Project Sidewalk: A Web-based Crowdsourcing Tool for Collecting Sidewalk Accessibility Data at Scale

Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, Jon Froehlich

University of Washington; University of Maryland, College Park; Montgomery Blair High School; University of Michigan; Singapore Management University

{manaswi, saugstad, aileenz, jonf}@cs.washington.edu

Figure 1: In an 18-month deployment study of Project Sidewalk, we collected 205,385 sidewalk accessibility labels, including curb ramps, missing curb ramps, sidewalk obstacles, and surface problems. Each dot above represents a geo-located label rendered at 50% translucency. Try out the tool at http://projectsidewalk.io.

ABSTRACT

We introduce Project Sidewalk, a new web-based tool that enables online crowdworkers to remotely label pedestrian-related accessibility problems by virtually walking through city streets in Google Street View. To train, engage, and sustain users, we apply basic game design principles such as interactive onboarding, mission-based tasks, and progress dashboards. In an 18-month deployment study, 797 online users contributed 205,385 labels and audited 2,941 miles of Washington DC streets. We compare behavioral and labeling quality differences between paid crowdworkers and volunteers, investigate the effects of label type, label severity, and majority vote on accuracy, and analyze common labeling errors. To complement these findings, we report on an interview study with three key stakeholder groups (N=14) soliciting reactions to our tool and methods. Our findings demonstrate the potential of virtually auditing urban accessibility and highlight tradeoffs between scalability and quality compared to traditional approaches.

CCS CONCEPTS

• Human-centered computing → Accessibility systems and tools: Interactive systems and tools; • Information systems → Crowdsourcing.

KEYWORDS

Crowdsourcing, accessibility, mobility impairments, GIS

ACM Reference Format:

1 INTRODUCTION

Geographic Information Systems (GIS) such as Google Maps, Waze, and Yelp have transformed the way people travel and access information about the physical world. While these systems contain terabytes of data about road networks and points of interest (POIs), their information about physical accessibility is commensurately poor. GIS websites like Axsmap.com, Wheelmap.org, and AccessTogether.org aim to address this problem by collecting location-based accessibility information provided by volunteers (i.e., crowdsourcing).
While these efforts are important and commendable, their value propositions are intrinsically tied to the amount and quality of data they collect. In a recent review of accessibility-oriented GIS sites, Ding et al. [15] found that most suffered from serious data sparseness issues. For example, only 1.6% of the Wheelmap POIs had data entered on accessibility. One key limiting factor is the reliance on local populations with physical experience of a place for data collection. While local users who report data are likely to be reliable, the dependence on in situ reporting dramatically limits scalability—both who can supply data and how much data they can easily supply.

In contrast, we are exploring a different approach embodied in a new interactive tool called Project Sidewalk (Figure 2), which enables online crowdworkers to contribute physical-world accessibility information by virtually walking through city streets in Google Street View (GSV)—similar to a first-person video game. Rather than pulling solely from local populations, our potential pool of users scales to anyone with an Internet connection and a web browser. Project Sidewalk extends previous work in streetscape imagery auditing tools like Canvas [4], Spotlight [7], BusStop CSI [23], and Tohme [26], all which demonstrate the feasibility of virtual auditing and, crucially, that virtual audit data has high concordance with traditional physical audits. However, this past work has focused on small spatial regions, relied on specialized user populations such as public health researchers [4, 7] and paid crowdworkers [23, 26], and has not been publicly deployed.

In this paper, we present an 18-month deployment study of Project Sidewalk in Washington DC. In total, 797 users contributed 205,385 geo-located accessibility labels and virtually audited the entirety of Washington DC (1,075 miles of city streets; see Figure 2). As the first public deployment of a virtual auditing tool, our research questions are exploratory: How can we engage, train, and sustain crowd workers in virtual accessibility audits? Are there behavioral and/or labeling quality differences between paid crowd workers and volunteers? What are some common labeling mistakes and how may we correct them in future tools? Finally, how do key stakeholder groups react to crowdsourcing accessibility and what are their concerns?

To address these questions, we analyzed interaction logs from our DC deployment, performed a semi-controlled data validation study, and conducted semi-structured interviews with three stakeholder groups (N=14): government officials, people with mobility impairments (MI), and caretakers. In our deployment study, we found that registered volunteers completed significantly more missions, on average, than our anonymous volunteers (M=5.8 vs. 1.5) and that our paid workers—who were compensated per mission—completed more than both (M=35.4 missions). In the data validation study, paid workers also significantly outperformed registered and anonymous volunteers in finding accessibility problems (recall=68% vs. 61% and 49%, respectively) but precision was roughly equivalent for all groups (~70%). Our findings also show that the number of found issues significantly increases with the number of labelers per street—with five labelers, recall rose from 68% to 92%.

To complement these findings, our interview study asked about perceptions of and experiences with urban accessibility and solicited reactions to Project Sidewalk and the idea of crowdsourcing accessibility in general. All three stakeholder groups were positive: while government officials emphasized cost-savings and civic engagement, the MI and caregiver groups focused more on personal utility and enhanced independence. Key concerns also arose, including data reliability, maintenance, and, for the MI participants, whether labels properly reflected their accessibility challenges (the latter echoes findings from [24]).

In summary, the contributions of this paper include: (i) Project Sidewalk, a novel web-based virtual auditing tool for collecting urban accessibility data at scale; (ii) results from an 18-month deployment and complementary data validation study exploring key behavioral and labeling quality differences between volunteer and paid crowd workers; (iii) findings from semi-structured interviews with three stakeholder groups soliciting reactions to Project Sidewalk and identifying key concerns and design suggestions; (iv) and our large, open-source sidewalk accessibility dataset¹, which we believe is the largest of its kind. By scaling up data collection methods for sidewalk accessibility, our overarching aim is to enable the development of new accessibility-aware mapping tools (e.g., [24, 32]), provide increased transparency and accountability about city accessibility, and work with and complement government efforts in monitoring pedestrian infrastructure.

