A pedestrian-centered approach for equitable access to built Environments

Nick Bolten, Sumit Mukherjee, Veronika Sipeeva, Anissa Tanweer, Anat Caspi

Abstract

Equitable pedestrian access is crucial for a barrier-free city, where people with different abilities
can independently access streets and services using current and relevant routing information.
Pedestrians present heterogeneous information requirements about the urban street environment,
consisting of both static and transient information. Similarly, governmental stakeholders such as
municipalities, transportation agencies, and city planners require an accurate base layer representing the built environment, and additional static/transient data types. However, such data
are generally unavailable in a user-consumable format, which leads to several challenges in
developing pedestrian-facing applications. In this paper, we describe these challenges in the
context of developing a customizable routing solution for pedestrians with disabilities. Because
existing routing solutions primarily optimize for short street distances and are unaware of
barriers such as steep inclines, a pedestrian-centric trip planning solution is essential. As part of
this case study, we demonstrate that a data model for equitable pedestrian wayfinding must
flexibly support an annotated pedestrian network in an open, shared spatio-temporal data
commons that can adequately support the data provenance needs of all stakeholders. Finally, we
generalize our experience to a model of a community-mediated data commons that can contribute
to a better functioning of the public sector.

Introduction

The pedestrian spatiotemporal information gap

Pedestrian travel on sidewalks is a primary mode of travel that supports nearly all others,
facilitating professional, recreational and community activities. Nearly everyone uses pedestrian

ways, and there are multiple stakeholders and institutions with an interest in having a detailed awareness of the pedestrian built environment. Analytics of sidewalk routes have been used in the fields of public health [1], urban infrastructure investment and transportation [2,3], and emergency planning [4]. Integrated pedestrian data can greatly facilitate infrastructure planning, use by the public, and navigability of pedestrian paths, and is a fundamental requirement of automated, accessible pedestrian trip planning.

Ubiquitous features of the pedestrian built environment such as sidewalk surface type, evenness of terrain, stairs, curbs, elevation changes, and construction pose severe outdoor barriers for individuals with limited mobility, as demonstrated by numerous studies [5,6]. While some criteria such as elevation or sidewalk location may change infrequently, others are transient, such as construction or weather conditions. In a formative study, Hara et al. determined the mobility assistive device they use. Because the importance of this information varies from individual to

individual, pedestrian informational needs go well beyond what established automotive trip planning services, which assume little heterogeneity in access by motor vehicles, can currently provide. For example, whereas grassy surfaces or curbs can be negotiated by cane users, they pose a significant barrier to wheelchair users, who seek out detours. Among wheelchair users, those with motorized wheelchairs can negotiate surfaces or slopes that are impossible for manual wheelchair users [5]. Diverse factors like strength and weight of the user, familiarity with the assistive device, and personal comfort are all considerations for determining the most appropriate pedestrian path for a given individual. In addition, features of the built environment, like curb ramps that encompass an entire sidewalk corner, can work well for wheelchair users, but pose a challenge for someone with a visual impairment seeking heading cues for crossings at an intersection.

Heterogeneous pedestrian information requirements suggest the need for detailed, specific data on the built environment. While geodata for automobile road networks are widely available in the U.S., digital sidewalk geodata are scarce, have no standard format, and when available are produced and maintained by local municipalities and not coordinated at the regional level [7].
Existing trip planning services cannot leverage pedestrian information that is not available, and often produce routes that include potential mobility barriers (Figure 1). Kasemsuppakorn et al. studied the routing preferences of 5 participants with mobility impairments. Although subjects had to travel an average of 14.64% farther, they rated the personalized routes to have better path quality than shortest-distance [8]. While this study demonstrates the benefits of having a personalized routing solution, it is necessarily of limited scale due to the lack of relevant data and subsequent need to manually curate routes for a limited geographic region.

The requirements of automated trip planners: challenges with pedestrian geodata

Optimal route-planning algorithms interpret linear geodata as a connected graph, where each edge is assigned a cost based on difficulty of travel using a cost function and the optimal path between vertices has the lowest summed cost. For a transportation routing graph to be useful, it must (1) have nearly complete edge coverage, so that no essential paths are missing, (2) be spatially accurate so that graph vertices can be generated where paths meet, and (3) data relevant to calculating edge costs must be relational such that they can be associated with a specific path (e.g., speed limits for streets). For automotive routing, street centerlines constitute graph edges and street intersections constitute graph vertices. Street data is ubiquitous and curated at the national level and focused on transportation network analysis, and so is consequently near-complete and spatially accurate. Relational data for calculating edge costs need only be related to streets spatially or street metadata, and is most often in the form of travel estimates for a small number of travel modes (driving a car and public transportation), using information like speed limits or wait times from aggregated GPS data.

