Creating Accessible PDFs for Conference Proceedings

Erin Brady
University of Rochester
Rochester, NY
brady@cs.rochester.edu

Yu Zhong
University of Rochester
Rochester, NY
zyu@cs.rochester.edu

Jeffrey P Bigham
Carnegie Mellon University
Pittsburgh, PA
jbigham@cmu.edu

ABSTRACT

A responsibility we have as researchers is to disseminate the results of our research widely. A primary way we do this is through research publications. When these publications are not accessible to everyone, some readers will be excluded and the impact of our research limited. In this paper, we explore this problem in two ways. First, we report on the accessibility of 1,811 papers in the technical program of several top conferences related to accessibility and human-computer interaction. Second, we reflect on our experience making papers accessible for any CHI 2015 author who requested it. We offer thoughts on research challenges and future work that may make our community's research more accessible.

1. INTRODUCTION

Though Portable Document Format (PDF) was created so documents would be readable across platforms, the content of PDFs is not as inherently accessible as other publishing formats, like raw text or HTML. PDFs are often unreadable by screen readers if incorrectly annotated, which excludes readers with disabilities from accessing their content. Not all authors are familiar with accessible authoring practices for PDFs, so their scientific documents are inaccessible.

In this communications paper, we discuss the accessibility of PDFs in general. We then go on to present the results of an automatic analysis of PDF accessibility for four years' worth of proceedings from CHI, ASSETS, and W4A, and a manual analysis of the accessibility of technical papers published at ASSETS 2014 and W4A 2014. We discuss our experiences working on making other authors' submissions to CHI 2015 more accessible, and present discussion on how to improve conference accessibility in the future.

2. RELATED WORK

2.1 Portable Document Format (PDF)

Portable Document Format (PDF) is a file format used so documents look the same across different systems, regardless of their software or hardware configurations. While PDFs could be read across multiple devices, their content was not accessible with screen readers. From 1993 when PDFs were introduced, until the release of version 1.4 in 2003, there were no structural tags on elements in PDFs, meaning that automated tools could not access the underlying data.

An extension of PDF format is PDF/Universal Accessibility standard (PDF/UA)[1]. Described in International Standard ISO 14289, PDF/UA provides concrete guidelines for creating accessible PDFs. The specification include requirements for the content creators to add to their documents (e.g., accurate tags, alternative text), requirements for the software that displays the PDFs (e.g., making all content from the document available to the screen reader), and requirements for compliant assistive technologies (e.g., ability to recognize and output all tags in a document) [2].

Despite the usefulness of this specification, the details of the requirements for content creators do not lend themselves to being checked automatically. Automatic checkers can see if documents meet the technical and syntatical requirements, but are unable to know the true structure of a document, and cannot verify that sections are tagged correctly or that figures are accurately described [2]. This limitation means that checking to see if a document meets PDF/UA standards involves both an automatic and manual check, which may be time-consuming and be hard for non-experts to do.

2.2 Guidelines for Creating Accessible PDFs

For authors unfamiliar with accessibility, their first introduction to the concept may come from conference-provided guidelines to making an accessible PDF. Indeed, both W4A and CHI offer guidelines for adding accessibility information to PDFs, though these guides are recent additions - the first appearance of W4A's guidelines was on the conference website for 2011, with CHI's guidelines following by the 2014 conference. While these guides are useful as a high-level introduction to accessibility, they are intentionally kept short to avoid inundating novices, instead providing overviews of certain topics and supplementing them with links to the full specifications created by Adobe, WCAG, and WebAIM.

Adobe has provided their own guidelines for verifying and correcting PDF accessibility[2]. Other groups, like WCAG [6] and WebAIM [7], have created their own guidelines to help document editors add accessible tags and metadata to their documents, while individual authors have also created

W4A '15 May 18 - 20, 2015, Florence, Italy.

http://dx.doi.org/10.1145/2745555.2746665 ...$15.00.

[1]http://www.iso.org/iso/catalogue_detail?csnumber=64599
[2]http://www.adobe.com/accessibility/products/acrobat/training.html

digital books [5]. These guidelines are often complex and extremely detailed - the Adobe XI accessibility guidelines, when totaled, are 188 pages long, and McCall's digital book is over 800 pages. For a novice, this volume of information may be overwhelming.

