Guiding Novice Web Workers in Making Image Descriptions Using Templates

Valerie S. Morash, University of California, Berkeley
Yue-Ting Siu, University of California, Berkeley & San Francisco State University
Joshua A. Miele, The Smith-Kettlewell Eye Research Institute
Lucia Hasty, Rocky Mountain Braille Associates
Steven Landau, Touch Graphics, Inc.

This paper compares two methods of employing novice web workers to author descriptions of science, technology, engineering, and mathematics images to make them accessible to individuals with visual and print-reading disabilities. The goal is to identify methods of creating image descriptions that are inexpensive, effective, and follow established accessibility guidelines. The first method explicitly presented the guidelines to the worker, and then the worker constructed the image description in an empty text box and table. The second method queried the worker for image information and then used responses to construct a template-based description according to established guidelines. The descriptions generated through queried image description were more likely to include information on the image category, title, caption, and units. They were also more similar to one another, based on Jaccard distances of q-grams, indicating that their word usage and structure were more standardized. Lastly, the workers preferred describing images using queried image description and found the task easier. Therefore, explicit instruction on image-description guidelines is not sufficient to produce quality image descriptions when using novice web workers. Instead, it is better to provide information about images, and then generate descriptions from responses using templates.

Categories and Subject Descriptors: K.4.2 [**Social Issues**]: Assistive technologies for persons with disabilities; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms: Design, Human Factors

Additional Key Words and Phrases: Accessibility (blind and visually impaired), image description, access technology, human computation, crowdsourcing

ACM Reference Format:
Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau, 2014. Guiding novice web workers in making image descriptions using templates. *ACM Trans. Access. Comput.* , , Article (2015), 21 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

There is a growing trend of integrating human workers into accessibility tools [Sato et al. 2010; Takagi et al. 2013]. This approach, referred to as human-powered access technology [Bigham et al. 2011], uses humans, alone or together with computer sys-

This work is supported by the Department of Education, Office of Special Education Programs, Cooperative Agreement H327B100001, under sub-contract to Benetechs DIAGRAM Center initiative.
Author's addresses: V. S. Morash, Psychology Department, University of California, Berkeley; Y.-T. Siu, Education Department, University of California, Berkeley & Special Education Department, San Francisco State University; J. A. Miele, Smith-Kettlewell Eye Research Institute; L. Hasty, Rocky Mountain Braille Associates; S. Landau, Touch Graphics, Inc..

DOI: http://dx.doi.org/10.1145/0000000.0000000

tems, to create accessible materials. An example of such a system is VisWiz, a smartphone application that enables blind and low-vision users to submit photos from their phone, and receive a near real-time textual description of the photo from a sighted web worker [Bigham et al. 2010a; Bigham et al. 2010b]. Another example is the Tactile Graphics Assistant (TGA), which transcribes some parts of a visual image into a tactile graphic using machine vision algorithms, which is then evaluated and revised by a tactile graphics expert [Jayant et al. 2007; Ladner et al. 2005]. In general, these technologies facilitate the production of accessible materials by non-expert workers, who are more available and lower-cost than accessibility specialists [Lasecki et al. 2012]. This paper examines how to best use non-expert workers, in particular web workers, to improve the accessibility of images for individuals with visual disabilities (i.e., those who are blind or have low vision).

Making images accessible to individuals with visual impairments requires the construction of either a tactile graphic or textual image description. Ideally, both of these options are made available to people with visual impairments to satisfy differing needs and preferences. The current report focuses on the creation of textual image descriptions (alt-texts), which can be accessed using a text-to-speech program or braille display. Image descriptions often contain both prose text descriptions and tabular data. These text-based descriptions are also of benefit to students with other print-reading disabilities (e.g., dyslexia) who may use auditory methods for reading textbooks.

Human workers provide a means to generate image descriptions, while machine-vision approaches remain an unsolved, and ongoing area of advanced research [Takagi et al. 2013]. Until computer algorithms can offer a complete solution, untrained workers can provide descriptions for images such as those in textbooks and on the internet [Dardailler 1997]. Using non-expert web workers for image description will significantly expand the available work force and reduce costs [Bigham et al. 2011]. However, it is unclear how these workers should be prompted to describe images to ensure that these descriptions are of high quality and follow established guidelines.

This research focuses on images that appear in the context of science, technology, engineering, and mathematics (STEM), such as charts and graphs. Image accessibility in STEM is particularly important, because STEM topics often rely heavily on images to convey information that is not presented in accompanying text [Jayant et al. 2007; Ladner et al. 2005]. Guidelines for describing STEM images, with specific examples, are provided by the National Center for Accessible Media (NCAM) on their website [Gould et al. 2008]. These guidelines were based on a multi-study project, which involved two rounds of a web-based Delphi survey, taken by over 30 expert describers and individuals with vision loss, to establish approaches to the description of STEM images. A follow-up 60 person end-user study, with participants who had visual impairments, confirmed that the description guidelines produced quality image descriptions, with high clarity and efficiency [Gould et al. 2008].

Non-expert image describers are unlikely to be familiar with the NCAM description guidelines. In the past, most image descriptions have been provided by braille transcribers and teachers of the visually impaired who have had an opportunity to intensively study the NCAM image-description guidelines. However, extensive training of web workers is inapt because of their inherent transience [Bigham et al. 2011], making real-time guidance more appropriate.

This research compares two methods of providing non-expert workers guidance in creating image descriptions: Free-Response Image Description (FRID) and Queried Image Description (QID). In Free-Response Image Description (FRID) the worker is provided with the image to be described along with the relevant NCAM image-description guidelines. An open-ended text box and empty table are provided for the worker to enter descriptive prose and data respectively. This method is similar to

the current approach used by the open-source image-description tool POET [Benetech 2012], although there are plans for the current research to be incorporated into POET. FRID is also similar to the description method in the Phetch game, in which a player generates a text description sufficient for another player to find the described image online, stimulating the creation of image text descriptions within a video game format [von Ahn et al. 2006]. In contrast, Queried Image Description (QID) presents the image without the NCAM guidelines. Instead, QID asks a series of questions specific to the image category, often with conditional sets of questions that may or may not be asked depending on responses to earlier questions. Responses are then used to generate a text and table description that conforms to NCAM guidelines.

The conditional questions used in QID allow the system to adjust to different image content. For example, if the worker indicates that the image has a title, he/she is then asked to enter the text of the title. If the worker indicates that there is no title, then the question asking for the text of the title does not appear. Similarly, if the worker indicates that there are 5 wedges in a pie chart, he/she is then asked 5 sets of questions, one for each wedge.