By contrast, pedestrian path geodata is curated by municipalities for inventory purposes and is published less frequently than street geodata. Because the purpose of curating data is not focused on treating pedestrian ways as a transportation network, when pedestrian path geodata are

available, they are stored in a variety of incompatible formats, including the metadata of streets
derived from aerial orthoimagery. Furthermore, they can contain frequent breakages or inaccuracies that prevent direct derivative creation of a routing graph. Recent work has shown
that actual sidewalk adjacency is an important factor that pedestrians consider when choosing a
walking route [9]. Furthermore, full and accurate coverage of sidewalk data is not sufficient for a
complete pedestrian graph in an urban environment, as sidewalks must be connected by supplemental paths such as crosswalks, bridges, and separate footways. The heterogeneity of
potential barriers impacting accessibility highlights the need for a cost function that determines
the best balance between several mobility criteria, and is broadly discussed in the literature
[5,6,8,10]. The data needed for a pedestrian routing cost function, such as curb ramps, construction information, and steepness, are often curated separately from sidewalk datasets, and
require significant post-processing to associate with pedestrian paths. The problem of relating
other pedestrian-relevant geodata to sidewalks is amplified by the high spatial precision requirements of sidewalks: the distance between sidewalks (the width of a street) is typically less
than that between streets (the width of a block), so relating independent geodata to a unique path
is more challenging. Wide availability of accessible pedestrian wayfinding therefore depends on
a data integration effort to use standardized data structures that describe the pedestrian built
environment as a transportation network, improve spatial accuracy in existing datasets, and
facilitate collaboration between the many stakeholders involved in producing and consuming
pedestrian geodata, including municipal agencies.

Data integration in the social realm

In many instances, institutional hurdles are so great that data sharing and standardization is
difficult and generally avoided despite the fact that all entities stand to gain by sharing fundamental data layers and, at least in the U.S., sharing information across organizational
boundaries has been a central objective in efforts to improve government operations and services

[11]. Data integration, the ability to effectively share data among entities and applications and integrate the data in efficient and flexible manner [12,13], continues to be a key problem in the innovation of private and public services. Using the case study of pedestrian-focused mapping and routing, this work illustrates the need for a data commons approach to data integration solutions, specifically within the subspace of data problems benefitting from public feedback that, when solved, benefit the social good. We discuss established and nascent techniques and tools for shared frameworks that enable sharing and integration of data across multiple private and public data providers and consumers in an open community-mediated data commons.

Progress in integrating pedestrian geodata may be hindered by the lack of technical and organizational infrastructure to facilitate cost-effective data sharing between stakeholders, namely stakeholders must currently maintain independent data layers in addition to a shared data commons. This may be due, in part, to the numerous independent stakeholders that maintain and consume sidewalk data, in addition to the technical challenge of requiring much more detailed and accurate data than the street network. Municipal agencies have an interest in maintaining pedestrian data for city planning as well as compliance purposes, for example compliance with

the Americans with Disabilities Act (ADA) in the United States [14]. Regional and national agencies have an interest in consuming and maintaining pedestrian data to offer accessible public transportation options and to enforce accessibility legislation. The public has an interest in maintaining and consuming sidewalk data for wayfinding applications and as a means to report local problems, such as sidewalk failures and barriers, and to engage with their communities.
Corporations have an interest in providing analytics and services based on pedestrian activities, such as walkability scores. Compounding the issue, the governmental agencies providing

services based on these data must be able to provide guarantees of accuracy to address liability concerns [15].

Despite data custody and provenance requirements for each stakeholder, an accurate, standardized, open, shared pedestrian geodata layer can greatly benefit all stakeholders, and emerging technologies may be able to satisfy both data validation and data integration. In the following sections, we introduce a case study in data science for social good that motivates a generalize-purpose model for a next generation geodata commons ecosystem that can serve a larger class of social good challenges. We present a set of established and projected tools and techniques, addressing the problem of data inaccuracy through heuristics, the crowd, and machine learning; user-facing applications for channeling crowdsourced knowledge into structured pedestrian data; the use of a public data commons map; and tools that can be used to branch this data commons into selectively-synchronized copies. Together, these components lay the foundation for a general-purpose data commons that serves stakeholder needs, coordinates use of shared data, and mitigates problems caused by data diversity by redistributing the responsibility for data wrangling and governance among data providers and consumers.

AccessMap: a case study for addressing the pedestrian information gap through data science

The AccessMap project aggregates rich, up-to-date information about sidewalks, including steepness, closures due to construction, curb ramps and other pedestrian-interest information from numerous sources, and integrates those data into one web map. While the main focus is on facilitating urban street traversal for people with mobility impairments, the underlying geodata can be analyzed with graph analytic and transportation planning methods to highlight and understand important transportation corridors for the public at large. At the time of writing, AccessMap supports the Seattle, Washington region, but uses an architecture that is region-agnostic. All software developed as part of AccessMap is open source and is developed using open source software, leveraging cutting edge visualization and UI technologies to deliver highly

interactive user experiences, such as client-side rendering of vector tiles.

