Utilizing Volunteered Information for Infectious Disease Surveillance

Shaun A. Langley, Department of Geography, Michigan State University, East Lansing, MI, USA

Joseph P. Messina, Department of Geography, Michigan State University, East Lansing, MI, USA

Sue C. Grady, Department of Geography, Michigan State University, East Lansing, MI, USA

ABSTRACT

With the advent of Web 2.0, the public is becoming increasingly interested in spatial data exploration. The potential for Volunteered Geographic Information (VGI) to be adopted for passive disease surveillance and mediated through an enhanced relationship between researchers and non-scientists is of special interest to the authors. In particular, mobile devices and wireless communication permit the public to be more involved in research to a greater degree. Furthermore, the accuracy of these devices is rapidly improving, allowing the authors to address questions of uncertainty and error in data collections. Cooperation between researchers and the public integrates themes common to VGI and PGIS (Participatory Geographic Information), to bring about a new paradigm in GIScience. This paper outlines the prototype for a VGI system that incorporates the traditional role of researchers in spatial data analysis and exploration and the willingness of the public, through traditional PGIS, to be engaged in data collection for the purpose of surveillance of tsetse flies, the primary vector of African Trypanosomiasis. This system allows for two-way communication between researchers and the public for data collection, analysis, and the ultimate dissemination of results. Enhancing the role of the public to participate in these types of projects can improve both the efficacy of disease surveillance as well as stimulating greater interest in science.

Keywords: *African Trypanosomiasis, Disease Surveillance, Geographic Information Science (GIS), Kenya, Spatial Databases, Volunteered Geographic Information (VGI)*

INTRODUCTION

Recent publications surrounding Volunteered Geographic Information (VGI) broadly represent the belief among some in the academic community that non-scientists can be engaged in and benefit from spatial data analysis (Connors et al., 2011; Flanagin & Metzger, 2008; Goodchild, 2007a, 2010), a field previously reserved exclusively for academics. Focus on VGI represents a paradigm shift from viewing science as having a single authority (the scien-

DOI: 10.4018/jagr.2013040104

tist) to a model where authority is relative and expressed contextually. Abundance, repetition, and the collective assessment of data (as well as the ability to correct) convey credibility to information that would not necessarily exist otherwise (Connors et al., 2011). In this sense, a non-scientist plays a role in validating data collected by others, and collectively assessing data quality (Craglia, 2007).

The concept of Web 2.0 incorporates bi-directional collaborations in which users collectively collate spatial data, stored in a central cloud repository and accessible by anyone for whatever purpose deemed worthy. The Web 2.0 paradigm is represented widely through web projects such as Wikimapia, OpenStreetMap, and even Google Earth. Within the context of these volunteered GISystems (VGIS), users contribute information to develop a collective knowledge base. Recent advances in mobile technology have furthered the applicability of Web 2.0 projects, enabling easier access to the information, and even allowing for novel uses of crowd-sourced information (Rosenberg, 2011). Sui (2008) extends the paradigm to include "the wikification of GIS", a notion which he defines as being the shift in perception that only people who are specifically trained to "do GIS" should interact with spatial data and perform analysis. It is upon this notion, specifically, that VGIS endeavors to enhance the role of the user in the collection and analysis of spatial data.

The use of volunteered information for disease surveillance draws upon themes in the participatory GIS (PGIS) literature in suggesting that GIS technologies can operate in concert with volunteered information and local knowledge (Boroushaki & Malczewski, 2010b; Connors et al., 2011; Elwood, 2010; Flanagin & Metzger, 2008). The key distinction between classical PGIS methods and VGIS involves the role of the scientist. We refer here to McCall's (2005) discussion of good governance through improving dialogue, legitimizing and using local knowledge, the redistribution of resources access and rights, and new skills training in geospatial methods. These concepts support the idea that a PGIS or VGIS approach can contribute to the adoption of new technologies for disease surveillance.

BACKGROUND

Traditional Paradigm

The traditional paradigm in GIScience partitions individuals into experts versus non-experts. In an academic context, this treats scientists as the experts and citizens as non-experts. Under this traditional paradigm, public participation in the research process is hindered by a number of factors. Most importantly, the traditional roles of experts (scientists) versus the public leaves little room to consider alternative knowledge bases (i.e., local knowledge). Furthermore, there is limited opportunity for citizens to become informed, equal participants; thereby limiting the potential applicability of any results/understanding gleaned from the research process (Boroushaki & Malczewski, 2010a).

Under the traditional GIS model, technology and software are not readily accessible, requiring either a specific skill set or simply being priced beyond the consumer market. Therefore, citizens are relegated to operating as consumers of information exclusively, or as indirect producers, mediated by communication to researchers in small group projects. Their interaction with the data in this regard is strictly as a provider of information, not as producers of spatial data products. Finally, the traditional GIS model treats data validation as achieved largely through reputation (Flanagin & Metzger, 2008). Scientists and researchers are perceived as producers of reliable data due to past training in data collection and analysis. Furthermore, the peer review process adds credibility by requiring outside researchers to assess quality. Broadly though, data collected by researchers are assumed to be reputable because it is collected within the context of academic endeavors, and done by trained individuals. Information of this sort is generally accepted to be true until shown to be otherwise. With

few exceptions, the vast majority of GIS data products are produced under the traditional GIS model. Citizens are largely excluded from the process of data collection and analysis (Connors et al., 2011). We do not, however, suggest that the traditional model must be replaced. Instead, we propose the standard model be extended to facilitate collaboration between citizens and researchers.