QID generates image descriptions by slotting worker answers into a template: a set of pre-designed sentences and table with appropriate "blanks" to be filled by worker answers. Additionally, worker answers determine whether certain sentences are included or omitted from the template's text description, and determine the number of rows and columns in the table description.

By removing the onus of reading and understanding the NCAM guidelines, we hope that fewer description mistakes will be made in QID compared to FRID. Specifically, the worker does not need to consider what image details to describe and what to omit, whether to put information in text or a table, and what order and with what language to present information. Thereby, resulting QID image descriptions may be more complete, use text and tables correctly, and use standardized language and content ordering. The description task may also be faster and/or easier with QID, which could increase the willingness of web workers to describe images [Takagi et al. 2013]. Templates have been used to generate image descriptions for a limited number of image categories before, e.g., simple bar charts and line graphs by automated image describing systems, which do not involve human input for creating descriptions [Demir et al. 2010; Ferres et al. 2013]. However, the quality of template-based descriptions has not been compared to that from other image description methods.

STEM images fall into a number of image categories, such as bar charts, line graphs, and pie charts [Jayant et al. 2007], which require the development of distinct questions and templates. One contribution of the current research is the development of questions and templates for six image categories, which can be used as is, or can be used to inform the development of questions and templates for additional image categories. The second, more significant contribution is a user study that demonstrates that our QID task results in better image descriptions than a comparable FRID task.

2. METHODS

2.1. Images & Templates

Six images, each from a different image category, were selected from the NCAM website on "Guidelines for Describing STEM Images" [Gould et al. 2008]. These included the examples of horizontal and vertical bar charts, line graph, Venn diagram, scatter plot, and pie chart. For these images, we created questions and templates based on the NCAM guidelines. The templates included both text and table descriptions. All images and templates are shown in Appendix A.

2.2. Participants

Twenty-two participants (web workers) completed the study remotely using their personal computers. Their average age was 34.81 years (SD = 13.06 years) and six were male. The participants were recruited through email lists for individuals willing to participate in any research or only image-description research. Each participant contacted a member of the research team for a unique survey code, which he/she used to access the survey online. This approach to enrolling participants in the study ensured that each participant completed the survey only once. Participants were not paid or otherwise compensated for their participation. The study procedures were approved and provided exempt status from obtaining informed consent by the Smith-Kettlewell Eye Research Institute's Institutional Review Board.

The study began with the participants answering questions about their characteristics. The majority of participants (59%) reported having no experience describing images (for someone with a visual impairment) before this study. The other participants reported having described images in a casual context (14%), at a non-expert level for books (18%), or at an expert-level for books (9%). Most participants were non-teachers (45%), and a smaller number had taught informally as a volunteer or in the community (14%), as a classroom teacher (27%), as a special-education or paraprofessional teacher (9%), or as a teacher of students with visual impairments (5%). The participants mostly had background in STEM as non-educators (36%) or had no background in STEM (32%). The rest of the participants were interested and knowledgeable about STEM with no formal background (23%), or taught STEM in a K-12 (5%) or post-secondary (9%) setting. Lastly, twenty of the twenty-two participants were native English speakers (90%), and the other two were living and working in the United States, and reported being competent (5%) or fluent (5%) in English.

2.3. Procedure

Participants provided responses online through a free open-source survey tool called LimeSurvey [LimeSurvey Project Team / Carsten Schmitz 2012], which provides conditional scripting support. Some examples of the LimeSurvey interface that participants used during the study are shown in Appendix B. An email was sent to each participant with a list of tasks for them to complete. The email instructed the participants that, while it was okay to take breaks between the tasks, once a task had been started the participant should not take a break until that task had been completed and closed, allowing an accurate estimate of task completion time.

The first task each participant completed was a survey on the participant's demographic information (results described above). The last task was an exit survey. The intervening tasks, two through seven, were the image-description tasks. Participants were informed that they were to create descriptions to make images accessible to an individual who is blind.

Each participant described all six images in random order. Of the six images, three were randomly selected to be described using FRID, and three using QID. Across all participants, each image was described an equal number of times with FRID and QID.

2.3.1. Free-Response Image Description (FRID). During free-response image description, the participant was presented with the image to be described and the NCAM guidelines for describing that category of image, the specific guidelines are shown in Appendix C. The participant was given an empty text box and an empty table to create their description. The empty table was sized 7 columns by 17 rows, to accomodate the specific images used in this study. Future implementations could allow the web worker to add/remove columns and rows to construct a custom-sized table.

2.3.2. Queried Image Description (QID). Queried image description presented the image, and then asked a series of questions about the image. A text description and table were then auto-generated from participant answers to create descriptions according to the NCAM guidelines. This was accomplished using the templates shown in Appendix A. The participant was given an opportunity to review and edit the template-generated text and table as the last part of the task.

2.3.3. Exit Survey. After completing all image descriptions, the participants were asked whether they preferred FRID, QID, or had no preference. They were also asked to rate the difficulty of describing images using FRID and QID methods on a 1-4 scale, 1 being easy and 4 being difficult.

2.4. Analysis

2.4.1. Descriptions. Participants' descriptions were exported into a spreadsheet program from LimeSurvey. These were output into a text file, one for each description, that contained the participant's description text and table. Each description (text and table) was then evaluated with several metrics that were chosen to reveal expected differences between FRID and QID descriptions. Unless otherwise noted, average metrics for FRID and QID descriptions were compared using planned paired t-tests, a pair for each image. All statistical tests were planned a priori, and were therefore not adjusted for multiple comparisons [Keppel and Wickens 2004]. Analyses were done in R [R Core Team 2013].

Evaluation metrics included the amount of overall time it took to complete an image description, as well as separate metrics for the descriptions' text and table parts. Text descriptions were evaluated for having been included or omitted; word count; inclusion of the image category, title, caption, units, and data trend, when applicable; similarity to other text descriptions (explained in the next section); presence of syntactic errors (spelling, punctuation, or capitalization); and presence of content errors (unit, number, or label). Table descriptions were evaluated for completeness (non-completed tables resulted from QID when the participant did not answer all of the question prompts), as well as for the presence of syntactic and content errors.