2.3 Academic Research on PDF Accessibility

Outside of web-based guidelines, other research has looked at the accessibility of PDFs and the impact of inaccessible documents on screen reader users.

Hewson and Tonkin used automated tools to evaluate accessibility of a repository of academic documents and found that 10% of the documents had used PDF tags to provide structure [3]. However, to our knowledge our paper is the first to discuss PDF accessibility specifically within the context of ACM conferences.

Other work has examined how the accessibility of PDFs impact screen reader users. A study with 100 blind screen reader users found that inaccessible PDFs were one of the major causes of frustration when browsing the web [4]. Lazar et al. point out in the paper that the problem with PDFs is not a lack of solutions for accessibility problems in the format, but instead a lack of knowledge or prioritization of accessibility by content authors.

3. ACCESSIBILITY OF RECENT CONFERENCES

To examine the accessibility of recent conference proceedings, we performed a two-fold analysis: (i) a large-scale automated check for accessibility on 1811 papers from the last four years of W4A, ASSETS, and CHI conferences, and (ii) a manual examination of accessibility for 26 papers from last years' W4A and ASSETS conferences. This dual analysis is suggested by the PDF/UA compliance checks (Section 2.1).

3.1 Automated Accessibility Check

For the automated tests, we selected all the papers from 2011 to 2014's conferences of CHI, ASSETS, and W4A (both the technical and communications tracks). We generated metadata from the conference proceedings' PDFs using PDF Accessibility Checker[3], a tool which allows easy access to the metadata generated by the accessibility process. Excluding a negligible number of files which failed to work with the Accessibility Checker, we had a collection of 1811 PDFs.

Using metadata from the PDFs, we were able to perform automated checks to see if papers were tagged at all, if any structural tags were present (specifically, H1 and H2 tags, which all conference-format papers should use), and if the document's language was specified as English. These elements of accessibility, specifically whether documents have been tagged or not, are among the simplest indicators of a document's accessibility. We did not perform any checking of the correctness of the tags in this analysis - for example, the presence of an H1 tag does not mean that the tag is used on the title of the document. Instead, this type of analysis was reserved for the manual accessibility check (Section 3.2).

The results of the automated accessibility check we performed are available in Tables 1, 2, 3, and 4. Each conference shows a different trend in accessibility of the proceedings. CHI has slowly increased in having documents tagged from 2011 to 2014, but even in 2014 only a quarter of the documents were tagged. ASSETS, which in 2011 and 2012 had extremely high rates of document tagging, has now slightly decreased from a high of 92% of documents being tagged in 2012 to only 71% in 2011. This may be the result of the growth of the ASSETS community, as new researchers join who are less familiar with how to make their documents accessible. W4A has greatly improved, going from having no tagged documents in 2011 (the year the guidelines were introduced) to 100% tagging over both communications and technical papers last year.

3.2 Manual Accessibility Check

After running automated tests, we examined the accessibility of papers from the W4A technical track and from ASSETS manually, in order to determine how well the automated accessibility represents the true accessibility of the documents. We hoped that these papers would be the most accessible, given that they come from communities of researchers who care about accessibility. We performed this analysis only on papers which had tags present, as the absence of tags would also mean the absence of most of the remaining accessibility indicators.

We limited our analysis to the W4A and ASSETS papers from 2014 to make this analysis feasible while still being able to get a sense of the accessibility of papers from communities where people are familiar with accessibility. Using only the papers from these conferences that were tagged, we analyzed 26 papers (20 ASSETS papers and 6 W4A technical papers).

We began by running a full accessibility check in Adobe Acrobat on each paper. The accessibility check passed for 16 (61.5%) of the papers. This check, recommended as one of the first things to perform by almost all PDF accessibility guides, can catch typical accessibility problems - that a document has not been tagged, images without alternative text, or missing tab order for the page.