The AccessMap architecture is based on a client-server web architecture, where users engage
pedestrian information through a web browser (Figure 2). When a user interacts with the
Accessmap web app, server-side functionalities such as trip planning requests are handled by
microservices that communicate with a postgres database for which the PostGIS and pgRouting
extensions have been enabled, and on which a routing graph has been built. In contrast to many
open source trip planning frameworks, pgRouting does not require computationally expensive

precalculation of edge costs, requiring only that the cost function be expressed as an SQL statement (including executing a stored procedure), a technology that enables individually-
customizable pedestrian routes. All pedestrian information displayed on AccessMap originates
from this database and is derived from an extensive processing pipeline that consumes municipal
open data (data.seattle.gov) in a variety of formats and United States Geological Survey surface
elevation information. As the municipal data sources update on a nightly basis, AccessMap
employs a microservice to regularly retrieve and reprocess every dataset into a new, annotated
routing graph, embodying a one-way flow of information from data providers to AccessMap to
end users.

Data integration and de-noising challenges
A previously-highlighted challenge in routing for pedestrians is the availability of ingestible
sidewalk geodata. Our first urban center target, Seattle, WA, provides open data on sidewalks in
the form of estimated sidewalk centerlines, an ideal format for adapting to a routing graph.
Despite that, there are still many data staging and denoising problems to address. At the time of
writing, the sidewalk data are not maintained as descriptions of independently-maintained
sidewalk locations, but are instead algorithmically derived by the Seattle Department of Transportation (SDOT) based on street metadata, and are actually parallel offsets of street lines

that have been trimmed by a constant length. Consequently, the sidewalk endpoints do not meet
end-to-end, either undershooting their true connection point, or intersecting to produce dangles.
trimming back of all sidewalk lines by a constant distance. Despite these inaccuracies, the SDOT
open sidewalk data is of relatively high quality, and only one of five cities that could be identified, at the time of writing, to offer sidewalk location information that is not planimetric
polygons from orthoimagery (the others being Bellingham, WA, Savannah, GA, Denver, CO,
Vernon, CA, and St. Augustine, FL). The AccessMap project developed multiple strategies to
increase the accuracy of parallel-offset sidewalk data. Three strategies were used: decision tree,
learning from crowdsourced data and an algorithm which uses inferred street metadata to redraw
offset lines that meet end-to-end by leveraging geometric buffers of street lines. The latter
method proved most useful and is currently running nightly to create the sidewalk graph for
AccessMap Seattle. The redrawing method requires only an estimate of offset distance and a flag
for whether a sidewalk is present on the left or right side of the street (or both), and could
therefore be used to leverage datasets that store sidewalk information in street metadata. These
efforts are a means for translating geodata in a variety of formats to one that is consistent and
routable, and represent an intermediate solution in the path to a longer-term, shared data layer for
pedestrian geodata.

Pedestrian ways that connect sidewalks are necessary for routing, but the centerlines for legal
crossing locations in the Seattle metropolitan region, and in most regions, are not explicitly
catalogued. In the case of Seattle, any intersection is a legal pedestrian crossing unless otherwise
marked, and laws regarding crossings vary regionally and internationally, including allowing
crossing at nearly any point on a residential street in many parts of Europe. As such, explicit,
long-term storage of centerlines for all crossing locations may not be suitable for all crossing

generate street crossing centerlines at intersections that connect around the intersection, which

are then annotated based on the existence of curb ramps at both ends. The result is a well-
connected sidewalk-and-crossing pedestrian graph covering all of the Seattle area.

To enable accessibility-focused routing for pedestrians we needed to assign attributes to each
sidewalk and street crossing that can inform our routing cost function. To calculate the mean
incline for each sidewalk, we use the 1/3 arc-second National Elevation Dataset (NED). Curb
ramp information is derived from latitude-longitude coordinates published by SDOT and used to
flag whether street crossings have curb ramps at both ends. However, these data suffer from
spatial inaccuracy: curb ramp coordinates are extrapolated from SDOT sidewalk locations,
whose endpoints are incorrect. As a result, AccessMap assigns curb ramp metadata to crossings
based on a generous proximity of 5 meters. Construction information is inferred from a street use
permit database that requires applicants to indicate sidewalk impact. Seattle street use permits
can include information impacting both streets and sidewalks, but the only geodata recorded are
the street affected and a building address, i.e., it includes no direct geodata on sidewalks, only
whether the sidewalk is blocked. To infer whether a sidewalk is impacted, we must find the
sidewalk(s) in our dataset that are located between the street centerline and street address.

The data processing challenge for a single city is significant, even with the advantage of a
municipal government dedicated to open transportation data, and is symptomatic of information
about the built environment being collected and disseminated for non-transportation purposes.
Furthermore, there are no technical or organizational frameworks in place to facilitate low-cost
facts.

Generating connected sidewalk graph segments with machine learning from crowdsourced information

In previous work [16], we developed heuristic algorithms to solve the sidewalk connectivity
classification problem, and currently use a redrawing algorithm that targets data issues specific to
Seattle sidewalk data. In the process of developing those heuristics, we also explored a

crowdsourced, machine-learning based method to tackle the problem of generating a connected graph for sidewalks. While this method is no longer in use by AccessMap for the city of Seattle, it represents a potential path forward for many other questions and particularly for addressing data issues in other cities.