Metrics were chosen based on a priori hypotheses that they would differ for FRID and QID descriptions. For example, we anticipated that there would be more syntactic errors produced with QID, due to participants answering queries using incorrect formatting, e.g., incorrectly capitalizing an answer that was then placed in the middle of a sentence. We also anticipated that salient and explicit image information, such as titles, data/trend summaries, and captions, would be included in both FRID and QID descriptions, but less salient and nonexplicit information, such as units and image types, would be forgotten during FRID.

2.4.2. Text Description Similarity. To evaluate the consistency of text descriptions we selected q-grams [Ukkonen 1992] for comparing the similarity of two text passages from a number of available options, see [Boytsov 2011] for review. We compared every pair of text descriptions generated for an image using a particular method, e.g., every possible pairing of the 11 FRID text descriptions for the bar chart. These were averaged to create an average similarity for each image's FRID and QID descriptions.

Using q-grams, each description was stripped of white space and decomposed into overlapping sections of q sequential characters or words (grams). In comparing two descriptions, the grams from the first description were compared to those of a second using the Jaccard similarity distance [Chaudhuri et al. 2003]. The Jaccard similarity distance was equal to one minus the ratio of shared q-grams to the total q-grams across both descriptions (one minus the Jaccard coefficient). The closer the Jaccard distance was to zero, the more similar the descriptions, and the closer to one, the less simi-

lar. We computed the Jaccard distance for several values of q, 1-6, using q sequential characters and, separately, q sequential words. We present results where the text was cleaned, stripped of punctuation and capitalization, and the symbols were removed or changed to words (= to equal, & to and, and % to percent). However, comparing cleaned versus original descriptions did not appreciably change the results or significance levels. Two separate Greenhouse-Geisser corrected ANOVAs were run, one for character similarity and one for word similarity, with within-image factors of q (6 levels, 1-6) and description method (2 levels, FRID versus QID). Of interest was the main effect of description method, and its interaction with q, on description similarity.

2.4.3. Exit Survey. Difficulty ratings, which were on a 1-4 scale, were compared using a paired Mann-Whitney U test, due to the fact that rankings were ordinal (ordered, 1 through 4), but unlikely to be interval (have equal spacing between 1 and 2, 2 and 3, and 3 and 4).

Preference ratings, which indicated whether participants preferred FRID, QID, or had no preference, were evaluated by a Thurstonian model. This model estimated the underlying distributions of preference for QID and FRID methods and then tested whether the difference in means scaled by standard deviation, d prime, was significantly different from zero [Bojesen Christensen et al. 2012]. A significant difference would indicate that FRID or QID was preferred over the other. This analysis was done in R using the sensR package [Bojesen Christensen et al. 2012].

3. RESULTS

Some randomly selected example descriptions are shown in table I. Only the texts from these descriptions are shown, omitting any tabular components. The mean results for QID and FRID descriptions (standard errors in parentheses) and associated statistical tests are shown in table II.

4. DISCUSSION

This research compared the quality of image descriptions created through a free-response method (FRID) to those created using a querying approach (QID). FRID provided the web worker with relevant description guidelines and prompted for an image description in an empty text box and table. QID created image descriptions by querying the worker about image properties, and then constructing a description from the worker's answers using a template, without explicitly informing the worker about description guidelines. Our results indicate that the QID image descriptions were at least equal to, and in many ways superior to FRID image descriptions.

FRID and QID descriptions took equally long to create, each taking, on average, 10-12 minutes per image. Some descriptions contained only text or only a table, and others had tables that were incomplete, but this was not more or less common in FRID or QID descriptions. The number of words in the text descriptions, when present, were also similar for FRID and QID descriptions. Lastly, descriptions created with FRID and QID were equally likely to include syntactic and/or content errors in both text descriptions and tables. It's possible that participants' exposure to QID items affected their FRID responses, possibly encouraging them to style FRID descriptions to be similar to those they made using QID. This may have reduced our ability to detect real differences between FRID and QID descriptions in the aforementioned metrics.

Descriptions created with FRID and QID *did* differ in the content they contained. Descriptions created with FRID were more likely to omit the image category, title, caption, and units. These are serious omissions, especially considering that these details are those immediately asked about by visually impaired users exploring a STEM image [Ferres et al. 2013]. Given that descriptions created with FRID were likely to omit

Table I. Example Text Descriptions

Image Category	FRID Example	QID Example
Bar Chart	This is a horizontal double bar chart providing a count of support provided by various categories for individuals for those with a physical disability. The first set of bar charts represents data related to individuals who are Deaf/HOH and the second set represents data related to individuals who are Blind/LV. See the table below for specific data.	This is a horizontal bar chart with 16 clusters of bars. The data range from 0 to 110, represented in increments of 10. TV is has the highest value for deaf/HOH; Radio has the highest value for blind/LV.
Scatter Plot	A scatter plot graph represents points that chart data for the y-axis labelled "Plant Tissue Production," measured in grams per m squared, compared to the x-axis labeled "Average Rainfall" measured in millimeters per year. Most of the points are in the lower-left corner of the graph and decrease in quantity as the ranges increase with only two outlying points in the upper right-hand corner. The data range of the y-axis starts at zero and ends at 3,000. The data range of the x-axis starts at zero and ends at 4,000.	This is a scatter plot, titled Rainfall and Plant Growth. A caption reads: "The graph below shows the relationship between annual rainfall and plant tissue growth rates in an ecosystem." The vertical axis is labeled Plant Tissue Production (g/m2 per year) and is in g/m2 per year, ranging from 0 to 3000 in increments of 500. The horizontal axis is labeled Average Rainfall (mm/year) and is in mm/year, ranging from 0 to 4000 in increments of 1000. The graph has approximately 100 points scattered in a pattern There is a positive correlation between the average rainfall and the plant tissue production with a regression slope larger than one. The data gets more scattered as x and y values increase..
Venn Diagram	This graph shows a survey of travelers. In a survey of 250 European travelers, 93 have traveled to Africa, 155 have traveled to Asia and 70 have traveled to both of these continents.	This is a venn diagram, titled Survey of Travelers, showing 2 circles. A caption reads: "In a survey of 250 European travelers, 93 have traveled to Africa, 155 have traveled to Asia, and 70 have traveled to both of these continents, as illustrated in the Venn Diagram above." The circles are labeled Africa, value equals 93 people, and Asia, value equals 155 people. There is 1 area of intersection that equals 70 people.

information, it's surprising that the lengths of text descriptions created with FRID and QID were similar. This may be explained by the similar inclusion of data/trends in FRID and QID text descriptions, which may have compensated, in word count, for omitted information. However, an image description that contains information on data/trends, without indicating the image category, tile, caption, or units, as was more often the case with FRID descriptions, has questionable utility.

