives and false negatives, which would be useful for examining the corectness. Moreover, the variability of the metric has not been explored to the end. The changes in metric's behaviour caused by small changes to the content and sampling of the pages would provide valuable information.

It has been observed that the quality of the human imput has a considerable impact on the outcomes produced by the metric. Even though the UTT bookmarklet tool is designed to be utilized by even non-experts within the accessibility field, some level of accessibility knowledge is strongly advised. As the study has shown, quality of the answers provided by the testers does matter. Likewise the completeness of the responses. The metric may be prone to the evaluator effect, a term coined for the usability assessment, which comes from a situation where the

evaluators in similar conditions identify substantially different observations. This leads to the accessibility problems and may result in changes to the accessibility scores.

Visual presentation of the results

Although the APC succeeds in presenting graphically the outcomes of the evaluation, some design ideas could be revisited. One possible aspect to consider would be to look closer at the part of the pie chart that indicates the percentage of the WCAG 2.0 that has not been addressed by the testing tools. At the moment, the 'not covered' slice is of a fixed size, i.e., 42.1% indicating the number of SC that the tools do not implement. However, in a situation when a website containing a sheer, plain text, that is very simple but fully accessible, gets evaluated, a diagram showing the 'not tested' part may not convey thoroughly its accessibility. It may even present it as a lower than it is in reality. The web page may obtain a lower accessibility score due to the fact that some SC may not have been checked even though they were not applicable for that particular web page and automatically limited the maximal accessibility score that a web page can achieve to 57.9%. Therefore it should be exchanged with a more responsive diagram, where, the 'not tested' part would indicate the percentage of the applicable SC that have not been checked. In that case the 'not tested' part would vary from page to page. The question however would be: how to calculate the number of applicable SC or tests of the website regardless of testing limits of the checking tools?

6. Conclusion

The research confirmed the hypothesis that outcomes from manual and automated web accessibility evaluation can be combined on the grounds of implemented guideline. The accessibility of the website can be summarized using a single, quantitative accessibility score. The study has shown that there is no need to choose either manual or automated method of accessibility testing. Both methods can be deployed in a synergistic way and complement each other. Thanks to the integration of manual and automated evaluation, WTKollen testing tools can cover up to 57.9% SC from WCAG 2.0 on Level AA. Previously, the maximal coverage for testing using only one method (either ATT or UTT) was equal to 42.1% and 28.9%, respectively.

The aim of this research was to create a method that would help to uncover existing barriers on the websites. The metric was supposed to comply with the quality framework developed by W3C [9]. The novel score function has proven to be applicable for larger implementation. Additionally, different ways of graphical presentation of the integrated results of the accessibility testing were explored. Finally, an Accessibility Pie Chart combined with a quantitative score – the Union Score, were proposed as a way of expressing the outcomes of the website's accessibility checking. Both graphical and numerical methods have been designed in a way to convey the report results. Moreover, impact of the dynamic content on the integration results of the evaluation was observed through the experiments and discussed with accessibility experts during the interviews. The study has shown that the challenge of the evaluation of dynamic content is complex and requires more attention in this context. A separate study is encouraged as a follow up action.

With the proposed, broader approach to accessibility evaluation, the process of accessibility assessment can be taken a step further and contribute in the long run to a more accessible Web for all.

6.1. Future Work

The outcomes of this thesis point out a few interesting directions to follow in the future. Although the study may answer important questions, other questions related to the subject remain unanswered due to the time and scope limitations.

First and foremost, the findings of this study can be utilized in implementation of the Union Score for the integrated approach. The study has laid theoretical foundations for the implementation and emphirically investigated the properties of the suggested approach. Implementation of the concepts in the WTKollen project would generate more data and provide essential feedback from the users, which could later be used for refinements in the next iteration.

More research is needed on the suggested metric for the integrated approach. It is proposed to consider a deeper evaluation of the metric from the mathematical perspective. A study of false positive and false negatives would be certainly beneficial to the approach. More testing is necessary to explore all of the properties of the metric.