VGIS Paradigm

Volunteered GIS represents a paradigm shift from viewing science as having a single authority (the scientist) to a model where authority is relative and can be expressed contextually. Information abundance, repetition, and the collective assessment of data convey plausibility to data that would not otherwise necessarily exist. In this sense, a non-scientist plays a role in validating data collected by others; collectively assessing data quality (Oreskes et al., 1994). This concept is explored further by Craglia (2007) in his assessment of individuals as geosensors; empowering them to validate global models using their own perceptions or impressions of the data. Volunteered GIS therefore represents the broad interest by non-scientists to be engaged in and to benefit from spatial data analysis.

Although Goodchild coined the term "volunteered GIS", the movement towards a new paradigm really began a decade earlier with the desire, on the part of scientists, to engage citizens directly in the research process. Sara Elwood, through her work with PGIS, exemplifies this desire and her work has been instrumental in the evolution of the VGIS paradigm (Elwood, 2006). Other contributions have included work by Elmes (2005) with his description of a "community integrated GIS" Turner's (2006) "Neogeography" Balram and Dragicevic's "collaborative GIS" (2006), and Sieber's "public-participation GIS" (2006). Collectively the work of these individuals demonstrates the broader goal of direct community engagement in the research process.

However, researchers have also made significant strides towards integrating components of a VGIS into their own projects, including studies in environmental sensing, decision-making, resource management, and community risk assessment. Project GLOBE, OakMapper, and Audubon's Christmas Bird Count (Connors et al., 2011; Goodchild, 2007a; House et al., 2001; Yaukey, 2010) are long running projects for the purpose of monitoring spatial and temporal distributions of resources and phenomena. By employing citizens to collect data, researchers are able to more effectually analyze spatial processes by generating much larger quantities of data. While data quality remains a concern, the large quantity of data collected diminishes the influence of inaccurate data (Flanagin & Metzger, 2008).

The use of VGIS to answer questions of decision making draw upon the PGIS literature in suggesting that GIS technologies and implementations can assist in conflict resolution and multiple-criteria decision making (Boroushaki & Malczewski, 2010b). Flanagin and Metzger (2008) make reference to these types of questions in using GIS for collective community efforts. The key distinction between classical PGIS methods and VGIS involves the role of the scientist. PGIS seeks to improve dialogue between actors for the purpose of legitimizing and using local knowledge, the redistribution of resources access and rights, and new skills training in geospatial methods (McCall & Minang, 2005). However, the researcher plays a limited role as teacher. VGIS builds on this by leveling the authority between actors; scientists and non-scientists are viewed as having [almost] equal authority, allowing both actors to communicate more freely with each other and to share expertise. Our prototype supports the idea that a PGIS or VGIS can contribute to addressing questions of decision-making, and later resource management and conflict resolution.

Volunteered GIS alters the standard GIS paradigm by substituting a producer-user model instead of the traditional expert-user archetype. Under this framework, researchers and citizens

can act as either producers or consumers of spatial data, depending on the context within which they are interacting with the data. Producers in this case need not necessarily be experts in all areas of GIS or the broader research context. Rather the producer's role is given to any individual with information to contribute to the aggregate knowledge base (Boroushaki & Malczewski, 2010a; Flanagin & Metzger, 2008). User roles are given to individuals who consume a spatial data product for any purpose. Under the new paradigm, roles are not fixed and not exclusive.

Finally, under the traditional GIS model, data were perceived to be trustworthy because of the perceived authority of the scientist. However, with changing roles, we need a new model for data error assessment (Flanagin & Metzger, 2008; Goodchild, 2007a, 2007b). Possibly the single largest barrier to the utilization of volunteered information is the uncertainty surrounding its credibility (Connors et al., 2011; Flanagin & Metzger, 2008; McKnight et al., 2011). Under the VGIS model, the credibility of volunteered information is achieved through volume. Intuitively, we understand that if multiple individuals report similar information, the reports are likely credible representations of the truth. The larger volume of data collected through a VGIS, albeit repetitive, can achieve the same threshold for credibility as data collected under the traditional model (Flanagin & Metzger, 2008).

Related, there's a significant degree of uncertainty as to the nature of volunteered information with respect to the types of error. McKnight (2011) explore the relation of volunteered information to assess spatial distribution of West Nile virus in Michigan. In their analysis, he raised the issue of uncertainty with regards to types of error that influence the data. For example, users may reliably report positive observations (i.e. in this case, observations made of dead birds), but reports are likely to indicate the absence of data. Therefore, volunteered information is heavily biased towards the observation of an outcome, and should not be interpreted as a metric of prevalence. Utilization of volunteer information must be cognizant of the nature of uncertainty.