It is interesting that participants omitted image type, title, caption, and units significantly more often during FRID, even though they could have been stimulated to include this information during interspersed QID items (FRID and QID items were randomly ordered for each participant). It's possible that longer exposure to QID could improve the quality of participants' FRID descriptions. Further research is needed to know whether passive non-instructive exposure to QID could effectively turn inexperienced web workers into expert image describers.

QID descriptions were also more similar to one another than were FRID descriptions. Description similarity was assessed using the Jaccard distance for different size q-grams of characters and words. Words and word fragments (clusters of adjacent char-

Table II. Results

Metric	Description Method: Free-Response (FRID)	Description Method: Queried (QID)	Test Statistic	p-Value
Text & Table				
Time to Describe (sec)	738.26 (155.21)	613.19 (111.05)	t(5) = 1.20	0.282
Any Text Description (%)	96.97 (3.03)	100.00 (0.00)	t(5) = -1.00	0.363
Completed Table (%)	60.00 (10.98)	89.09 (3.40)	t(4) = -2.14	0.099
Text Description				
Word Count (count)	63.09 (9.28)	82.17 (20.23)	t(5) = -1.19	0.289
Syntactic Error (%)	37.04 (7.49)	48.48 (13.42)	t(5) = -0.78	0.469
Content Error (%)	11.28 (4.62)	12.12 (5.59)	t(5) = -0.22	0.835
Data/Trend Summary (%)	61.28 (11.42)	75.76 (16.87)	t(5) = -1.05	0.343
Image Type (%)	72.05 (6.49)	100.00 (0.00)	t(5) = -4.31	0.008**
Title (%)	34.55 (15.05)	83.64 (16.36)	t(4) = -2.90	0.044*
Caption (%)	18.18 (5.25)	93.94 (3.03)	t(2) = -9.45	0.011*
Units (%)	36.20 (6.59)	100.00 (0.00)	t(5) = -9.67	<0.001***
Text Description Similarity				
Average Jaccard Distance (q = number of characters)				
$q = 1$	0.17 (0.03)	0.06 (0.02)		
$q = 2$	0.50 (0.04)	0.20 (0.06)		
$q = 3$	0.70 (0.03)	0.33 (0.09)		
$q = 4$	0.79 (0.02)	0.39 (0.10)		
$q = 5$	0.83 (0.02)	0.42 (0.11)		
$q = 6$	0.86 (0.02)	0.44 (0.11)		
Main Effect q			F(1, 25) = 0.00	1.000
Main Effect Method (FRID vs. QID)			F(1, 5) = 12.19	0.017*
Interaction q x Method			F(1, 25) = 0.00	1.000
Average Jaccard Distance (q = number of words)				
$q = 1$	0.73 (0.03)	0.35 (0.09)		
$q = 2$	0.91 (0.01)	0.48 (0.11)		
$q = 3$	0.96 (0.01)	0.53 (0.12)		
$q = 4$	0.98 (0.01)	0.57 (0.12)		
$q = 5$	0.99 (0.00)	0.60 (0.12)		
$q = 6$	0.99 (0.00)	0.63 (0.12)		
Main Effect q			F(1.12, 5.58) = 60.38	<0.001***
Main Effect Method (FRID vs. QID)			F(1, 5) = 11.83	0.018*
Interaction q x Method			F(1.10, 5.50) = 4.30	0.086
Table Description				
Syntactic Error (%)	12.66 (3.58)	8.44 (3.81)	t(4) = 1.64	0.177
Content Error (%)	14.52 (6.65)	10.22 (5.48)	t(4) = 0.64	0.554
Participant Feedback				
Difficulty (1-4)	2.95 (0.18)	1.95 (0.14)	U = 174	0.001**
Preference (count)	3 (3 no pref)	16	d' = 1.20 (0.40)	0.002**

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

acters, lengths 1 through 6) and sentences and sentence fragments (clusters of adjacent words, lengths 1 through 6) were more frequently shared between pairs of QID descriptions than FRID descriptions. For example, an average of 37% of 6-word segments were shared between pairs of QID descriptions. In contrast, only 1% of 6-word segments were shared between pairs of FRID descriptions. Therefore, image descriptions created using QID were more standardized in word use and content ordering. Regularity in QID descriptions could be further improved by creating question scripts and templates for trend descriptions, which, in the current implementation, were filled in by

participants without specific queries, causing much of the QID description irregularities. Future work to improve the similarity of QID descriptions would be worthwhile because previous research has shown that it is easier to understand and use image descriptions that have consistent language and present information in a systematic and predictable format [Ferres et al. 2013].

Description of data trends may be informed by research on image summaries and captions, which has the goal of expressing the high-level composition or intention of an image that appears in popular media. The goal of such research is to not only convey the meaning of an image to individuals with visual impairments, but to also facilitate summarization of multimodal documents and their processing by computer systems [Demir et al. 2012; Demir et al. 2013; Elzer et al. 2005; Feng and Lapata 2010]. This is typically accomplished in two stages. The first stage is to access image information through machine vision, human processing (possibly web workers), or extracting embedded digital information, and organize this information in, e.g., XML or a relational ontology [Demir et al. 2012; Dumontier et al. 2010; Berners-Lee et al. 2001; Fasciano and Lapalme 1996]. Once the image information is extracted and stored in some representation, it can be used, amongst other possibilities, to create an image summary similar to what a human may generate, achieved using natural language generation techniques, or allow the user to interactively query for image information [Demir et al. 2012; Dumontier et al. 2010; McCoy et al. 2001; Carberry et al. 2012; Wu et al. 2010]. Determining the intended message of the image can be accomplished using probabilistic inference with the image's semantic information, appearance, and possibly accompanying text [Vinyals et al. 2014; Wu et al. 2010; Elzer et al. 2005; Feng and Lapata 2010; Carberry et al. 2012]. As mentioned, our system could make use of such image summaries in place of asking web workers to describe data trends. Although an image summary is less specific and detailed than what is often needed by people with visual impairments, our results motivate these parallel research efforts, especially the creation of automated means of extracting and storing image information. Our results indicate that using image data stored within an ontology such as those designed for the Semantic Web [Berners-Lee et al. 2001] to create image descriptions using our templates will produce higher quality image descriptions than a web worker using free description.