One subject that remains to be explored is to investigate more deeply the impact of the dynamic content. As a research avenue it is suggested to revisit the work done so far and conduct an experiment focused solely on the role of the dynamic content on the accessibility scores. One possible experiment could be to perform an evaluation twicely on the same dataset containing dynamic pages. First the pages would be evaluated as there are and secondly the dynamic content should be swiched off to investigate how the accessibility scores vary and their impact on the integration. Another iteresting approach to follow for further research on dynamic content is the idea of analysis of web page's behaviours, mentioned by G. Brajnik.

The accessibility evaluation should also leave an open door for a feedback from the community. In such an approach of crowdsourcing, users can report errors and accessibility divides, providing another view of the site. Possibly the gap that remains unaddressed after the combination of the manual and automated accessibility assessment could be covered with the feedback from the website's users. An example of such a solution could be implemented in form of a button that would serve as a way to report the barriers to the developers/website owners.

A. Appendix A

Table A.1

Detailed information about the SC coverage (Level AA) by studied AEMs with consideration of particular Principles from WCAG 2.0.

Principle	Guideline	# SC Guideline	# SC ATT	# SC UTT	# SC Union Approach
Perceivable	G1.1	1	1	1	1
	G1.2	5	-	2	2
	G1.3	3	1	-	1
	G1.4	5	1	1	2
total P:		14	3	4	6
Operable	G2.1	2	1	2	2
	G2.2	2	2	-	2
	G2.3	1	-	-	0
	G2.4	7	4	4	6
total O:		12	7	6	10
Understandable	G3.1	2	2	-	2
	G3.2	4	1	-	1
	G3.3	4	1	-	1
total U:		10	4	0	4
Robust	G4.1	2	2	1	2
total R:		2	2	1	2
Sum all:		38	16	11	23

B. Appendix B

Table B.1

UTT id indicates question id. *Question text* presents the question that the user is asked. *Answer* lists out possible user responses. *Result* displays the outcome in terms of *pass/fail* values, whereas the Quantification column shows a binary value that was assigned to the outcome.

UTT id	Question text	Answer	Result	Quantification
q_1	Is the text on this page easy to understand?	yes	pass	1
		no	fail	0
q_2	Are there difficult words on this page?	yes	fail	0
		no	pass	1
q_3	Is the information on this page up to date?	yes	pass	1
		no	fail	0
q_4	Does the text have an introduction or summary part?	yes	pass	1
		no	fail	0
q_5	Does the heading describe the content and purpose of	yes	pass	1
	this web page?	no	fail	0
q_6	Is the website well displayed on a small screen?	yes	pass	1
		no	fail	0
q_7	Can you navigate this page using only the keyboard?	yes	pass	1
		no	fail	0
q_8	Is the focus of active elements visible?	yes	pass	1
		no	fail	0
q_9	If you resize the text, is it then necessary to scroll horizontally?	yes	fail	0
		no	pass	1
q_{10}	Does the video on this page have captions?	yes	pass	1
		no	fail	0
q_{11}	Does the video on this page have audio descriptions?	yes	pass	1
		no	fail	0
q_{12}	Does the alt text “” describe describe the image below sufficient?	yes	pass	1
		no	fail	0
q_{13}	Does the web page title describe the content	yes	pass	1
	and purpose of this web page?	no	fail	0

Bibliography

[1] "Frequency of internet use, 2016 (% of individuals aged 16 to 74)." http://ec.europa.eu/eurostat/statistics-explained/images/c/cb/Frequency_of_internet_use%2C_2016_%28%25_of_individuals_aged_16_to_74%29_YB17.png. Accessed: 2018-03-19.

[2] "World Report on Disability." http://www.who.int/disabilities/world_report/2011/report.pdf, 2011. Geneva.

