PDF Accessibility of Research Papers: What Tools are Needed for Assessment and Remediation?

Aravind Jembu Rajkumar[1], Jonathan Lazar[1], J. Bern Jordan[1], Alireza Darvishy[2], Hans-Peter Hutter[2]
[1]Trace Center, HCIL, College of Information Studies, University of Maryland, USA
[2] ICT-Accessibility Lab, Zurich University of Applied Sciences, Switzerland
(aravind@umd.edu; jlazar@umd.edu; jbjordan@umd.edu; dvya@zhaw.ch; huhp@zhaw.ch)

Abstract

Trillions of documents online are in PDF format, but only a small amount of these PDF documents include the necessary markup to make them accessible for people with disabilities. This paper presents the results of three related data collection efforts: a survey (with 61 participants), interviews (with 6 participants), and usability testing (with 6 participants), to learn more about what tools are needed for content contributors, to assist them in the assessment and remediation of accessibility in PDF documents. The paper provides suggested features and usability needed for software tools to support PDF document accessibility, as well as implications for content creators, scientific publishers, as well as the creator of the PDF format, Adobe.

1. Introduction

The topic of web accessibility has received much attention in the news recently, due to the large number of lawsuits on the topic [16]. A related topic, which has received much less attention, is the accessibility of PDF files. While many scholars, researchers and practitioners are aware of Web Content Accessibility Guidelines WCAG 2.1, few are aware of the PDF accessibility guidelines, known as the PDF Universal Accessibility (U/A) Guidelines, AKA the "Matterhorn protocol."

Our focus in this article is on PDF accessibility for STEM (Science, Technology, Engineering, and Mathematics) practitioners and researchers, because of their unique needs for complex graphics, tables for representing data, and mathematical formulae. Furthermore, many of the same needs for presenting complex data graphics and financial formulae also occur in the financial and business areas. However, unlike in the more business-oriented fields, there is already strong interest in various STEM fields in making research papers accessible for people with disabilities [13]. For instance, some of the Association for Computing Machinery (ACM) Special Interest Groups (e.g. SIGACCESS and SIGCHI) include accessibility information in their research paper templates and have accessibility chairs for conferences. Many technology companies are increasing the availability of accessibility features in their tools [9], including those involved in STEM education. A recent report states that about 10% of employed scientists and engineers report one or more disabilities, including limitations in hearing, vision, cognitive ability, or independent living [8]. For all of these reasons, this paper will focus on understanding the needs of content creators in STEM fields, related to software tools for making their PDF documents more accessible for people with disabilities.

2. Previous work

2.1 About PDF

PDF was first created for printing purposes by Adobe in the early 1990s. PDF is not really designed to be a content presentation tool. It is a graphical layout model, analogous to using CSS but without the HTML semantic markup. There are over 2.5 trillion PDF documents online [11], and we do not know how many of them are accessible. According to the World Health Organization, there are 36 million people who are blind and 217 million people who have moderate to severe visual impairments, which may restrict them from using computer screens [12]. Individuals with low vision [19] and/or learning disabilities [17,18] also frequently use screen readers and may find inaccessible PDF documents hard to use.

In two previous research studies, blind users mentioned that inaccessible PDF files were a barrier, but in both studies, the main focus of the research was on the broader topic of web accessibility, which often gets more attention than PDF accessibility [20],[21]. PDF accessibility is often the "step-child" of digital accessibility. For instance, a hot topic in the news currently, is the concept of "open access" journals, so that the public has free access to scientific publications. This attention to open access often focuses on the

URI: https://hdl.handle.net/10125/64254
978-0-9981331-3-3

benefits to the public, and the financial costs imposed by large publishers of scientific publications such as Elsevier. Yet rarely in discussions of open access, do you hear the question of whether these publications are really open to all users. If a scientific research article is free to the general public, but is not in an accessible format, it is not actually "open" to a portion of the population.

It is necessary to understand the various steps involved in creating a PDF file. Many authors start with word processing documents, such as MS-Word format. When a word processing document is "printed" to a PDF file, much of the semantic markup (which is needed for accessibility) present in the word processing file disappears. Some content creators print to a PDF file within MS-Word, others use free online tools, and some actually own the full paid copy of Adobe Acrobat Pro, within which you can load a MS-Word document and create a PDF document. These various scenarios create PDF documents with various levels of accessible markup present. For instance, the least accessible PDF file would be a word processing document that was printed to paper and then graphically scanned (100% graphic with no accessibility markup). The PDF file with the most accessibility markup present (although still not fully accessible) would be a PDF document created natively in Adobe Acrobat Pro, but only in the Windows version, since the Mac version does not include as much of the original markup. Regardless, additional enhancements to the file to make it accessible will still be necessary (discussed in later sections). This is one factor that makes PDF accessibility different from web accessibility (where you can proactively design using standards and existing tools to make fully accessible web content). PDF accessibility generally requires remediation after the content has been created, even when created in the right way.