When developing heuristic methods for improving sidewalk data, it became apparent that individuals could readily predict sidewalk connectivity from visual cues in web maps. This motivated two hypotheses: that (1) untrained users have the potential to provide a reliable ground truth dataset for comparing heuristics, and (2) given this labeled data and a set of features, we could apply supervised learning to predict sidewalk connectivity on a city-wide scale. One advantage of this approach is scalability: errors in municipal data can systematically vary from city to city, foiling a static heuristic, but by learning the weights of universal sidewalk attributes

from crowdsourced data, we could learn to fix sidewalk data on a per-city basis from relatively few examples.

We posed the sidewalk classification problem for supervised machine learning thusly: given any two sidewalk endpoints of different sidewalks, should the two endpoints be connected? The goal was therefore to predict a label (1, connected, or 0, disconnected) for new examples based on a relatively small training set. Each learning example constitutes a relationship between two sidewalk endpoints, and includes and a set of potentially relevant features such as the Euclidean distance between the sidewalks, whether the sidewalks intersect, the azimuth between the line drawn between the sidewalk ends intersects with a street. Non-spatial features included both sidewalks having the same sidewalk width or surface material. To reduce the problem size, we limited sidewalk pairs to those within 10 meters of one another.

Ground truth data were obtained for sidewalk pairs by developing an online crowdsourcing web

application through which a group of non-expert users could submit responses. The crowdsourcing web application presented users with a bird's eye view of two sidewalk segments
over a basemap, and asked them to decide whether the sidewalks ends should be connected (1) or
not (0), collecting more than 15,000 unique labels. To validate the consistency and accuracy of
non-expert responses, we compared their responses to 578 labels contributed by experts
previously familiar with the dataset (AccessMap developers). We found that individual crowdsourced labels disagreed with the expert labels in less than 3.5% of cases.

We separately applied a set of supervised learning classifiers to evenly split training:test datasets
using the scikit-learn library: logistic regression, support vector machines with linear and radial
basis functions, decision trees, and a naïve Bayesian classifier. Cross validation was done to
ensure robustness of learned models. Every classifier gained at last 5% improvement in accuracy
over the heuristic algorithm and the naïve approach of joining the nearest sidewalk segments.
Among the classifiers, logistic regression with second order polynomial features and an SVM
with radial basis function had the highest accuracy of 95.58%, both with precision scores over
91% and recall greater than 84%. Spatial features received the strongest regression coefficients,
with metadata-based features trending to zero due to regularization, suggesting that this approach
may scale well to other cities. Moreover, these algorithms achieved high (>90%) accuracy with
resampled training set sizes of 1000 data points, demonstrating that crowdsourced learning can
be a low-cost, rapid method for addressing pedestrian data problems.

Customizable routing with Accessmap
The architecture of AccessMap allows for on-the-fly recalculation of arbitrary routing cost
a linear combination (or weighted sum) of terms in the the form $w_i f_i(x_i)$, where w_i is a tunable
scaling factor and $f_i(x_i)$ is an arbitrary rational-valued function on column value x_i (which can be
arbitrarily generated at runtime). In most cases, f_i is the identity function and describes a slight
cost for crossing the street (to avoid time delays), a high cost (or indicator) for encountering curb

ramps when they are required for the user, a high cost (or indicator) for construction, or a moderate cost for segment length, to discourage overly long trips. The steepness cost is more complex for people with limited mobility, however, and exhibits a U shape as can be seen in Figure 2. The cost of elevation change is generally ignored for automotive routing and is typically modeled monotonically when optimizing for low-difficulty bicycle routes: the more uphill a bicycle climbs, the more costly the path, and downhill is less costly. For a manual wheelchair user, uphill climbs quickly become strenuous above 4-5% grade downhill descents also present a risk of losing control and potentially entering street traffic. This produces a U-shaped cost function (x-axis is steepness, y-axis is cost) with the easiest (the minimum) being a slightly downhill grade (e.g. -1%). An optimal cost function for wheelchair users is also likely convex, as constant changes in steepness create increasingly difficult climbs or descents. Systematic user studies are necessary to report reliable, optimal, data-driven cost functions for people with limited mobility, but approximations exemplifying likely variations are listed in individual and their assistive device: athletic, experienced manual wheelchair users and powered wheelchair users may allow for steeper climbs, while novice manual wheelchair users may prefer more flat areas.