[3] S. L. Henry, "Introduction to Web Accessibility." https://www.w3.org/WAI/intro/accessibility.php, 2015. Accessed: 2018-03-28.

[4] S. Abou-Zahra, *Web Accessibility Evaluation*, pp. 79–106. London: Springer London, 2008.

[5] T. Lang, "Comparing website accessibility evaluation methods and learnings from usability evaluation methods," *Peak Usability*, 2003.

[6] M. Vigo, J. Brown, and V. Conway, "Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests.," in *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, p. 1, ACM, 2013.

[7] K. Andreasson and D. Alarcon, "Web accessibility and the European Internet Inclusion Initiative," tech. rep., DAKA Advisory AB, February 2016.

[8] "Directive (EU) 2016/2102 of the European Parliamen and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies.," tech. rep., The European Parliament and the Council of the European Union, 2016. In Official Journal of the European Union: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.327.01.0001.01.ENG&toc=OJ:L:2016:327:TOC, Accessed: 2018-03-27.

[9] W. W. Research and D. W. G. (RDWG), "Research report on web accessibility metrics," in *W3C WAI Symposium on Website Accessibility Metrics* (M. Vigo, G. Brajnik, and J. O. C.

eds., eds.), W3C WAI Research and Development Working Group (RDWG) Notes, W3C Web Accessibility Initiative (WAI), first public working draft ed., August 2012.

[10] M. Goodwin, D. Susar, A. Nietzio, M. Snaprud, and C. S. Jensen, "Global web accessibility analysis of national government portals and ministry web sites," *Journal of Information Technology & Politics*, vol. 8, no. 1, pp. 41–67, 2011.

[11] M. Snaprud, K. Rasta, K. Andreasson, and A. Nietzio, "Benefits and challenges of combining automated and user testing to enhance e-accessibility – the european internet inclusion initiative," in *Computers Helping People with Special Needs* (K. Miesenberger, D. Fels, D. Archambault, P. Peňáz, and W. Zagler, eds.), (Cham), pp. 137–140, Springer International Publishing, 2014.

[12] A. Nietzio and F. Berker, "D4.1 Integration Methodology." European Internet Inclusion Initiative (EIII) Project Deliverable. (`http://eiii.eu/`), 11 2015.

[13] C. Bühler, H. Heck, O. Perlick, A. Nietzio, and N. Ulltveit-Moe, "Interpreting results from large scale automatic evaluation of web accessibility," in *Computers Helping People with Special Needs* (K. Miesenberger, J. Klaus, W. L. Zagler, and A. I. Karshmer, eds.), (Berlin, Heidelberg), pp. 184–191, Springer Berlin Heidelberg, 2006.

[14] M. Goodwin and M. Snaprud, "eGovMon - a user driven project for benchmarking accessibility, transparency, effiency and impact.." `http://www.mortengoodwin.net/publicationfiles/NOKIOS2008.pdf.1`, October 2008. Norsk Konferanse for IKT i offentlig sektor 2008.

[15] A. Nietzio, C. Strobbe, and E. Velleman, "The unified Web evaluation methodology (UWEM) 1.2 for WCAG 1.0," in *International Conference on Computers for Handicapped Persons*, pp. 394–401, Springer, 2008.

[16] E. Velleman and S. Abou-Zahra, "Website accessibility conformance evaluation methodology (WCAG-EM) 1.0," *W3C Working Group Note. http://www.w3.org/TR/WCAG-EM*, 2014.

[17] "Ergonomics of human-system interaction – Part 171: Guidance on software accessibility.," iso 9241-171:2008 Standard, International Organization for Standardization, Geneva, July 2008.

[18] "Information technology – W3C Web Content Accessibility Guidelines (WCAG) 2.0," ISO 40500:2012 Standard, International Organization for Standardization, Geneva, October 2012.