2 RELATED WORK

We present background on sidewalk accessibility, survey existing methods for collecting street-level accessibility data, and review volunteer geographic information (VGI) systems.

Street-Level Accessibility

Accessible infrastructure has a significant impact on the independence and mobility of citizens [1, 40]. In the U.S., the Americans with Disability Act (ADA) [53] and its revision, the 2010 ADA Standards for Accessible Design [52], mandate that new constructions and renovations meet modern accessibility guidelines. Despite these regulations, pedestrian infrastructure remains inaccessible [18, 28]. The problem is

¹http://projectsidewalk.io/api
not just inaccessible public rights-of-way but a lack of reliable, comprehensive, and open information. Unlike road networks, there are no widely accepted standards governing sidewalk data (though some recent initiatives are emerging, such as OpenSidewalks.com [41]). While accessible infrastructure is intended to benefit broad user populations from those with unique sensory or physical needs to people with situational impairments [58], our current focus is supporting those with ambulatory disabilities. In Project Sidewalk, we focus on five high-priority areas that impact MI pedestrians drawn from ADA standards [51, 52, 55] and prior work [35, 37]: curb ramps, missing curb ramps, sidewalk obstacles, surface problems, and the lack of a sidewalk on a pedestrian pathway.

Collecting Street-Level Accessibility Data
Traditionally, collecting data on street-level accessibility has been the purview of local and state governments; however, with widespread access to the Internet and smartphones, three alternatives have emerged: in situ crowdsourcing where a user explicitly captures and reports data [11, 15, 36, 38], automatic or hybrid reporting using sensors [8, 29, 31, 44, 48], and remote crowdsourcing using streetscape imagery [19, 23, 25, 26]. Each approach has benefits and drawbacks—e.g., in terms of data type, maintenance, and coverage—and should be considered complementary. While in situ crowdsourcing relies on local knowledge and is likely to produce high-quality data, both academic and commercial tools have struggled with data sparsity [15], perhaps because of high user burden and low adoption. Automatic reporting tools lower user burden by implicitly capturing accessibility data using smartphone- or wheelchair-based sensors; however, accurately converting these quantitative measurements (e.g., accelerometer data) to useful sidewalk assessments is still an open research area. Moreover, these tools are limited to capturing where wheelchair users already go, not where they are unable to go (though [30] is attempting to address this limitation, in part, by combining sensor data with continuous video recording).

Most related to our work are virtual auditing tools of street-level accessibility using streetscape imagery. While initial research focused on establishing the reliability of GSV-based audits compared with traditional, physical-based methods [5, 9, 47, 57], more recent work has introduced and evaluated web-based tools in controlled studies [19, 23, 25, 26]. Project Sidewalk builds on these systems by gamifying the user experience and supporting open-world exploring via missions—similar to first-person video games. Additionally, we present the first public deployment study, which enables us to uniquely compare user behavior and labeling performance across user groups and contributes the largest open dataset on sidewalk quality in existence.

Volunteered Geographic Information (VGI)
Project Sidewalk is a new type of volunteered geographic information (VGI) system [21]. In VGI, non-experts contribute GIS-related data through open mapping tools—e.g., Wikimapia, Mapillary, CycloPath [42], and most notably, OpenStreetMap (OSM). In comparison to more authoritative sources, VGI data quality and spatial coverage are key concerns [3]. While some studies have shown comparable quality between VGI and government maps [20, 22, 34], recent work has identified strong biases in contributions correlated with population density [33, 45]. We address this limitation by combining both volunteer and paid crowd workers and by eliminating the need to have physical access to a place to contribute data. Our work contributes to VGI by analyzing contribution patterns and labeling quality differences between these two user groups.

3 PROJECT SIDEWALK
To use Project Sidewalk, users visit http://projectsidewalk.io on a laptop or desktop (touchscreens are not currently supported). The landing page provides a brief description of the tool—both its purpose and how to use it—along with basic statistics and visualizations to encourage participation. Upon clicking the ‘Start Mapping’ button, new users are greeted by a multi-stage interactive tutorial to learn both about the user interface and basic accessibility concepts. Once the tutorial is completed, users are auto-assigned a neighborhood in DC and given their first mission. Missions guide users through specific neighborhood streets: as the user walks virtually along their route, they are asked to find, label and rate street-level accessibility issues. After completing a mission, a “mission complete” screen is displayed and a new mission is assigned. Users can choose to contribute anonymously or to register and login. We prompt anonymous users to register after finishing their first street segment. Registered users can resume missions and check their contribution activity on an interactive dashboard. Currently, however, there is no way to view or compare performance to others (e.g., a leaderboard).

Training users. Training crowdworkers is difficult, especially for subjective judgment tasks like classifying entities [2]. While a wide range of training approaches are possible—from ground truth seeding with real-time performance feedback to qualification tasks that ensure proficiency [46]—our current training strategy is three-pronged. First, new users are presented with an interactive tutorial, a technique common to modern video games called onboarding [43]. We on-board users through an initial guided mission that explains the UI and key game mechanics, provides information about street-level accessibility concepts, and monitors and helps
Figure 2: In Project Sidewalk, users are given missions to explore city neighborhoods and find accessibility problems. The UI is comprised of four parts: (center) GSV-based exploration and labeling pane; (top) button menu bar; (right) mission pane with progress tracking and navigation; (left) and settings menu. See the Supplementary Video for a demonstration.

the user correct mistakes. As users step through the onboarding experience, their mission status pane is updated just like a normal mission. In total, there are 37 onboarding parts, which are designed to take less than four minutes.

Second, after completing onboarding, initial missions include pre-scripted help dialogs that are triggered based on user behavior. For example, after panning 360° around their first street intersection, Project Sidewalk helps the user use the top-down mission map to take a step in the right direction. These help dialogs are complementary to onboarding: there is an inherent tradeoff between building skills and knowledge through initial training time, and actually having the user begin using the tool in earnest.