PDF U/A (PDF Universal Accessibility), also known as the "Matterhorn protocol," is a set of 31 checkpoints comprised of 136 failure conditions. This is a technical specification (adopted by ISO), intended for developers implementing PDF writing and processing software. PDF/UA provides definitive terms and requirements for accessibility in PDF documents and applications [10]. Conformance with PDF/UA ensures accessibility for people with disabilities who use assistive technology such as screen readers.

Adding alternative text to images often exposes weird bugs in Acrobat [4]. For instance, it is quite common for an image to change size in the PDF upon adding alternative text, or for tables to lose their position when tagged. Acrobat Pro does not include an undo function for any of this functionality, and so if users make a mistake, they must start over. Inserting alternative text for images also frequently introduces a new font that is not embedded in the document (probably a vestige of PDF's past life as a printing format). Making a table accessible in Adobe Acrobat often messes up its position in the paper. But with no 'undo' feature in Acrobat, many authors simply give up in frustration.[4]

Many people use Microsoft Word for word processing, and the Windows-based version helps in creating accessible documents, for example, Word has a feature that allows rich charts and figures (with markup) to be copied and pasted into Word documents from other Microsoft Office programs like Excel (but MS-Word for Mac has less accessibility support). [4]

Assistive technologies work well with text but are less reliable when dealing with non-textual components like graphs and formulae. Though formulae can be presented using 8-dot braille, there is no unified standard braille code for presenting formulae. Graphs and formulae are inaccessible in word documents if not properly marked up. Armano (2014) points out that LaTeX is the most widely used markup language by the scientific community for producing high-quality documents with mathematical contents since the screen and braille readers can access the raw LaTeX documents containing the formulae [5]

If a PDF document is simply a scanned image, a screen reader cannot access any information from it. Nazemi et al (2014) proposed an idea to make a scanned image as useful as a normal PDF by performing PDF Layout Analysis, using OCR to extract text from images. It also involves multiple layers of segmentation to deal with complex scientific equations. [6] However, this approach is still likely to require additional remediation.

Different tools are used to create a PDF document and have different ways of handling the metadata transfer from one file format to PDF file format. Microsoft Word has made many changes recently to make word documents and PDF documents created from MS-Word, more (although not fully) accessible. Using various packages in LaTeX, metadata like title and authors are transferred over while generating PDF documents from LaTeX. Using Action Wizard for PDF in Adobe's InDesign tools, it is possible to include some semantic markup needed for accessibility in a PDF document. While the metadata transfer from these tools to PDF is somewhat successful in the Windows operating system, the same cannot be said for the Mac OS [4].

2.2 PDF format within STEM fields

PDF is currently the preferred format for conference proceedings in many STEM fields. For instance, the Association for Computing Machinery (ACM) generally follows the two-column layout, PDF format

approach. It is argued that this two-column approach is not ideal for screen readers [1], and even for people without disabilities, it's hard to read a two-column PDF document on a tablet or a mobile phone [2]. The two-column approach followed by ACM (and many other digital libraries) frustrates blind users when the PDF file is inaccessible, as screen readers may jump from one column to the other across the page, thus breaking the reading order.

Nganji (2015) studied the accessibility of papers of four major disability-related publications between 2009 to 2013, sampling 200 papers. The author based his study on the 11 criteria set by the WCAG 2.0 (although PDF U/A would be the ideal way to evaluate a PDF document, as the strategies in WCAG may not map directly to PDF documents). Of the documents that Nganji evaluated, 95.5% were not tagged, and 97% had no alternative text for the images. Only 13.7% of the articles evaluated had meaningful title metadata. Of the four journals, Elsevier's *Research in Developmental Disabilities* was the most accessible [3]. Another study compared the PDF accessibility of the CHI, ASSETS, and W4A conference proceedings. CHI was the least accessible of the three conferences in 2014, although ironically 2014 was the highest level of paper accessibility for CHI over a 6-year period, due to specific actions taken by the program committee. ASSETS consistently had a higher percentage of paper accessibility, as it is a requirement for authors. After W4A introduced guidelines to make PDFs accessible in 2011, the number of tagged documents rose to 100% in 2014 [1]. There are many different approaches that can be used to encourage or enforce PDF accessibility of papers.

2.3 Tools for PDF evaluation and remediation

Most of the tools available for PDF remediation are not open source and usually are paid versions. Even paid applications like Adobe Acrobat Pro are not necessarily user-friendly and bug-free. While there are some tools and websites that assist in evaluating PDF documents (like PAC 3, CommonLook PDF Accessibility Software, European Internet Inclusion Initiative's PDF checker, WebAIM's WAVE) and can indicate where problems exist, there are limited tools that can remediate the accessibility issues identified. The most commonly used tool is Adobe Acrobat Pro/DC. The accessibility toolkit is a part of the paid version (not the free Adobe Acrobat Reader), which very few people have access to.