[19] Mueller, Mary Jo and Jolly, Robert and Eggert, Eric, "Web Accessibility Laws and Policies." `https://www.w3.org/WAI/Policy/`, 2018. Accessed: 2018-03-27.

[20] Bundesministerium der Justiz und für Verbraucherschutz, "Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung - BITV 2.0)." https://www.gesetze-im-internet.de/bitv_2_0/BJNR184300011.html, 2011. Accessed: 2018-03-21.

[21] Le portail de la modernisation de l'Etat, "Référentiel Général d'Accessibilité pour les Administrations (RGAA) Version 3)." https://references.modernisation.gouv.fr/rgaa-accessibilite/, 2017. Accessed: 2018-03-21.

[22] Kommunal- og moderniseringsdepartementet, "https://lovdata.no/dokument/sf/forskrift/2013-06-21-732." https://lovdata.no/dokument/SF/forskrift/2013-06-21-732, 2013. Accessed: 2018-03-22.

[23] Asociación Española de Normalización y Certificación, "Norma UNE 139803:2012: Requisitos de accesibilidad para contenidos en la Web." http://http://www.aenor.es/aenor/normas/normas/fichanorma.asp?tipo=N&codigo=N0049614#.WrO1K9bA9hE, 2012. Accessed: 2018-03-22.

[24] Swedish Government, "Discrimination Act (2008:567)." http://www.government.se/information-material/2015/09/discrimination-act-2008567/, 2008. Accessed: 2018-03-22.

[25] General Services Administration (GSA), "Section 508 of the US Rehabilitation Act of 1973, as amended." https://www.section508.gov/section-508-of-the-rehabilitation-act, 1998. Accessed: 2018-03-22.

[26] US Access Board, "Americans with Disabilities Act of 1990." https://www.ada.gov/pubs/adastatute08.htm, 2009. Accessed: 2018-03-22.

[27] US Access Board, "Equality and Human Rights Commission." http://www.legislation.gov.uk/ukpga/2010/15/contents, 2010. Accessed: 2018-03-22.

[28] Commission of the European Communities, "Mandate 376: Towards an accessible information society." http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52008DC0804, 2008. Accessed: 2018-03-22.

[29] G. Brajnik, "Web accessibility testing: when the method is the culprit," in *International Conference on Computers for Handicapped Persons*, pp. 156–163, Springer, 2006.

[30] G. Brajnik, "Ranking websites through prioritized web accessibility barriers," in *Technology and Persons with Disabilities Conference, Los Angeles*, Citeseer, 2007.

[31] G. Brajnik, "A comparative test of web accessibility evaluation methods," in *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pp. 113–120, ACM, 2008.

[32] A. P. Freire, C. Power, H. Petrie, E. H. Tanaka, H. V. Rocha, and R. P. Fortes, "Web accessibility metrics: Effects of different computational approaches," in *International Conference on Universal Access in Human-Computer Interaction*, pp. 664–673, Springer, 2009.

[33] A. P. Freire, R. P. Fortes, M. A. Turine, and D. Paiva, "An evaluation of web accessibility metrics based on their attributes," in *Proceedings of the 26th annual ACM international conference on Design of communication*, pp. 73–80, ACM, 2008.

[34] I. S. Baazeem and H. S. Al-Khalifa, "Advancements in web accessibility evaluation methods: how far are we?," in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, p. 90, ACM, 2015.

[35] S. Laurin, A. Cederbom, J. Martinez-Usero, L. Kubitschke, A. Moledo, B. Simons, and S. Abou-Zahra, "Monitoring methodologies for web accessibility in the European Union – Final report." `http://www.ec.europa.eu/newsroom/dae/document.cfm?doc_id=19274`, 2016.

[36] S. Harper and Y. Yesilada, *Web accessibility: a foundation for research*. Springer Science & Business Media, 2008.

[37] M. Rowan, P. Gregor, D. Sloan, and P. Booth, "Evaluating web resources for disability access," in *Proceedings of the Fourth International ACM Conference on Assistive Technologies*, Assets '00, (New York, NY, USA), pp. 80–84, ACM, 2000.

