THE CASE FOR NEXT GENERATION TEXT DESCRIPTION SOLUTIONS FOR VISUAL INFORMATION ACCESSIBILITY

Dennis B. Tomashek, MS, Keith D. Edyburn, BSE, Rachael Baumann, BS,
Roger O. Smith, PhD
R_2D_2 Center, University of Wisconsin Milwaukee

ABSTRACT

People with vision impairments are often put at a distinct disadvantage when important information is presented in graphic or visual form. Standards (World Wide Web Consortium (W3C), guidelines, and even laws (Section 508) exist requiring the implementation of text descriptions, yet few websites even meet the suboptimal standards that currently exist. Further, several institutions have developed protocols and guidelines for writing thorough, useable text descriptions, but they are heavily dependent on the author to write, and require much time and expertise. Current advances in computer science, including computer vision and object recognition, and natural language generation provide optimism that in the near future, text description generation will become a mostly automated process, allowing for better access of information for people with disabilities. This paper documents the need and key components of a new approach.

BACKGROUND

Text descriptions are considered to be one of the most important accessibility aids on the internet (Nielson, 2005). In WebAIM's Screen Reader User Survey #4 (WebAIM 2012), only 35% of respondents thought web content accessibility had improved over the previous year. The survey also found that "images with missing or improper descriptions" were the 4th most problematic item for screen reader users (behind Flash content, CAPTCHA, and ambiguous links). However, several major issues prevent text descriptions from being widely used. First, there are billions of images on the internet. For example, in 2010, Flickr hosted 5 billion photos increasing to 6 billion in 2011. In 2010, Facebook users uploaded over 3 billion photographs per month increasing to more than 100 billion photographs monthly in 2011. Instagram users uploaded 3,600 images per minute in 2011 (Pingdom, 2012). These numbers exemplify the enormity of having text descriptions available for all images. Second, useful text descriptions are not easy to write, and can be time consuming. One study found that it takes about 15 minutes for a novice to write a text description for a single image (Maggard, 2008). Third, recently conducted focus groups of blind and visually impaired individuals found sometimes contradictory preferences for the type and amount of information contained in a text description (Baumann & Smith, 2012).

CURRENT STANDARDS AND METHODS OF IMPLEMENTATION

There are several mechanisms for authors to provide text descriptions for images on a HTML webpage. The most widely known and implemented is the `alt` attribute on `<img>` tags which dates back to the first formal specification of HTML, HTML 2.0 (Berners-Lee & Connolly, 1995). However, it was included due to technical issues (e.g. the slow connections of the time), rather than for accessibility. HTML 4.0 made alt required for valid HTML, which helped increase awareness. Since screen readers default to vocalizing the text from the `alt` attribute ("`alt` text") when an image with the attribute is encountered, best practices have developed to improve the experience of users. These best practices include setting blank `alt` text for non-content/decorative images, and keeping `alt` text short and focused.

In order to accommodate longer and more detailed descriptions, the `longdesc` attribute, which indicates a link to a long description of the image, was added in HTML 4.0 (W3C, 1997). Unfortunately, even though the HTML 4.0 specification was published in 1997, `longdesc` is still not widely known or implemented. In fact, an analysis in 2007 found that out of 1 billion images only 0.13% had a `longdesc` attribute (Pilgrim, 2007). Additionally, many images with the `longdesc` attribute did not implement it correctly. From the user side, WebAIM's Screen Reader User Survey #3 found that over 20% responded "I don't know" when asked about the usefulness of `longdesc` (WebAIM, 2011), implying they either did not know `longdesc` existed or had not experienced using it.

Due to the low adoption of `longdesc` the World Wide Web Consortium (W3C) proposed that it be removed from the HTML5 specification. There was a significant response from the web accessibility community, and a list was assembled of 182 real websites using `longdesc` (Carlson, n.d.). The W3C has not yet reached a final decision on `longdesc` in HTML5 (W3C, 2008). Additional methods of text description implementation have been made possible Accessible Rich Internet Applications (WAI-ARIA) standard (W3C, 2011), such as a hybrid `aria-describedby` and `longdesc` (Lembree, 2011).

It is important to note that, by default, web browsers do not inform sighted users of the existence of `alt` or `longdesc` attributes. Thus, sighted users often do not realize that even if tooltips are displayed when hovering over an image (which are specified by the web page author using the `title` attribute), the web page may not be accessible to screen reader users.