Another tool called PAVE (PDF Accessibility Validation Engine) is a free web-based software tool that supports remediating accessibility issues. PAVE was developed by the ICT-Accessibility Lab of the Zurich University of Applied Science. It allows manual tagging in addition to automatic tagging. For example, users can draw rectangular boxes to group the paragraphs together [7]. The learning curve for PAVE can be steep and from our pilot study (described below), we have heard that it can be overwhelming to users, hard to understand, and is not intuitive.

3. Research methodology

We had two overarching research questions: in general, what are the needs of STEM researchers and practitioners when it comes to the development of PDF accessibility tools that would be useful to them, and more specifically, what improvements do STEM researchers and practitioners need in a prototype tool for remediating PDF accessibility called PAVE. While it has been documented in the research literature that PDF accessibility is a problem, there are no previous studies examining potential solutions in terms of software tools. We decided to use a multi-method approach which consisted of simultaneous surveys, interviews, and usability testing with the one existing free PDF accessibility remediation tool, PAVE. For the study, we wanted to recruit participants representing the diversity of STEM fields. The recruitment emails were sent to the various University of Maryland and University of Washington email listservs and to some of the Association for Computing Machinery (ACM) Special Interest Groups (e.g. SIGACCESS and SIGCHI). The surveys were also posted in the Facebook groups and social circles so that it reached a wider and more diverse set of STEM researchers and practitioners. All the surveys and usability testing were conducted online. Interviews were conducted in person and over the phone.

We had a goal of diversity across STEM fields as well as diversity in terms of experience and expertise on pdf accessibility. We were more successful in meeting both goals in the interviews and usability studies, than in the survey. All the participants were unique across all the methods except for one participant who participated in both a survey and an interview.

3.1. Development of research materials

3.1.1 Surveys. To reach a wider (and geographically distributed) set of participants, we created an online survey, to understand how familiar individuals are with the concepts of assistive technologies and PDF accessibility. In the survey, we had 5 sections – Awareness about PDF accessibility, Tools, Guidelines, Suggestions and needs for future development and Contact information. The survey asked how frequently individuals remediated their PDF documents before submitting for publication, which tools they used most

often for evaluating and remediating accessibility issues, and the time taken per paper to correct all the accessibility issues. Sample questions from the survey include:

1. Of the scientific papers that you have submitted for review in the past five years, how many of them were submitted in PDF format?
2. How frequently do you evaluate your PDF paper submissions for accessibility (like reading order, alt text, captions, etc.) before you submit them?
3. If you find accessibility problems in your paper, how often do you fix (remediate) them before submitting the paper?
4. If you do not make all of your PDF files accessible, what is the biggest reason for not doing so?
5. If you fix (remediate) the PDF accessibility problems in your file, which tool do you use?

3.1.2 Interviews. Unlike surveys, interviews are a qualitative method which allows for deeper understanding of issues, as well as follow-up questions when interesting patterns emerge [15]. The interview questions were similar to the survey questions but were more detailed. We wanted to probe into what content authoring tools the authors use, their process of evaluating and remediating their PDFs, the various guidelines that they have used to follow accessibility templates, and generally, the "why" of their process, rather than just the "what." Some of the interview questions included:

- "What is the bare minimum that you do to make PDFs accessible"
- "If you find any accessibility problem (like improper reading order, missing alt text and captions etc), how do you remediate it before submitting? If not, why not?"
- "Who should be responsible for making documents accessible if they are not accessible? Authors or publication/journal?"
- "Have you been informed by the publishers that you need to make your PDF-formatted papers accessible?

These questions helped with understanding the participant's awareness of accessibility and their interaction with the publishers.

3.1.3 Usability Studies. PAVE is currently the only free online tool that helps in evaluating and remediating PDF accessibility issues. We wanted to get a better understanding of the usability of the tool, to understand what functionality and usability is needed for tools in general, and to provide feedback specifically to the PAVE tool developers. The task list for the usability testing was developed, keeping in mind the various elements (table, figures, lists) and document properties (title, language and author of the papers) required by the Matterhorn protocol to be considered as an accessible PDF. As is typical for usability testing, both task and time performance data were collected. The usability testing was all done in-person.

3.2. Pilot studies

We conducted pilot studies with 5 researchers and practitioners in STEM fields, for each of the survey, interview scripts, and usability testing materials. The following were our observations and resulting modifications:

1. A consistent 3-point scale (e.g. *Not at all familiar*, *somewhat familiar* and *Very familiar*) was used wherever we wanted to understand the knowledge of the participant. (
2. Since many people were unaware about PDF accessibility, we re-framed our questions in a way that they were specific about PDF accessibility issues like reading order, captions, alternate texts and language of the PDF, rather than just mentioning PDF accessibility in general.
3. Based on the pilot interviews and feedback from experts in the field of accessibility, we reworded some questions and included additional insightful questions:
   a) *"Who should be responsible in making the PDF accessible – Authors or Publishers?".*
   b) *"What is the bare minimum you do to make your PDF document accessible?"*
4. We observed that pilot participants who watched the video tutorial performed the tasks better than those who did not, so it was decided to have all participants watch the video tutorial before attempting any tasks.