Finally, throughout every mission, our tool continuously observes user behavior and provides brief, transient usage tips to encourage proper labeling behavior and increase user efficiency. For example, if we observe that a user is not providing severity ratings, we provide a friendly reminder. If we observe only mouse clicks, we encourage keyboard shortcuts. These one-line tips auto-disappear and can also be explicitly dismissed. Importantly, we cannot provide corrective labeling feedback because we do not know about a label's correctness a priori.

Exploring and labeling. Similar to [23, 26], Project Sidewalk has two modes of interaction: explorer mode and labeling mode. In explorer mode, users follow turn-by-turn directions to explore their assigned mission routes using GSV's native navigation controls. If users get lost exploring, they receive reminders to return to their mission routes, which can be clicked to auto-jump back. At any location, the user can pan (360° horizontally and 35° vertically) and zoom to assess sidewalks more closely. The user's FOV is 89.75°.

Users enter the labeling mode by clicking on a labeling button. There are five primary label types: curb ramp, no curb ramp, obstacle, surface problem, and no sidewalk (Figure 3). In this mode, all interactions for controlling movement and the first-person camera view (e.g., pan, pitch) are disabled and the mouse cursor changes to a circular icon representing the selected label. To place a label, the user clicks directly on the accessibility target in the GSV image. A context menu then appears, which asks the user to rate problem severity on a 5-point scale where ‘5’ represents an impassable barrier for
someone in a manual wheelchair. The user can also enter additional notes in a description text field or mark a problem as temporary (e.g., due to construction). After closing the context menu, Project Sidewalk automatically reverts to explorer mode.

Project Sidewalk seamlessly repositions applied labels in their correct location as the user pans or zooms—thus, labels appear to “stick” to their associated target. However, once a user takes a step, their labels are no longer visible in the GSV interface (unless they return to their original labeling location). This is due to GSV API limitations. Instead, previously placed labels can be viewed on the top-down mission map.

Missions. Missions serve a two-fold purpose: first, as a game mechanic, they provide an easy-to-understand and engaging narrative for directing data collection tasks. Second, from a system design perspective, missions provide a flexible approach to discretize, assign, and distribute work. Though we envision a variety of future mission types—e.g., data validation missions, labeling user supplied imagery—our current system focuses on encouraging exploration and labeling in the GSV interface. Users are assigned a high-level goal of auditing a neighborhood and then routed on missions of increasing length and complexity within that neighborhood. Mission lengths increase from 500ft to a maximum of 0.5mi (2,640ft). Mission feedback is provided via a mission status pane, completion screens, and, for registered users, an interactive dashboard. If a user gets stuck during a mission, they can choose to “jump” to a different part of their assigned neighborhood or manually choose a new neighborhood. For finishing a mission or completing a neighborhood, users are rewarded with mission completion screens and sound effects.

4 IMPLEMENTATION, DATA, AND API
Creating a robust, usable, and publicly deployable system required a significant human-centered design and engineering effort. Our open-source GitHub repository\(^2\) has 2,747 commits from 20 team members and 43,898 lines of developed code (excluding comments). Project Sidewalk’s backend is built in Scala and PostgreSQL with the PostGIS spatial extension, and the frontend is in JavaScript and HTML/CSS. Below, we describe four key implementation areas: preparing a city for deployment, work allocation algorithms, triangulating and clustering labels, and our API.

Preparing a City
Project Sidewalk has two data prerequisites for deployment: GSV and OSM availability. To construct a street network topology, we extract OSM `<way>` elements marked with street-related tags within a city’s geographic boundary. We also extract `<node>` and `<nd>` elements for metadata (e.g., lat-long coordinates) and links between nodes and edges. Because `<way>` polylines can extend multiple city blocks, we create smaller units, called `street segments`, by partitioning streets at each intersection. For DC, this resulted in 15,014 street segments with a total length of 1,164 miles. We filtered 892 segments that contained highways and/or where GSV imagery was unavailable due to government security precautions. In total, we were left with 14,037 segments over 1,075 miles (Figure 4).

Allocating and Distributing Work via Missions
Allocating and distributing work is a two-step process consisting of assigning neighborhoods then street segments. We use the `mission` construct to do both. We iterated on these task allocation algorithms throughout our deployment as we discovered inefficiencies or mistakes. Below, we present our
current approach, which was used for the last three months of our deployment, and briefly mention old approaches.

Our current version is based on a "work quality" threshold determined by analyzing labeling behavior from our research group and informal manual reviews of end-user contributions. We define a "good" user as someone who contributes a minimum of 3.75 labels per 100 meters on average. While labeling frequency is an imperfect proxy for worker quality, it is easy to implement and fast to compute. We integrate this quality metric to prioritize street segments:

\[
\text{priority}_{\text{street}} = \begin{cases} 
1, & \text{if } \text{cnt(‘good’ users)}>0 \\
\frac{1}{1+x}, & \text{otherwise}
\end{cases}
\]

where, \(x = \text{cnt(‘good’ users)} + 0.25 \times \text{cnt(‘bad’ users)}\). This algorithm prioritizes street segments inversely proportional to the number of previous audits with a weight penalty assigned for "bad" users.

**Allocating neighborhoods.** Users are given missions to explore and label assigned neighborhoods. Neighborhoods are allocated at two points: after a user completes onboarding and after they complete a previously assigned neighborhood. In earlier versions of Project Sidewalk, we randomly assigned users to neighborhoods within the top ten lowest completion rates. This approach, however, treated all previous work equivalently. In the current version, we incorporate street segment priority by first calculating the mean priority of all street segments for each neighborhood and then randomly assigning neighborhoods from a list with the top five highest means. Users can also choose their own neighborhoods; however, this feature was somewhat hidden and not prominently used in our deployment.

