This article is a preprint. It has been submitted to Frontiers in Computer Science for review.

Monitoring Systems for Checking Websites on Accessibility

Andreas Burkard[1, *], Gottfried Zimmermann[1], Bettina Schwarzer[1]

[1]Competence Center for Digital Accessibility, Hochschule der Medien, Stuttgart, Germany

*** Correspondence:**
Andreas Burkard
andreas_burkard@gmx.de

Keywords: accessibility (for disabled), monitoring, gamification, User Experience (UX) evaluation, user testing and evaluation, comparative study.

Abstract

Web accessibility monitoring systems support users in checking entire websites for accessibility issues. Although these tools can only check the compliance with some of the many success criteria of the Web Content Accessibility Guidelines, they can assist quality assurance personnel, web administrators and web authors to discover hotspots of barriers and overlooked accessibility issues in a continuous manner. These tools should be effective in identifying accessibility issues. Furthermore, they should motivate users, as this promotes employee productivity and increases interest in accessibility in general. In a comparative study, we applied four commercial monitoring systems on two of the Stuttgart Media University's websites. The tools are: (1) The *Accessibility* module of *Siteimprove* from Siteimprove, (2) *Pope Tech* from Pope Tech, (3) *WorldSpace Comply* (now called *axe Monitor*) from Deque, and (4) *ARC Monitoring* from The Paciello Group. The criteria catalogue consists of functional criteria that we gleaned from literature and user experience criteria based on the *User Experience Questionnaire*. Based on a focus group consisting of experts of Stuttgart Media University, we derived individual weights for the criteria. The functional evaluation criteria are: Coverage of the website and the guidelines, completeness, correctness, support in locating errors, support for manual checks, degree of implementing gamification patterns, support for various input and report formats, and methodological support for the *Website Accessibility Conformance Evaluation Methodology* 1.0 and for the German procurement law for public authorities "Barrierefreie Informationstechnik-Verordnung" 2.0. For determination of the user experience criteria, we conducted exploratory think-aloud user tests (n=15) using a coaching approach. Every participant tested all tools for 15 minutes (within-subject design). The participants completed post-test questionnaires, including the *User Experience Questionnaire*. According to our results, Siteimprove turned out to be the best tool for our purposes.

1 Introduction

The topic of digital accessibility is becoming increasingly important. As of 2016, an estimated 100 million persons in Europe had disabilities (European Disability Forum 2019). In Germany, at the end of 2019, around 7.9 million severely disabled people were living (Statistisches Bundesamt (Destatis) 2020). The Federal Statistical Office in Germany further reports that this was around 1.8% more than at the end of 2017. Companies often tend to ignore these facts because they do not consider it profitable. However, accessibility for the web is becoming more and more enforced, with lawsuits and fines threatening those who do not bother. It should also be noted that, if care is taken to ensure accessibility, all users will benefit because even users without disabilities will have advantages from

accessibility features (Schmutz 2016). Moreover, people are getting older on average, which means that they will also have poorer eyesight and other ailments. Very likely, at some point in life, people will be grateful for accessible technologies and websites.

Unfortunately, the importance of accessibility is often not recognized and a general awareness for digital accessibility is missing, even for universities and other educational institutions (Solovieva 2014), and for the European parliaments (Siteimprove 2019). Another problem is that accessibility guidelines are often difficult to understand and often people lack the motivation to learn and follow them. To counteract the mentioned problem, the use of tools that automatically check websites for accessibility is useful. Such tools automatically identify barriers and many of them explain how these barriers affect people with disabilities and how to fix them. Nevertheless, these tools cannot replace manual accessibility checks (Vigo 2013). The reason is that many of the guidelines cannot be checked automatically, at least not according to the current state of the art, without Artificial Intelligence or similar. On the other hand, it is difficult even for experts to manually find all accessibility errors, so the support of automated tools is still essential (Abascal 2019). It is often assumed that accessibility can be built in retrospectively, but in reality, this is much more complex and expensive than paying attention to accessibility right from the development stage.

For these reasons, we were looking for a suitable solution on monitoring the accessibility of websites at the *Hochschule der Medien* (HdM) in Stuttgart (in English it would be translated as *Stuttgart Media University*). This solution should support the website administrators at HdM to optimize the websites in their own responsibility for accessibility. We investigated how far commercial accessibility monitoring systems (AMS) are able to promote accessibility, which functionality they offer, how motivating they are to use and how well they perform compared to other systems. We reached out to six vendors and received trial versions for four tools: *Siteimprove* from Siteimprove, *Pope Tech* from Pope Tech, *WorldSpace Comply* (now called *axe Monitor*) from Deque, and *ARC Monitoring* from The Paciello Group. We included the browser extensions of the respective companies in our study: *Siteimprove Accessibility Checker* by Siteimprove, *WAVE Evaluation Tool* for Pope Tech (even though this tool was not developed by the company Pope Tech itself but by WebAIM.org[1], it is still the tool used internally in Pope Tech), *axe Expert* by Deque and *ARC Toolkit* by The Paciello Group. We analyzed these tools, applied them to two websites of Stuttgart Media University and compared them with each other using our own set of evaluation criteria.

The remainder of this paper is structured as follows: In chapter 2, we describe the relevant background for our study. Chapter 3 introduces the four AMS we used in our study, while chapter 4 explains our methodology. In chapter 5, we present the results which are followed up by a discussion (chapter 6). In chapter 7, we note some limitations of this study and possible future work for remedy. We finally provide our conclusions in chapter 8.

2 Material and Methods

In this chapter, we introduce the relevant guidelines, prior work, existing methodologies for checking websites and gamification patterns.

[1] https://webaim.org/

2.1 Guidelines

2.1.1 Web Content Accessibility Guidelines 2.1

The Web Content Accessibility Guidelines (WCAG) 2.1 (World Wide Web Consortium and others 2018) are a collection of guidelines that define digital accessibility requirements on three conformance levels: A, AA and AAA (highest accessibility). They are an international standard for digital accessibility that is mandated for public sector websites in the European Union by the Web Accessibility Directive (European Commission and others 2016).

2.1.2 Barrierefreie-Informationstechnik-Verordnung 2.0 and EN 301 549 v3.1.1

The „Barrierefreie-Informationstechnik-Verordnung“ (BITV) 2.0 is a national legislation in Germany and is intended to ensure a comprehensive and basically barrier-free design of modern information and communication technologies for the public sector (German federal ministry of labour and social affairs 2019). Section 3 of BITV 2.0 mentions the technical standards to be applied, which have to be “harmonized standards” by the European Union. The EN 301 549 (European Telecommunications Standards Institute 2019) is such a harmonized standard (European Commission 2018). Its section 9 contains a complete set of the technical requirements of WCAG 2.1 on the conformance level AA. In addition, BITV 2.0 recommends that WCAG 2.1 AAA criteria be met for portal pages, web forms and login pages. It should be noted that BITV 2.0 includes more requirements beyond WCAG 2.1 AAA, but these are not relevant for this publication.

2.2 Prior Work

Abascal et al. (2019) discuss the different types of accessibility evaluation tools and show how they can be used to support manual reviews. Currently, there are few studies on the comparison of tools for automatic web accessibility evaluation. The study of Vigo et al. (Vigo 2013) shows the problems of relying only on automated web evaluation tools. It empirically demonstrates the capabilities of the following tools: AChecker, SortSite, Total Validator, TAW, Deque and AMP. In this work, WCAG 2.0 was used as the underlying guidelines. Abduganiev (2017) analyzed and compared the following eight tools, using various evaluation criteria: AChecker, Cynthia Says, EIII Checker, MAUVE, SortSite, TAW, Tenon, and WAVE. The chosen guidelines for the review was WCAG 2.0. The most recent comparison was made by Pădure and Pribeanu (Pădure 2019). They tested the following tools: AChecker, Cynthia Says, TAW, Wave, and Total Validator. Some of the evaluation criteria used in these studies were applied directly or in a modified form in our work. Unlike previous studies, we have analyzed and compared mainly commercial automatic web accessibility evaluation tools, whose focus is not only on finding accessibility errors, but also on monitoring the progress of a complete website over a longer time period in terms of accessibility. In the context of this work, we will refer to these as AMS (short for “Accessibility Monitoring Systems”). In contrast to previous studies that relied on WCAG 2.0 AA, we based our evaluations on WCAG 2.1 and included its requirements of the highest conformance level AAA.

2.3 Methodologies for Checking the Accessibility of Websites.

Velleman and Shadi have developed the Website Accessibility Conformance Evaluation Methodology (WCAG-EM) 1.0, a methodology to establish good practices for evaluators to check websites for accessibility with respect to WCAG guidelines (Velleman 2014). A similar methodology is the “BITV-Test” of the project series "barrierefrei informieren und kommunizieren", in English: "accessible information and communication" (BIK-Projekt 2019). Apart from the fact that the BITV-Test refers to BITV 2.0 and therefore indirectly to EN 301549, there are other differences between

WCAG-EM and the BITV-Test. For example, in the BITV-Test, the qualification of the testers is an important factor, whereas in the WCAG-EM it is rather a recommendation. Unlike WCAG-EM, the BITV-Test does not use random sampling.

2.4 Gamification Patterns

Gamification is a concept that has existed for many years, although it has not always been referred to as such. In their work, Deterding et al. (2011) unify these concepts under the term *gamification*. In their literature review of empirical studies on gamification, Hamari et al. (2014) indicated that gamification has a positive influence on users. But how strong this influence is, depends on the particular application and user. Majuri et al. (2018) compiled probably the most comprehensive literature review to date with 128 empirical studies they examined. Their results serve as a basis for the weights of the individual gamification patterns used in our study.

2.5 Ethics Statements

2.5.1 Studies involving animal subjects

No animal studies are presented in this manuscript.

2.5.2 Studies Involving Human Subjects

The studies involving human participants were reviewed and approved by Gottfried Zimmermann, member of the commission of good scientific conduct at "Hochschule der Medien". The patients/participants provided their written informed consent to participate in this study.

2.5.3 Inclusion of identifiable human data

No potentially identifiable human images or data is presented in this study.