4. Results

A total of 61 people responded to the survey (all remotely), 6 people took part in the interviews, (5 of the participants were remote and one in-person), and 6 people took part in the usability testing (all in-person) on the PDF accessibility remediation tool called PAVE. It is important to acknowledge an inherent bias in the response to the call for participation—most people who took part had at least heard of PDF accessibility and felt comfortable enough to respond. It is unlikely that someone who had never heard of PDF accessibility or never considered it, would have been willing to

participate, even though we attempted to get a diverse sample in terms of PDF knowledge. That means that our participants, while representing a diverse set of STEM disciplines, is probably more educated about PDF accessibility than the general population. Table 1 presents the diverse set of participants from various STEM disciplines who took part in the 3 data collection methods:

STEM Discipline	Survey	Interview	Usability Testing
Computer Science	30		
Information Science	17		2
Engineering	9		
Chemical Engineering		1	1
Civil Engineering		1	
Biochemical Engineering			1
Mechanical Engineering			1
Physics	1	2	
Chemistry		1	1
Biology	1		
Psychology	1		
Environmental Science	1		
Other	1	1	

Table 1 – Participant distribution across the three data collection methods

4.1 Survey Results

A total of 61 participants responded to the survey. Of participants who responded to the survey, 85.2% of the participants responded that they are aware of or have heard about PDF accessibility. 88.5% of participants submitted at least some scientific papers in PDF format in the last 5 years.

Of the 54 participants who submitted their work in PDF format, only 16.6% of them "always" evaluated their PDFs for accessibility problems. Of the participants who "always" check for accessibility errors, 66.6% "always" *fix* those accessibility issues (meaning that, even when aware of accessibility problems and always checking, 33.3% of participants who check for accessibility problems don't fix them). This data is represented in figure 1, a funnel chart showing the number of participants who actually fix their accessibility problems.

Adobe Acrobat Pro is used by 55.7% of the participants making it the most popular tool to fix accessibility issues, but 21% of the participants who said they use Adobe services have mentioned that they are not satisfied with the tool. Apart from this, 6.5% of the participants said that they use Microsoft Word's accessibility checker to fix accessibility issues (which may improve the resulting accessibility in a PDF file if one is using Windows, but not Mac).

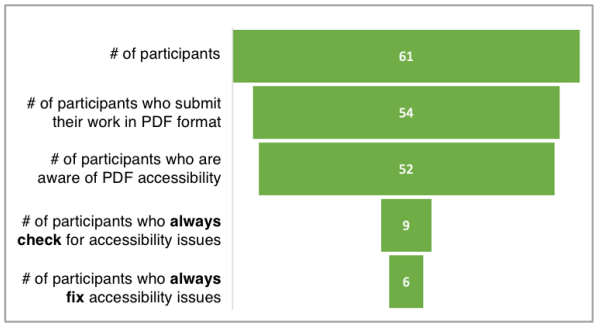

Figure 1 – Funnel chart showing the number of participants who fix their accessibility issues.

Only 37% of the participants said that they have been contacted by publishers to correct accessibility issues. Of those participants, 81.8% of them said they either "always" or "often" fix these issues. Out of all participants, 70.5% want the publishers to take responsibility for making the PDF documents accessible. LaTeX, Microsoft Word, MathType extension for MS Word, ChemDraw, Chem 3D, Microsoft Equation Editor are popular content authoring tools used by the authors.

4.2 Interview Results

A total of 6 participants were interviewed for this study. Some of the data collected during the interviews echoed the results from the survey, related to lack of awareness and knowledge on PDF accessibility. For instance, half of the interview participants said that adding alt text makes the document accessible (in actuality, alt text is just one part of document accessibility). One participant said, *"I always provide alt text to my images. I don't do anything more than that. Unless the publishers ask me to do so, which they don't. But I always insert alt text whenever possible"* [IP1].

Five interviewees from engineering, physics and chemistry, said that they usually do not submit their work in PDF file format to their publishers. They either use Microsoft Word or LaTeX file format. The authors who submit their work in MS-word format get a PDF rendered from the publishers for confirmation. If there are any errors or updates, the authors will provide an updated Microsoft word or LaTeX file to the publishers.

Therefore, these participants stated that they have no control over creating an accessible PDF. One participant said, *"I cannot make the PDFs accessible because I only submit Word document[s]. The PDF is generated by the publishers"* [IP3].

Other interviewees noted that even if the authors provide a PDF with accessibility features marked up to the publishers, the final PDF generated by the publishers often are inaccessible. *"I don't make it accessible. Firstly, I am not required to. Once I tried to make it accessible and when it was published, my colleague was not able to read it with a screen reader. I felt too bad."*[IP4].