Beyond issues of description quality, the QID method provides several additional benefits in a web-worker scenario. One benefit is that it makes accuracy checking by comparing responses from redundant web workers easier. Comparing one- or few-word answers about image properties is easier than comparing multi-sentence descriptions and a table. Previous work has shown that such redundancy is a viable approach to ensure accuracy from web workers [Von Ahn and Dabbish 2004], and this is a less costly method than using vetted or expert workers to check accuracy [Bigham et al. 2011].

Image-description data, for book and/or website images, can be indexed and stored on a remote server, as proposed by [Dardailler 1997]. This enables many people to access the image descriptions and reuse the work of image describers [Takagi et al. 2013]. The QID method enables image information to be stored, not as a finished description, but in labeled parts. Thereby, users could tailor the description to their needs through the design of custom templates. For example, a user may prefer a template that provides a brief overview of the image, color information to be omitted, or a table of data only. Given that blind and sighted individuals may navigate the information in image descriptions differently [Ferres et al. 2013], preferences in image description styles may depend on extent of vision loss and time since onset. Alternatively, image data could be used to render tactile graphics or audio/tactile graphics where appropriate. Storing image data in an object-oriented database or the Semantic Web, rather than

a table-based relational database, may be more extensible, facilitating the addition of new image categories, custom templates, and new information types, e.g., color, while also maintaining economy of representational space [Dumontier et al. 2010].

Putting aside alternative uses of our system, there are several ways in which the current image descriptions could be improved. Text processing could be applied to remove spelling and grammatical errors, or at least make these errors more apparent to the worker. Also, visual information may be added to some of our templates. For example, information was not collected on the colors or visual patterns of the bars in bar charts. This may be important information, e.g., for a student when the teacher asks "what trends do you see in the red bars?"

The current research develops description templates for six STEM image categories. Before description, a web worker would need to specify the image category or otherwise select a description template. This task could be accomplished prior to image description, possibly within a separate web worker task that determines the nature of the image, e.g., photo or graph, and whether the material in the image would be most accessible in tactile format, text description, or should be omitted [Benetech and Touch Graphics 2014]. Unusual images, such as those that contain features of bar charts and line graphs, may require the web worker to create custom templates or flag the image for expert assistance.

It is important to note that our present research only offers metrics related to the production of image descriptions with FRID and QID. This work does not investigate any metrics related to the differential usability of descriptions created with FRID and QID. We did not have users with visual impairments assess the usefulness or usability of image descriptions. Future research is needed to understand what properties of image descriptions promote comprehension and usability by individuals with visual impairments in different contexts, and whether these properties are more easily obtained using FRID or QID methods.

In conclusion, the current work demonstrates that the QID method produced better image descriptions from novice web workers than the FRID method. Image descriptions created with QID had less omitted information, were more similar to one another, and were preferred by and easier to create for describers. Therefore, our research suggests that web workers tasked with producing image descriptions, and possibly other accessible materials, cannot be expected to successfully follow accessibility guidelines, even with explicit instruction. Instead, it is better to use web workers to extract image, or other inaccessible information, and then use this information to construct accessible materials according to established guidelines.

APPENDIX A: Images & Templates

Horizontal Bar Chart

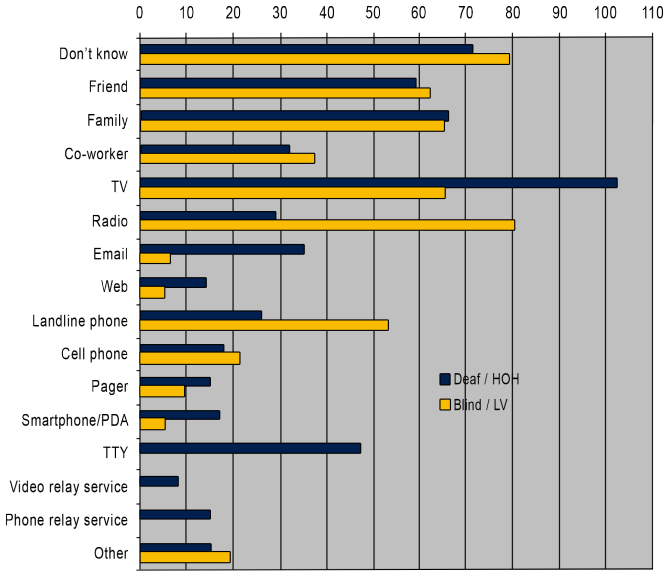


Fig. 1. Horizontal bar chart image.

Horizotal Bar Chart Text Description Template.

This is a horizontal bar chart, titled *title* (if titled), measuring *x-axis units* (if x-axis has units)

for *number* *y-axis label* (if y-axis labeled) *bars / clusters of bars* (if y-axis not labeled).

A caption reads: "*caption*." (if captioned) The data range from *x-axis min* to

x-axis max in increments of *x-axis increments* *x-axis units* (if x-axis has units).

Description if data trends. (if trends described)

Horizontal Bar Chart Table Description Template

number of columns equal to number of sub-bars per cluster

	Sub-Bar 1 Label	*Sub-Bar 2 Label*
Bar/Cluster 1 Label	*x-value*	*x-value*
⋮	⋮	⋮
Bar/Cluster 16 Label	*x-value*	*x-value*

number of rows equal to number of bars or bar clusters

Line Graph

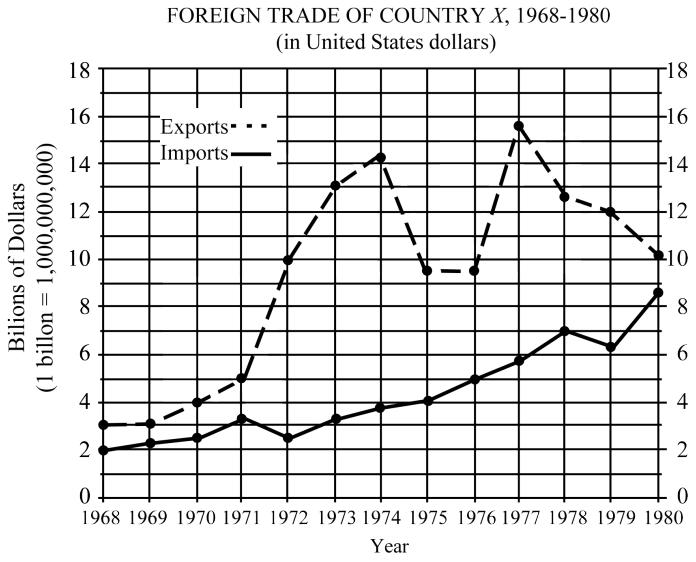


Fig. 2. Line graph chart image.