The prevalence of mobile devices that have GPS capabilities, including cell phones, tablets and laptop computers, has increased the accessibility of spatial data. Hardware is no longer priced outside the realm of ownership for many people in the world, meaning that users can now directly engage with spatial data in ways that they simply could not do before. The interaction of the public with spatial data is now so prevalent that most users have developed sufficient technological skills and spatial cognition (through interaction with online mapping tools) to enable them to interact with spatial data in an intelligent manner, precluding the need for training prior to participation in GIScience research. Paradoxically, it would appear that spatial cognition is unrelated to global geographic awareness. Although people are able to position themselves abstractly on the landscape, they remain illiterate as to the broader geographic context in which they live.

Software interfaces fall into two broad classes: traditional desktop products and web-based applications. For the purposes of interacting with a VGIS, citizens are most likely to use a web application since this does not require a specific platform or license to run. Desktop applications, on the other hand, can be distributed to certain groups, allowing for a more targeted interaction with the spatial data. The open-source software movement is most directly credited with making GIS software accessible to the public, removing financial and hardware restrictions for many GIS products. Most notable among these are GRASS and QuantumGIS, free GIS packages modestly equivalent to ESRI's ArcGIS®. Interfaces for spatial data analysis have been developed with R and Python, interacting directly with Grass and QuantumGIS. Increasing familiarity on behalf of the public in spatial tools, geospatial technologies, and mapping increases the likelihood that they will be able to act as producers of high quality information. While the increasing

availability of mobile technologies has spurred public interested in GIS, the cost of adopting new technologies remains a principle challenge, particularly in developing countries.

Disease Surveillance

Disease surveillance systems are established for the purpose of collation, analysis, and dissemination of information so as to facilitate the allocation of resources in handling disease outbreaks (Thacker et al., 1983). Broadly, surveillance programs are categorized as either passive or active. Passive disease surveillance programs rely on reporting by healthcare providers to public health authorities when specific signs and symptoms are observed, a diagnosis is made and/or a diagnostic test is confirmed. Public health authorities collate these reports and assess the need for a coordinated response. Upon making a determination, the authorities communicate back to the local health care providers, and the necessary recommendations are set forth to address the disease outbreak. An example of passive surveillance in the United States involves the reporting of certain communicable diseases by health care workers after a diagnosis is made. These reports are received by public health officials who are tasked with ensuring the disease does not pose a threat to the welfare of the public.

Passive disease surveillance systems are hierarchical in nature with space and time important factors. Knowledge of highly infectious diseases may be reported up and information on the control of those diseases may be reported down the hierarchy very rapidly; whereas more common diseases may be reported and intervened on slower schedules such as monthly or semiannually. The management and coordination of communication within a passive surveillance system therefore, needs to be agreed upon by all parties (i.e., levels within the hierarchy in order to ensure the protection of population health). However, passive surveillance systems are widely criticized for underreporting diseases (Thacker et al., 1983). When passive surveillance systems break down and mandated reportable disease(s) are not communicated from the local to central levels, there is a need to respond by implementing an active surveillance program.

Active surveillance programs directly address the underreporting of disease by utilizing teams to assess local conditions. Such programs begin with the recognition at the central level that the expected communication in space and/or time has not been received and in response actively reach out to that location for the information. During these visits, retrospective data are collected and the management of the passive surveillance system is revived (e.g., manpower, technology). The operations of disease surveillance systems are therefore highly dependent upon the cooperation of all participants at level of the hierarchy. One well-cited example of active surveillance is Snyder and Merson's (1982) meta-analysis of diarrheal disease prevalence and mortality throughout the developing world. Here they review 24 studies where data was actively collected (either through home visits or other means) by trained personal. In contrast to a passive surveillance program, workers were employed for the sole purpose of collecting disease prevalence data.

CASE STUDY

Purpose

African Trypanosomiasis (AT) is a zoonotic disease transmitted by the tsetse fly. In Kenya, the two most common forms of AT are *Trypanosoma brucei* (Nagana), the form of the disease that affects cattle, and *Trypanosoma rhodesiense* (Sleeping Sickness) that affects humans. While sleeping sickness is relatively rare in Kenya, Nagana is widespread and represents a major threat to the livelihood of pastoralists (Baird et al., 2009; Tarimo-Nesbitt et al., 1999; Waller, 1990). The prevalence of Nagana has increased in recent decades due to a decline in control regimes, climate change, and anthropogenic