Five of the interview participants said that they have not been asked by publishers to make the PDFs accessible. Many of the templates provided by the publishers have no accessibility component involved. One participant said "*We don't have templates. That's the problem."* [IP2] Participants have mentioned they do not see the necessity of accessible PDFs. *"I don't see the need for making the PDFs accessible because I feel that my audience are not visually challenged, at least that is my assumption"* [IP4].

Participants who knew about PDF accessibility often did not make their PDF documents accessible because they lacked the knowledge to make them accessible. One participant said that they "google" to figure it out, but it does not always work [IP4]. Out of the 6 interview participants, only 2 of them make their PDF documents accessible (one always and the second sometimes makes PDFs accessible).

When trying to make PDF documents accessible, interview participants reported difficulties. Lack of knowledge and tools were the biggest hurdles to make PDFs accessible. Participant IP2 said, *"Adobe Acrobat Pro is a poorly designed tool. It is not at all user friendly…[Time taken to remediate] depends on the type and complexity of the PDF."* Most of the participants are unhappy with the tools available currently for remediating accessibility issues. Four interview participants felt that publishers should be responsible for making accessible PDFs. Two out of 6 participants said that they need content authoring tools to include accessibility features and they should not require an additional tool for PDF accessibility. One participant said, *"It is a shame we need tools to make PDFs accessible. Why aren't the word processing tools natively [making documents] accessible? The government has to threaten to sue these tech companies if they don't make the tools natively accessible."*[IP6]

When asked about the amount of time it takes for them to make the PDF accessible, many factors came into play. The content of the PDF file, complexity of the PDF file and the level of knowledge (document properties, alt text etc.) affected the perceived time taken to make accessible PDFs. For example, one participant noted that they strictly follow the PDF accessibility guidelines, and so it takes close to 2 hours, especially if the document has long tables.

4.3 Usability Testing Results

We conducted usability testing with 6 participants, involving the PDF accessibility remediation tool PAVE. We wanted to get a better understanding of 1) what improvements need to be made specifically to the PAVE tool, and 2) what functionality and usability is generally needed in a tool for those responsible for making PDF files accessible? We also wanted to understand what task areas are most challenging and might require new research and development.

Each participant was asked background questions to understand their knowledge on PDF accessibility before the usability testing commenced. All of the participants used their own machines. Out of the 6 participants, 2 of them used desktops. The other 4 participants used their laptops. Both the participants who used desktops were running Windows OS (Google Chrome browser). Of the participants who used laptops, 3 used Windows (2 Google Chrome and one Microsoft Edge browser) and one used the Mac OS (running the Safari browser).

We asked the participants to complete 11 tasks to improve the accessibility of an existing PDF file of a scientific paper (provided to them) using PAVE. The document which users received, had no accessibility markup. The tasks are summarized below:

Task 1 – Change language of the tool to English
Task 2 – Find the meaning of the PAVE acronym
Task 3 – File Upload process
Task 4 – Edit/update document properties
Task 5 – Review instructions on how PAVE works
Task 6 – Make a formula accessible
Task 7 – Make a figure accessible
Task 8 – Correct a wrongly tagged element
Task 9 – Make a table accessible
Task 10 – Make a list accessible
Task 11 – Set elements to be ignored by screen readers (Decorative artifacts or elements that can be skipped by the screen readers).

For the usability study, we selected both general tasks like finding help or FAQ-related (task 1 and 2) and accessibility related tasks (task 4-11).

Standard usability testing metrics of task performance (appears in table 2) and time performance (appears in table 3) were utilized. The task performance table shows how many participants performed the task against a set of completion criteria like Successful, Partly successful, Unsuccessful, Skipped and Not Applicable (that only occurred in task 1).

	Not Applicable	Successful	Partly successful	Un-successful	Skipped
T1	3	2	0	1	0
T2	0	6	0	0	0
T3	0	6	0	0	0
T4	0	6	0	0	0
T5	0	5	0	1	0
T6	0	5	0	0	1
T7	0	0	5	1	0
T8	0	0	4	0	2
T9	0	1	0	4	1
T10	0	5	0	1	0
T11	0	5	1	0	0

Table 2 – Task performance results

None of the participants were able to successfully complete task 7 (Making a figure accessible) and task 8 (Correct a wrongly tagged element). All of the participants tried to complete all of the tasks except for participants P1, P2 and P4. Task 1 was not applicable to P1, P2 and P5 since the page loaded in English language and not in the default German language (it was unclear why that happened).