2.6 Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

3 Accessibility Monitoring Systems

An Accessibility Monitoring Systems (AMS) is a tool to automatically check a website including its subpages for accessibility problems. It is possible to schedule these scans so that they are performed once a week or once a month, for example. The decisive factor here is not only the detection of errors, but also the progress of the website with regard to accessibility, whether the website gets more barriers, through new content that is added or whether it becomes more accessible. This development can be traced in most of these tools using diagrams, reports, and archived scans. The user can also specify that an issue cannot be fixed. In the next scan, this issue will be ignored. It is important to note, however, that these tools are not yet able to cover all WCAG 2.1 guidelines, as some of them cannot be automatically checked. Therefore, such tools will not replace manual checks in the foreseeable future (Vigo 2013), but can only be used as a support tool. Although the names are different for each AMS, each of these tools has two or three different categories into which the problems found are sorted. For the sake of simplicity, we will unify these names into "errors", "warnings", and "best practices":

- *Errors* are accessibility problems that the respective AMS reports as actual errors.
- *Warnings* are potential problems that require the user to manually check whether these are really errors.
- *Best practices* are minor problems that may make the site more difficult to use for people with disabilities, but according to WCAG 2.1 are not errors that must be corrected for a website to be considered accessible.

3.1 Siteimprove

Siteimprove's AMS "Accessibility"[2] is one of several modules of the Siteimprove tool. The findings of the other modules can also help to improve accessibility in some cases but for the sake of comparability, only Siteimprove's "Accessibility" module is analyzed and compared. The dashboard of the "Accessibility" module of Siteimprove is shown in Figure 1.

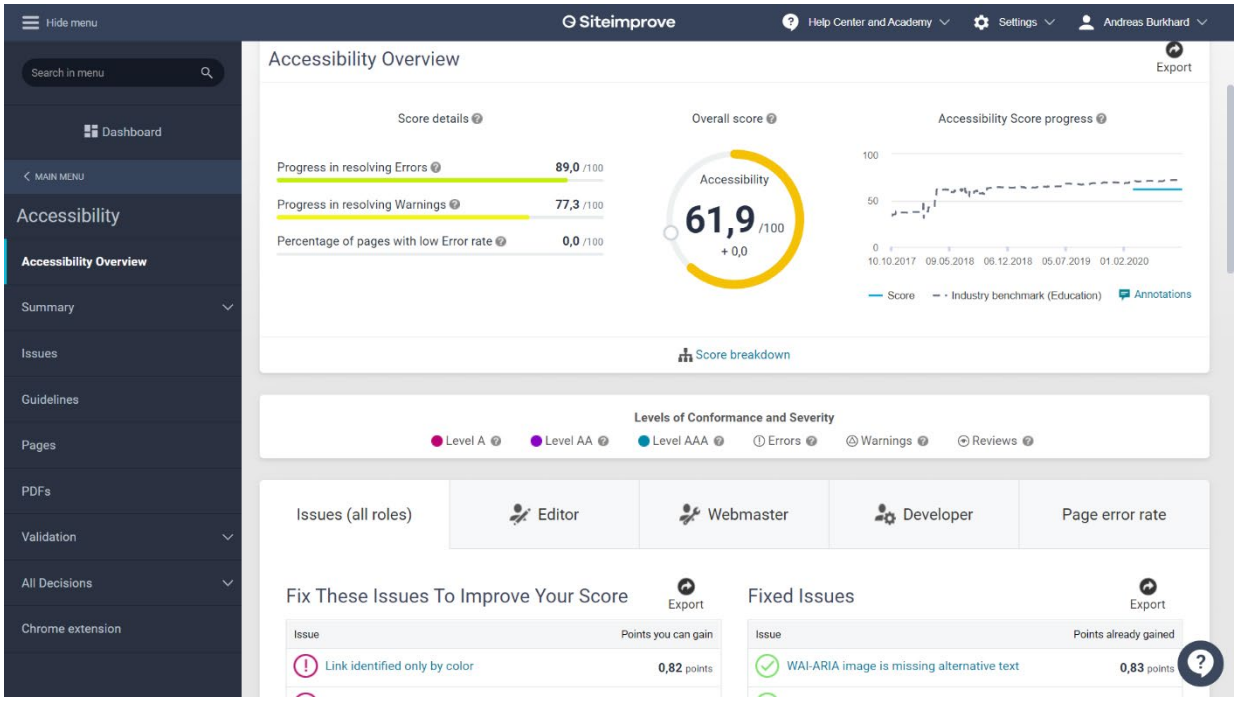

Figure 1. Siteimprove **"Accessibility"** module dashboard[3].

3.1.1 Siteimprove – Gamification

One of the biggest differences between Siteimprove and the other AMS is that Siteimprove relies heavily on gamification patterns to increase user motivation. Right on the Accessibility module dashboard, various gamification patterns can be found. The dashboard is modifiable, but the default settings are assumed in our description. As illustrated in Figure 1, there are different progress bars on

[2] The version of the Siteimprove "Accessibility" module was continually updated automatically during the study. The last update happened on August 27th, 2020.

[3] This screenshot is described in the following YouTube video: https://youtu.be/incmOBu_19E

the left, which show how many accessibility issues have already been fixed, how many warnings have been checked, and how many pages with only one or no errors are present.

The *overall score* in the middle of the screen shows a progress bar in the form of a ring, which indicates how high the calculated accessibility score is. The accessibility score is made up of the various accessibility problems and how much of them have already been solved in terms of percentage. These values are weighted according to importance and then aggregated to one value. At the edge of the ring, there is a small circle. This represents the industry benchmark for the relevant area. In this study, this is the area of education. This industry benchmark was composed of the average values of this area and the competitive comparison also increases the motivation of the users.

The industry benchmark can be found once again in the diagram on the right, represented by the dotted grey line, while the blue continuous line represents the HdM. In this diagram, the user can see the progress or regression of the HdM Accessibility Score including that of the industry. Another gamification pattern are points obtained by fixing problems. Points are used in two modules: *Fix These Issues to Improve Your Score* - this is a list of the most serious accessibility problems displayed, sorted by points; *Fixed Issues* - this list shows the already fixed accessibility issues.

3.1.2 Siteimprove – Locating Accessibility Problems

Siteimprove provides several ways to search for issues. The first one is using the menu item *guidelines*. Here, the user can view all errors sorted by violated WCAG 2.1 guidelines on all crawled pages for the respective individual guideline. Another possibility is the menu item *pages* which displays all crawled pages and shows how many errors are present at which conformance level and at which page level they are located. The third possibility is via the menu item *issues* where all errors found in the crawled webpages are displayed, sorted by how many different webpages these errors occur on. No matter which of these methods the user chooses, they are given the opportunity to click on one of the webpages or errors and an archive of the last scan of the corresponding webpage opens with Siteimprove toolbars on the left and at the top.

In a toolbar, all error categories are listed sorted by guidelines. It offers various filter settings like role assignment, severity, conformance level, and decisions made. If the user clicks on one of these error categories, all occurrences of the corresponding error are listed. For errors that affect images, small image previews are also displayed. When one of the occurrences is selected, the website automatically scrolls to that error and marks it with a red frame that flashes a few times. Alternatively, the location can be displayed and marked in the source code. Information about this error is given in the toolbar. These details include, which WCAG success criterion was violated, what the problem is, how to fix it, and a link to the various techniques suggested by WCAG 2.1.

In another toolbar, the user can display the HTML code instead of the visual presentation of the website. In addition, it is also possible to show and hide the CSS and to turn the JavaScript on and off. With the *Accessibility Explorer* it is possible to simulate total color perception deficiency, red-green deficiency, or blue-yellow deficiency for the webpage.

3.1.3 Siteimprove – Accessibility Checker Browser Extension

The browser extension *Accessibility Checker*[4] is similar to the tool described in the chapter "Siteimprove – Locating Accessibility Problems" with the toolbars. Some of the functions like the

[4] Siteimprove Accessibility Checker was used several times in the study, at latest with version 126.

Accessibility Explorer and the possibility to enable or disable JavaScript and/or CSS are missing in the browser extension, but the browser extension is free of charge unlike the AMS itself.

3.1.4 Siteimprove – Assignment and Handling of Issues

In Siteimprove, errors are not assigned to individuals, but to roles. These roles are *Editor*, *Webmaster*, and *Developer*. Accessibility errors found are automatically assigned to one of these roles based on the error category, but the user can manually adjust this assignment. With "Accessibility Policies", the user can easily define their own guidelines, for which the web pages are also checked. Siteimprove can check PDFs for accessibility issues. The PDF documents are opened in a separate program by Siteimprove. If the user chooses one of the findings, the document automatically scrolls to the correct location and frames it in red. Siteimprove has built-in validation for all scanned pages in HTML and CSS. It provides a connection to the project management and task management software *Jira* and *Azure DevOps Connector*, but this was out of scope for the study.

3.1.5 Siteimprove – Included Additional Services

Other additional free benefits of Siteimprove include the Siteimprove Academy with many IAAP-certified accessibility courses, a Help Center and FAQs.

3.2 Pope Tech

The tool "Pope Tech"[5] from Pope Tech focuses on a clear user interface, a visualization with consistent use of colors and symbols, and on ease of use. The findings are divided into six different categories. Errors and contrast errors have red symbols. Warnings and best practices have yellow icons. *Structural elements* are indicated by blue symbols and are not considered errors. They provide information about the structure of the site. *Features* with green symbols are common accessibility features found on the site. They improve accessibility when used correctly. This category is also purely informative. Purple symbols indicate information on *Accessible Rich Internet Applications* (ARIA) (Craig 2009) features, in particular about the location of elements that use ARIA. Even if these are not errors, the user should check ARIA elements manually to make sure they are used correctly. Every type of finding has its own unique symbol. The user can click on these symbols and get information about them including: the symbol in large, the name of the error, an description of the problem, and an explanation of how this problem can represent a barrier for people with disabilities. The user also gets instructions on how to fix the issue, a description on how the algorithm that found the problem works and which success criteria are affected, including a link to the corresponding success criteria.

In Pope Tech, the viewport of the scan is customizable. Thus, it is possible to scan the mobile view of a webpage. The reports can be configured precisely for every single finding type and/or category and which selection of them to be included in the report. As illustrated in Figure 2, in the module *Result Types*, the user can find unique symbols divided into the six mentioned categories and see how many findings were found in the last scan of the websites. In the *Scan Details* module, a list of all current scans of the selected website (or all websites) is shown. The numbers of errors and warnings are listed. The *Details* button leads the user to the listing of the findings. The module *Most Common Issues* gives an overview of the most common types of errors (or optional warnings) across all websites and the percentage of errors they account for in total.

[5] The version of Pope Tech was continually updated automatically during the study. The last update happened on June 27.