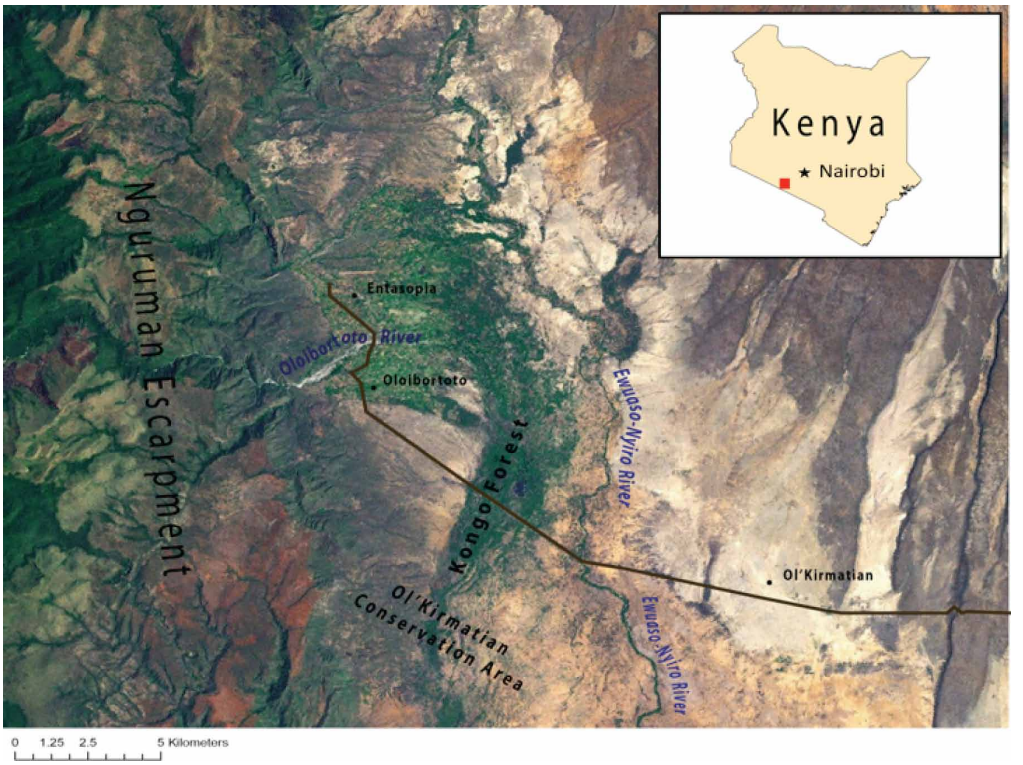
factors (Batchelor et al., 2009; Bauer et al., 1992; WHO, 2005). Our case study describes the prototyping of a VGIS for the purpose of surveillance of an infectious disease vector.

Site Description

Nguruman (Figure 1) is located at the base of the Rift Valley in southern Kenya, just east of the Nguruman Escarpment. Formally, Nguruman is the local Maasai name for the settlement, which occupies the area west of the Ewuaso-Nyiro River and the Kongo Forest to the Oloibortoto water intake; it is bounded to the south by the Ol'Kirmatian Conservation Area and to the north by the Oloibortoto River. Nguruman is also referred to locally as Oloibortoto. North of the Oloibortoto River, and broadly included in our study area, is Entasopia, the largest settlement in the area. The political "capital" of the Nguruman area is Ol'Kirmatian, a settlement 6.5 km west of the Ewuaso-Nyiro River, and home to the District office for the Kenya government as well as the office of the local governor for the Ol'Kirmatian group ranch, the political arm of the Maasai in this area.

From the base of the Rift to Oloibortoto, the predominant land use is smallholder agriculture. Streams dissect the region and are maintained by the community as means to irrigate their farms. Dominant agricultural crops throughout the region include tomatoes, vegetables destined for South Asian markets, and fruit trees (e.g., bananas, mangos) (Langley, 2010). Southeast along the road from Oloibortoto to the Kongo forest, vegetation density rapidly increases. The area is extremely rocky and dominated by herbaceous and woody shrub vegetation, most abundant of which are *Acacia tortillis*, *Salvandora persca* (toothbrush tree), *Grewia tembensis*, and *Cordia sinensis* (Maitima, 2012; Morris et al., 2009). Dominant

Figure 1. Study area



grasses throughout the region include *Sporobolus spp.*, *Setaria spp.*, and *Cynodon dactylon* (Morris et al., 2009). The Kongo forest is the area of densely vegetated land between Oloibortoto and the Ewuaso-nyiro River. It is within this zone that we find abundant tsetse; unfortunately this zone is often the only option available to the community for grazing their animals during the dry season. Moving east from the Ewuaso-Nyiro River, the landscape dries quickly, resulting in a rapid decrease in vegetation density. During the dry season, the area is devoid of most vegetation; however after a short period of rains, the grasses in the area of Ol'Kirmatian re-emerge with vigor. Across the entire region, these eco-zones are highly dynamic and respond rapidly to local climatic shifts and the occurrence of precipitation.

Global climate change is dramatically influencing the local environment within our study area (Moore et al., 2012). In past decades, annual precipitation in southern Kenya has remained relatively constant despite significant increases in annual mean temperatures, however the variance of the magnitude of precipitation events have increased and the seasonality of total precipitation has become less predictable (Altmann et al., 2002; Moore et al., 2012; Moore & Messina, 2010). The observed climate change and uncertainty in precipitation will undoubtedly threaten the livelihood of farmers and pastoralists (Fischer et al., 2005). Indeed, these concerns were conveyed to us in the course of our work; many farmers have already found it difficult to determine the right time for planting due to changes in local weather and precipitation events (Langley, 2010).