**Calculating mission routes.** Mission routes are composed of street segments, which are dynamically selected when a user reaches an intersection (i.e., the end of a segment). To enhance immersion and limit user confusion, the routing algorithm attempts to select contiguous segments whenever possible. In older versions of Project Sidewalk, the segment selection algorithm simply chose a randomly connected segment that the current user had not already audited. However, this failed to incorporate work completed by other users, which was inefficient. In our current implementation, for each neighborhood, we maintain a discretized list of unaudited street segment priorities (bin size=0.25). When a user reaches an intersection, we randomly select any unaudited connected street segment with the same discretized priority as the highest one in the neighborhood list. If none exist, we inform the user that they have completed this part of the neighborhood and automatically transport them to the highest priority remaining neighborhood street. We use a similar process for positioning users when they first begin

a new neighborhood—we place them at the beginning of the highest priority street segment.

**Project Sidewalk Data**

In Project Sidewalk, users label streetscape panoramas projected into 3D space [17]. We need to convert these 3D-point labels to 2D lat-lng coordinates and then aggregate multiple labels for the same target into a single cluster.

**3D to 2D.** To obtain geo-located labels from the 3D projection, we use: (i) the panorama’s 3D-point cloud data, which is obtained by LiDAR on the GSV cars; (ii) the \(\text{lng, lat}\) coordinate of the GSV car; and (iii) the \(x_{\text{img}}, y_{\text{img}}\) position of the label on the panorama. More specifically:

\[
\frac{\ln_{\text{target}}}{\text{lat}_{\text{target}}} = \left(\frac{\ln_{\text{GSV-car}}}{\text{lat}_{\text{GSV-car}}} + \left(\frac{\Delta \text{lng}}{\Delta \text{lat}}\right)\right)
\]

where, we compute \(\Delta \text{lng}, \Delta \text{lat}\) by using the \(x_{\text{img}}, y_{\text{img}}\) label position on the panorama and the 3D-point cloud data to obtain the offset \(dx, dy, dz\) at \(x_{\text{img}}, y_{\text{img}}\). The offset is in meters, which we convert to \(\Delta \text{lng}, \Delta \text{lat}\) and plug into the equation. See the function `imageCoordinateToLatLng(imageX, imageY, lat, lng)` in MapService.js (Line 1275) in the GitHub repo.

**Raw label data.** For each label, we record three sets of information: who provided the label and when, how the data was collected in GSV (the user’s POV, heading, source panorama id), and information about the label itself, such as label type, lat-long position, x.y.position on panorama, severity rating, textual description, and a temporary flag.

**Clustering.** Because users can find and label the same accessibility problem from different panoramas, we needed to develop an algorithm to aggregate labels for the same target together. We do this by clustering. Each cluster refers to a single found problem (and may contain one or more raw labels). We use a two-stage clustering approach: single-user clustering followed by multi-user clustering. First, we consolidate raw labels for each individual user into intermediate clusters—this is necessary because some users choose to label a single problem from multiple viewpoints. Second, we combine these individual user clusters together to create our final cluster dataset. Both stages use the same hierarchical agglomerative clustering approach: the Vorhees clustering algorithm with the haversine formula to compute distances between labels and clusters.

For stage one, we cluster raw labels of the same type that are within a certain distance threshold. Because some label types are often legitimately close together—e.g., two curb ramps on a corner—we use two different thresholds: 2 meters for curb and missing curb ramps and 7.5 meters for other label types. These thresholds were determined empirically by iteratively computing clusters at different threshold levels from 0 to 50 meters (step size=1 meter) and qualitatively analyzing the results. Stage two clustering is similar but
uses the centroids of stage one clusters with slightly looser thresholds (7.5 and 10 meters, respectively).

Public API

To enable the use and broader study of our collected data, we developed and released an initial public REST API (http://projectsidewalk.io/api). The API has three endpoint types: *labels* for obtaining raw label data, *clusters* for obtaining label clusters, and scores, which provide computed scores for street and neighborhood accessibility. Each API requires a lat-long bounding box to specify an area of interest for input and returns data in the *GeoJSON* format. For the score APIs, we developed a simple scoring model that incorporates the number of problem labels and returns an accessibility score between 0 and 1. Providing a robust, personalizable, and verifiable scoring algorithm is ongoing work.

5 DEPLOYMENT STUDY

In August of 2016, we launched an 18-month deployment study of Project Sidewalk. Washington DC was selected as the study site because of its large size (158 km$^2$), diverse economic and geographic characteristics, and substantial commuter population—many of whom take public transit and use pedestrian infrastructure [54]. Additionally, as the nation’s capital, which draws ~20m visitors/yr [56], there is increased pressure to follow and model ADA guidelines.

We recruited two types of users: *volunteers* through social media, blog posts, and email campaigns, and *paid crowd workers* from Amazon Mechanical Turk (turkers). We further divide volunteers into *anonymous* and *registered* groups; the former was tracked by IP address. For comparison, we also show data from 28 members of our research lab, who voluntarily contributed to help test the tool and received in-person training on how and what to label. We paid turkers a base amount for completing the tutorial and first mission ($0.82) and a bonus amount for each mission completed thereafter ($4.17/mile). These rates were based on US federal minimum wage ($7.25/hr), assuming an expected labeling rate of 1.74 miles/hr, which was drawn empirically from our data. In practice, our turkers earned $8.21/hr on average (SD=$5.99), which increased to $12.76 (SD=$6.60) for those 69 turkers who audited at least one mile. Turkers could see their earnings in real-time via the mission panel. We posted a total of 298 assignments over a 6-month period.

Results

Overall, Project Sidewalk had 11,891 visitors to the landing page, of which 797 (627 volunteers; 170 turkers) completed the tutorial and audited at least one street segment in the first mission. In total, these users contributed 205,385 labels and audited 2,941 miles of DC streets (Table 1). Below, we analyze user behavior, contribution patterns, and responses from a pop-up survey given to turkers. We examine worker and data quality in a separate section.