AccessMap provides interactive visual feedback when tuning the parameters of the steepness cost function, allowing users to use familiar locations to adjust their preferences. For example, decreasing the maximum allowed uphill incline will result in more sidewalks turning to the coloration indicating sidewalks to avoid, which at the time of writing is red, enabling a workflow where a user can tune their preferences based on familiar sidewalks. This workflow requires no working knowledge of the inner workings of AccessMap, including the cost function or its form.
Once a route has been found using common presets or a customized cost function, the expected difficulty of traversal is color-coded along each segment of the predicted route so that any potential difficulties can be anticipated. This level of immediate visual feedback during routing preference customization is unique to AccessMap to the best of our knowledge. Future work in

this area will be targeted at providing convenient, data-driven default cost functions, and

Factors that limit scaling AccessMap and similar services to other metropolitan areas

AccessMap currently operates based on a one-way flow of information, processing municipal data on a nightly basis into a routable graph. Difficulties in expanding to new cities and sharing data between stakeholders are symptomatic of this unidirectional flow of information, as neither AccessMap nor the data producers need to communicate in a common data format. With unidirectional information flow, every new metropolitan area will pose unique data cleaning challenges for a service like AccessMap, as data storage and dissemination formats vary from city to city. Sidewalk data is more frequently available in the form of planimetric polygons, and are challenging to translate into centerlines, as they must first be processed into centerlines using approaches like skeletonization. Additionally, planimetric sidewalk polygons for a single sidewalk are often segmented incorrectly, producing sets of small, disjoint polygons that would produce inaccurate and disconnected centerlines when skeletonized. Municipal data sources are

not perfect and would benefit from reports of inaccuracies by motivated end users, but this raises another challenge: there is no system in place for systematically communicating data back to any city. To communicate the full extent of inaccuracies would overburden municipalities with significant technical overhead, given a lack of data standards and tooling. For example, sidewalks are often inventoried as part of constrained legacy asset management systems, and the incorporating new geometries or changes would imply significant time (and infrastructure) overhead for cities. This leaves the problem of deriving value from data integration up to individual stakeholders, such as AccessMap. Furthermore, municipalities open themselves to liability based on the data they publish, and have no system in place to verify data submitted by users. Finally, there are multiple overlapping agencies that all have an interest in accurate sidewalk data, but also need to store private information, such as personally identifiable information on citizens, which prevents submission of data back to municipalities and regional authorities. Thus, there is an unmet need for a system that coordinates these data across municipalities and other stakeholders.

To these ends, the AccessMap project was again supported by the DSSG program with UW eScience in 2016, with a focus on developing a sustainable data structure for pedestrian geodata, as well as tooling to facilitate conversion of data into that structure. This work led to the creation of the OpenSidewalks project, which is dedicated to advocating functional standards for pedestrian routes in OpenStreetMap (OSM), a crowd-based mapping platform. As part of this work, OpenSidewalks developed and improved tooling to facilitate the import of municipal data into the more maintainable and flexible OpenStreetMap data format. As a data commons, OpenStreetMap is uniquely positioned to act as a framework for coordinating pedestrian geodata.

An emerging community-moderated data commons model for collecting, managing, and sharing multi-stakeholder pedestrian data

In recent years, there has been growing interest in the creation of digital data commons, in which datasets and supporting software platforms are shared between various stakeholders. Data commons vary in function and domain, ranging from scientific repositories for research [17] to civic directories for the provision of social services [18] to data pools for predictive business analytics [19]. Regardless of this diversity, organizations involved in the formation of digital data commons face a number of shared technical and social challenges. Technical challenges include interoperability, storage, latency, security, and reliability.

With regard to social challenges, early discussions of natural resource commons focused people are disincentivized to invest in the maintenance of resources that can be freely obtained, and therefore squander those resources. Although much work has largely displaced notions of the tragedy of the commons with nuanced studies of the multiple modes of investment in common resources that can and do actually exist in the world [20–22], challenges certainly remain. For example, Frischmann et al. [20] articulate a number of “social dilemmas” that accompany the creation of knowledge commons, of which data commons can be considered a subset. These include: the production of resources to be shared in the commons; coordination of resource

navigation of the entanglement of nonrivalrous data resources with rivalrous resources like time,
labor, and hardware.

A number of these challenges - such as the definition of community boundaries and the navigation of rivalrous resources - are specific to the constellation of actors involved in any
given data commons, and have local solutions that are uniquely contingent upon their patterns of
data commons, through our experience developing AccessMap we have observed several design
elements that we believe can serve as a model to partially address some of the aforementioned
challenges in a nascent data commons, particularly one that faces the prospect of handling data
that is primarily in the public domain, while allowing stakeholders some degree of control over
their own datasets. Our preliminary model (Figure 3) includes the following components, which
will be further articulated in the remainder of this paper: (1) a stable data commons base layer to
facilitate interoperability, (2) processes for community-governed verification of contributed data,
and (3) a capacity for customization and reservation of private data layers from the commons.
We discuss how the intersecting parts of this model are coalescing within the OpenStreetMap
community, and how our experience using OSM to build AccessMap contributes to an understanding of the challenges and opportunities involved with developing this model.
Examples of situations that may fit these criteria are nautical maps, building usage, and emergency resource coordination.

A Data Commons Base Layer
OpenStreetMap has developed a robust community of individuals, private companies, NGOs,
and agencies that contribute data and tools to a sustainable, current global estimate of the
geographical state of the world. Essential to the success of OpenStreetMap is the reliance on an
extensible, layer-less data model used in a versioned, global record of the current state of the
OpenStreetMap database: if an individual notices something incorrect, it can be edited and nearly

immediately fixed using community contribution and validation tools.