Trypanosomiasis (Nagana) in cattle is a major threat to the livelihood of Maasai pastoralists in Nguruman. The risk of infection is chief among their concerns to the health and well being of their cattle herds. An important consideration for the community is the management of grazing for cattle herds among the members of the group ranch. A committee of elders, whose chief aim is to maximize utilization of the limited resources (while advocating sustainability) for the benefit of the community, manages the patchwork of grazing areas. Of particular interest to the grazing committee (as expressed through interviews[1]) is the ability to work with our research lab to incorporate predictions of the spatial and temporal trends in tsetse populations and models of risk aversion.

DeVisser et al. (2010) developed a species distribution model for tsetse (TED) that predicts tsetse presence/absence every 16 days based on the habitat requirements and movement rates of the fly. The precision of the model predictions is limited spatially by the resolution of the inputs (250m), and temporally by the availability of MODIS LST and NDVI data products (8 and 16 days respectively). It is well established that tsetse are highly responsive to microclimatic conditions supported by local variations in vegetation (Terblanche et al., 2008). The spatial resolution of the TED model predictions limits consideration of such local configurations, thereby increasing the likelihood of errors of omission. The TED model was designed to identify endemic tsetse and does not model transient tsetse populations. By incorporating volunteered information from citizen reporters, TED could better illustrate the distribution of the flies over space and time by reducing errors of omission and reporting transient populations. Volunteered information may also be used (to an extent) to confirm (Oreskes et al., 1994) the TED model predictions by giving us a means to estimate model uncertainty.

Conceptual Model

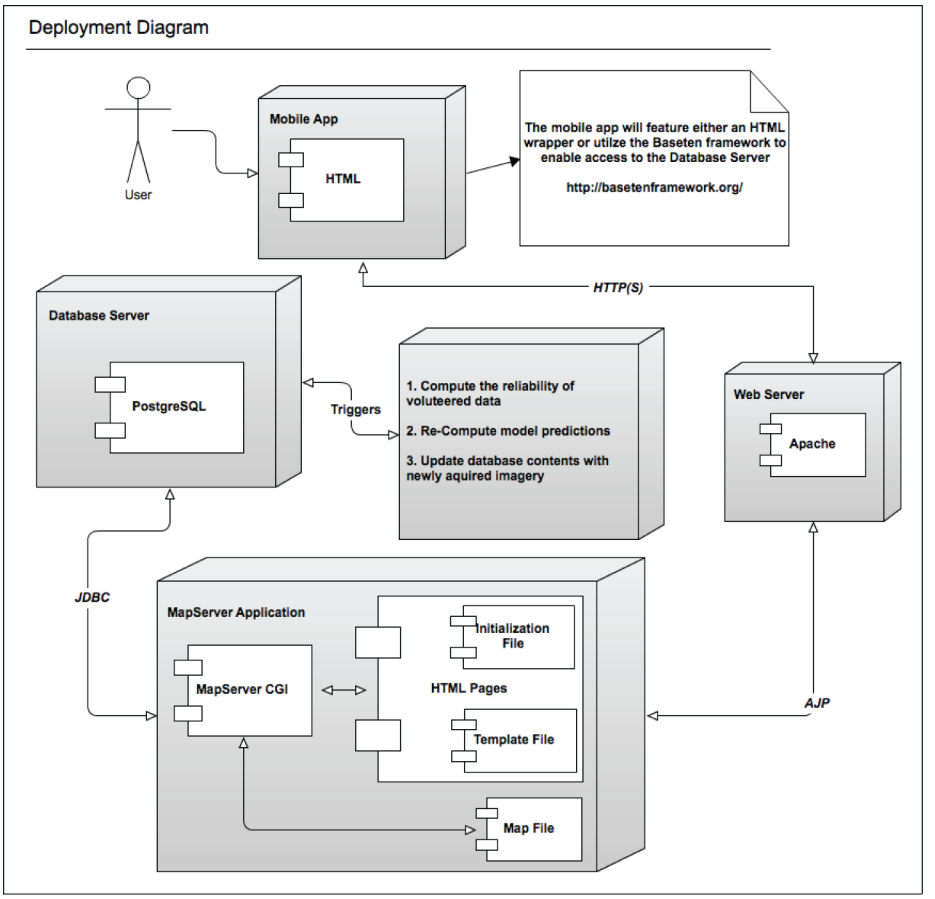
Here we elaborate on the previously published framework for a VGIS (Langley & Messina, 2011) by illustrating the construction and deployment of a working prototype. Furthermore, we discuss potential challenges and limitations of our implementation and propose strategies to address these issues. Figure 2 outlines the basic implementation of the proposed VGI

and the methods by which users will be able to interact with the spatial database (sDBMS), specifically through mobile devices. The core of the proposed implementation is a Postgres database server that stores both the spatial data as well as scripts to compute predictions of tsetse distributions, process volunteered data (submitted first to a reliability assessment), and automatically retrieve and process remotely sensed imagery as it becomes available. Users interact with the database through an Apache server and an HTML interface. A MapServer[2] implementation provides functionality for visualization of spatial data. All components are open-source and platform independent so as to convey maximum portability.

