# Artificial Intelligence and Accessibility for Administrative Applications

Sara Frug
Legal Information Institute
Cornell University
Ithaca NY United States
sara@liicornell.org

Thomas Bruce
Legal Information Institute
Cornell University
Ithaca NY United States
tom@liicornell.org

## ABSTRACT

In this paper, we suggest that accessibility is an emerging, underfulfilled legal requirement that presents not only a potential locus for activity but also an avenue for research. We describe a proof-of-concept use of machine-learning-based image classification as a managerial support tool for accessibility enhancement, and suggest directions for further research. Although this discussion focuses on the government information landscape in the United States, the adoption of the Web Content Accessibility Guidelines in the European Union extends its applicability.

## CCS CONCEPTS

• Accessibility • Assistive technologies • People with disabilities

## KEYWORDS

Accessibility, Artificial Intelligence, Regulations

**ACM Reference format:**

In: Proceedings of the First Workshop on AI in the Administrative State, June 17, 2019, Montreal, QC, Canada.


Published at http://ceur-ws.org



## 1 Information Accessibility and Government Administration

The availability of government information is well accepted as a requirement for efficient public administration. Machine-readability of administrative information, although frequently acknowledged as a goal, is often neglected. As a basis for accessibility for the disabled, it receives even less attention. This discussion focuses on web accessibility, although it views web accessibility as a consequence of document accessibility. Although this discussion focuses on the United States, the adoption of the Web Content Accessibility Guidelines in the European Union [1] extends its applicability.

### 1.1 Regulatory Requirements

In the United States, the 1998 amendments [2] to The Rehabilitation Act of 1973 [3] explicitly require that federal electronic and information technology (EIT) be accessible to people with disabilities. The regulations promulgated under the 1998 amendments required adoption of standards consistent with (but not identical to) the Web Content Accessibility Guidelines Version 1.0 Level A. [4] In 2017, the regulations were refreshed to incorporate by reference the Web Content Accessibility Guidelines Version 2.0. [5]

### 1.2 Document Accessibility and Web Accessibility Content Guidelines (WCAG)

The Web Content Accessibility Guidelines provide both specific requirements and a general framework for understanding what makes a document accessible. The acronym "POUR" (Perceivable, Operable, Understandable, Robust) summarizes these requirements, the most fundamental of which ensure that information (e.g., words) not be locked in a medium (e.g., a picture PDF) that cannot be perceived by a person with a disability (e.g., blindness). [6]

### 1.3 Non-Compliance

In 2008 (ten years after the 1998 amendments), the Digital Communications Division of the Department of Health and Human Services (HHS) wrote:

> "Section 508 requires that Web sites and associated content created with federal funding, whether internal or external, government- or contractor-hosted, are accessible to persons with disabilities. The law has been in effect since June 21, 2001. Federal compliance – including that of HHS -- has lagged." [7]

By that point, the 2.0 version of the Web Content Accessibility Guidelines was about to be released. HHS's compliance timetable put project completion at 2013.

In 2018, WCAG 2.0 became the standard for Federal websites. The safe harbor provision, however, protected legacy content.

> "This safe harbor provision applies on an "element-by-element" basis in that each component or portion of existing ICT is assessed separately. In specifying "components or portions" of existing ICT, the safe harbor provision independently exempts those aspects of ICT that comply with the existing 508 Standards from mandatory upgrade or modification after the final

> rule takes effect. This means, for example, that if two paragraphs of text are changed on an agency Web page, only the altered paragraphs are required to comply with the Revised 508 Standards; the rest of the Web page can remain "as is" so long as otherwise compliant with the existing 508 Standards." [5]

As of this writing, even Section508.gov and 18F's Accessibility Guide yielded accessibility errors.

Beyond the protection of the safe harbor, government agencies persist in publishing new, non-accessible content. Most prominently, on April 18, 2019, the U.S. Department of Justice released the much-anticipated so-called Mueller Report as an image-PDF, downloadable from a web page that displayed the following notice:

> "The Department recognizes that these documents may not yet be in an accessible format. If you have a disability and the format of any material on the site interferes with your ability to access some information, please email the Department of Justice webmaster. To enable us to respond in a manner that will be of most help to you, please indicate the nature of the accessibility problem, your preferred format (electronic format (ASCII, etc.), standard print, large print, etc.), the web address of the requested material, and your full contact information, so we can reach you if questions arise while fulfilling your request." [8]

Although the most high-profile, this is far from the only example of new, non-compliant content published on federal agency websites.