User behavior. On average, registered users completed more missions (5.8 vs. 1.5), contributed more labels (171.1 vs. 33.7), audited faster (1.93 mi/hr vs. 1.22), and spent more time on Project Sidewalk (55.8 mins vs. 18.3) than anonymous users (Table 2). Registered users also took longer on onboardings (6.9 mins vs. 3.8) and left more open-ended descriptions (10.0 vs. 1.6). Paid workers, however, did significantly more work on average than either volunteer group: 35.4 missions, 887.3 labels, and spent 4.4 hrs using the tool. If we examine only those users who passed our “good” user heuristic, we filter 28.2% paid, 23.7% anonymous, and 22.6% registered workers; however, relative user behaviors stay the same. Similar to [30], user contribution patterns resemble a power law distribution: the top 10% anonymous, registered, and paid workers contributed 56.7%, 86.6%, and 80.2% of the labels in their group, respectively. By the top 25%, contribution percentages rise to 77.4%, 93.6%, and 94.8%.

User dropout. To examine user dropout, we analyzed interaction logs for the last eight months of our deployment (after we added comprehensive logging to the tutorial). User dropout was steep. While 1,110 users started the tutorial, only 568 finished it (51%), 479 (43.2%) took one step in their first mission, and 328 (29.5%) completed at least one mission. Of
those 328, a majority, went on to finish their second mission (59.8%; 196 users) and then dropoff dampened substantially. For example, 74.0% of the users who completed Mission 2 also completed Mission 3. When splitting the 1,110 users by group—846 volunteers and 264 turkers—we found different patterns of behavior. While only 43.9% of volunteers finished the tutorial and only 19.1% finished the first mission, turkers were far more persistent: 74.6% finished the tutorial and 62.9% completed the first mission.

**Pop-up survey.** To begin exploring why users contribute to Project Sidewalk, we developed a 5-question survey shown to users after their second mission. The first three questions asked about task enjoyment, difficulty, and self-perceptions of performance via 5-point Likert scales while the last two questions were open-ended asking about user motivation and soliciting feedback. A single researcher analyzed the two write-in questions via inductive coding. Though the survey is now given to all user groups, it was only available to turkers during our deployment study—which we analyze here.

In all, 123 turkers completed the survey. Of those, 110 (89.4%) stated that they enjoyed using Project Sidewalk ($M=4.4$; $SD=0.7$). For task difficulty, the responses were slightly more mixed: 83 turkers (67.5%) selected *easy* or *very easy* and 5 selected *difficult* ($M=3.9$; $SD=0.9$). When asked to self-rate their performance, 81 turkers (65.9%) felt that they did at least a *very good* job and none reported *poor* ($M=4.0$; $SD=0.9$). For the first open-ended question (required) about user motivation, 74 (60.2%) mentioned that the task was interesting or fun—“It was an interesting and unique change to my day” (U111); 48 (39.0%) felt that the task was important/helpful—“I think it is important for those who are using wheelchairs to be able to safely navigate streets.” (U223); and 20 (16.3%) mentioned money—“It was interesting work and good pay” (U61). The last question was optional and asked for feedback: 68 turkers chose to answer, mostly to thank us for the task (55 of 68): “Good & interesting task. Thank you” (U96). Six suggested features, five asked questions about labeling, and two reported bugs.

### 6 DATA VALIDATION STUDY

To investigate data quality and compare performance across user groups, we performed a data validation study using a subset of DC streets. This study occurred approximately halfway into our public deployment. Because pedestrian infrastructure can differ based on neighborhood type (e.g., commercial vs. residential), age, and density, we first divided DC into four quadrants based on official geographic segmentation data [12]. We then sub-divided each quadrant into land-use zones using DC’s open zoning regulation dataset [14]. Finally, we randomly selected the first two or three mission routes completed by individual volunteer users. This resulted in a test dataset of 44 miles (625 street segments) from 50 registered and 16 anonymous users across 62 of the 179 DC neighborhoods. We then verified that the selected routes had similar geographic and land-use distributions compared to all streets in DC.

To compare volunteer vs. paid worker performance, we posted the selected missions in our test dataset to Amazon Mechanical Turk. Other than payment, we attempted to carefully mimic the volunteer work experience: individual turkers completed onboarding and then were implicitly assigned either an anonymous user’s mission set (two) or a registered user’s mission set (three). To control for experience and learning effects, we did not allow deployment turkers to participate. We paid workers based on US federal minimum wage drawn from median volunteer completion times: $2.00 for the tutorial + two missions (~2,000ft) and $3.58 for the tutorial + three missions (~4,000ft). Unlike the deployment study, turkers could not choose to complete additional missions for bonus payment. To examine the effect of multiple labelers on performance, we hired five turkers per mission set for a total of 330 turkers.

To create ground truth, we first developed a labeling codebook based on ADA guidelines [51, 52, 55], which was then vetted and refined by a person who has used a wheelchair for 20 years. Following iterative coding [27], three researchers began labeling the same subset of data: one randomly selected mission set for an anonymous user and one for a registered user. For each round, the researchers met, resolved disagreements, and updated the codebook accordingly. After seven rounds, the average Krippendorff alpha score was 0.6 ($range=0.5-0.8$) and raw agreement: 85.4% ($SD=4.1$%). The three researchers then split the remaining 52 mission sets equally and a final review was performed. In total, ground truth consists of: 4,617 clusters, including 3,212 curb ramps, 1,023 surface problems, 295 obstacles, and 87 missing curb ramps. Though laborious, we note that this ground truth approach allows us to more deeply examine labeling performance compared with verifying placed labels—as the latter does not allow us to calculate false negatives.