Line Graph Text Description Template

This is a line graph, titled *title* (if titled). A caption reads: "*caption*." (if captioned) The horizontal axis is labeled *x-axis label*, and ranges from *x-axis min* to *x-axis max* in increments of *x-axis increments* *x-axis units*. The vertical axis is labeled *y-axis label*, and ranges from *y-axis min* to *y-axis max* in increments of *y-axis increments* *y-axis units*. There *is / are* *number* *line / lines* on the graph. Line 1 is labeled *line 1 label* and appears *line 1 trend*.

Line *number* is labeled *line label* and appears *line trend*. (repeat to number of lines)

Line Graph Table Description Template

number of columns equal to number of lines

number of rows equal to number of x values (x min, max, & increment)

x-axis units	*line 1 label (y units)*	*line 2 label (y units)*
x-value	*y-value*	*y-value*
⋮	⋮	⋮
x-value	*y-value*	*y-value*

Pie Chart

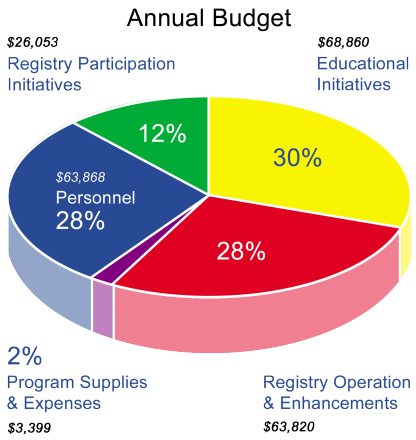


Fig. 3. Pie chart image.

Pie Chart Text Description Template

This is a pie chart, titled *title* (if titled) . A caption reads: " *caption* ." (if captioned) The chart has *number* wedges, labeled in units (if units) and (if both) percentages (if percentages) . The data are summarized in the following table.

Pie Chart Table Description Template. A table is only presented if there are 10 or less data points.

		if labeled in units	if labeled in percentages
	Color / Pattern	*units*	Percentage
Wedge 1 Label	*color / pattern*	*value*	*value*
⋮	⋮	⋮	⋮
Wedge 5 Label	*color / pattern*	*value*	*value*

number of rows equal to number of wedges

Scatter Plot

The graph below shows the relationship between annual rainfall and plant tissue growth rates in an ecosystem.

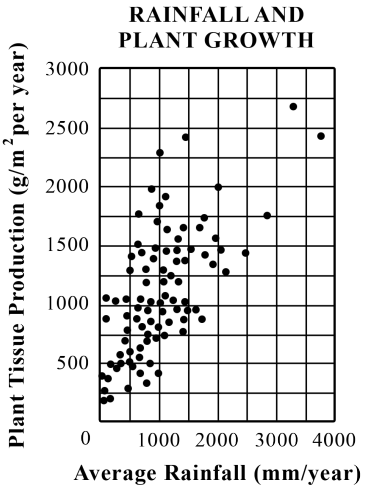


Fig. 4. Scatter plot image.

Scatter Plot Text Description Template.

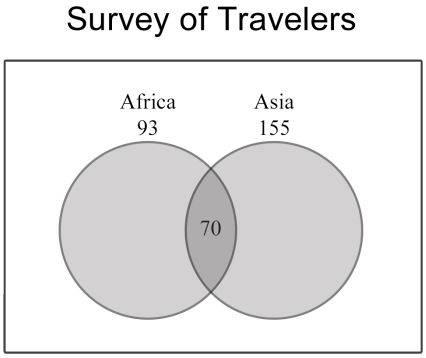
This is a scatter plot, titled *title* (if titled) . A caption reads: " *caption* ." (if captioned) The horizontal axis is labeled *x-axis label* , and ranges from *x-axis min* to *x-axis max* in increments of *x-axis increments* *x-axis units* (if x-axis has units) . The vertical axis is labeled *y-axis label* , and ranges from *y-axis min* to *y-axis max* in increments of *y-axis increments* *y-axis units* (if y-axis has units) . The graph has approximately *number* points scattered in a pattern *pattern* (if there's a pattern) . The data are summarized in the following table. (if number points ≤ 10)

Scatter Plot Table Description Template. A table is only presented if there are 10 or less points.

x-axis label (*x units*)	*y-axis label* (*y units*)	Point Type (if # of point types > 1)
x-value	*y-value*	*type*
⋮	⋮	⋮
x-value	*y-value*	*type*

(number of rows equal to number of points)

Venn Diagram



In a survey of 250 European travelers, 93 have traveled to Africa, 155 have traveled to Asia, and 70 have traveled to both of these continents, as illustrated in the *Venn diagram* above.

Fig. 5. Venn diagram image.

Venn Diagram Text Description Template.

This is a Venn diagram, titled *title* (if titled), showing *number* circles.

A caption reads: "*caption*." (if captioned) The circles are labeled *circle 1 label*, value equals *value* *units*, circle *number*, values equals *value* *units* (repeat to number of circles).

There is *number* *area / areas* of intersection that *equal / equals* *number* *units*, between circles *list*. (repeat to number of intersections)

Venn Diagram Table Description Template.

number of rows equal to number of circles and intersections

	Value (*units*)
Circle 1 Label	*value*
Circle 2 Label	*value*
Intersection 1 Label	*value*

Vertical Bar Chart

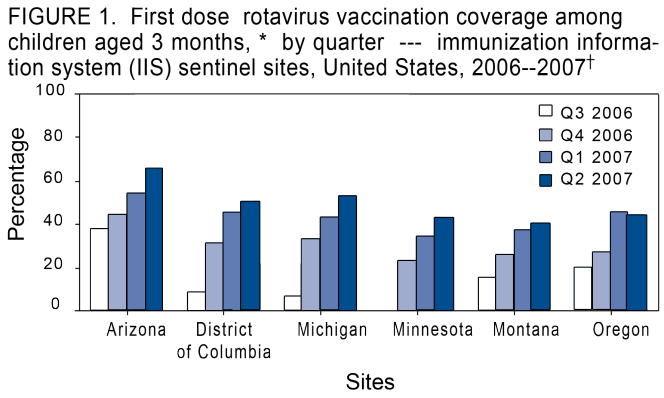


Fig. 6. Vertical bar chart image.