[38] W3C, "Wcag 2.0 Appendix B: Checklist (Non-Normative)." `https://www.w3.org/TR/2006/WD-WCAG20-20060427/appendixB.html`, 2006. Accessed: 2018-03-27.

[39] J. L. Fuertes, R. González, E. Gutiérrez, and L. Martínez, "Hera-FFX: a Firefox add-on for semi-automatic web accessibility evaluation," in *Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A)*, pp. 26–35, ACM, 2009.

[40] J. L. Fuertes, E. Gutiérrez, and L. Martínez, "Developing Hera-FFX for WCAG 2.0," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, p. 3, ACM, 2011.

[41] M. Naftali and O. Clúa, "Integration of Web Accessibility Metrics into a Semi-automatic Evaluation Process," in *W3C Online Symposium on Web Accessibility Metrics*, 2011.

[42] A. Nietzio, M. Eibegger, M. Goodwin, and M. Snaprud, "Towards a score function for WCAG 2.0 benchmarking," in *Proceedings of W3C Online Symposium on Website Accessibility Metrics*, 2011.

[43] M. Vigo, M. Arrue, G. Brajnik, R. Lomuscio, and J. Abascal, "Quantitative metrics for measuring web accessibility," in *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, pp. 99–107, ACM, 2007.

[44] M. Vigo, G. Brajnik, M. Arrue, and J. Abascal, "Tool independence for the web accessibility quantitative metric," *Disability and Rehabilitation: Assistive Technology*, vol. 4, no. 4, pp. 248–263, 2009.

[45] M. Vigo and G. Brajnik, "Automatic web accessibility metrics: Where we are and where we can go," *Interacting with Computers*, vol. 23, no. 2, pp. 137–155, 2011.

[46] S. Song, C. Wang, L. Li, Z. Yu, X. Lin, and J. Bu, "WAEM: A Web Accessibility Evaluation Metric Based on Partial User Experience Order," in *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, W4A '17, (New York, NY, USA), pp. 21:1–21:4, ACM, 2017.

[47] N. Fernandes, R. Lopes, and L. Carriço, "A template-aware web accessibility metric," in *W3C/WAI Research and Development Working Group (RDWG) Website Accessibility Metrics Symposium*, 2011.

[48] J. Bailey and E. Burd, "Towards more mature web maintenance practices for accessibility," in *Web Site Evolution, 2007. WSE 2007. 9th IEEE International Workshop on*, pp. 81–87, IEEE, 2007.

[49] J. Brewer, "Web accessibility highlights and trends," in *Proceedings of the 2004 international cross-disciplinary workshop on Web accessibility (W4A)*, pp. 51–55, ACM, 2004.

[50] G. Brajnik and R. Lomuscio, "SAMBA: a semi-automatic method for measuring barriers of accessibility," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pp. 43–50, ACM, 2007.

[51] J. Mucha, M. Snaprud, and A. Nietzio, "Web page clustering for more efficient website accessibility evaluations," in *International Conference on Computers Helping People with Special Needs*, pp. 259–266, Springer, 2016.

[52] H. Petrie, A. Savva, and C. Power, "Towards a unified definition of web accessibility," in *Proceedings of the 12th Web for all Conference*, p. 35, ACM, 2015.

[53] R. Rutter, P. H. Lauke, C. Waddell, J. Thatcher, S. L. Henry, B. Lawson, A. Kirkpatrick, C. Heilmann, M. R. Burks, B. Regan, *et al.*, *Web accessibility: Web standards and regulatory compliance*. Apress, 2007.

[54] Y. Yesilada, G. Brajnik, M. Vigo, and S. Harper, "Exploring perceptions of web accessibility: a survey approach," *Behaviour & Information Technology*, vol. 34, no. 2, pp. 119–134, 2015.