**Analysis.** We examine accuracy at the street-segment level. We first cluster all labels from anonymous, registered, and paid workers using single-user clustering. We then use haversine distance to associate label clusters to their closest street segment. To compute our accuracy measures, we sum the number and type of label clusters for each segment and compare the result to ground truth. This produces counts of true/false positives and true/false negatives at each segment, which we binarize for final analysis. In total, 89.6% (560/625) of the street segments contained accessibility labels in ground truth. Unlike the four other label types, the no sidewalk label is not used for single-point targets but rather targets that extend multiple panoramas. Thus, we exclude this label from our analysis.
We examine overall performance across user groups, the majority we hired five turkers per mission set. We examine was 71.7% (labeled) compared to those in ground truth while precision measures the correctness of those labels. Ideally, each measure would be 1.0; however, similar to other crowdsourcing systems (e.g., [26]), we prefer high recall over precision because correcting false positives is easier than false negatives—the former requires verification while the latter requires users actually re-explore an area. Except for the multiple labelers per segment analysis, we use only the first hired turker for each mission (rather than all five). For statistical analysis, we use binomial mixed effects models with user nested in mission route id and a logistic link function with accuracy, recall, and precision modeled as binomials. We assess significance with likelihood-ratio (LR) tests and use post-hoc Tukey’s HSD tests to determine statistical orderings. Our analysis was performed using the R statistical language.

Results
We examine overall performance across user groups, the effect of label type, label severity, and multiple labelers on accuracy, and common labeling mistakes.

User performance. The overall average raw accuracy was 71.7% (SD=13.0%) with all three user groups performing similarly (~70%). Because of the high true negative rates in our data—that is, most panoramas do not have accessibility issues and were correctly labeled that way—recall and precision are more insightful measures (Figure 5). Turkers found significantly more issues than registered and anonymous users (recall=67.8% vs. 61.4% vs. 48.8%, respectively) at similar precision levels (68.8% vs. 72.2% vs. 74.5%). With an LR test, user group had a statistically significant association with recall ($lr=21.6$, $df=2$, $n=132$, $p<0.001$) and precision ($lr=7.1$, $df=2$, $n=131$, $p=0.028$) but not raw accuracy. Pairwise comparisons for recall were all significant but none were for precision.

To explore the effect of multiple labelers on performance, we hired five turkers per mission set. We examine majority vote for each group size (3, 5) as well as treating each contribution individually (e.g., Turk3maj vs. Turk3all). We expect that Turkmaj will result in higher precision but lower recall as it requires more than one user to label the same target and just the opposite from Turkall (i.e., higher recall, lower precision). Indeed, this is what we found: from Turk1 (baseline) to Turk5all, recall rose from 67.8% to 91.7% but at a cost of precision (from 68.8% to 55%). In contrast, for majority vote, recall fell from 67.8% to 59.5% for Turk1 to Turk5maj but precision rose from 68.8% to 87.4%. We found turker group had a statistically significant association with recall ($lr=498.96$, $df=4$, $n=330$, $p<0.001$) and precision ($lr=374.88$, $df=4$, $n=330$, $p<0.001$). All pairwise comparisons for recall and precision were significant except for Turk5maj < Turk3maj—for recall only.

Label type. To examine accuracy as a function of label type, we analyzed labeling data across users (Table 3). Curb ramps were the most reliably found and correctly labeled with recall=86.0% and precision=95.4%. In contrast, while no curb ramps had reasonably high recall at 69.3%, precision was only 20.5% suggesting an incorrect understanding of what justifies a no curb ramp label. The other two label types, obstacle and surface problem, had lower recall (39.9% and 27.1%) but comparatively higher precision (47.5% and 72.6%), which mirrors our experience with ground truth—these accessibility problems are hard to find and require diligent exploration. In addition, these two label types can legitimately be switched in some cases (e.g., a patch of overgrown grass could be marked as either an obstacle or surface problem). We explore labeling mistakes in more detail below.

Effect of severity. We hypothesized that high-severity problems would be easier to find. To explore this, we partitioned ground truth labels into two groups: low severity ($<=$ 2 rating) and high severity ($>$ 3 rating). The low severity group contained 1,053 labels and the high 352 labels. As expected, we found that high-severity labels had significantly higher recall ($M=83.3%$; $avg=69.8%$; $SD=35.5%$) than low-severity labels ($Mdn=56.3%;$ $M=57.0%$; $SD=32.3%$). To determine significance, we created a binomial mixed effect model with severity (high or low) as the fixed effect and user nested in mission route id as random effects. Result of LR test ($lr=10.6$, $df=1$, $n=246$, $p=0.001$).

Common Labeling Errors
To better understand labeling errors and to contextualize our quantitative findings, we conducted a qualitative analysis of labeling errors. We randomly selected 54 false positives and 54 false negatives for each label type, which resulted in 432
Figure 6: An overview of false positive and negative labeling mistakes ordered by frequency (taken from 432 error samples in the data validation study).

In analyzing false positives, we observed that most mistakes were understandable and either easy to correct with better training or innocuous. For example, 66.6% of incorrect curb ramp labels were applied to driveways, nearly half of obstacles and surface problems were potentially legitimate issues but not on the primary pedestrian route (e.g., middle of street vs. crosswalk), and almost 30% of incorrect missing curb ramps were on extended residential walkways. Moreover, 32.7% of surface problems and 9.3% of obstacles were correctly labeled as problems but with a different label type from ground truth—e.g., a surface problem marked as an obstacle.

For false negatives (i.e., a user did not label a problem when one exists), it is harder to discern clear patterns—at least for some label types. For obstacles and surface problems—both of which had the lowest recall and thus can be considered hardest to find—salience appears to be a contributing factor: 50% of missed obstacles were only partially blocking the pedestrian path and nearly 30% of surface problems were grass related. For missing curb ramps, 46.3% of missed labels were at a corner where at least one other curb ramp exists though the second most common error was more egregious: a pedestrian path to a street had no curb ramp and no alternative accessible route (37.0%). We discuss potential solutions to addressing labeling errors in the Discussion.

7 SEMI-STRUCTURED INTERVIEW STUDY

To complement our deployment and data validation studies and to solicit reactions to Project Sidewalk from key stakeholders, we conducted an interview study with three DC-area groups (N=14): six government officials (G), five people with mobility impairments (MI), and three caregivers (C). G included state and city transportation employees with oversight of pedestrian infrastructure, MI participants used a mobility aid such as a wheelchair or cane, and caregivers took care of a person with a MI either as a professional, family member, or friend. Participants were recruited via mailing lists, word-of-mouth, and social media.