Vertical Bar Chart Text Description Template.

This is a vertical bar chart, titled *title* (if titled), measuring *y-axis units* (if y-axis has units) for *number* *x-axis label* (if x-axis labeled) *bars / clusters of bars* (if x-axis not labeled).

A caption reads: "*caption*." (if captioned) The data range from *y-axis min* to *y-axis max* in increments of *y-axis increments* *y-axis units* (if y-axis has units).

Description if data trends. (if trends described)

Vertical Bar Chart Table Description Template.

number of columns equal to number of sub-bars per cluster

number of rows equal to number of bars or bar clusters

	Sub-Bar 1 Label	*Sub-Bar 2 Label*
Bar/Cluster 1 Label	*y-value*	*y-value*
⋮	⋮	⋮
Bar/Cluster 4 Label	*y-value*	*y-value*

APPENDIX B: Interface Examples

Line Graph QID Input Example

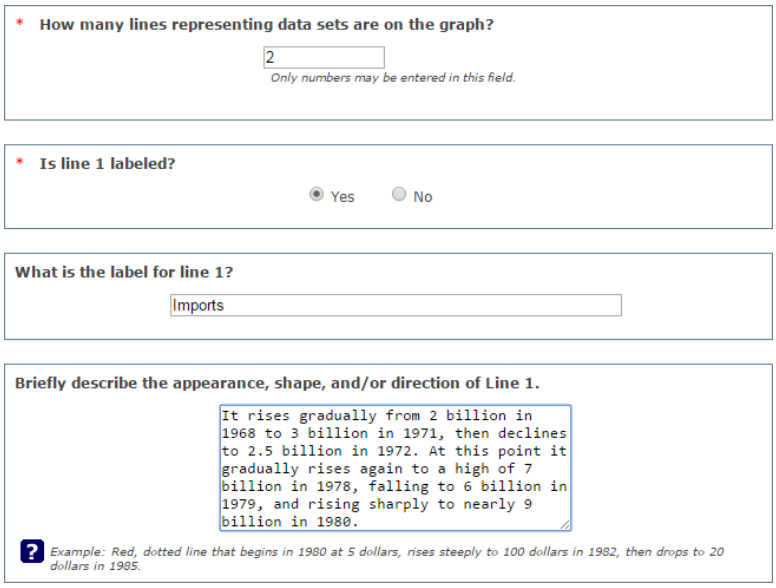


Fig. 7. Example interface to input some line information in line graph QID.

Pie Chart QID Input Example

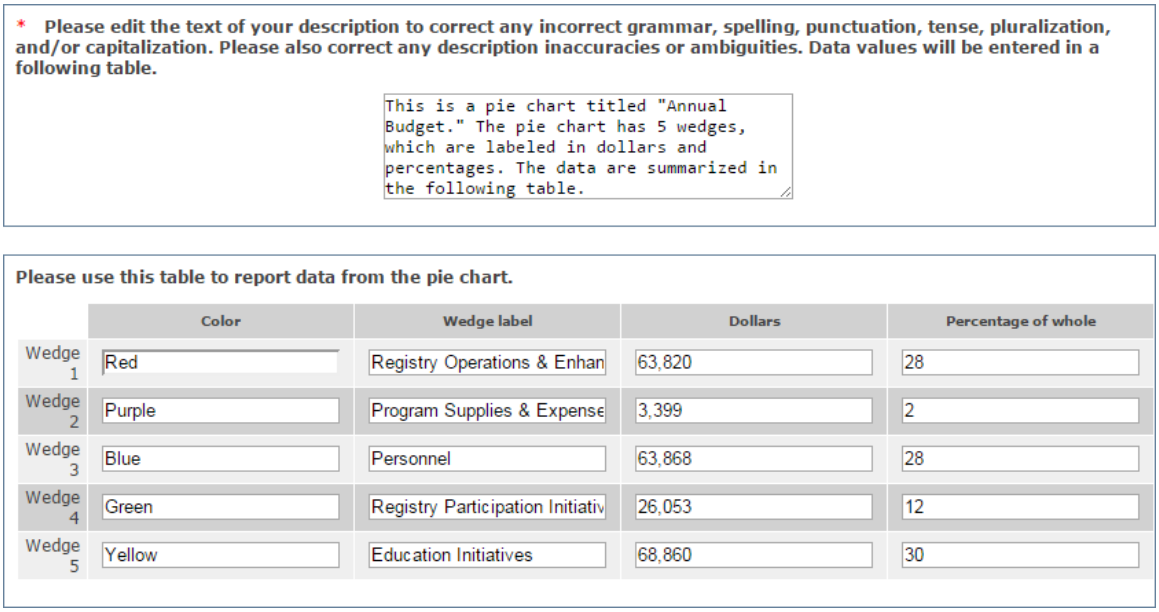


Fig. 8. Example interface to edit the final text description and enter tabular data in pie chart QID.

APPENDIX C: NCAM Guidelines

The NCAM guidelines [Gould et al. 2008] specific to the images used in the current study, presented to the participants during FRID, are as follows.

Horizontal Bar Chart

— Most bar charts should be converted into accessible tables, but simple charts can be presented as text in a list.
— Provide the title and labels.
— It is not necessary to describe the visual attributes of the bars unless there is an explicit need.

Line Graph

— Line graphs should be converted into accessible tables.
— Briefly describe the chart and give a summary if one is immediately apparent.
— Provide the title and axis labels.
— It is not necessary to describe the visual attributes of the lines, e.g. solid, dashed, unless there is an explicit need such as an exam question referring to these attributes.

Pie Chart

— Pie charts should be converted into accessible tables.
— It is not necessary to describe the visual attributes of the charts, e.g., red wedge, blue lines, etc., unless there is an explicit need such as an exam question referring to these attributes.
— It is helpful to list the numbers from smallest to largest, regardless of how they are presented in the image.

Scatter Plot

— Scatter plots are among the more difficult graphs to describe, especially if there is a need to make specific data point accessible.
— Provide the title and axis labels.
— Identify the image as a scatter plot and focus on the change of concentration.
— If it is necessary to be more specific, and there are 10 or less points, convert the data into an accessible table.

Venn Diagram

— Focus on the data in the Venn diagram, not on its appearance.
— Provide the data in brief statements.
— Give a summary if one is immediately apparent.
— Include the caption only if it is not accessible from elsewhere in the text.