[55] "Introduction to Web Accessibility. Web Accessibility Initiative (WAI), W3C." https://www.w3.org/WAI/fundamentals/accessibility-intro/#what. Accessed: 2018-05-27.

[56] G. Brajnik, "Beyond conformance: the role of accessibility evaluation methods," in *International Conference on Web Information Systems Engineering*, pp. 63–80, Springer, 2008.

[57] S. Luján-Mora and F. Masri, "Evaluation of web accessibility: A combined method," in *Information Systems Research and Exploring Social Artifacts: Approaches and Methodologies*, pp. 314–331, IGI Global, 2013.

[58] W3C, "Web Content Accessibility Guidelines (WCAG) 2.0, World Wide Web Consortium Recommendation 11 December 2008." https://www.w3.org/TR/WCAG20/, 2008.

[59] "Introduction to Understanding WCAG 2.0." https://www.w3.org/TR/UNDERSTANDING-WCAG20/. Accessed: 2018-05-27.

[60] G. Sreedhar, "Identifying and evaluating web metrics for assuring the quality of web designing," *Design Solutions for Improving Website Quality and Effectiveness*, pp. 1–23, 2016.

[61] W3C, "WCAG Techniques for dynamic content." https://www.w3.org/WAI/GL/wiki/WCAG_Techniques_for_dynamic_content, 2014. Accessed: 2018-03-21.

[62] N. Fernandes, R. Lopes, and L. Carriço, "On web accessibility evaluation environments," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, p. 4, ACM, 2011.

[63] M. Denscombe, *The Good Research Guide: For Small-Scale Social Research Projects: for small-scale social research projects.* Open UP study skills, McGraw-Hill Education, 2010.

[64] J. Creswell and V. Clark, *Designing and Conducting Mixed Methods Research.* SAGE Publications, 2007.

[65] J. Schoonenboom and R. B. Johnson, "How to construct a mixed methods research design-wie man ein mixed methods-forschungs-design konstruiert," *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, vol. 69, no. 2, pp. 107–131, 2017.

[66] A. Tashakkori and C. Teddlie, *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Applied Social Research Methods, SAGE Publications, 1998.

[67] R. B. Johnson and A. J. Onwuegbuzie, "Mixed methods research: A research paradigm whose time has come," *Educational Researcher*, vol. 33, no. 7, pp. 14–26, 2004.

[68] T. D. Jick, "Mixing qualitative and quantitative methods: Triangulation in action," *Administrative Science Quarterly*, vol. 24, no. 4, pp. 602–611, 1979.

[69] N. Denzin, *The Research Act: A Theoretical Introduction to Sociological Methods*. Methodological perspectives, Aldine Publishing Company, 1973.

[70] A. O'Cathain, E. Murphy, and J. Nicholl, "Three techniques for integrating data in mixed methods studies," *BMJ*, vol. 341, 2010.

[71] U. Flick, *Introducing Research Methodology: A Beginner's Guide to Doing a Research Project*. SAGE Publications, 2011.

[72] "WTKollen Page Checker - Find Barriers in your Web Page." `http://checkers.wtkollen.se/`, 2017. Accessed: 2018-03-15.

[73] "WTKollen User Testing Tool." `https://www.accessiblecheck.com/`, 2018. Accessed: 2018-03-15.

[74] Post-och Telestyrelsen (PTS), "Webbriktlinjer - Vägledning för webbutveckling." `https://webbriktlinjer.se/riktlinjer/`, 2015. Accessed: 2018-03-19.

[75] D. T. Larose and C. D. Larose, *Exploratory Data Analysis*, pp. 51–90. John Wiley Sons, Inc., 2014.

[76] C. C. Martínez, L. Martínez-Normand, and M. G. Olsen, "Is it possible to predict the manual web accessibility result using the automatic result?," in *International Conference on Universal Access in Human-Computer Interaction*, pp. 645–653, Springer, 2009.