Stakeholder Verification of Contributed Data

Despite the apparent risk in trusting anonymously-submitted map data, stakeholders in the OpenStreetMap community have developed practical verification frameworks that work around this issue. A common solution is to create a local copy (or branch) of the OpenStreetMap database and systematically screen all incoming changes. Companies offering driving directions employ this approach, as does the public transportation trip planner offered by the TriMet transportation agency, responsible for public transportation in the greater metropolitan area of Portland, OR. This model of periodic review and synchronization promotes contribution to the OpenStreetMap data commons while allowing stakeholders in the to offer assurances of accuracy and verification. The left portion of Figure 3 illustrates a use case for government institutions to keep a locally vetted copy, not merely a view, that contains only data validated within their organization.

The Portable OpenStreetMap (POSM) project developed for the American Red Cross takes the concept of a local copy a step further by spawning large portions of the OpenStreetMap software ecosystem as local instances. Intended for gathering data in low resource areas where electricity and internet connectivity is scarce, the POSM model for data contribution is a delayed synchronization: a subset of OpenStreetMap data is downloaded to the local instance, numerous methods for inputting new data to the server are employed, and the new data can be uploaded to the global OpenStreetMap. Due to the time delay, conflicts can appear and must be resolved. This model is similar to branching in version control of software, and the developers of POSM have posited a git-like comparison method for use in their synchronizations [24].

Customization and Private Layers

An essential functionality for many stakeholders is the ability to reserve a portion of data from synchronization. For example, a transportation authority may wish to keep survey data about

individuals private for ethical and liability reasons. Or, a stakeholder may want to store information that is inappropriate for OpenStreetMap, such as data on temporary disaster conditions, but still need to synchronize other information with the data commons. Therefore, these stakeholders require a systematic method to prevent synchronizing potentially arbitrary categories of information. While there are many potential methods to achieve this goal (unique tags for reserved data, a tracking table for private data), there is currently no readily available system for selectively synchronizing OpenStreetMap data. In addition, many of the data producers do not have the resources to perform stringent data processing, which, according to some estimates, consumes 70% of the time spent on analytic projects [25].

Challenges and opportunities in the emerging data commons model

From the experience of AccessMap, it is clear that mutually beneficial integration of pedestrian data among many stakeholders has challenges, some that are shared across all data integration problems and others that are unique. We propose that the pedestrian information gap problem is just one instance of a class of data challenges relating to social good that dictate the need for a community-mediated data commons approach (Figure 3).

Work in data federation and data integration highlight some of the common challenges that enterprise data integration efforts experience, particularly as they attempt to extract enterprise-specific value from analyzing massively heterogeneous available data [13,20,26]. A recently data kept in its original format and dataset affiliation, available on demand [25]. But "raw repository that does not address data staging only defers the problems of dealing with data quality, provenance and governance to the entity accessing the data from the data lake. In an enterprise solution setting, Terrizzano propose a team of curators can provide up-front services in both data cleaning and arbiter licensing and permissions guidelines. In the scenario we described, the central repository (the data commons) has no employees or resources to task a group to vet data contributions [27].

If stakeholders can easily maintain private copies of data, there is a risk of creating an ecosystem

of data silos with no incentive for sharing information. This risk is an open question for the plausibility of using a data commons model, but there are mechanisms inherent in the process that may promote sharing. Any stakeholder wishing to receive new information from the data commons will need to redundantly address conflicts, promoting fixing the conflict upstream to save time. Thus, the incentive of receiving data promotes contribution.

We suggest that the data commons can impose governance functions that distributes the burden of many data integration tasks among community members. The data commons itself provides a centralized arbiter that is readily available for access and queries anytime, and can provide some guarantees on fast access to accurate and updated data. Importantly, rather than being a dumping ground for raw data, the data commons must must impose standardized structure on data entry.

Conclusion

In this paper, we have described the case study of how AccessMap addresses the informational gap faced by pedestrians with disabilities. Implementing AccessMap required data-intensive methods and algorithms highlighted data challenges faced by stakeholders in pedestrian information, suggesting the need for a data commons framework, where multiple interested stakeholders can selectively share information with a global ground truth using a common interoperability standard.

Because the use case of automated pedestrian routing for people with disabilities has been data-limited, there are many open user experience questions. For example, AccessMap currently relies on the U.S.-centric traffic light metaphor (green, yellow, red) to communicate elevation change, but this does not translate to all cultures and does not serve many colorblind users. Therefore,

visualization questions are of immense interest to future work. Much of the data relevant to
providing trip plans for people with limited mobility such as the location of sidewalks and
marked crossings and the existence of construction information could be applied to provide trips
for those with visual impairments.

Existing and emerging technologies increasingly promote large-scale, verifiable gathering of
passive data by non-experts, particularly pedestrian geodata. Smartphones can leverage GPS
data, a magnetometer, gyroscope, and accelerometer, and in combination used this information to
measure the incline of sidewalks (steepness at a location in a particular heading). Phone GPS
combined with a series of images can be used to reconstruct 3D models using structure from
motion software. Finally, advances in low-cost lidar technology and increasingly common VR
applications could soon mean that high-resolution 3D scanning technology is ubiquitous in urban
areas, making it possible to carry out a detailed survey of the built environment by non-experts.