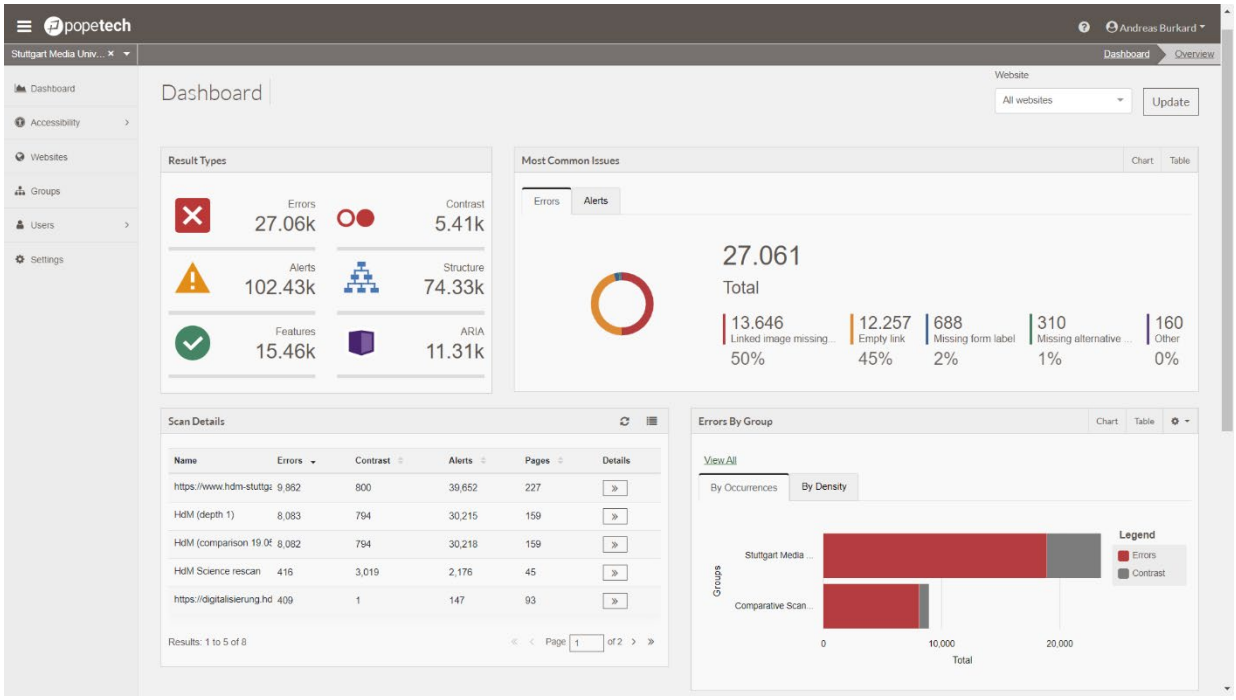

Figure 2. Pope Tech dashboard[6].

If the user clicks on one of these, a list of the webpages on which this error occurs is displayed, sorted by frequency. In the module *Errors by Group*, the user can see the relation between errors, contrast errors, and warnings per website group. *Issues Over Time* shows how many errors, contrast errors, optionally warnings, and checked webpages have been added or removed over the course of the scans in a diagram.

3.2.1.1 Pope Tech - Locating Accessibility Problems

Pope Tech offers several ways to examine accessibility errors and other findings. By clicking on one of the categories of the findings, all corresponding finding types can be viewed separately. For each finding type, the total number of occurrences and the number of occurrences on the individual webpages are displayed. As a variation hereof, *Most Common Issues* lists the most common findings, but focusing on true accessibility errors, contrast errors, and warnings. A third option is about the details of the individual scans. All scanned webpages are displayed and the details of the occurring finding types. Regardless of the selected variant, the user has the following options per webpage or finding type:

- "Page": The user simply visits the webpage.
- "Code": Toolbar including a code view opens and appears at the bottom. It jumps to the code of the first occurrence of the finding. In this code, all issues are displayed in the form of symbols at the corresponding code line.
- "Dismiss": The user can declare an error as "dismissed" and indicate the reason why this error should be excluded from future scans in an input field.

[6] This screenshot is described in the following YouTube video: https://youtu.be/2vZC0XWVRY8

- "WAVE" opens the webpage with the Web Accessibility Evaluation Tool (WAVE) browser plugin.

3.2.1.2 Pope Tech – Web Accessibility Evaluation Tools

WAVE[7] is the browser extension used by the AMS of Pope Tech. The symbols for the different finding types in Pope Tech are the same as those used by WAVE. All findings on the examined webpage are marked with symbols of the finding types. Within the toolbar, the user can switch off the styles for the webpage. That allows either to view certain hidden contents or to better understand the programmed structure. If the user clicks on the "code" button in the middle at the bottom of the screen, another bar extends with the source code of the webpage. Within the code view, the findings are marked with the respective symbols of the finding type. In the toolbar, different views can be selected using the tabs:

- "Summary" shows an overview of the number of occurrences divided into result type categories.
- "Details": The user can see several symbols for each finding type corresponding to the number of occurrences. If one of these symbols pressed, the visual display of the web page scrolls directly to the corresponding occurrence or code location and the symbol flashes briefly. Icons that are half transparent indicate hidden contents. In many cases, it is possible to make them visible by switching off the styles of the web page.
- "Reference": The user gets information about a selected finding.
- "Navigation" is an overview of the nesting of the structural elements on the webpage.
- "Contrast": Input fields for the foreground color and one for the background color. Underneath each field, there is a color picker. It is indicated for every conformance level and normal or large text whether the color contrast is sufficient. A "Desaturate page" link shows how the page would look like without colors in grayscale.

3.3 axe Monitor

At the time of the study, the AMS was called *WorldSpace Comply*, but has since been renamed to *axe Monitor*[8]. We use the name *axe Monitor* in this study. The progress and decline of accessibility can be read from a score set by Deque and its visualization, as shown in Figure 3. *axe Monitor* is part of a product suite from Deque. Note that only the AMS *axe Monitor* and the browser extension *axe Expert* were evaluated in this study.

3.3.1.1 axe Monitor – Assignment and Handling of Issues

In axe Monitor, the issues found in the scan are recorded in the form of code sections containing the affected elements, but without line numbers. When selecting an error, additional information is given on how to correct the error and a selector to find the element on a website. Issues can be assigned to team members. The user can write *issue comments* or *suggested remediation* on every issue to provide the team with information. The user can also add an "Issues Label" to an issue, which acts as a keyword.

[7] https://wave.webaim.org/extension/

[8] The version was continually updated automatically during the study; the last version available was "WorldSpace Comply" v6.5.0.62138. The ruleset used for the comparative scan was WCAG 2.1 Level AA with "Attest Version" 3.3.1, according to the initial configuration by Deque. We never changed the rulesets of any AMS after the initial scan.

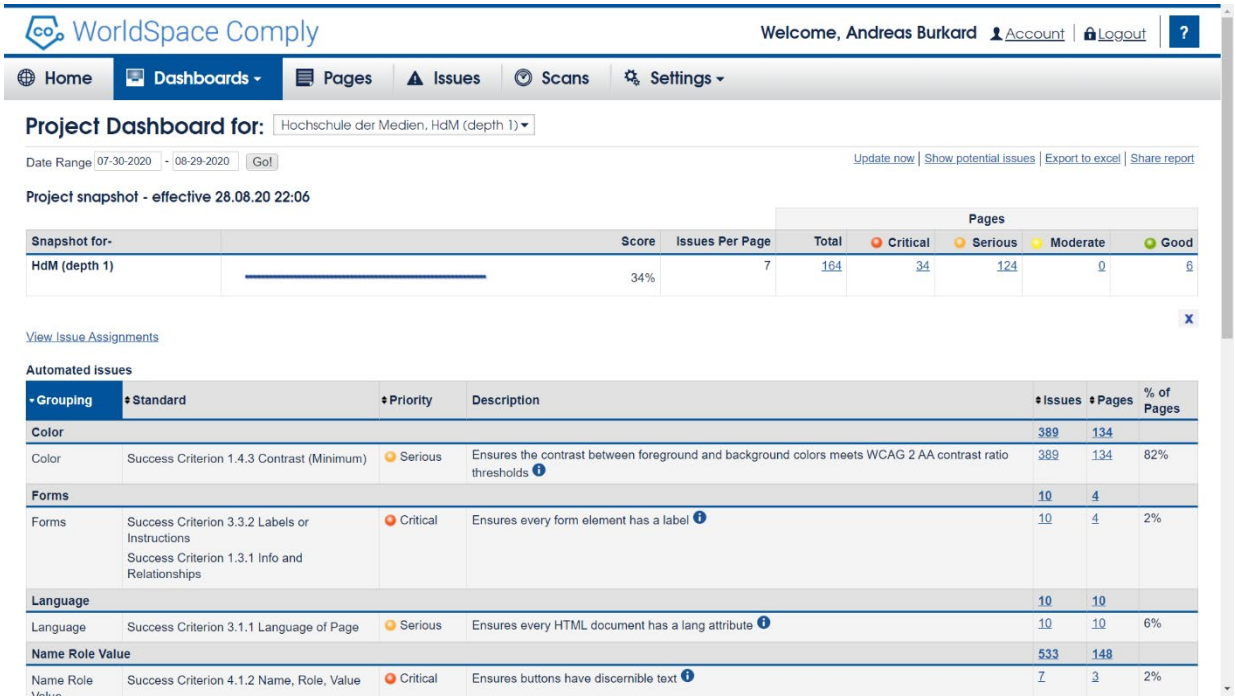

Figure 3. axe Monitor (formerly *WordSpace Comply*) project dashboard[9].

A connection of axe Monitor to project management and task management software such as *Jira*, *HP Quality Center v11.0*, and *HP ALM v12.2* can improve team collaboration and facilitate management. However, this functionality was not tested in the study. The errors can be assigned directly to team members via tickets, for example in Jira. A connection can be made between the issues of axe Monitor and the tickets of the different project management tools.