Another problem with both `alt` text and `longdesc` are the lack of standards for content. There is no set length of an `alt` text or `longdesc`, and it is up to the author to provide the information.

SURVEY OF ALT TEXT AND LONGDESC ON TOP WEBSITES

The home pages of the internet's top 500 websites, as ranked by Alexa (Alexa, n.d.), were analyzed on January 8, 2013 for their implementation of `alt` and `longdesc` attributes. Out of the 20,138 images found, 4,924 (24.5%) had no `alt` attribute. Only 196 images (0.97%) had a `longdesc` attribute, but all were implemented incorrectly, containing URLs of images instead of URLs of webpages with text descriptions. Out of 2508 images with title text, 1980 (78.9%) had `alt` text which exactly matched the title text, and another 120 (4.8%) had title text but no `alt` attribute.

CURRENT METHODS OF WRITING TEXT DESCRIPTIONS

The Diagram Center has developed the Poet Image Description Tool (Diagram Center, n.d.), to add image descriptions to Daisy audio books, including a step by step process and basic guidelines, including language, context of the image, how the image is being used, and audience level.

The R_2D_2 Center at UW-Milwaukee has created a three-tiered structure Equivalent Text Description (EqTD) protocol for writing text descriptions (R2D2 Center, n.d.), which include "brief", "essential" and "detailed" descriptions. The brief description can be thought of as a title, with enough information to allow the reader to decide whether to move to the next level or skip it. The essential description includes the meaning of the picture, and what is its purpose within the context it is being used. The detailed description contains detailed visual descriptions of the image that could be used by a person who is blind to describe the image to a sighted person.

Both of these protocols are heavily dependent on the author, provide no specific length guidelines, and require significant training and practice for authors to become proficient in following the prescribed guidelines. Given the nearly incalculable number of images being uploaded by millions of users, it is unreasonable to think that enough authors can be trained using these techniques to even begin attacking the problem. Additionally, it is not known if these guidelines actually address the needs of the consumers.

END USER VIEWPOINTS

Two focus groups were conducted to allow blind and individually impaired individuals to comment on what type of content they would like in a text description. The individuals were presented with text descriptions and asked to comment on those specifically. Several themes were evident. All participants agreed on the desire and need for good text descriptions, but differed on the content. Participants commented that too much information was not good, but would rather have too much than none at all. Also, the participants stated that they would like to be able to choose the amount and level of information. Several individuals stated that they did not want to be told the meaning of the image, but would rather form their own interpretation, and thus would prefer more basic description. This was mitigated by the current and past level of vision. For individuals with developmental blindness, visually descriptive words are often meaningless, but they may give important information to someone who has an acquired visual impairment. The context of the image was seen as vitally important. Participants stated that they preferred text descriptions only in cases in which the information was not redundant to the text. Also, images text descriptions were considered essential for images such as maps, graphs, and math and science images, pictures of products for online shopping, and icons and on-screen buttons.

From the hypothetical `alt` text descriptions in Table 1 it is clear that inaccurate and incomplete descriptions of images can impose significant barriers

to understanding. Figure 1 is also an example of an image which should have a long description (unless a table listing results by county is provided).

Table 1: Sample short descriptions for Figure 1

Type	Short Description
Inaccurate	Texas
Incomplete	Wisconsin
Complete	Map of Wisconsin 2012 presidential election results by county

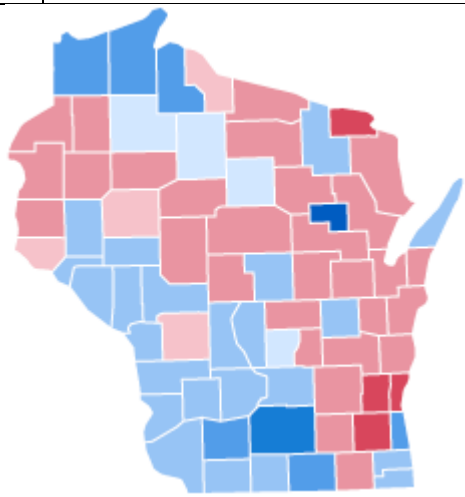

Figure 1: Sample image - Map of Wisconsin 2012 presidential election results by county (Wikipedia, 2012)