The three-part study began with a semi-structured interview about participants’ current perceptions of and problems with urban accessibility. We then asked participants to use Project Sidewalk while “thinking aloud.” Finally, we concluded with a debrief interview about the tool, including its perceived utility, concerns, and design ideas. Sessions lasted between 60-65 minutes, and participants were compensated $25. One government session was a group interview with three participants (coded G3); all other interviews were individual. Sessions were audio- and screen-recorded, which were transcribed and coded to find emergent themes using peer debriefing [10, 49]. Using deductive coding, one researcher created an initial codebook for the interviews, which was refined with the help of a peer. A randomly selected transcript was then coded, which was reviewed by a second researcher using peer debriefing. To resolve conflicts and update the codebook, the two researchers met after each review process. The final codebook was produced after three iterations (with one transcript coded per stakeholder group) and 46 conflict resolutions over 305 excerpts and 1,466 applied codes. The remaining data was then coded by the initial researcher.

Results

We describe findings related to the perceived value and usability of Project Sidewalk as well as design suggestions and concerns. For quotes, we use (participant group + id).

Perceived value. Overall, all three stakeholder groups felt that Project Sidewalk enabled rapid data collection, allowed for gathering diverse perspectives about accessibility, and helped engage citizens in thinking about urban design. Government officials emphasized cost savings and community involvement envisioning Project Sidewalk as a triaging tool before sending out employees to physically examine...
areas: “It’s really good for a starting point. This is a first observation, and when you send somebody out in the field, they can see those observations and pick up more information. It’s just neat” (G4). The MI and caregiver groups focused more on personal utility, envisioning accessibility-aware navigation tools that could incorporate Project Sidewalk data: “I might take advantage of more opportunities knowing that, okay, if I could rely on the data and knew I could anticipate how difficult it was going to be for me to get around” (M11). Six of the seven MI and caregiver participants mentioned that Project Sidewalk data could enhance their independence, give them confidence to explore new and unfamiliar areas, and/or help them achieve the same pedestrian rights as everyone else.

**Usability.** Participants across groups felt that the tool was easy-to-learn and fun to use. G3, for example, stated: “I think it’s awesome. […] It’s a lot of fun” and reported “feeling good” contributing data to a social purpose while also being motivated by the game design elements: “we’re looking at the 71 percent complete, and we’re pretty excited!” Three participants appreciated relying on a familiar technology like GSV, “You’re not introducing like yet another platform that somebody has to relearn—that was helpful” (G3). Almost everyone (13/14) found the labeling system comprehensive as captured by M13: “the labeling is pretty all-inclusive.”

**Concerns.** Key concerns included outdated GSV imagery or labels (N=6), data reliability (3), and conflicting data (4). Towards outdated imagery and labels, C1 asked “if a street light was marked as an obstacle and if it was replaced or moved, would the labels reflect that?” While this is one limitation of our virtual auditing approach, four participants mentioned that they would rather be aware of a potential issue even if it no longer existed. For example, C2 stated: “if there was a label, I’d rather be aware of it.” For data reliability, G4 suggested that each road be audited by multiple people: “I would have more confidence if different people did it, did the same street.” Four participants (2 Cs, 2 MIs) were concerned about how labelers may differ in interpreting problems compared with former, prior work has found high agreement between virtually verified labels [6, 46], allocates the number of assigned labelers per street based on inferred performance, and integrates other datasets (e.g., top-down imagery). Second, though not explored in this paper, our mission-based architecture supports a large variety of diverse mission tasks—e.g., verification missions and ground truth seeding missions, both which will enable us to more reliably identify poor-quality workers. Third, Project Sidewalk currently relies solely on manual labeling; we are experimenting with deep learning methods trained on our 240,000+ image-based label dataset to detect problems automatically (building on [26, 50]), triage likely problem areas, and/or aid in verification missions. Finally, our results suggest that many false positives could be corrected via improved training (e.g., a driveway is not a curb ramp) and by using simple automated validation (e.g., check for labels in unlikely areas).

**Suggestions.** Participants suggested developing mechanisms to keep information up-to-date (4)—for example, by adding a complementary smartphone-based data collection app, adding verification interfaces (3), and surfacing data age (2). All government officials were interested in ways to export and visualize the data; one suggested integrating directly into their service request backend. At a more detailed tool level, seven participants suggested adding new label types, including for crosswalks, the presence of sidewalks, access points (such as driveways), and construction.

## 8 DISCUSSION AND CONCLUSION

Through a multi-methods approach, our results demonstrate the viability of virtually auditing urban accessibility at scale, highlight behavioral and labeling quality differences between user groups, and summarize how key stakeholders feel about Project Sidewalk and the crowdsourced data. Below, we discuss worker and data quality, future deployment costs and worker sources, and limitations.

**Label quality.** Our data validation study found that, on average, users could find 63% of accessibility issues at 71% precision. This is comparable to early streetscape labeling work by Hara et al. [25], where turkers labeled at 67.0% and 55.6% for recall and precision, respectively; however, our tasks are more complex, contain more label types, and are evaluated at a larger scale. Like [25], we also show how assigning multiple labelers can improve results and describe tradeoffs in aggregation algorithms—e.g., by combining labels from five turkers per street, recall rose to 92%; however, precision fell from 69% to 55%. We believe our findings represent a lower bound on performance and provide a nice baseline for future work.

To improve quality, we envision four areas of future work: first, a more sophisticated workflow pipeline that dynamically verifies labels [6, 46], allocates the number of assigned labelers per street based on inferred performance, and integrates other datasets (e.g., top-down imagery). Second, though not explored in this paper, our mission-based architecture supports a large variety of diverse mission tasks—e.g., verification missions and ground truth seeding missions, both which will enable us to more reliably identify poor-quality workers. Third, Project Sidewalk currently relies solely on manual labeling; we are experimenting with deep learning methods trained on our 240,000+ image-based label dataset to detect problems automatically (building on [26, 50]), triage likely problem areas, and/or aid in verification missions. Finally, our results suggest that many false positives could be corrected via improved training (e.g., a driveway is not a curb ramp) and by using simple automated validation (e.g., check for labels in unlikely areas).