3.3.1.2 axe Monitor – axe Expert Browser Extension

As part of the study, Deque also provided us with a free trial version of *axe Expert*. The *axe Expert* (originally called *WorldSpace Attest*) is a browser extension by Deque. In our study, the browser extension was integrated into the developer tools of the browser. With the menu item *analyze*, the webpage can be scanned for accessibility problems. The following information can be found for each occurrence: Why it is (possibly) an error, what kind of barriers it causes, where to find it, and the corresponding element as code snippet. Additionally, a description of how to fix the issue, a sequence of nested elements leading up to the affected element, the error category, which guidelines have been violated and whether the finding was found through the *Best Practices* or *Experimental* options. The button *inspect node* can be used to jump directly to the corresponding code location in the developer tools. With *Highlight*, the user can jump directly to the problem in the visual display of the page in which it is surrounded by a dashed frame. Under the menu item *page insights,* a variety of tools can be found: *Headings*, *Links*, *Lists*, *Images*, *Focus*, *Frames*, *Objects*, *Landmarks* and *Autocomplete*. These tools can each mark, list, and output information about the respective name-giving elements that support manual checks.

A special feature of axe Monitor is that it allows to record scripts together with the browser extension axe Expert. The user first configures what is to be recorded, then starts the recording, and performs

[9] This screenshot is described in the following YouTube video: https://youtu.be/TUa5mu5S1z8

the actions that are to be recorded. This script can be used to check processes for barriers or to perform automatic authentication. A *scope* (or sometimes called a *template*) is an accumulation of elements that remain the same across multiple web pages of a website. Typical examples are the navigation bar, headers, and footers. These scopes can be set by the browser plugin axe Expert via CSS selectors and XPath technologies. The benefit of the scopes is that errors that are common between webpages are not listed on every webpage. Instead, they will appear on the Project Dashboard as a *Common Issue* and a second time in the *Automated Issues* table, and only once in the Issues Report.

3.3.1.3 axe Monitor – Scan Settings

axe Monitor offers a wide range of options for scan settings. The settings specific to axe Expert are as follows:

- *User Agent*: Webpages may have different versions based on the browser, whether it is a mobile device or a computer, a certain browser, and other factors. In a dropdown menu, the user can choose between different agents or input a custom one.
- *Single Sign-On* (SSO): The user can define whether multiple browser sessions may be used for the scan. It is faster but some websites refuse to be scanned in parallel by multiple sessions.
- *Session Establishment Tasks*: The user can name a task, further selecting one of four options:
    - *User authentication*: The user can enter a username and password, which will allow the scan to automatically log in to the website to access protected areas (e.g. intranet). There are three different authentication types: *Basic*, *Digest*, and *NTLM*.
    - *Client certificate* requires a public and private keystore file; the formats used are JKS or PKCS12.
    - *Select a script from below*: The recorded scripts can be selected and uploaded to axe Monitor to be used in the scan.
    - *Choose support for responsive design*: The support of this scan for responsive design can be turned on and off. More detailed configurations are specified in the menu item *iterate the scan*.

3.4 ARC Monitoring

As shown in Figure 4, ARC Monitoring[10] integrates its knowledge base, its newsfeed (like upcoming webinars) and its courses (TPG Tutor) prominently in the dashboard. *WCAG density* represents the average number of WCAG violations with the conformance level A or AA in the respective scanned domain. ARC Monitoring uses many diagrams with which the evolution of each error can be observed. The errors are sorted and prioritized according to the impact they have on the page. With ARC Monitoring, websites can be checked with various rulesets. At the time of the study, the following rulesets were available: *ARC Rules 3.2.1* developed by The Paciello Group, and *axe Core v3.3.1* developed by Deque. The user can use both to capture a larger number of errors. In the comparative scan, the *ARC Rules* 3.2.1 with WCAG 2.1 were used as the *default rules engine* since this was the initial setting. Unique to ARC Monitoring is the *Initial Domain Analysis* (IDA).

[10] The version of ARC Monitoring was continually updated automatically during the study; the last version available was release 4.2.1.

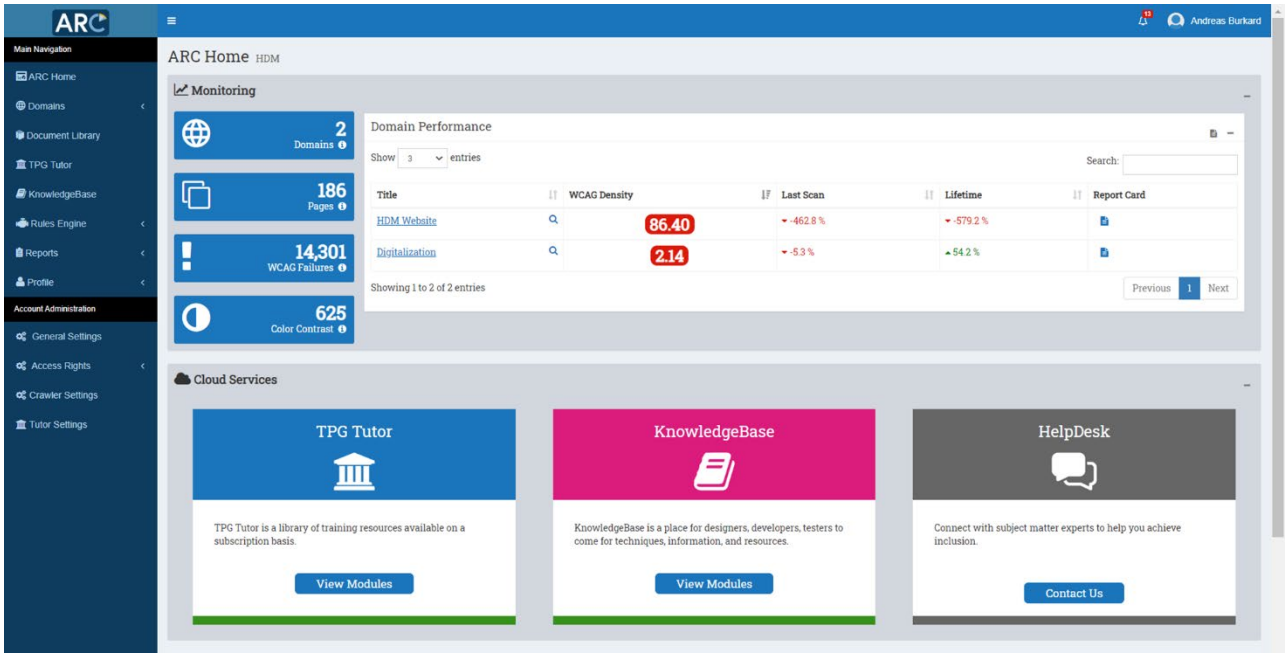

Figure 4. ARC Monitoring dashboard[11].

When a domain is scanned for the first time, the first 25, then 50, and then 100 webpages are scanned in succession. This way, it can be determined how accessible the pages at the surface of the website are compared to the webpages located deeper. It also shows how accessible the scanned domains are compared to the average of the other domains scanned by ARC Monitoring.

The collected data is clearly presented in diagrams. In the IDA, conclusions are drawn from the scanned pages, which are automatically converted into tips and useful information as part of a report. An example from a report on the domain of the HdM is: "A crawl of the first 100 pages of a website typically includes a sampling of pages from more than one sub-area of the website. These pages are 276.1% less accessible when compared to the first 50 pages of the website. This variance is significant, suggesting that the accessibility of the website is likely to degrade further along the user journey." In the historical progress report, a checklist shows for a domain which steps have already been taken to make it more accessible and which should be taken in the future.

3.4.1.1 ARC Monitoring – Locating Accessibility Errors

The detected errors or warnings are recorded as code snippets containing the affected elements without line numbers. This includes the name of the error, a description of how the error affects people with disabilities, which WCAG success criteria were violated, which rule set was used to find the error, and links to knowledge base materials related to the error. In addition, a comment can be added to the finding in the comment field.

3.4.1.2 ARC Monitoring - Policy-driven Test Initiatives

Policy-driven test initiatives can be used to set a goal, e.g., the elimination of a specific type of error. The user can define what needs to be achieved, e.g., the type of error to fix and an improvement by a

[11] This screenshot is described in the following YouTube video: https://youtu.be/-k1KaYtNqo8

certain percentage, and the time frame for achieving the goal. This goal is displayed to all team members and allows them to work on a problem in a coordinated manner.

3.4.1.3 ARC Monitoring – ARC Toolkit Browser Extension

We used "ARC Toolkit"[12], a free browser extension offered by The Paciello Group. There are other browser extensions available by The Paciello Group, but we did not include them in our study. ARC Toolkit is integrated into the developer tools; we used it in the Google Chrome browser. When the website is automatically checked for accessibility with ARC Toolkit, the findings are classified in a table. This table contains accessibility errors, possible errors, and findings, divided into visible and invisible elements. When selecting one of the table columns containing an error category, all elements in the extension are listed and the affected elements are visually marked on the page and colored black if there is no problem, and red if there is a problem. Each error or warning is given a name in camel case (e.g. "emptyAltWithTitle"), a description of the problem and a suggestion how to solve it. For findings concerning pictures, a small preview of the picture is shown.

ARC Toolkit supports manual checking in the following ways. The tab order can be displayed visually with red lines. The WCAG success criterion "1.4.10: Reflow" can be easily checked with the "check page reflow" function by setting the viewport to 1280 CSS pixels and zoom to 400%. With "check text spacing", the browser is set to the conditions required for the WCAG success criterion "1.4.12: Text Spacing", which include: Line height (line spacing) to at least 1.5 times the font size, spacing following paragraphs to at least 2 times the font size, letter spacing (tracking) to at least 0.12 times the font size, and the word spacing to at least 0.16 times the font size (World Wide Web Consortium and others 2018). The "show and track focus" function clearly emphasizes the focus with a thick red border when marked by the tab key, even if the border is turned off by the webpage. The tool also offers a validation of the webpage based on the DOM or URL.

4 Parameters of the Comparative Study

4.1 Evaluation Criteria

Our evaluation criteria are partly based on evaluation criteria used in other studies (Abduganiev 2017, Pădure 2019, Vigo 2013) for comparing automatic web accessibility evaluation tools. In addition, we identified new evaluation criteria, which were discussed and weighted in an expert meeting. In the following, the total number for an evaluation criterion means the number of findings of all AMS including the manual check, which meet the conditions of the evaluation criterion. Our complete set of evaluation criteria is: Coverage of webpages, coverage of success criteria, completeness, correctness, support for localization of errors, support for manual checks, User Experience (empirical), gamification patterns, input formats, output formats, Methodology Support for WCAG-EM, and Methodology Support for the BITV-Test. These are described in detail in the evaluation section.