Our implementation of a VGIS seeks to achieve three goals: 1) facilitate user interaction with the VGIS and model results so as to allow for the reporting of information that may correct otherwise inaccurate data (defined as those predictions that contradict ground-level reports); 2) assess the reliability of volunteered information; and 3) incorporate volunteered information to calibrate a model of tsetse distribution and reduce errors of omission. To assess the functionality of the VGIS to achieve these goals, we have developed a working prototype of the system to illustrate our approach.

We incorporate a variety of software packages, including GRASS and QGIS (for visual

Figure 2. This deployment diagram illustrates the interaction of the separate components of the VGIS and the flow of information between each component



GIS support), Python and R (for statistical and modeling tasks), each of which provides the user with statistical, visual, and geoprocessing capabilities; the user can interact with these packages through a GUI or through a command line interface. Our selected DBMS is PostgreSQL 9.1 (Postgres), an advanced, readily available, open-source, object-relational database management system. Using standard SQL syntax, Postgres allows for complex query capabilities, including spatial queries, and facilitates strict rule and primary key enforcement. Postgres is also extensible, allowing for the addition of new functionality (Stonebraker & Kemnitz, 1991; Stonebraker & Rowe, 1986). In contrast to previous implementations of MySQL, Postgres, and other common spatial databases, modern DMBS models facilitate the combined storage of spatially explicit data and corresponding metadata together in the database (Elmasri & Navathe, 2008; Watson et al., 2004). The proposed spatial computing environment uses open source, community-supported software and standards, providing a solution to the data-management problem that is temporally extensible. Of critical importance to us is the improved functionality available in PostGIS 2.0, which adds support for raster data types. PostGIS is an extension to the Postgres language that adds functionality for the storage and retrieval of spatial data. PostGIS is, at its core, a suite of tools that serves as the back end for spatial functionality in Postgres.

Data Collection and Interface

Users and producers alike are able to interact with the VGIS in many ways, each according to their own skills, interests, and available hardware. In our case study, we outline the mechanisms whereby participants (either researchers or community members) can interact with the VGIS.

Figure 3 illustrates the broad deployment strategy for mobile device interaction. Related to the deployment of a mobile interface, the web interface sports a comprehensive suite of tools available to all participants from any web browser, with functionality dependent upon the credentials supplied to the system.

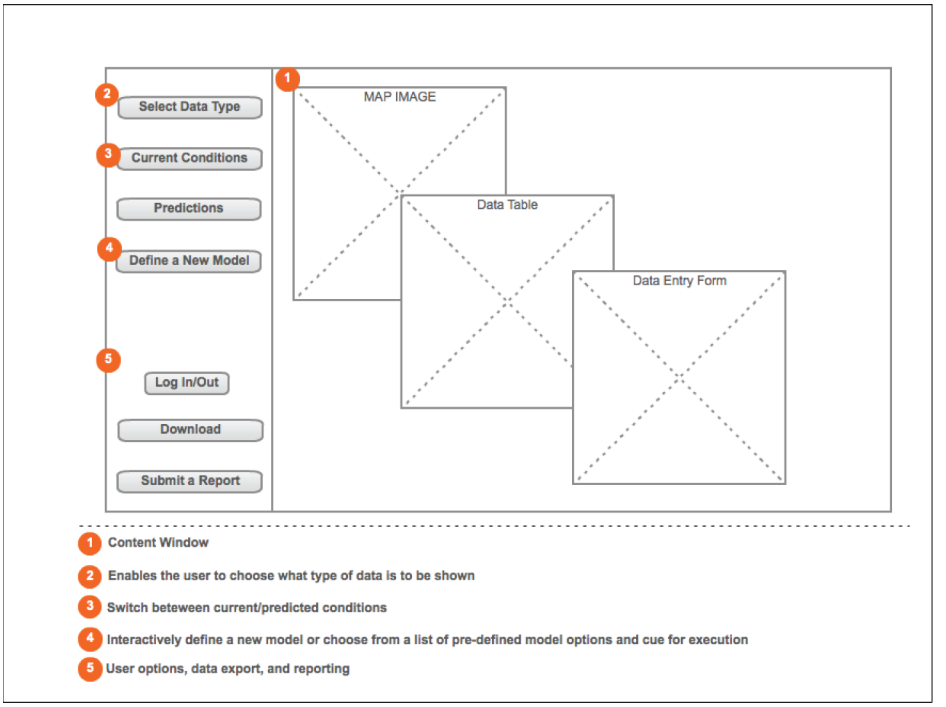
Most participants will find themselves interacting with the VGIS primarily through their mobile devices. For this purpose, we propose an iOS application that, for the most part, simply employs an HTML wrapper allowing the user to interact with a MapServer application. Users are able to query the database for specific, albeit limited, types of data, even define a specific range of times over which to aggregate the data; the application is geographically aware, so it is able to return information for a user's specific coordinates by passing the current longitude and latitude to the server. Most importantly, a user is able to volunteer information, with regards to the distribution of tsetse, through the application. Simply, a user can use this function to report that tsetse flies are present at their current location. A user's unique device ID is logged with the report and serves as a surrogate measure to distinguish between users.