**Data age.** Our interview participants raised two concerns about data age: GSV image age and label age. Towards the former, prior work has found high agreement between virtual audit data of pedestrian infrastructure compared with traditional audits [5, 9, 23, 26, 47, 57]. Google does not publish how often their GSV cars collect data; however, in a 2013 analysis of 1,086 panorama sampled across four North American cities, the average age was 2.2yrs (SD=1.3) [26]. In our dataset, workers labeled 74,231 panoramas, which at the time of first label, were also M=2.2yrs old (SD=1.5). As a comparison, the official opendata.dc.gov curb ramp dataset [13] was captured in 1999 and last updated in 2010 (nine
years ago) but this only covers curb ramps (no other label
types are included). Our general approach should work with
any street scene imagery dataset, including Mapillary [30],
CycloMedia, or Bing StreetSide—many of which are explor-
ing high-refresh methods via automated vehicles and crowds
contributions. In terms of maintaining labels over time, one
benefit of our scalable approach is that streets can be periodi-
cally re-audited and old labels can be used to study historical
change (e.g., as initially explored in [39]).

Cost. While future deployments could rely solely on paid
workers, ideally Project Sidewalk would also engage on-
line and local communities who are concerned with urban
accessibility. Based on our deployment study, we estimate
that auditing DC with 100 paid workers alone would cost
$34,000 and take 8 days (assuming five labelers/street, 8hrs of
work per day, and that 72 of 100 met our “good” user quality
threshold). If one-third of DC was audited by volunteers,
costs fall below $25,000. However, DC is a large city and
has a reasonably well-resourced transportation department
with full-time ADA compliance staff; small-to-medium sized
cities often lack ADA budgets and could particularly benefit
from Project Sidewalk. Indeed, we have been contacted by
more than a dozen cities in the US and Canada about future
deployments.

Increasing user engagement. While ~63% of turkers
who started the tutorial went on to complete one mission,
this value was 3x lower—19.1%—for volunteers. To increase
user engagement, we plan to explore: (1) supporting smartphones,
which will increase the reach of the tool and allow
any-time access (e.g., users can complete missions while on
the bus or subway). This will hopefully result in more re-
peated visits and higher mission completion rates (our web
logs show nearly 25% of traffic is mobile); (2) providing users
with visual feedback about the impact of their contributions
(e.g., via accessibility visualizations like [32]); (3) incorporat-
ing more gamification principles such as additional mission
types (e.g., rapid data validation mini-games, scavenger hunt
missions), badges, and leaderboards—all of which have been
shown to improve retention in VGI systems [16]; and (4) and
better engaging the local community through outreach ef-
forts to pedestrian and accessibility advocacy organizations.

Limitations. There are three main limitations with crowd-
sourcing virtual audits: panorama age, label quality, and the
ability for crowdworkers to see and assess sidewalks from
GSV. We addressed the former two points above. Towards
the latter, users could mark areas as occluded in our tool (e.g.,
a truck blocking a sidewalk); however, occlusion constituted
only 0.4% of all applied labels in our deployment suggesting
that most sidewalks are visible. For study limitations, we
employed a multi-methods approach to mitigate the effects
of any one study technique. Still, longitudinal deployment
studies are messy and ours is no exception: we lost over
two months of deployment time due to changes in the GSV
API, maintenance upgrades to our servers, and personnel
changes. For the data validation study, we were unable to
consistently reach high agreement for obstacles and surface
problems during our seven iterative rounds of coding; these
label types are challenging and can be legitimately conflated
(e.g., marking overgrown grass as a surface problem vs. an
obstacle). Our performance results for these label types may
have been impacted.

Finally, while our studies take place in the US, accessible
infrastructure is a global problem. Project Sidewalk should,
ostensibly, work wherever GSV and OSM are available. That
said, Project Sidewalk’s label types were drawn from US ADA
standards [51, 52, 55], prior work [35, 37], and our previous
experience working with US-based stakeholders. While we
believe that these label types constitute primary accessibility
barriers for people with mobility impairments and are
likely relevant to most North American and European cities,
more work is necessary to explore mobility barriers in other
regions. As we plan future deployments, we will work with
local stakeholders to better understand regional contexts,
socio-cultural concerns, and unique, localized infrastructural
accessibility issues. Project Sidewalk can be updated per re-
gion to, for example, add specific label types or instructions
for a city.

9 ACKNOWLEDGEMENTS

This work was supported by an NSF grant (IIS-1302338), a
Singapore MOE AcRF Tier 1 Grant, and a Sloan Research
Fellowship. We thank Soheil Behnezhad, Daniil Zadorozhnyy,
and Johann Miller for their code contributions and Rachael
Marr for designing our landing page and logo.

REFERENCES

www.leagle.com/decision/1993S17069FD110671900
Ahmed Loai Ali, Nuttha Sirilertworakul, Alexander Zipf, and Amin
Mobasher. 2016. Guided Classification System for Conceptual Over-
lapping Classes in OpenStreetMap. ISPRS International Journal of
Geo-Information 5, 6 (Jun 2016), 87. https://doi.org/10.3390/ijgi5060087
Quality: An Overview. ISPRS Annals of the Photogrammetry, Remote
et/II-3/W5/345/2015/
han, Kathryn M. Neckerman, Andrew G. Rundle, and Julien O. Teitler.
2015. Development and deployment of the Computer Assisted Neigh-
borhood Visual Assessment System (CANVAS) to measure health-
related neighborhood conditions. Health & Place 31 (Jan 2015), 163–172.
https://doi.org/10.1016/J.HEALTHPLACE.2014.10.012
[5] Hannah M Badland, Simon Opit, Karen Witten, Robin A Kearsns, and
Suzanne Mavoa. 2010. Can virtual streetscape audits reliably replace