Acknowledgments
We would like to acknowledge the eScience Institute at the University of Washington and
participants of the 2015 and 2016 Data Science for Social Good projects for their contributions,
without whom most of this work would be impossible.

References
1. McCormack GR, Shiell A, Giles-Corti B, Begg S, Veerman JL, Geelhoed E, et al. The association
between sidewalk length and walking for different purposes in established neighborhoods. Int J
Behav Nutr Phys Act. BioMed Central; 2012;9: 1.
2. Lee C, Moudon AV. The 3Ds+ R: quantifying land use and urban form correlates of walking. Transp
Res Part D: Trans Environ. 2006;11: 204––215.
3. Ewing R, Cervero R. Travel and the built environment. J Am Plann Assoc. 2010;76: 265––294.
4. Carroll JM, Mentis M, Convertino G, Rosson MB, Ganoe C, Sinha H, et al. Prototyping collaborative geospatial emergency planning. Proceedings of ISCRAM. 2007; 13–16.
5. Beale L, Field K, Briggs D, Picton P, Matthews H. Mapping for wheelchair users: Route navigation
in urban spaces. Cartogr J. Maney Publishing; 2013;

6. Hara K, Chan C, Froehlich JE. The Design of Assistive Location-based Technologies for People
with Ambulatory Disabilities: A Formative Study. Proceedings of the 2016 CHI Conference on
Human Factors in Computing Systems. ACM; 2016. pp. 1757–1768.
7. Kang B, Scully JY, Stewart O, Hurvitz PM, Moudon AV. Split-Match-Aggregate (SMA) algorithm:
integrating sidewalk data with transportation network data in GIS. Int J Geogr Inf Sci. 2015;29: 440–
453.
8. Kasemsuppakorn P, Karimi HA, Ding D, Ojeda MA. Understanding route choices for wheelchair
navigation. Disabil Rehabil Assist Technol. Taylor & Francis; 2015;10: 198–210.
9. Gallimore JM, Brown BB, Werner CM. Walking routes to school in new urban and suburban
neighborhoods: an environmental walkability analysis of blocks and routes. J Environ Psychol.
2011;31: 184–191.
10. Völkel T, Weber G. RouteCheckr: personalized multicriteria routing for mobility impaired
pedestrians. Proceedings of the 10th international ACM SIGACCESS conference on Computers and
accessibility. ACM; 2008. pp. 185–192.
11. (Jochen) Scholl HJ, Klischewski R. E-Government Integration and Interoperability: Framing the
Research Agenda. International Journal of Public Administration. 2007;30: 889–920.
12. Doan A, Ramakrishnan R, Halevy AY. Crowdsourcing systems on the world-wide web. Commun
ACM. ACM; 2011;54: 86–96.
13. Halevy A, Rajaraman A, Ordille J. Data Integration: The Teenage Years. Proceedings of the 32Nd
International Conference on Very Large Data Bases. Seoul, Korea: VLDB Endowment; 2006. pp. 9–
16.
14. Percy SL. ADA, Disability Rights, and Evolving Regulatory Federalism Articles. Publius: The
Journal of Federalism. CSF Associates; 1993;23: 87–106.

14
15. Kitchin R. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.
SAGE; 2014.
16. Bolten N, Amini A, Hao Y, Ravichandran V, Stephens A, Caspi A. Urban sidewalks: visualization
and routing for individuals with limited mobility. Proceedings of the 1st International ACM
SIGSPATIAL Workshop on Smart Cities and Urban Analytics, UrbanGIS@SIGSPATIAL 2015,
Bellevue, WA, USA, November 3-6, 2015. 2015. pp. 122–125.
17. Contreras JL. Constructing the Genome Commons. Governing Knowledge Commons. 2014. pp. 99–
136.
18. Bloom G. Towards a Community Data Commons. Beyond Transparency: Open Data and the Future
of Civic Innovation. Beyond Transparency; 2013.
19. Walters L. Small commercial data pool project: Leveraging partnership. Insurance Issues. 2016;Nov.
20. Frischmann BM, Madison MJ, Strandburg KJ. Governing Knowledge Commons. Oxford University
Press; 2014.
21. Brady GL, Ostrom E. Governing the Commons: The Evolution of Institutions for Collective Action.
South Econ J. 1993;60: 249.
22. Rose C. The Comedy of the Commons: Custom, Commerce, and Inherently Public Property. Univ
Chic Law Rev. 1986;53: 711.