4.2 Weighting of the Evaluation Criteria - Meeting of Experts

Six experts from HdM (professors working in appropriate fields and accessibility experts) were invited in the context of this work to discuss and vote on the weighting of the evaluation criteria. The criteria were presented to them in detail and they discussed them. In a questionnaire, the experts would vote on each criterion on a scale from 0 (not important) to 100 (very important). In a similar

[12] ARC Toolkit was used several times in the study, at latest with version "ARC Toolkit" v3.3.2.0.

way, the voted on the weighting of the six scales of the *User Experience Questionnaire* (UEQ) to calculate the *Key Performance Indicator* (Laugwitz 2008), except that here a Likert scale of 1 (not important) to 7 (very important) was used, as defined in the UEQ manual. For every evaluation criterion, the sum of weights obtained from the experts was normalized to a range from 0 to 1 in relation to the other AMS.

4.3 Website Samples

Since we were looking for a suitable solution for monitoring website accessibility, a selection of five webpages from two HdM websites was used as sample in this study:

- The main HdM website with the following pages:
  - https://www.hdm-stuttgart.de
  - https://www.hdm-stuttgart.de/hochschule/profil/qm
  - https://www.hdm-stuttgart.de/science

- The digitization website of the HdM with the following pages:
  - https://digitalisierung.hdm-stuttgart.de
  - https://digitalisierung.hdm-stuttgart.de/barrierefreiheit/barrieren-melden

This sample was selected based on importance and page type. The selected pages include a variety of layouts, navigation menus, tables, and elements that only appear on specific subpages. For the comparison, the pages were scanned by all AMS at the same time as far as possible. Each of the websites was scanned with a crawl depth of 1; this includes the start page and the first level of all associated subpages. For the calculation of the evaluation criteria results, the main website and the digitization website each contributed 50% of the rating, even though they contain different numbers of subpages.

4.4 Methodology for Carrying Out the Comparison

To verify the accuracy of the AMS results, we conducted a manual evaluation of the audited websites. The following definitions are important for understanding the evaluation criteria, which have already been used in prior studies (Abduganiev 2017, Vigo 2013):

- True positive (TP): An error reported by an AMS which was found to be an actual accessibility error by the manual check.
- False positive (FP): An error reported by an AMS which was not found to be an actual accessibility error by the manual check.
- False negative (FN): An accessibility error identified by the manual check, but not recognized as an error by an AMS.

The scan of May 19, 2020[13] was used to compare the AMS regarding the success criteria that concern the comparison of a single scan. We tested against WCAG 2.1 conformance level AAA which is relevant for portal, login and form pages of government agencies according to the German

[13] Since we used a live website for our study, some content changes affecting the accessibility of the website may have been applied from the time of the initial scan to the comparative scan on the sites. However, none of them were tailored towards a specific AMS.

regulation BITV 2.0. Each of the AMS was set to its highest possible conformance level, but without the need for defining custom rules.

All results from an AMS with their corresponding browser extensions were added up. An error is only considered an error if it violates a WCAG success criterion. If an error was reported as accessibility error by an AMS, but did not violate any WCAG success criterion, we counted it as FP. If several success criteria are violated by a single error, then this error counts as an error for every success criterion in violation. The AMS often make suggestions as to which success criteria have been violated by a single error, which we have manually checked and adjusted. If the same error occurs in the same place on different webpages, for example in a consistent navigation, then it will only count as TP or FP on one of the webpages.

5 Results

Table 1 shows the normalized results of our evaluation criteria. The evaluation criteria are listed as rows. The "weighting" column shows the weighting by the expert meeting for the individual evaluation criterion. The values of the results of the evaluation criteria are normalized to be in a range from 0 to 1. An AMS' total result is determined by the sum of its evaluation criteria, each multiplied by its weighting. In summary, Siteimprove achieved the highest score with 0.87 points, followed by axe Monitor with 0.71 points, then ARC Monitoring and Pope Tech tied with 0.69 points for the third place. All raw data for the evaluation criteria are provided as supplementary materials in an Excel file called "raw data.xls".

Table 1. Evaluation criteria results of all accessibility monitoring systems (conformance level AAA)[14]. The result of every data cell except in the column *weighting* is normalized. The *result* row contains the results of the individual evaluation criteria multiplied by their respective weighting and then summed up.

Evaluation criterion	Weighting	Siteimprove	axe Monitor	ARC	Pope Tech
Coverage of webpages	10.49%	0,74	1,00	0,75	0,75
Coverage of success criteria	10.84%	1,00	0,67	0,87	0,83
Completeness	9.42%	0,97	0,62	1,00	0,70
Correctness	9.59%	0,85	1,00	0,65	0,83
Support for localization of errors	10.49%	1,00	0,60	0,80	0,80
Support for manual checks	9.20%	0,23	1,00	0,88	0,36
User Experience (empirical)	14.96%	1,00	0,09	0,07	0,73
Gamification patterns	5.62%	1,00	0,24	0,48	0,12

[14] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/W3Cxec9d5f9C92T/download

Evaluation criterion	Weighting	Siteimprove	axe Monitor	ARC	Pope Tech
Input formats	5.50%	1,00	1,00	0,50	0,50
Report formats	3.54%	1,00	1,00	1,00	1,00
Methodology Support for WCAG-EM	4.61%	0,75	1,00	0,80	0,70
Methodology Support for the BITV-Test	5.73%	1,00	1,00	1,00	0,87
Result	**100%**	**0,87**	**0,71**	**0,69**	**0,69**

In the following subsections, we describe the evaluation criteria of the study and how the AMS performed on them.

5.1 Coverage of Webpages

The evaluation criterion *coverage of webpages* covers the number of pages that an AMS can crawl in relation to the number that all AMS can crawl together. Only webpages that contain unique content are counted for this criterion. For example, if the same web page is accessible via several different Uniform Resource Locators (URLs), it is still counted as one web page only. However, if the same page with different URL parameters calls up different contents, then each variant is considered a separate web page. With some AMS, it is possible to automatically authenticate and crawl non-public pages such as web pages in the intranet. This requires valid access data. Since, due to data protection rules, it was not possible to use the login data of a real student account to check the intranet of the HdM website and no fake account could be created in time, non-public pages in the sample were not included. Instead, bonus points from 0 to 1 were awarded for the following functions that enable the crawling of additional webpages:

- + 0.25 / 0.5 points: The AMS scores 0.25 points if automatic authentication is possible but must be enabled and set up by the company. If the user themselves can set up automatic authentication with the AMS, the tool receives 0.5 points instead.
- + 0.5 points: Complete pre-defined processes can be evaluated with the AMS, which may result in more pages being available.

All AMS were able to crawl a similar number of pages, axe Monitor scored best due to the additional features, see Table 2. With the help of recorded scripts, axe Monitor / axe Expert can scan and evaluate processes. Each of the tools provides automatic authentication so that the scan can access pages that require a login.

Table 2. Coverage of websites results[15]. For the evaluation components in the rows, "w" indicates the weight. Every data cell provides the absolute test result (e.g. "160 pages"), followed by the normalized result in the second line (e.g. 1.0).

	axe Monitor	ARC Monitoring	Pope Tech	Siteimprove
Pages crawled (HdM) w: 25%	**160** pages **1.0**	**158** pages **0.99**	**157** pages **0.98**	**155** pages **0.97**
Pages crawled (Digitization) w: 25%	**19** pages **1.0**	**19** pages **1.0**	**19** pages **1.0**	**19** pages **1.0**
Automatic Authentication? w: 25%	Yes **1.0**	Yes **1.0**	Yes **1.0**	Yes **1.0**
Result (normalized)	**1**	**0.75**	**0.75**	**0.74**

5.2 Coverage of Success Criteria

The evaluation criterion *coverage of success criteria* (Abduganiev 2017, Vigo 2013) is the number of violated WCAG success criteria (SC) indicated by at least one TP divided by the total number of violated SCs. Siteimprove has the highest coverage of success criteria, as shown in Figure 5. One reason for this may be that Siteimprove is the only one of the four AMS that is able to check websites for conformance level AAA.

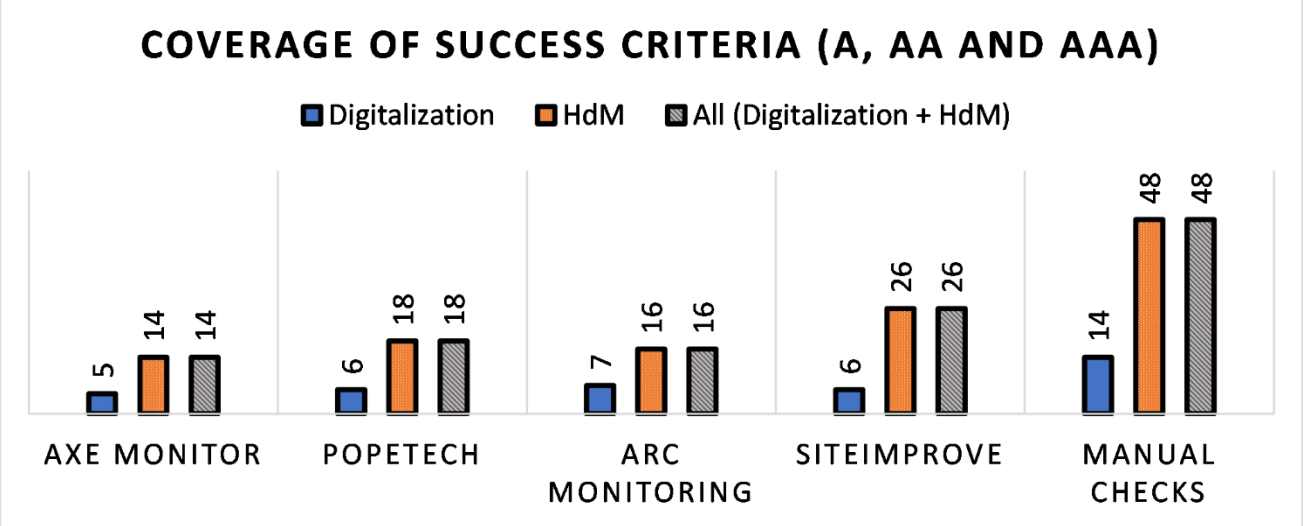

Figure 5. Coverage of success criteria based on WCAG 2.1 AAA[16].

Note that we have made an alternative evaluation against level AA which did not change the ranking of the AMS regarding coverage of success criteria (see 5.13).