73.1% of papers had alternative text for all figures provided, and 84.6% had the proper tab order specified. These high levels of compliance indicate that most authors understand the importance of this accessible information or navigation aids, and will take the time to add them in. However the use of structural tags was haphazard - for example, only 11.5% of papers had the title tagged with an H1. This shows that, while documents may appear accessible due to the presence of tags, they are not always correctly applied, and thus cannot be used as a good indicator of accessibility.

4. MAKING CHI 2015 MORE ACCESSIBLE

As part of our exploration of the accessibility of PDFs, a group of us volunteered to make the camera-ready versions of technical papers for authors[4]. Authors were asked to email us their PDFs and we promised to email them back quickly with an accessible version. We wanted authors to feel free to send us their papers without worry of confidentiality, and so we promised to delete these papers after making them accessible - instead, we discuss below overall themes observed during the tagging process. Overall, we processed 25 PDFs during the first two weeks of January 2015.

While the process of making other people's documents accessible was unfamiliar to us, it quickly became routine. Structural tags are evident from the formatting used in ACM

[3]https://github.com/pdfae/PDFAInspector

[4]http://accessibility.cs.cmu.edu/chi2015/

Accessibility Features of Conference Proceedings from 2014				
Conference	*# Papers and Notes*	*Documents Tagged*	*Heading Tags*	*Language Specified*
CHI	459	26.8%	23.3%	10.9%
ASSETS	28	71.4%	64.3%	35.7%
W4A (technical)	6	100%	100%	83.3%
W4A (communications)	18	100%	77.8%	77.8%

Table 1: Results of the automatic accessibility check for conference proceedings from 2014.

Accessibility Features of Conference Proceedings from 2013				
Conference	*# Papers and Notes*	*Documents Tagged*	*Heading Tags*	*Language Specified*
CHI	393	20.6%	19.3%	0.3%
ASSETS	23	78.3	60.9%	34.8%
W4A (technical)	6	50.0%	50.0%	33.3%
W4A (communications)	15	53.3%	46.7%	33.3%

Table 2: Results of the automatic accessibility check for conference proceedings from 2013.

Accessibility Features of Conference Proceedings from 2012				
Conference	*# Papers and Notes*	*Documents Tagged*	*Heading Tags*	*Language Specified*
CHI	369	17.1%	16.5%	0.0%
ASSETS	24	91.7%	83.3%	53.8%
W4A (technical)	6	0.0%	0.0%	0.0%
W4A (communications)	14	7.1%	7.1%	7.1%

Table 3: Results of the automatic accessibility check for conference proceedings from 2012.

Accessibility Features of Conference Proceedings from 2011				
Conference	*# Papers and Notes*	*Documents Tagged*	*Heading Tags*	*Language Specified*
CHI	409	9.3%	7.8%	0.0%
ASSETS	23	91.3%	82.6%	43.5%
W4A (technical)	6	0.0%	0.0%	0.0%
W4A (communications)	12	0.0%	0.0%	0.0%

Table 4: Results of the automatic accessibility check for conference proceedings from 2011.

templates (e.g., section headers are in small caps and should be tagged H2). Image descriptions were not difficult to generate from the images provided, even without in-depth knowledge of the subject domain. However, the process did require an extensive knowledge of the intricacies of Adobe Acrobat, and much of the time of the initiative was spent learning how Acrobat works or how to work around it's automated approaches, which occasionally modified documents to differ from the original, or made them less accessible.

5. DISCUSSION

Our evaluation of the accessibility of papers from recent conferences (Section 3) demonstrates that only a small fraction of research papers, even in the conferences most related to accessibility, are accessible themselves. We believe that our experience with the process that we followed (as accessibility professionals ourselves) to make CHI 2015 papers accessible shows that it is likely unrealistic to expect authors to make their PDFs accessible given current tools.

We believe it is then natural to ask what approaches we might use as a research community to address this problem. As many of the problems encountered seem to result as a result of PDF being the chosen format, a relatively simple action that conferences could take is to require an alternative format in addition to or instead of PDF. Given the availability of many robust tools for creating accessible HTML content, HTML seems like an excellent candidate. We note that WWW has for many years required HTML versions of papers, and that recent efforts have been made to create CSS stylesheets that mimic the existing ACM formats[5], although it is often argued that the two column format is non-ideal for reading.