Users are able to interact with the system in different environments, including a web browser, a desktop application, and on a mobile device (Figure 3). Users ultimately will be able to conduct a range of operations, such as obtaining spatially contextual information and model predictions, defining new model runs, exporting data, and submitting volunteered information or reporting map/model errors; however for the purpose of our prototype, functionality is limited to the data querying, visualization, and reporting of tsetse occurrences.

Information Reliability

The traditional model of data reliability emphasizes the authority of the researcher and our belief that trained individuals will generate reliable, trustworthy data (Craglia, 2007). Within

Figure 3. We propose an iOS application (for iPhone or iPad) that allows users to interact with the VGIS, explore the model predictions, volunteered data, or to contribute their own observations



the context of a VGIS, we relax this assumption; instead qualifying reliability through data volume; the idea being that credible information will tend to be generated independently by more than one user (Flanagin & Metzger, 2008).

There are two fundamental approaches to assessing the reliability of crowd-sourced information. In the simplest case, information is assumed either credible or not until confirmed or rejected by a subsequent report. Under this model, all participants are treated equally with respect to their prior knowledge/skills; reliability is assessed by their peers through the creation of informal social networks [of trust] (Bishr & Kuhn, 2007; Bishr & Mantelas, 2008; Flanagin & Metzger, 2008; Metcalf & Paich, 2005).

The second model takes a more nuanced perspective of the user, taking into account the skill set of the person filing a report and their prior credibility. This approach is best approximated as a Bayesian model of data quality where the reliability of a report is dependent on the prior assessment of the user and previous reports made to the system (Crosetto, 2001). If a number of prior reports are rejected, the individual is given a low reliability score that may lead to automatically rejecting any subsequent reports made (unless of course those reports are later confirmed independently). However, if the user has a history of high quality submissions that are routinely confirmed, they may be given a high credibility score, leading to automatic accepting of the report into the database. Prior experience with the user is the crux of this model approach to data quality. Conati (2004) demonstrated this approach in evaluation of models of user affect; their study required that

they be able to assess the reliability of self-reporting of emotional states.

In our case study, we employ a simple decision model (Figure 4), that integrates social trust networks (e.g. Bishr & Mantelas, 2008; Metcalf & Paich, 2005) and Bayesian methods (e.g. Conati, 2004; Crosetto, 2001), to assess the accuracy of data and the reputability of volunteers who report on the presence of tsetse flies. To demonstrate our approach, we evaluate the reliability of volunteered information under two scenarios. In the first, a single reporter volunteers on multiple occasions. In this scenario, the reliability of the information is determined and the rating (Equation 1) can be associated with the reporter's ID. Subsequent reports are evaluated on the merits of the information as well as the reliability score of the reporter. In short, a reliable reporter is likely to submit reliable information. A record can also be approved if a user is deemed trustworthy under the model. This value is calculated over time as a measure of the number of reports that are confirmed versus contradicted.

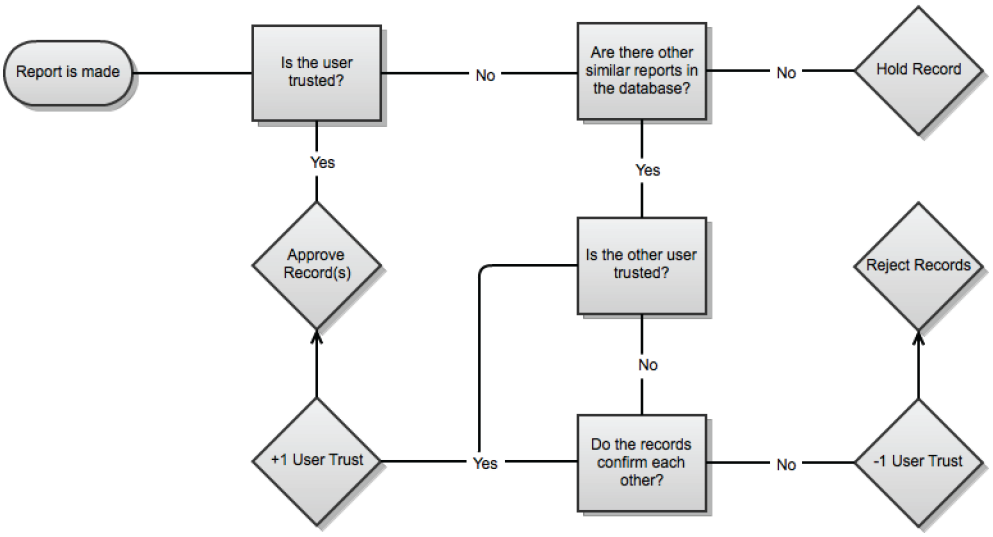
$$User\ Rating = \alpha + \Delta \quad (1)$$