Vertical Bar Chart

— Bar charts should be converted into accessible tables.
— Briefly describe the chart and give a summary of any trends that are immediately apparent.
— Provide the title and axis labels.
— It is not necessary to describe the visual attributes of the bars, e.g. dark blue, light blue, unless there is an explicit need such as an exam question referring to the colors.

REFERENCES

Benetech. 2012. POET image description tool. http://diagramcenter.org/development/poet.html. (2012).

Benetech and Touch Graphics. 2014. Decision tree: Image sorting tool. http://diagramcenter.org/decision-tree.html. (2014).

Tim Berners-Lee, James Hendler, Ora Lassila, and others. 2001. The semantic web. *Scientific American* 284, 5 (2001), 28–37.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010a. VizWiz: nearly real-time answers to visual questions. In *23nd Annual Symposium on User Interface Software and Technology*. ACM, 333–342.

Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010b. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In *Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 65–72.

Jeffrey P Bigham, Richard E Ladner, and Yevgen Borodin. 2011. The design of human-powered access technology. In *13th International Conference on Computers and Accessibility (SIGACCESS)*. ACM, 3–10.

Rune Haubo Bojesen Christensen, Hye-Seong Lee, and Per Bruun Brockhoff. 2012. Estimation of the Thurstonian model for the 2-AC protocol. *Food Quality and Preference* 24, 1 (2012), 119–128.

Leonid Boytsov. 2011. Indexing methods for approximate dictionary searching: Comparative analysis. *Journal of Experimental Algorithmics* 16 (2011), 1–1.

Sandra Carberry, Stephanie Elzer Schwartz, Kathleen Mccoy, Seniz Demir, Peng Wu, Charles Greenbacker, Daniel Chester, Edward Schwartz, David Oliver, and Priscilla Moraes. 2012. Access to multimodal articles for individuals with sight impairments. *ACM Transactions on Interactive Intelligent Systems* 2, 4 (2012), 21.

Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, and Rajeev Motwani. 2003. Robust and efficient fuzzy match for online data cleaning. In *International Conference on Management of Data (SIGMOD)*. ACM, 313–324.

Daniel Dardailler. 1997. The ALT-server ("An eye for an alt"). http://www.w3.org/WAI/altserv.htm. (1997).

Seniz Demir, Sandra Carberry, and Kathleen F McCoy. 2012. Summarizing Information Graphics Textually. *Computational Linguistics* 38, 3 (2012), 527–574.

Seniz Demir, David Oliver, Edward Schwartz, Stephanie Elzer, Sandra Carberry, Kathleen F Mccoy, and Daniel Chester. 2010. Interactive SIGHT: textual access to simple bar charts. *New Review of Hypermedia and Multimedia* 16, 3 (2010), 245–279.

Seniz Demir, Stephanie Elzer Schwartz, Richard Burns, and Sandra Carberry. 2013. What is being Measured in an Information Graphic? In *Computational Linguistics and Intelligent Text Processing*. Springer, 501–512.

Michel Dumontier, Leo Ferres, and Natalia Villanueva-Rosales. 2010. Modeling and querying graphical representations of statistical data. *Web Semantics: Science, Services and Agents on the World Wide Web* 8, 2 (2010), 241–254.

Stephanie Elzer, Sandra Carberry, Ingrid Zukerman, Daniel Chester, Nancy Green, Seniz Demir, and others. 2005. A probabilistic framework for recognizing intention in information graphics. In *International Joint Conference On Artificial Intelligence*, Vol. 19. 1042.

Massimo Fasciano and Guy Lapalme. 1996. Postgraphe: a system for the generation of statistical graphics and text. In *8th International Workshop on Natural Language Generation (INLG)*. ACL SIGGEN, 51–60.

Yansong Feng and Mirella Lapata. 2010. How many words is a picture worth? automatic caption generation for news images. In *48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1239–1249.

Leo Ferres, Gitte Lindgaard, Livia Sumegi, and Bruce Tsuji. 2013. Evaluating a tool for improving accessibility to charts and graphs. *ACM Transactions on Computer-Human Interaction* 20, 5 (2013), 28.

Bryan Gould, Trisha OConnell, and Geoffrey Freed. 2008. Guidelines for describing STEM images. http://ncam.wgbh.org/experience_learn/educational_media/stemdx/guidelines. (2008).

Chandrika Jayant, Matt Renzelmann, Dana Wen, Satria Krisnandi, Richard Ladner, and Dan Comden. 2007. Automated tactile graphics translation: in the field. In *9th International Conference on Computers and Accessibility (SIGACCESS)*. ACM, 75–82.

Geoffrey Keppel and Thomas D Wickens. 2004. *Design and analysis: A researcher's handbook* (4 ed.). Upper Saddle River, NJ: Pearson Education.

Richard E Ladner, Melody Y Ivory, Rajesh Rao, Sheryl Burgstahler, Dan Comden, Sangyun Hahn, Matthew Renzelmann, Satria Krisnandi, Mahalakshmi Ramasamy, Beverly Slabosky, and others. 2005. Automat-

ing tactile graphics translation. In *7th International Conference on Computers and Accessibility (SIGACCESS)*. ACM, 150–157.

Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *25th Annual Symposium on User Interface Software and Technology*. ACM, 23–34.

LimeSurvey Project Team / Carsten Schmitz. 2012. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany. http://www.limesurvey.org

Kathleen F McCoy, Sandra Carberry, Tom Roper, and Nancy Green. 2001. Towards generating textual summaries of graphs. In *International Conference on Universal Access in Human-Computer Interaction*. 695–699.

R Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Daisuke Sato, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. 2010. Social accessibility: the challenge of improving web accessibility through collaboration. In *2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. ACM, 28.

Hironobu Takagi, Susumu Harada, Daisuke Sato, and Chieko Asakawa. 2013. Lessons learned from crowd accessibility services. In *Human-Computer Interaction–INTERACT 2013*. Springer, 587–604.

Esko Ukkonen. 1992. Approximate string-matching with q-grams and maximal matches. *Theoretical computer science* 92, 1 (1992), 191–211.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555* (2014).

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *SIGCHI Conference on Human factors in Computing Systems*. ACM, 319–326.

Luis von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum. 2006. Improving accessibility of the web with a computer game. In *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 79–82.

Peng Wu, Sandra Carberry, Stephanie Elzer, and Daniel Chester. 2010. Recognizing the intended message of line graphs. In *Diagrammatic Representation and Inference*. Springer, 220–234.

Received February 2007; revised March 2009; accepted June 2009