[15] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/xkkSLomppq8awHx/download

[16] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/B9NSs9zaGjJJt9s/download

5.3 Completeness

While *coverage of success criteria* indicates the width of the problems, the evaluation criterion *completeness* (Abduganiev 2017) indicates the depth. *Completeness* shows the real accessibility errors, or TPs in relation to FNs. The criterion is calculated by dividing the number of errors found by an AMS that contain at least one TP by the overall number of errors found (including manual checks). While ARC Monitoring found most errors on the digitization website with 40%, Siteimprove found most on the main website of the HdM with 36%. With both websites combined, ARC Monitoring has scored best in this evaluation criterion, see Table 3.

Table 3. Completeness results (conformance level AAA)[17]. For the evaluation components in the rows, "w" indicates the weight. Every data cell provides the absolute test result (e.g. "21 %"). The last row contains the normalized results (e.g. 1.0).

	axe Monitor	Pope Tech	Siteimprove	ARC Monitoring	Manual checks
Completeness (Digitization) w: 50 %	15%	21%	21%	**40%**	100%
Completeness (HdM) w: 50 %	21%	21%	**36%**	20%	100%
Result (normalized)	**0.62**	**0.70**	**0,97**	**1.00**	

5.4 Correctness

The evaluation criterion *correctness* (Abduganiev 2017, Vigo 2013) expresses how many of the TPs found by an AMS were actually TPs and not FPs. *Correctness* could only be ensured by manual checking. This value was calculated for every AMS by dividing the number of true errors found (TPs) by all errors reported by the AMS. Remarkably, axe Monitor did not find a single FP, as Figure 6 shows. The correctness of ARC Monitoring on the main HdM website was low. A possible explanation is that some of the errors found are no WCAG violations but problems of usability and are therefore considered FPs. Additionally, ARC Monitoring found accessibility errors on elements that were not visible to any user. Because these elements did not represent barriers, they were counted as FPs.

[17] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/a4byaoRD8gBRacw/download

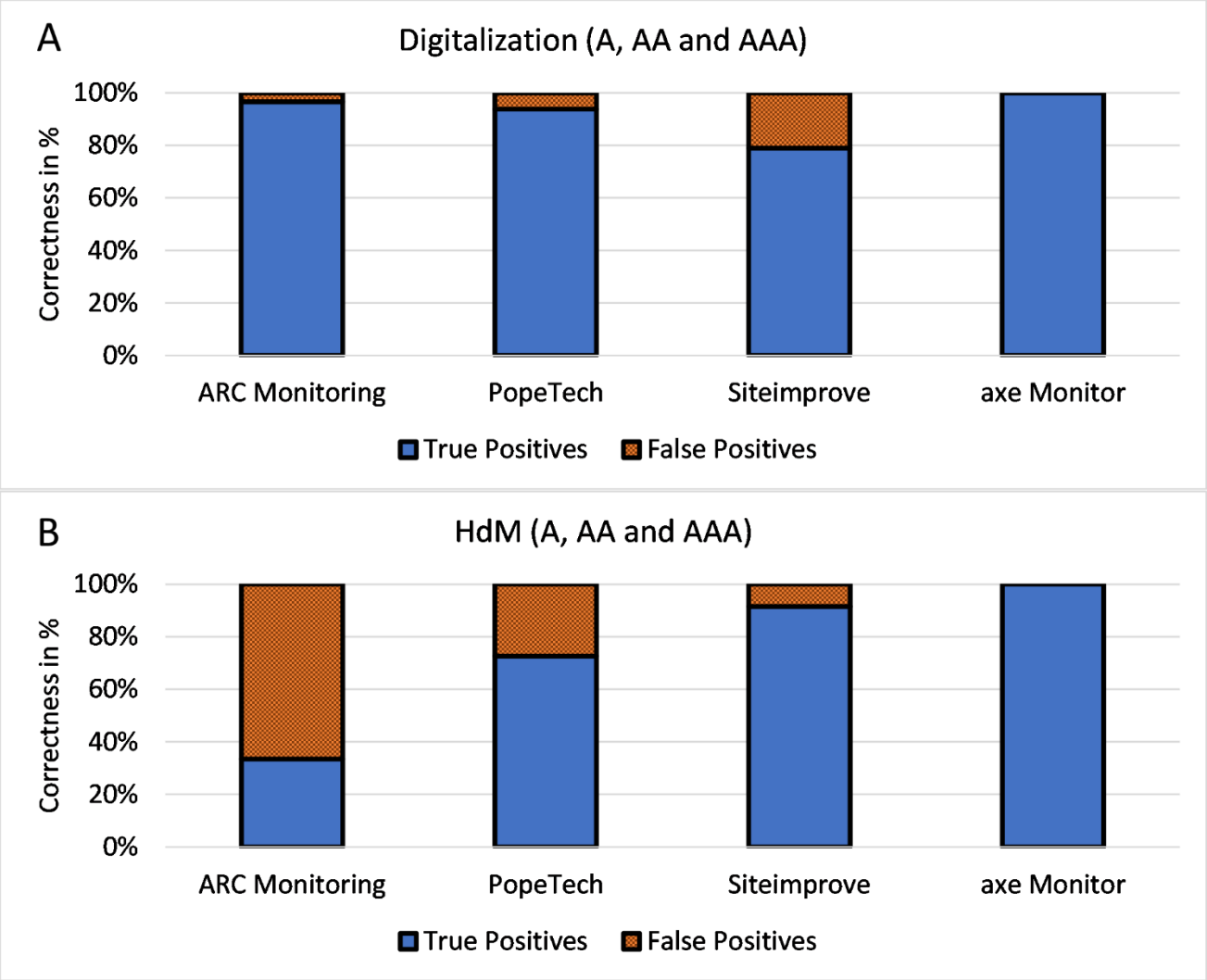

Figure 6. Correctness for conformance level AAA[18]. (**A**) Digitization website. (**B**) HdM website.

5.5 Support for Localization of Errors

The criterion *support for localization of errors* evaluates how well a tool supports a user in locating an error on a webpage. Points were awarded for various features in this regard, see Table 4. Table 4 indicates that Siteimprove offers the most support in locating an error. The user does not need to open a browser extension manually and is directed from the AMS to the error, which is also highlighted. In addition, a preview of the image where a problem has been found is always displayed.

[18] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/TooB59ZwLibekfN/download

Table 4. Support for localization of errors results[19]. Every data cell provides the awarded points for a certain feature. The last row is the normalized result (e.g. 0.8).

	Siteimprove	Pope Tech	ARC Monitoring	axe Monitor
AMS opens error page in browser extension	1	1		
Highlights errors	1	1	1	1
Scrolls to the error	1	1	1	1
Image preview	1		1	
Jump to error code	1	1	1	1
Sum	**5**	4	4	3
Result (normalized)	**1.0**	**0.8**	**0.8**	**0.6**

5.6 Support for Manual Checks

Warnings are possible errors, for which the AMS cannot automatically detect whether it is an actual accessibility error and therefore requires manual checking. The evaluation criterion *support for manual checks* includes how many warnings an AMS finds that turn out to be TPs after a manual check. Note that TPs are determined against WCAG 2.1 AAA; therefore, best practices did not count in this criterion. Although Siteimprove displays the most warnings, those from axe Monitor are the most accurate closely followed by ARC Monitoring, see Figure 7.

5.7 User Experience (Empirical)

For the evaluation criterion *User Experience (empirical)*, we conducted a user study with 15 participants. Each user test took 1.5 to 2 hours. The participants were website administrators at the HdM and students or graduates who attended at least one lecture on digital accessibility and one on web development. Every participant tested all four AMS including the pertaining browser extensions for 15 minutes each, according to a within-subject design (Nielsen 1994). Since AMS vary widely in their functionality and focus, it was not possible to design a user test with prescribed tasks in a fair way without discriminating against an AMS. For this reason, a free exploration test (Goodman 2012) was conducted. This means that the users should examine the respective AMS on their own and without concrete tasks. In line with the coaching approach (Nielsen 1994), the participants were allowed to ask questions at any time during the user test and were asked to think aloud as they interacted with the tools, as required by the Think-Aloud protocol (Nielsen 1994). To avoid fatigue or learning effects, counterbalancing (Albert 2013) was used, which in this case means that the order in which the users tested the tools was changed for each session. Care was taken to ensure that each tool was tested as often as possible on every position. After 15 minutes testing on a single tool, the participant filled out our own questionnaire and the UEQ (Laugwitz 2008). The UEQ allows to empirically determine the overall user experience of a product. The participants are presented with word pairs, whose results can be divided into six different rating scales: *attractiveness*, *perspicuity*, *efficiency*, *dependability*, *stimulation*, and *novelty*.

[19] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/xWGcf9qo9pHtCYb/download