CONCLUSIONS

The number of images being uploaded to the internet is immense. Even the current, insufficient standards of `alt` text and `longdesc` are not being implemented. The current strategies for writing text descriptions include extensive training by the authors, and take time. Even then, the text descriptions produced may not meet the needs or wants of the consumers of text descriptions. Thus, the need for a comprehensive, multidisciplinary approach to writing and implementing text descriptions that becomes less dependent on human authors and more automated. Current research in computer vision (Gupta, Srinivasan, Shi, & Davis, 2009; Siddiquie & Gupta, 2010) and natural language generation (Yao, Yang, Lin, Mun Wai & Song-Chun, 2010; Demir, Carberry, & McCoy, 2011).make recognition of objects within a picture and their relationship to each other more feasible. Natural language generation allows for intelligent learning systems that, when paired with computer object recognition, could allow for at least detailed visual descriptions of an image with minimal human input. Computer scientists and web designers must work with accessibility experts and members of the blind and visually impaired communities to develop methods of allowing consumers, including those who use screen readers quick and easy access to multiple levels of content, depending on their desire at the time. Accessibility experts must work with the blind and visually impaired communities to ensure that the content being provided is actually what is wanted.

REFERENCES

Alexa. (n.d.) The top 500 sites on the web. Retrieved January 8, 2012, from: http://www.alexa.com/topsites

Berners-Lee, T. & Connolly, D. (1995, November). Hypertext Markup Language - 2.0. Retrieved January 11, 2012, http://tools.ietf.org/html/rfc1866

Carlson, L. (n.d.) Research: longdesc. Retrieved January 11, 2012, http://www.d.umn.edu/~lcarlson/research/ld.html

Demir, S., Carberry, S., & McCoy, K.F. (2011). Summarizing information graphics textually. *Computational Linguistics*, 28(3), 1-48.

Diagram Center. (n.d.) Poet Image Description Tool. Retrieved January 11, 2013, http://diagramcenter.org/development/poet.html

Gupta, A., Srinivasan, P., Shik J., & Davis, L. (2009). Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2009*, 2012-2019.

Lembree, D. (2011, May 11). Longdesc & Other Long Image Description Solutions — Part 2 of 2: The Solutions. Retrieved January 11, 2012, http://designfestival.com/longdesc2/

Nielsen, J. (2005). Accessibility is not enough. Retrieved January 10, 2013 from: http://www.useit.com/alertbox/access

Maggard, K. R. (2008). *The efficacy of an equivalent text description protocol for increasing the accessibility of information*. University of Wisconsin-Milwaukee, Milwaukee, WI.

Pilgrim, M. (2007, September 14). The longdesc lottery. Retrieved January 9, 2012, from http://blog.whatwg.org/the-longdesc-lottery

Pingdom (2012, January) Internet 2011 in numbers. Retrieved August 17, 2012, from http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/

R2D2 Center. (n.d.) Writing Equivalent Text Descriptions (EqTDs) Posterette. Retrieved August, 23, 2012, from http://access-ed.r2d2.uwm.edu/Entry/2154

Siddiquie, B. and A. Gupta (2010). Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. *Computer Society Conference on Computer Vision and Pattern Recognition 2010*, 2979-2986.

W3C. (2011, January 18). Accessible Rich Internet Applications (WAI-ARIA) 1.0. Retrieved January 12, 2012, http://www.w3.org/TR/wai-aria/

W3C. (1997, December 18). HTML 4.0 Specification. Retrieved January 11, 2012, http://www.w3.org/TR/REC-html40-971218/

W3C. (2008, February 6). ISSUE-30: Should HTML 5 include a longdesc attribute for images. Retrieved January 9, 2012, from http://www.w3.org/html/wg/tracker/issues/30

WebAIM. (2011, February). Screen Reader User Survey #3 Results. Retrieved January 9, 2012, from http://webaim.org/projects/screenreadersurvey3/

WebAIM. (2012, May). Screen Reader User Survey #4. Retrieved January 9, 2013, from http://webaim.org/projects/screenreadersurvey4/

Wikipedia (2012, December). Wisconsin presidential election results 2012. Retrieved January, 11, 2013, from http://en.wikipedia.org/wiki/File:Wisconsin_presidential_election_results_2012.svg

Yao, B.Z., Yang,X., Lin, L., Mun Wai, L., Song-Chun, Z. (2010). I2T: Image parsing to text description. *Proceedings of the IEEE 2010*, 98(8), 1485-1508.