23. Kuhlen R. Knowledge is the water of the mind : how to structure rights in immaterial commons. In:
Bollier D, Helfrich S, editors. The wealth of the commons : a world beyond market and state.
Amherst, Mass.: Levellers; 2012. pp. 331–339.
24. Fitzsimmons S. Merging Offline Edits with the POSM Replay Tool [Internet]. 21 Aug 2016 [cited 8
Jan 2017]. Available: https://hi.stamen.com/merging-offline-edits-with-the-posm-replay-tool-
2f39a4410d2a#.bstdjy7h5
25. Chessell M, van der Starre R, Scheepers F, Nguyen N, van Kessel R. Governing and Managing Big
Data for Analytics and Decision Makers. IBM Redguides for Business Leaders [Internet]. 2014.
Available: http://www.redbooks.ibm.com/redpapers/pdfs/redp5120.pdf
26. Haas LM, Hernández MA, Ho H, Popa L, Roth M. Clio grows up. Proceedings of the 2005 ACM
SIGMOD international conference on Management of data - SIGMOD '05. 2005.
doi:10.1145/1066157.1066252
27. Terrizzano I, Schwarz P, Roth M, Colino JE. Data Wrangling: The Challenging Journey from the
Wild to the Lake. Available: http://cidrdb.org/cidr2015/Papers/CIDR15_Paper2.pdf

Bios

Nick Bolten, University of Washington, Seattle, USA 98195. Mr. Bolten is a PhD student in the
Department of Electrical Engineering at the University of Washington, a member of the Klavins
Lab, and founding member and lead developer for AccessMap. His work explores applications
of open source programmatic and data science methods to underserved domains, including
synthetic biology and accessibility technologies.
Anat Caspi, PhD University of Washington, Seattle, USA 98195. Dr. Caspi is Director of the
Taskar Center for Accessible Technology at the UW department of Computer Science and
can challenge and transform the current economics of assistive technology and incentivize rapid
development and deployment of access technologies and universally designed technology.
Sumit Mukherjee, University of Washington, Seattle, USA 98195. Mr. Mukherjee is a PhD student in the department of Electrical Engineering at University of Washington. He has received
a B.E. in Power Engineering from Jadavpur University, India and a M.S. in Electrical Engineering from Rensselaer Polytechnic Institute, NY, USA. His primary areas of research are
computational biology and machine learning.
Veronika Sipeeva, University of Washington, Seattle, USA 98195. Ms. Sipeeva is a Master's
student of Human Centered Design and Engineering department at the University of Washington. Prior to joining the program, she graduated as a journalist from Moscow State
Linguistic University in Russia. Her primary focus is on design for social change, accessibility,
and universal design.
Anissa Tanweer, University of Washington, Seattle, USA 98195. Ms. Tanweer is a PhD candidate in the Department of Communication at the University of Washington, a research
assistant in the Human-Centered Data Science Lab, and an ethnographer on UW's Data Science

Studies team. She is interested in the ways people organize with and around data, and how the
availability of increasingly large, heterogeneous data sets is transforming the way we construct
knowledge and make decisions.

Figures and tables

Table 1 Table of data sources, format, and required de-noising

Data type	Source	Format	Required De-noising
Sidewalks	Seattle DOT	Inaccurate centerlines	Redrawing sidewalk centerlines with better algorithm
Streets	SDOT	Accurate centerlines	None
Curb ramps	Sidewalks dataset	Points extrapolated from sidewalk endpoints	Extrapolate location from data, fuzzy spatial matching to
16			
sidewalk endpoints			
Elevations	National elevation dataset	Raster data in meters above sea level	None
Construction permits	data.seattle.gov	Street centerlines and street address points	Location inferred from street line and address

Figure 1 The image above highlights the informational gap faced by pedestrians, particularly
those with limited mobility. Sidewalk, crossing, curb ramp, and construction information is
usually missing from maps and online resources. The map below displays 1000 routes from the
is colored by whether the incline is greater than 8.33% (red, exceeds ADA ramp requirements) or
account for pedestrians, so more than 1 in 3 routes in this example included impassable routes.
17

Figure 2 How AccessMap addresses the pedestrian informational gap via user experience and
AccessMap and familiarize themselves with simple, consistent visual metaphors for pedestrian
accessibility. Users can plan trips by selecting presets for common use cases, or tweak the
routing settings manually to retrieve a trip plan customized to their needs. The workflow below
illustrates how one-way data integration processing powers the AccessMap user experience. Raw
data is processed and imported into a master database, after which it is derived into a routing
graph, and can then service custom routing requests.

18

Figure 3 The Community-Mediated Data Commons Data Integration Model addresses a number of challenges inherent in open sharing, maintaining and governing data integration
among multiple stakeholders with heterogeneous data. Specifically, the community decides on
an acceptable data standard for the shared data layer, but it is the responsibility of the individual
stakeholders to perform the up-front data denoising and reformatting to result in a fully consumable dataset that is synchronized to the shared data layer. In order for a synchronized set
to be accepted into the data commons, the commons provides tools for community members to
validate and ensure proper governance of the data. This is analogous to the Tasking Manager on

the OpenStreetMap paradigm. Finally, some stakeholders may be required or want to maintain
their own validated copies of portions of the data layer, which is available by “forking” the
database and potentially curating the local, proprietary copy. The data commons consumers may
be stakeholders, or other parties with analyses or applications built using the shared data layer.