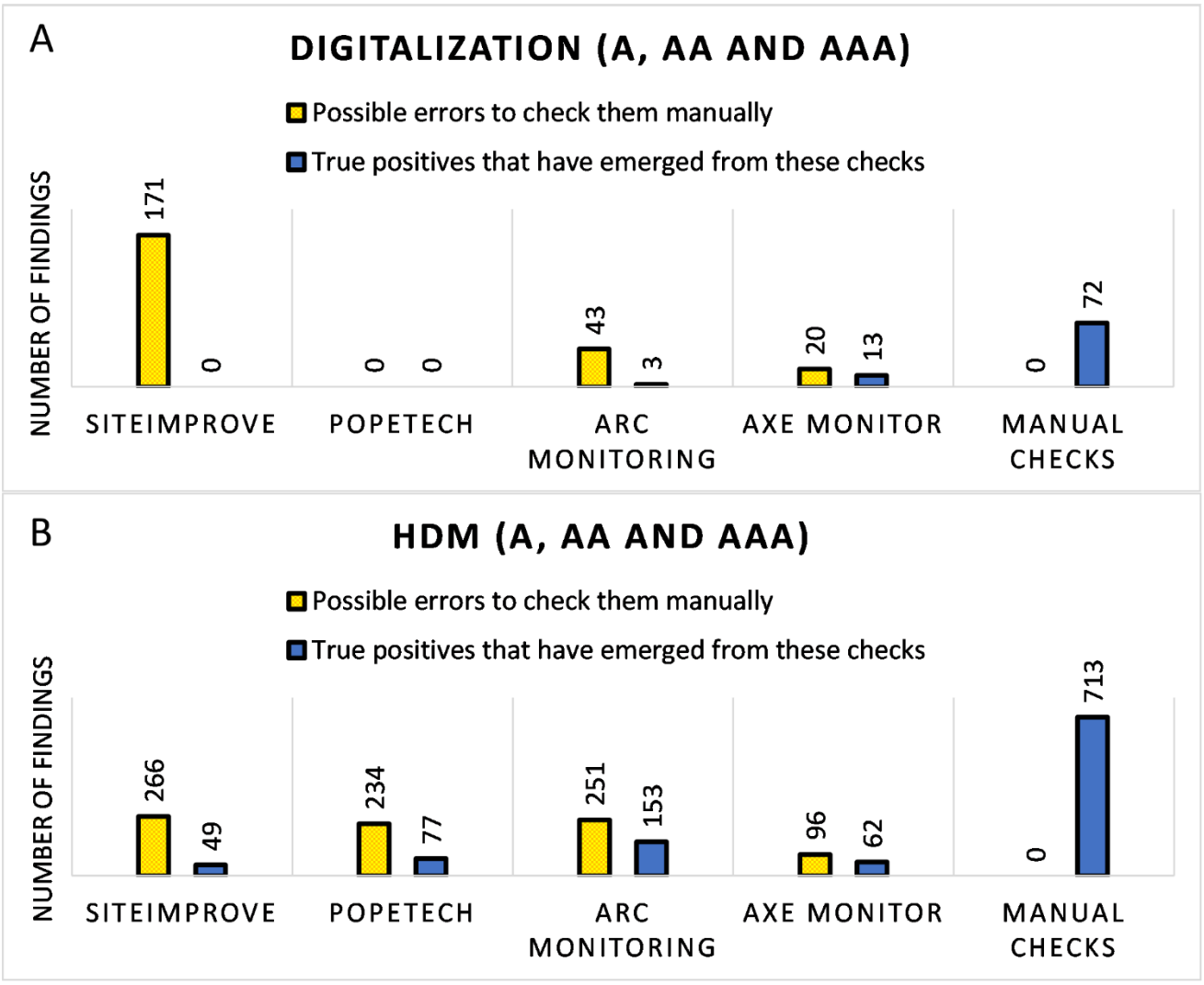

Figure 7. Support for manual checks[20]. (**A**) Digitization website. (**B**) HdM website.

Based on the Key Performance Indicator (KPI) (Laugwitz 2008), the six scales are combined into a single total value. The weighting of the scales was determined by the expert meeting (see supplementary Table 1). ANOVA (Fahrmeir 2015), an analysis of variance for three or more samples, was used to determine the significance between the results of the user study.

Siteimprove made the highest score in the user test, followed by Pope Tech, see Table 5. However, the only significance that the ANOVA test was able to determine is the difference between the first-placed Siteimprove and the last two AMS, axe Monitor and ARC Monitoring. For almost all word pairs, Siteimprove received more positive than neutral and negative responses combined (see supplementary Figure 1). Siteimprove scored mainly between above average and good, except for *novelty* in which it scored below average (see supplementary Figure 2). Pope Tech, the runner-up

[20] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/cy8EMBG8FyKoP8y/download

with a normalized result of 0.73, also received mainly positive values on the word pairs (see supplementary Figure 3). In the benchmark, most scales for Pope Tech are in the range slightly below average, except for *novelty* which is above average (see supplementary Figure 4). axe Monitor (third place) achieved a normalized value of 0.09, ARC Monitoring (fourth place) 0.07. According to the word pairs, axe Monitor was perceived as "consultative", "conventional" and "unattractive" (see supplementary Figure 5). axe Monitor's user experience was mostly rated bad, except for *dependability* (see supplementary Figure 6). During the user test, it was repeatedly referred to as a traditional working software, with little emphasis on motivation. In ARC Monitoring, as can be seen in the word pair distribution, almost everything was criticized in terms of user experience, except "secure" and "slow" (see supplementary Figure 7). In the benchmark, all scales are in the "bad" range (see supplementary Figure 8). Nevertheless, there were also responses that were positive.

Table 5. User Experience (empirical) results[21]. In the row *KPI* every data cell provides the absolute value of the Key Performance Indicator (e.g. 1.33). The last row is the normalized result (e.g. 0.73).

	Siteimprove	Pope Tech	axe Monitor	ARC
KPI	1.33	0.97	0.11	0.09
Result (normalized)	**-> 1.00**	**-> 0.73**	**-> 0.09**	**-> 0.07**

5.8 Gamification Patterns

The evaluation criterion *gamification patterns* evaluated the use of gamification patterns to increase motivation. For the evaluation, the occurrence of the gamification patterns is counted and multiplied by the respective weighting. The weighting was derived from the literature review by Majuri et al. (Majuri 2018). Every gamification pattern was categorized by positive, neutral, or negative results. We adopted the table of the work of Majuri et al. and added a column called *weighting* (see supplementary Table 2). For the purpose of our study, we determined a pattern's weight by dividing the number of its positive results by the total number of results.

According to this calculation, Siteimprove is the most motivating of the AMS. It includes many gamification patterns, such as points the user can achieve, a benchmark against sites in a similar area, and various progress bars, as shown in Table 6. Furthermore, the user can earn points by finishing tasks.

[21] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/dfG8Kec8jdXdHZ3/download

Table 6. Gamification patterns results[22]. Points are awarded for every used *gamification pattern* multiplied by its *weighting*. In the row *result* provides the combined absolute value (e.g. 3.15) and *result (normalized)* the normalized result of this criteria.

	Weighting	Siteimprove	ARC Monitoring	axe Monitor	Pope Tech
Points, score, XP	0.73	2		1	
Progress, status bars, skill trees	0.81	2	1	1	1
Competition	0.88	1	1		
Challenges, quests, missions, tasks, clear goals	0.73	1	1		
Performance stats, performance feedback	0.93	2	1		
Result		6.55	3.15	1.54	0.81
Result (normalized)		**1.00**	**0.48**	**0.24**	**0.12**

5.9 Input Formats

For the criterion *input formats*, points were assigned for the file formats that are relevant for accessibility and can be checked by the AMS (i.e. HTML and PDF). Table 7 shows that both Siteimprove and axe Monitor can check both file formats among the AMS. For PDF, an extra tool is even included that allows users to frame the errors and locate them.

Table 7. Input Formats results[23]. Every data cell provides the awarded points for a format that can be scanned by the respective tool, in the *result* row these points are summed up as an absolute value (e.g. 2). The last row is the normalized result (e.g. 0.8).

	Siteimprove	ARC Monitoring	axe Monitor	Pope Tech
HTML	1	1	1	1
PDF	1	0	1	0
Result	**2**	1	**2**	1
Result (normalized)	**1.00**	**0.50**	**1.00**	**0.50**

[22] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/WWxG9EbE5GD2ieM/download

[23] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/AETxm9SqzQ4tKA2/download

5.10 Output Formats

For the evaluation criterion *output formats*, points were awarded for each report format that an AMS can export. As shown in Table 8, all AMS have the same number of formats they can export to, a draw in this case.

Table 8. Output Formats results[24]. Every data cell provides the awarded points for a format as which the report can be exported, in the *result* row these points are summed up as an absolute value (e.g. 2). The last row is the normalized result (e.g. 0.8).

	Siteimprove	ARC Monitoring	Pope Tech	axe Monitor
HTML	1	0	1	1
PDF	1	1	1	0
XLSX	0	1	0	1
CSV	1	1	1	1
Result	**3**	**3**	**3**	**3**
Result (normalized)	**1.0**	**1.0**	**1.0**	**1.0**

5.11 Methodology Support for WCAG-EM

For the evaluation criterion *methodology support for WCAG-EM*, points were awarded on how many steps of the WCAG-EM methodology (Velleman 2014) were supported by the respective AMS. axe Monitor turned out to have the most features that support the user when using WCAG-EM (see supplementary Table 3). This was mainly due to the possibility to evaluate processes with recorded scripts and the possibility to adjust the user agents.

5.12 Methodology Support for the BITV-Test

The evaluation criterion *methodology support for the BITV-Test* evaluated how many methods and steps of the BITV-Test (BIK-Projekt 2019) are supported by the respective AMS (see supplementary Table 4). The evaluated BITV-Test steps are the following: "4.1. What belongs to the test item?", "6.1. Analysis of the web presence", "6.2.3. Cover all barriers", "6.2.4. Include different page types", "6.2.5. Include different page states" and "6.2.7. Include pages with different functions". Scores for each of these steps were evaluated on a theoretical level by looking at the functionalities of the tools, as reported by their companies. Most BITV-Test steps refer to the WCAG 2.1 guidelines. ARC Monitoring, axe Monitor and Siteimprove each offer the greatest support for the BITV-Test with their toolsets, all scoring 38 points, followed by Pope Tech with 33.

5.13 Comparative Study Without the Conformance Level AAA

Since not all AMS natively support the WCAG 2.1 conformance level AAA, we have alternatively computed the evaluation criteria for level AA (see supplementary Table 5). Siteimprove is still in first place with a value of 0.85, but the other ranks have changed. ARC Monitoring is second with 0.70, followed by axe Monitor and Pope Tech who are third with a tie of 0.68. Note that the

[24] The alternative Excel file can be downloaded here: https://cloud.mi.hdm-stuttgart.de/s/JGieqn7sAC9aykD/download

exclusion of conformance level AAA did not change the ranking for evaluation criteria *coverage of webpages*, *coverage of success criteria* (see supplementary Figure 9), *completeness* (see supplementary Table 6) and *correctness* (see supplementary Figure 10), *support for manual checks*, *user experience (empirical)*, *gamification patterns*, *input formats*, *report formats*, *methodology support for WCAG-EM* and *methodology support for BITV*. Only for the evaluation criterion *support for manual checks*, the ranking has changed, with ARC Monitoring now in first and axe Monitor in second place (see supplementary Figure 11).

6 Discussion

While prior studies (Abduganiev 2017, Pădure 2019, Vigo 2013) have mostly evaluated free tools for checking accessibility, this work compares some commercial AMS that can monitor entire websites and evaluate their accessibility over time. Our study did not only assess the tools' functionality and effectiveness, but also how user-friendly and motivating they are to use. In our opinion, this plays an important role for an organization's strategy on improving its websites' accessibility in the long-term. Note that website administrators often have little experience in the field of accessibility.

We appreciate that a single score, as determined in our study, cannot adequately describe and rate a tool's functionality and usefulness for a particular context of use. It is reasonable that one and the same tool may not be appropriate for one context, but quite useful in another context. In the remainder of this section, we give a supplementary qualitative assessment of the tools and their individual strengths and weaknesses, as observed in our study.