### 1.4 Publication Practices

The Mueller Report is a good example of a general data impoverishment phenomenon in government publishing, which deserves to be the object of attention from all communities that consume government information. The Mueller Report could not have been drafted as a set of pictures of words; rather, the original, machine-readable document had to have been converted for publication into a set of pictures. This data-impoverishment process is not unique to this document—it can be observed throughout the Code of Federal Regulations. Documents that had to have been authored electronically are converted to pictures for publication, leaving the data consumers to "unscramble the egg" and convert them back into machine-readable data formats.

## 2 Artificial Intelligence and Document Accessibility

Although there is promising work, notably from Rohatgi [9], Wu et al. [10], and Choi et al. [11], to support extraction of machine-readable data from images of charts, graphs, and other data artifacts, for researchers and application developers, common image types have not been addressed systematically.

### 2.1 Pilot Project: Workflows, Experimentation, and Decision Support

LII has begun a pilot project to establish a data conversion workflow and support automation efforts for data-de-impoverishment. The approach has been three-pronged: 1) manually sort and convert figures to SVG and images of equations to MML; 2) annotate SVG images with descriptions of their content; 3) research machine-readable data sources represented as pictures; 4) apply machine-learning techniques to provide decision support for human annotation and conversion.

The pilot project involved collaboration from a specialist in graphics conversion, law and computer science students, and LII's text specialist. The graphics conversion specialist analyzed 14,486 images from the Code of Federal Regulations and sorted them into categories, such as math (6255), diagrams (1410), data tables (1238), maps (3194), forms (1892), labels (351) and logos (77) (some outlier categories, such as photographs, were discovered in the process). Images transformed prior to this project (1149) were sorted into math (241) and non-math (908) and set aside for testing. The images were grouped according to which areas of the CFR they appeared in and prioritized according to how much web traffic each containing document (section or appendix) received on the LII website. As of this writing, the graphics conversion specialist has converted 2913 math elements to MML and 1005 diagrams to SVG format. Also as of this writing, law students have located alternate sources for 2706 images, most notably over 90 images of pages from the 1991 Standards for Accessible Design as Originally Published on July 26, 1991. The data that has been gathered and generated in this process will be reusable for other such endeavors.

In the process of planning our accessibility project, LII discovered the following problems. First, manual annotation of images has proceeded quite slowly compared with other tasks. As of this writing, fewer than 100 image annotations have been completed. Second, math conversion is much faster than SVG conversion. Third, sorting for the purposes of identifying good candidates for SVG conversion produces a different categorization than sorting for purposes of distinguishing similar content.

Because LII wished to deploy newly-accessible content as quickly as possible, we focused on techniques that would enable us to quickly prepopulate a queue with mathematical content, which is easy both to classify and convert. At the same time, the classification process provides additional clues to aid in re-sorting non-mathematical images for further treatment. Using Keras and OpenCV, we trained a classifier on the eCFR images for the purpose of identifying math. Initial results yielded precision 0.86 and recall 0.88. In practical terms, this approach immediately identified 215 out of 243 math images for conversion and incorrectly identified only 35 out of 875 non-math images. This enables us to speed deployment by prepopulating a work queue through automation.

## 2.2 Future Work

The initial proof-of-concept effort simplified the task to address identification of mathematical images and non-mathematical images. This pre-sorting is adequate for cost estimation purposes and makes it feasible to generate machine-readable data before comprehensive sorting is complete.

Because conversion projects frequently include tabular data, forms, and textual images, training the model using additional categories would be quite valuable. Because images may contain mixed content, feature identification and multi-label classification are natural areas for further work.

The initial proof-of-concept effort deliberately eschewed image preprocessing. Characteristics of the images suggest techniques for producing more robust and comprehensive models. For example, basic case-insensitive extraction detected image labels —variants of the terms "figure" (1395), "illustration" (19), "plate" (240), or "legend" (410)—in approximately 14% of the training-set images. Because the choice to annotate within the image rather than within the text surrounding the image should be arbitrary, and because images classified as equations almost never have a legend, it seems worthwhile to purge the image legend before training.

Finally, thus far, LII has not yet taken advantages of metadata external to the images themselves. Because the images in question are embedded within documents that are published on the web, several additional variables could be made available to the model. The training data could include the catchline for the section or appendix within which the image appears; the full structural location of that document; the text, if any, immediately preceding or following the image; terms assigned to the containing document from an unsupervised topic model; terms assigned to the containing Part by the Office of the Federal Register; even variables such as co-location within a single document or volume of web-traffic to the containing document could prove relevant to image type and could be worth testing.