Another direction is to make the tools for making PDFs accessible better. Open source tools for manipulating PDFs are underdeveloped, perhaps because the PDF format was initially proprietary. The commercial tools in popular use can only be modified by the companies responsible for them, who have shown little external interest in doing so. The missing features, bugs and usability problems that lead to them being difficult to use have been persistent over several years. Some authors use LaTeX to create their papers. In the past it had been impossible to natively make PDFs created this way accessible, but recent efforts have led to a workable (if imperfect) accessibility package designed to work with the CHI LaTeX template[6]. The regular pattern and repetitive actions could perhaps be automated, using existing tools, *e.g.* Acrobat, in combination with pixel-level macro tools like [8] to perform tagging. If sufficient tools are developed for understanding and manipulating PDFs, many of the accessibility features could be added automatically, *e.g.*, automatically labeling headings [1].

Finally, our time estimates from our CHI 2015 experience suggest that making the (mostly) 10-page CHI papers accessible required approximately 15-20 minutes on average. Making the nearly 500 papers in the CHI technical program accessible would therefore require an estimated 160 hours. If we assume a person (or group) with this expertise could be hired for $30 USD/hour, making all of CHI's technical program accessible would cost less than $5000 USD. For smaller conferences such as W4A (24 papers last year), the program could be made completely accessible for less than $250 USD. For publications with open-access fees, this cost could be included into the cost per author, at a bulk rate rather than the authors needing to hire their own accessibility consultant. It seems that an open question for the community is whether we see making papers accessible as dependent on author initiative, or more similar to other pre-publication steps that we require or that the conference pays publishers to perform.

[5]http://thomaspark.me/2015/01/pubcss-formatting-academic-publications-in-html-css/

[6]https://code.google.com/p/sigchi-latex/wiki/Accessibility

6. CONCLUSION

In this paper, we have explored the accessibility of research papers from several different perspectives. First, we explored the accessibility of papers in the technical programs of several recent conferences related to accessibility, and showed that even in these conferences most of the papers submitted are not created in an accessible way. Second, we explore the process of creating accessible PDFs documents (the format required by all of these conferences), and demonstrate that it is complex and the results opaque. Finally, we argue that without substantially better tools, it is unlikely that authors will be able to make their papers accessible on their own, and offer a number of alternative models that may work better, including using alternate formats with better tool support and hiring consultants to make the papers accessible for authors as part of the publishing process.

7. ACKNOWLEDGMENTS

This work was funded by National Science Foundation Award IIS-1149709, and an Alfred P. Sloan Foundation Fellowship. Thanks go to Anhong Guo, Kenneth Huang, and Luz Rello, who were part of the CHI 2015 accessibility effort.

8. REFERENCES

[1] Brudvik, J. T. and Bigham, J. P. and Cavender, A. C. and Ladner, R. E. Hunting for headings: sighted labeling vs. automatic classification of headings. Proceedings of ASSETS 2008.

[2] Drümmer, O., and Chang, B. PDF/UA in a Nutshell: Accessible documents with PDF. 2014. http://www.pdfa.org/publication/pdfua-in-a-nutshell/

[3] Hewson, A. and Tonkin, E. Supporting PDF accessibility evaluation: early results from the FixRep project. Proceedings of QQML2010.

[4] Lazar, J. and Allen, A. and Kleinman, J. and Malarkey, C. What frustrates screen reader users on the web: A study of 100 blind users. International Journal of Human-Computer Interaction, 2007.

[5] McCall, K. Accessible and Usable PDF: Techniques for Document Authors. Third Edition. 2010. ISBN 978-0-9868085-0-0.

[6] W3C. PDF Techniques for WCAG 2.0. 2012. http://www.w3.org/TR/2014/NOTE-WCAG20-TECHS-20140408/pdf.html.

[7] WebAIM. PDF Accessibility. 2014. http://webaim.org/techniques/acrobat/

[8] Yeh, T. and Chang, T. and Miller, R. C. Sikuli: using GUI screenshots for search and automation. Proceedings of UIST 2009.