Pope Tech is beginner friendly. It is intuitive to use, easy to learn, and offers an attractive user interface. This is reflected by its second rank in the user study concerning user experience. Pope Tech is visually appealing, and through the consistent use of meaningful icons and colors, users can quickly find their way around websites that need to be checked. By clicking these icons, they can see all the information they need about accessibility issues including where they are, how to fix them, and what impact they have on persons with disabilities. Even inexperienced users are led directly to the error. The structure of a website can also be easily displayed. However, it lacks the ability to assign issues to team members, which could make working difficult in large teams.

The advantages of *ARC Monitoring* are that archiving works well; it is possible to see what has changed at any time during each scan. The integration of the knowledge database and the newsfeed in the AMS is handy. It offers many visually appealing diagrams. The user can export insightful reports in which useful information and tips are extracted from the scans and added to a task list. The rulesets are customizable and can be chosen between different ones. When using the ARC ruleset, the user will find many errors, but these may contain FPs, while the axe ruleset does not find FPs, but fewer errors in general. These can also be used in combination, thus balancing the disadvantages of both rule sets. Together with the browser extension "ARC Toolkit", it offers some convenient features for manual checking, such as visually displaying the tab order and checking page reflow and text spacing. Clear goals can be defined for teamwork with a progress indicator and a deadline. However, no errors can be assigned to other users.

axe Monitor is rather designed for experts, who know the tool in detail. Such features include recordable scripts to scan processes, setting scopes to avoid checking the same errors on different pages, and configurable agents to scan different versions of a website, etc. It also offers a lot of functionality for working in a team, like assigning issues to team members with many possibilities to

specify them or connecting to management software like Jira. While axe Monitor does not find many errors, the ones found are certain to be TPs.

Siteimprove is the only tool that natively supports the AAA conformance level of WCAG 2.1. Through the skillful use of gamification patterns and an attractive user interface, Siteimprove was perceived as the most motivating tool in our user study. Finding issues was easy and efficient even for inexperienced users by directly jumping to them from the AMS, highlighting them and displaying a detailed description. The comparative study revealed that Siteimprove was able to check the most guidelines for violations among all tools in our use cases. It scored well on all evaluation criteria except "support for manual checks", where many possible errors were displayed, but hardly any TPs resulted from them. By assigning issues to roles and a possible connection to tools like Jira, Siteimprove is also well suited for working in large teams. With Siteimprove, it is possible to create custom policies. Errors can be found in various ways. The dashboard is highly customizable. Siteimprove and axe Monitor are the only AMS that can also check PDFs for accessibility and validate HTML. Siteimprove also offers many interfaces for extensions or connections of other programs. This makes Siteimprove a tool that is easy to use, but also offers many features for experts.

7 Limitations and Possible Future Work

While our study was quite comprehensive and conducted with a rigid methodology, there are some limitations that should be noted.

The number of websites and samples checked should be increased to obtain an even more representative result. However, this would require more experts to manually check these sites. Due to the pandemic caused by COVID-19, fewer participants were found for the user study than initially planned. With a larger number of participants, it would no longer be necessary for each participant to test each tool (within-subject), but each participant could test a single tool (between-subjects). Thus, it would also be reasonable to give a longer time per tool for each participant so that they can take a closer look at the tools. The fatigue and learning effects would also be lower, even if these were counteracted in this study by counterbalancing.

Additional AMS could be reviewed and additional criteria evaluated in future studies. The choice of the evaluation criteria and their weights were specifically tailored to our use cases at HdM. In future studies, the AMS could be evaluated specifically for other use cases, which could lead to different results.

Not all modules and tools were enabled by the respective companies, which also made the comparison more difficult. Unfortunately, no dummy user could be provided for automatic authentication, otherwise this feature could have been investigated in detail and not just theoretically.

The AMS are constantly being updated, which means that if this study had been conducted later, the results might have been different. We did not test the complete product suites offered by the companies, but rather equivalent parts of them, as much as possible. In further studies, other products or even the complete product suites could be tested.

8 Conclusions

In our study, Siteimprove scored best, based on our framework of evaluation criteria, our weights, and our use cases. Siteimprove was found to be motivating, easy to learn, and powerful in its functionality.

Even though AMS are powerful tools with many functions, our study has shown that manual checking is still essential. The ability of AMS to identify the most common errors across an entire website with subpages and to view the accessibility progress over time is crucial for large websites as well as the possibility to comment on errors, to declare them as non-solvable, and to mark them as non-errors.

Abbreviations

The following abbreviations are used in this paper:

- AMS: accessibility monitoring system
- ARIA: Accessible Rich Internet Applications
- BITV: Die Barrierefreie-Informationstechnik-Verordnung
- FN: false negative
- FP: false positive
- HDM: Hochschule der Medien (translated in English: “Stuttgart Media University“)
- TP: true positive
- UEQ: User Experience Questionnaire
- UX: User Experience
- WCAG: Web Content Accessibility Guidelines

Acknowledgements

We thank the vendors of Deque, Pope Tech, Siteimprove, and The Paciello Group for providing us with free trial versions of their tools and for their technical support. We thank Christophe Strobbe for his support on the validation of TPs, FPs and FNs in the study. We also thank the numerous voluntary participants in the user study who took part despite the difficult conditions of the COVID-19 pandemic.

A preprint of this paper in accessible format is available on the website of the Competence Center for Digital Accessibility (Andreas Burkard 2020).

References

Abascal, Julio and Arrue, Myriam and Valencia, Xabier. "Tools for Web Accessibility Evaluation." In *Web Accessibility: A Foundation for Research*, edited by Yeliz and Harper, Simon Yesilada, 79-503. Springer London, 2019.

Abduganiev, Siddikjon Gaibullojonovich. "Towards Automated Web Accessibility Evaluation: a Comparative Study." *Int. J. Inf. Technol. Comput. Sci. (IJITCS)*, no. 9 (2017): 18-44.

Albert, William and Tullis, Thomas. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics.* Newnes, 2013.

Andreas Burkard, Gottfried Zimmermann, Bettina Schwarzer. „Monitoring Systems for Checking Websites on Accessibility.“ *Competence Center for Digital Accessibility [Preprint].* 5. November 2020. https://digitalisierung.hdm-stuttgart.de/barrierefreiheit/2020/10/26/monitoring-systems-for-checking-websites-on-accessibility/ (Zugriff am 12. November 2020).

BIK-Projekt. *BIK BITV-Test.* 2019. https://www.bitvtest.de/bitv_test/das_testverfahren_im_detail/verfahren.html (accessed September 5, 2020).

Craig, James and Cooper, Michael and Pappas, L and Schwerdtfeger, R and Seeman, L. „Accessible rich internet applications (WAI-ARIA) 1.0.“ *W3C Working Draft*, 2009.

Deterding, Sebastian and Dixon, Dan and Khaled, Rilla and Nacke, Lennart. „From game Design Elements to Gamefulness: Defining Gamification.“ In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 9-15. 2011.

European Commission and others. *Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies (Text with EEA relevance).* 2016.

European Commission. „Commission Implementing Decision (EU) 2018/2048 of 20 December 2018 on the harmonised standard for websites and mobile applications drafted in support of Directive (EU) 2016/2102 of the European Parliament and of the Council.“ *EUR-Lex.* 21. 12 2018. https://eur-lex.europa.eu/eli/dec_impl/2018/2048/oj/eng (Zugriff am 17. 10 2020).

European Disability Forum. "How many persons with disabilities live in the EU?" *European Disability Forum (EDF).* 11 28, 2019. http://www.edf-feph.org/newsroom/news/how-many-persons-disabilities-live-eu (accessed 10 17, 2020).

European Telecommunications Standards Institute. *Draft EN 301 549 - V3.1.1 - Accessibility requirements suitable for public procurement of ICT products and services in Europe.* 2019.

Fahrmeir, Ludwig and Hamerle, Alfred and Tutz, Gerhard. *Multivariate statistische verfahren.* Walter de Gruyter GmbH \& Co KG, 2015.

German federal ministry of labour and social affairs. „Barrierefreie Informationstechnik-Verordnung (BITV).“ *German federal ministry of labour and social affairs.* 25. 05 2019. https://www.bmas.de/DE/Service/Gesetze/barrierefreie-informationstechnik-verordnung-2-0.html (Zugriff am 17. 10 2020).

Goodman, Elizabeth and Kuniavsky, Mike and Moed, Andrea. *Observing the user experience: A Practitioner's Guide to User Research.* Elsevier, 2012.

Laugwitz, Bettina and Held, Theo and Schrepp, Martin. "Construction and Evaluation of a User Experience Questionnaire." Edited by Springer. *Symposium of the Austrian HCI and usability engineering group*, 2008: 63-76.

Majuri, Jenni and Koivisto, Jonna and Hamari, Juho. "Gamification of education and learning: A review of empirical literature." *Proceedings of the 2nd international GamiFIN conference, GamiFIN 2018* (CEUR-WS), 2018.

Nielsen, Jakob. *Usability Engineering.* Edited by Morgan Kaufmann. 1994.

Pădure, Marian and Pribeanu, Costin. "Exploring the Differences Between Five Accessibility Evaluation Tools." 2019.

Schmutz, Sven and Sonderegger, Andreas and Sauer, Juergen. „Implementing Recommendations From Web Accessibility Guidelines: Would They Also Provide Benefits to Nondisabled Users.“ *Human Factors*, 2016, 58 Ausg.: 611-629.

Siteimprove. „Democracy, Digital Accessibility, and EU Member Parliament Websites.“ 18. 2 2019. https://siteimprove.com/media/7147/accessible-report-eu-democracy.pdf (Zugriff am 17. 10 2020).

Solovieva, Tatiana I and Bock, Jeremy M. „Monitoring for Accessibility and University Websites: Meeting the Needs of People with Disabilities.“ *Journal of Postsecondary Education and Disability*, 2014: 113-127.

Statistisches Bundesamt (Destatis). *7,9 Millionen schwerbehinderte Menschen leben in Deutschland.* June 24, 2020.

Velleman, Eric and Abou-Zahra, Shadi. „Website Accessibility Conformance Evaluation Methodology (WCAG-EM) 1.0.“ *W3C Working Group Note. http://www. w3. org/TR/WCAG-EM*, 2014.

Vigo, Markel and Brown, Justin and Conway, Vivienne. "Benchmarking Web Accessibility Evaluation Tools: Measuring the Harm of Sole Reliance on Automated Tests." *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, 2013: 1-10.

World Wide Web Consortium and others. "Web Content Accessibility Guidelines (WCAG) 2.1." World Wide Web Consortium, 2018.