	P1	P2	P3	P4	P5	P6
T1	NA	NA	9	60	NA	43
T2	20	30	20	74	35	39
T3	22	25	31	33	27	34
T4	57	111	184	140	102	81
T5	37	41	56	67	56	20
T6	S	60	251	185	215	233
T7	77	55	138	81	243	303
T8	130	S	340	S	204	273
T9	103	S	49	42	163	203
T10	142	117	38	45	20	200
T11	50	60	80	47	80	187

Table 3 – Time performance results

Table 3 depicts the amount of time taken by each participant to complete the tasks. Tasks 6 (Making a formula accessible), task 7 (Making a figure accessible) and task 8 (Making a wrongly categorized figure accessible) were the most time consuming. Task 8 (Making a wrongly categorized figure accessible) was skipped by P2 and P4.

5. Discussion and Summary

PDF accessibility is an important problem to solve, but it involves many different stakeholders and many different steps in the content production workflow. Our current research focuses primarily on the needs of the content creators. We split up our discussion into the following sections: knowledge of PDF accessibility, clearer responsibility and guidance, general needs related to PDF accessibility tools, and needs specific to the PAVE tool. We also address implications for scientific organizations and publishers, implications for content authoring tools, and implications for Adobe.

5.1 Knowledge of PDF accessibility

Using the three research methods, we learned much about the existing problem. While PDF is the most widely used file format for the authors to submit their work to the publishers, MS-Word and LaTeX are the preferred file formats in some STEM communities

In our study, we found there was a general lack of awareness about PDF accessibility and lack of knowledge about how to fix these accessibility issues. Some users did not know what PDF accessibility was about and confused "accessibility" with the methods used to "access" (open or download) PDF documents. This confusion may explain the high delta between the number of participants who said they were aware about PDF accessibility and those who check and fix the issues in the survey as shown in Figure 1. Another potential reason for the high delta value is that, most of the survey participants had computer science backgrounds which may have exposed them to the concepts of PDF accessibility. For those who try to make their PDF files accessible, lack of knowledge is a big hurdle. For instance, interview participant IP1 stated that providing alt text alone is enough to make a document accessible (it is not). Of the people who remediate their PDF files, 22.2% mentioned that they do not follow any specific guidelines. There is clearly a need for more awareness and knowledge.

5.2 Clearer responsibility and guidance

Across the three data collection methods, a number of people reported that they felt they do not/should not have responsibility for making the PDF documents accessible. There were multiple reasons given: 1) because their publishers did not require/inform them about accessibility in PDF, 2) they are not required to submit in PDF, and 3) because respondents felt that their target audience generally do not have disabilities. Respondents in our study who have been informed by their publishers about accessibility issues often make those changes. This suggests that with notification and encouragement from the publisher about PDF

accessibility, authors may put in some effort to make accessible PDFs, despite 70.5% of the respondents who feel that it is the publisher's responsibility to make the PDF document accessible.

5.3 General needs for PDF accessibility tools

Researchers and practitioners in STEM fields were generally unhappy with the existing tools to help them fix PDF accessibility issues. A few participants reported that limited access to such tools and the tools themselves are the reason why they avoid making accessible documents. *"I have used Adobe to make pdf accessible. The process was a bit difficult… it needs time to do trial and error and get something fixed."* [SP29]. The users seemed to be overwhelmed by the very tools they need to use in making PDFs accessible.

The most commonly used tool to fix accessibility issues among the respondents was Adobe Acrobat Pro. A few of the respondents from engineering have mentioned that, despite having knowledge on PDF accessibility, they are unable to create accessible PDF documents because Adobe Acrobat pro is not supported by Adobe on the Linux OS. *"I don't know of any other way to test them. I use Linux, so accessibility options that are built in to Office and Adobe products don't work for me."* [SP20].

Another factor reported for not making accessible PDF files is time: 24.5% of the survey participants mentioned time as a big demotivation when it comes to making the PDFs accessible. As an example, authors who use LaTeX usually recompile their paper after making minute changes and generate a PDF file. Therefore, they have to start fixing the accessibility issues from scratch again. This can be avoided by having a scriptable/reusable block that is natively accessible. A few participants mentioned this strategy. For example, the user can drag and drop predefined or custom defined placeholders that can be repeated or duplicated throughout the document. Thus, these blocks will be natively accessible with all the semantic markup built-in. Respondents mentioned that fixing the same accessibility issues repeatedly is time consuming and sometimes problematic as it messes up other formatting requirements for the paper.

5.4 Needs specific to the PAVE tool

Participants who took part in the usability testing of the PAVE tool found it to be challenging to use. Many users were unsuccessful in adding captions, alt text, and tagging tables. From table 2 it is clearly seen that participants struggled with tasks 7, task 8 and task 9. Users found it difficult to provide appropriate alt text, especially for papers which involve graphical figures. The PAVE interface to tag a table properly also was not intuitive. Five out of six participants were unsuccessful in task 9 (Table tagging) where users just highlighted the contents of the table together and unsuccessfully attempted to tag it as table instead of individually tagging each cell of the table first and then tagging the table as table element. This knowledge gap between the user's perception of accessible PDFs and the proper way of making the PDFs accessible using PAVE is a pressing problem. Even though all of the participants watched a 12-minute tutorial, most of the participants (5/6) did not remember how to perform the task of tagging a table. While quick, on-demand video tutorials could aid participants when they are stuck in fixing accessibility issues, ideally a tool should be easy enough to use without having to watch training videos.