## 3 Caveats and Conclusions

As mentioned earlier, in the pilot study, the greatest impediment to training a model proved to be some subtle and some not-so-subtle differences between the type of classification needed to support professional workflow and the type of classification that would support automated extraction. Because our preferences for populating the queue in this instance were determined by the volume of traffic and co-location of images within a section, several types of content were not distinguished in the initial sorting. For example, where multi-page forms appeared, images containing entirely textual content (such as full pages of instructions) were not distinguished from the form pages for which they provided guidance. Other images, such as tables, typically contained three sections: a caption, the data table, and a set of footnotes. In order to produce useful decision-support tools, training data would best be annotated granularly, identifying features within each image.

Law-and-AI researchers who work on public administration should be aware that the Access Board estimated day-forward web-accessibility compliance resources for the federal government at 5% of web development, software development, and audio-visual production costs, plus an additional 1.25% for evaluation. Should comprehensive conformance become a requirement, the costs will increase accordingly. The Office for Civil Rights of the U.S. Department of Education has, of late, included web accessibility in its enforcement of Section 504 of the Rehabilitation Act, which requires comprehensive equal access to educational services for recipients of federal funding; this means that, as a rule, universities are scrambling to bring their websites into conformance with WCAG 2.0 level AA. [12] Finally, the number of ADA lawsuits treating websites as public accommodations has increased dramatically during the past few years, and a public accommodations case is currently pending before the United States Supreme Court. [13] Reducing data impoverishment in the publication process should limit the need for such work to addressing the challenge of converting non-born-digital images. The combination of labor required and urgency of need makes AI-enhanced automation a timely and valuable avenue for research. Finally, an increased focus on document accessibility can create a virtuous circle in which artificial intelligence applications will both help create, and benefit from, the availability of more machine-readable data.

## ACKNOWLEDGMENTS

Our thanks to the LII development team, Sylvia Kwakye, Nic Ceynowa, Ayham Boucher, and Jim Phillips; to students Mason Roth, Evelyn Hudson, Charu Murugesan and Jiali Liu; to Point.B Studios and Public.Resource.Org; and to Justia, Inc.

## REFERENCES

[1] 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. ELI: http://data.europa.eu/eli/dir/2016/2102/oj.

[2] Electronic and information technology. 29 U.S.C. § 794d. Retrieved from https://www.law.cornell.edu/uscode/text/29/794d.

[3] Pub.L. 93–112, 87 Stat. 355, enacted September 26, 1973), codified as 29 U.S.C. § 701 et seq. https://www.govinfo.gov/content/pkg/STATUTE-87/pdf/STATUTE-87-Pg355.pdf.

[4] Architectural and Transportation Barriers Compliance Board. Electronic and Information Technology Accessibility Standards. 2000. https://www.federalregister.gov/documents/2000/12/21/00-32017/electronic-and-information-technology-accessibility-standards .

[5] Architectural and Transportation Barriers Compliance Board. Information and Communication Technology (ICT) Standards and Guidelines. (Final Rule). 2017. 82 FR 5790. https://www.federalregister.gov/documents/2017/01/18/2017-00395/information-and-communication-technology-ict-standards-and-guidelines.

[6] W3C. Web Content Accessibility Guidelines (WCAG) 2.0. 2008. https://www.w3.org/TR/WCAG20/#intro-layers-guidance.

[7] United States Department of Health and Human Services. 508 Web Compliance and Remediation Framework. 2008. Retrieved by the Internet Archive on 2/6/2018. https://web.archive.org/web/20180206161308/https://www.hhs.gov/web/section-508/compliance-and-remediation/framework/index.html .

[8] Special Counsel's Office. Report on the Investigation into Russian Interference in the 2016 Presidential Election. 2019. https://www.justice.gov/storage/report.pdf.

[9] Ankit Rohatgi. WebPlotDigitizer. Version 4.2. 2019. https://automeris.io/WebPlotDigitizer.

[10] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17). ACM, New York, NY, USA, 1180-1192. DOI: https://doi.org/10.1145/2998181.2998364.

[11] J. Choi, S. Jung, D.G. Park, J. Choo, and N Elmqvist. 2019. Visualizing for the Non-Visual: Enabling the Visually Impaired
to Use Visualization. Eurographics Conference on Visualization (EuroVis) 2019, Computer Graphics Forum, Vol. 38, No. 3. http://users.umiacs.umd.edu/~elm/projects/vis4nonvisual/vis4nonvisual.pdf .

[12] Lindsay McKenzie, Feds Prod Universities to Address Website Accessibility Complaints. 11/16/2018. Inside Higher Education. https://www.insidehighered.com/news/2018/11/06/universities-still-struggle-make-websites-accessible-all .

[13] Lindsay McKenzie, 50 Colleges Hit With ADA Lawsuits. 12/10/2018. https://www.insidehighered.com/news/2018/12/10/fifty-colleges-sued-barrage-ada-lawsuits-over-web-accessibility .