Overall, the participants were confused and frustrated while using the PAVE tool. A big reason for the frustration is that the participants did not know how to override the auto tagged elements by the tool, and often weren't even aware of what specifically was being auto-tagged. This may have been why most participants failed to tag the caption for a figure, since the tool automatically tagged the text as "paragraph" and users did not know how to override them. Currently, tags that were automatically tagged by the tool must be deleted and then tagged manually. Based on the usability testing, the following suggestions are made for improving the PAVE interface:

- Redesigning the tool's layout and the naming conventions would be helpful. 4 out of 6 users mentioned that the tool is not user-friendly as the interface is not robust and "cludgy" to use.
- 5 participants reported that the function pane (task, properties, issue details, reading order) and the PDF page often "bounces or jumps" when the user hovers over the contents of the functions pane. This frustrated the users.
- None of the participants were aware of the filter function for the various tags. This can be a very useful and helpful feature, so this feature should be made easier to discover.
- There needs to be increased transparency about which features were auto-tagged by the PAVE tool. Currently, the tool simply notes that issues were addressed, without stating which issues were addressed.
- The PAVE tool itself must be accessible to users who are low vision or Blind (currently it is not, and even participants with sight needed to zoom in due to the small fonts)
- From the post-usability study survey, the participants' perceptions of a tagged table are different from an accessible table. All 5 participants who unsuccessfully tagged a table

thought that they were successful. The tool should provide more information to assist in understanding how to properly tag a table.

- Instead of a full 12-minute video tutorial, having "How-to" videos on-demand may aid the users. 5 participants who watched the videos were not able to remember much of the process from the tutorial. One participant wanted to go back and forth to the tutorial video to perform the task.
- From an operability standpoint, the software or server cannot process many users or complex tasks. A few activities took a longtime to load and the web site was unable to perform or crashed a couple of times during the usability study.
- Including gestures (like drawing a line using the mouse or pointer) to define the reading order—especially in a two-column layout—might be more intuitive than the current interface.
- Showing the various errors or issues in the PDF along with how to remediate them would be useful in educating the users. Adobe Acrobat Pro provides this feature, but PAVE does not.

Overall, the users were happy to see that a tool like PAVE exists but were disappointed by the user experience with using the tool.

5.5 Implications for scientific organizations and publishers

The majority (70.5%) of the survey respondents felt that it is the publisher's responsibility to make the PDF accessible. There are different models for doing this, even within one professional organization (ACM): the ASSETS conference requires that all authors submit fully accessible PDF files. The CHI conference informed each author of the accessibility barriers in their PDF file in 2014, but didn't require that the authors fix them, and this still significantly increased the number of accessible papers. [13]. The UIST conference (a small conference) is having a student volunteer make all of the PDF files accessible. So, there is a need to evaluate the various models of responsibility for PDF accessibility (including, for example, incorporating accessibility information in paper templates), to determine which ones are most successful. There is also a need to increase information flows, and clear lines of responsibility, between content authors and professional organizations/publishers.

5.6 Implications for content authoring tools

Authors from diverse STEM fields use different tools as a part of their workflow. Chemists may use ChemDraw, biologists may use Protein Data Bank, and physicists and mathematicians may use MathType (usually an add-in for MS office) and Microsoft Word Equation editor to include mathematical formula and equations. The respondents, across the data collection methods, felt that there should be a way to integrate these third-party applications inside Adobe Acrobat Pro as an extension, or Adobe Acrobat Pro should be used as a content authoring tool which should natively include accessibility components. All content authoring tools should be improved to be natively accessible and any format they can be exported to, should also be accessible – especially PDFs. When asked what would actually motivate users to create accessible PDFs, one survey participant said, "Fear of being rejected over not complying with submission format requirements."

5.7 Implications for Adobe, creator of PDF format

There were many usability problems reported about the Adobe Acrobat Pro tool, including a poor user interface, no undo option to revert back their actions easily, complicated visualization of the various tags, and the cumbersome way to tag a table. There is also inconsistency in how different versions of Acrobat Pro, on different platforms, and using different source file formats, retain accessibility markup. But one of the largest problems is that these features for making a PDF accessible, are only available in the paid version of Adobe Acrobat Pro, not in the free Adobe Reader available to the public. There is evidence that the popularity of PDF files is decreasing, which may impact Adobe. For instance, Bookshelf, an e-book application with over 20 million users, reports that in 2015, 11 of their top 25 books and 419 of their top 500 books were in PDF format, but by the first half of 2019, none of Bookshelf's top 25 books, and only 137 of their top 500 books, are in PDF format. EPUB3, a fully accessible format, has become the prevalent format: As of May 1, 2019, all of the top 25 books, as well as 60% overall of their inventory, are in EPUB3 format. [14]. In addition, some disability advocates have called for a boycott of PDF format, due to the challenges in making it fully accessible. Unless it becomes easier to make PDF files accessible, there may be a continuing drop of the popularity of the PDF format.

6. Acknowledgements

The work reported in this publication was supported, in part, by grant number 90RE5027 (Universal Interface & Information Technology Access RERC) and 90REGE0008 (Inclusive ICT Rehabilitation

Engineering Research Center), from the National Institute on Disability, Independent Living, and Rehabilitation Research, U.S. Administration for Community Living, Department of Health and Human Services.

7. References

[1] Brady, E., Zhong, Y. and Bigham, J. (2015). Creating accessible PDFs for conference proceedings. Proceedings of the 12th Web for All Conference on - W4A '15, p.34.

[2] Nielsen, J. (2018). PDF: Unfit for Human Consumption. [online] Nielsen Norman Group. Available at: https://www.nngroup.com/articles/pdf-unfit-for-human-consumption/ [Accessed 13 Oct. 2018].

[3] Nganji, J. (2015). The Portable Document Format (PDF) accessibility practice of four journal publishers. Library & Information Science Research, 37(3), pp.254-262.

[4] Bigham, J., Brady, E., Gleason, C., Guo, A. and Shamma, D. (2016). An Uninteresting Tour Through Why Our Research Papers Aren't Accessible. ACM CHI 2016 Extended Abstracts, 621-631.

[5] Armano, Tiziana & Capietto, Anna & Illengo, Marco & Murru, Nadir & Rossini, Rosaria. (2014). An overview on ICT for the accessibility of scientific texts by visually impaired students. In Congresso Nazionale SIREM 2014 (pp. 119-122)

[6] Nazemi, A., Murray, I., and Mc Meekin, D. (2014). Practical Segmentation Methods for Logical and Geometric Layout Analysis to Improve Scanned PDF Accessibility to Vision Impaired. International Journal of Signal Processing, Image Processing and Pattern Recognition, 7(4), 23-36.

[7] Doblies, L., Stolz, D., Darvishy, A. and Hutter, H. (2014). PAVE: A Web Application to Identify and Correct Accessibility Problems in PDF Documents. Proc. Of International Conference on Computers for Handicapped Persons (pp. 185-192). Springer-LNCS.

[8] Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019 | NSF - National Science Foundation. (2019). Retrieved from https://ncses.nsf.gov/pubs/nsf19304/digest/employment

[9] Wait, P. (2018). As tech giants focus on accessibility tools, the equation changes for education | EdScoop. Retrieved from https://edscoop.com/learning-accessibility-tools-mainstream-microsoft-apple-google/

[10] Wikipedia contributors. (2019, May 29). PDF/UA. In Wikipedia, The Free Encyclopedia. Retrieved 18:38, June 12, 2019, from https://en.wikipedia.org/w/index.php?title=PDF/UA&oldid=899358198

[11] Do you know how many PDF documents exist in the world? (2019). Retrieved from https://itextpdf.com/en/blog/technical-notes/do-you-know-how-many-pdf-documents-exist-world

[12] Vision impairment and blindness. (2018). Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment

[13] Lazar, J., Churchill, E., Grossman, T., Van Der Veer, G., Palanque, P., Morris, J., & Mankoff, J. (2017). Making the field of computing more inclusive. Communications of the ACM, 60(3), 50-59.

[14] Personal communication with Rick Johnson, Vice President, Product Strategy, BookShelf.

[15] Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction (2nd ed.). Cambridge, MA: Morgan Kaufmann/Elsevier.

[16] Harris, E. (2019). Galleries From A to Z Sued Over Websites the Blind Can't Use. Retrieved from https://www.nytimes.com/2019/02/18/arts/design/blind-lawsuits-art-galleries.html

[17] Working Together: People with Disabilities and Computer Technology | DO-IT. (2019). Retrieved 6 September 2019, from https://www.washington.edu/doit/working-together-people-disabilities-and-computer-technology

[18] Learning Disorders. (2019). Retrieved 6 September 2019, from https://accessibility.psu.edu/accommodations/audience/learningdisorders/

[19] Low Vision. (2019). Retrieved 6 September 2019, from https://accessibility.psu.edu/accommodations/audience/lowvision/

[20] Tomlinson, S. M. (2016). Perceptions of accessibility and usability by blind or visually impaired persons: a pilot study. In Proceedings of the 79th ASIS&T Annual Conference, p. 120.

[21] Lazar, J., Allen, A., Kleinman, J., & Malarkey, C. (2007). What frustrates screen reader users on the web: A study of 100 blind users. International Journal of Human-Computer Interaction, 22(3), 247-269.