α = prior score

Δ = change in score output from Figure 4

In the second scenario, several reporters each volunteer information only once. In this case, a reliability score cannot be computed or used to evaluate the reliability of the information; a report made under this scenario must be evaluated solely on the merits of the content. There are two components in the report (in addition to the information itself) that are used to assess reliability, context and authorship. In this scenario, authorship is of limited value since each reporter submits only once; we cannot conceptualize an author profile. However, we can evaluate the content of the information in the context of current predictions (of tsetse distribution) as well as prior years' predictions for the same period. A reliability score (equation 2) is computed as a cumulative product of a user's rating, the number of times a cell is occupied in the previous time step in the current year, the number of times the cell is occupied

Figure 4. To assess the reliability of volunteered information, a report is evaluated in the context of a set of conditions. This figure presents a logical thought diagram for the application of the computation of reliability (Equation 2).



on the same date in previous years, and the number of neighbors occupied in the previous time step. A report is deemed credible if the score exceeds a certain threshold. This threshold is initially set to 5, but should be re-evaluated periodically to ensure data quality is maintained.

$$Reliability = \theta + \rho + \frac{\kappa}{4} + \gamma \quad (2)$$

Θ = user score
ρ = the number of times the cell was occu pied previously, including the previous time step in the same year and the same time step in the previous year (max = 2)
κ = number of neighboring cells that are oc cupied (max = 8)
γ = number of supporting reports

Volunteered information under both scenarios can also be evaluated in the context of TED model predictions. Model predictions that take into account volunteered reports are compared to predictions made 16 days prior as well as to the distribution of tsetse at the same time in the previous year. We can reasonably assume that pockets of tsetse should maintain connectivity. If a report is made of tsetse occurrence in an isolated area (as measured by number of neighbors, κ) where no tsetse are predicted to occur, the probability of this report being accurate is low. If we were to incorporate these data into the model, the resulting predictions might dramatically impact the local reliability of the model outputs. By incorporating volunteered information into our prediction of tsetse distribution, we can better represent fine scale variability, particularly with regards to our ability to represent real-time distributions.

Utilizing Volunteered Information to Reduce Model Error and Uncertainty

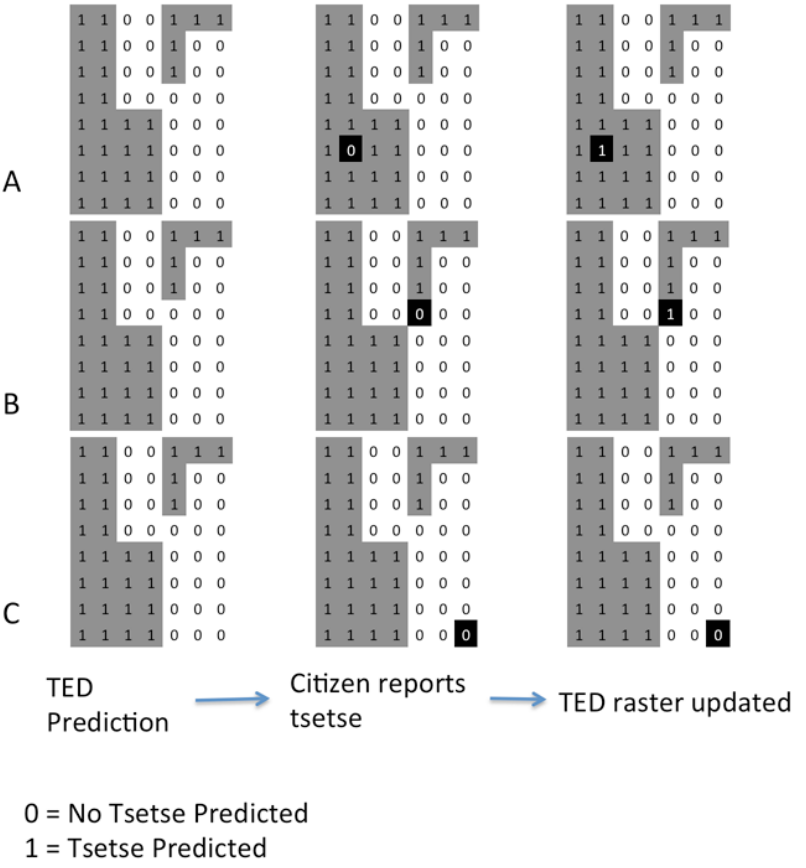
Previously, we detailed our approach to assessing the reliability of volunteered information in the context of our case study. To illustrate the performance of the VGIS in making this determination, we simulate the reporting of tsetse occurrences across the study area. The simulated reports are generated for each iteration of TED model prediction. Additionally, we can illustrate reliability assessment under each of the two scenarios we detailed earlier; multiple reports from a single user or single reports made from multiple users.

The TED model outputs a binary raster at 250 m pixels which represents the minimum mapping unit for the predicted distribution of tsetse on the date the latest MODIS data product was captured; the predictions are not real-time estimates (always 30-45 days past) of tsetse distribution and are designed to underestimate the maximum distribution. Incorporating volunteered data allows us to fill in the gap, providing more up-to-date predictions (Figure 5). If the data reports are deemed reliable and differ from TED predictions, the cell represented in the binary raster for the previous time step is updated to reflect tsetse presence. The next iteration of TED will build on the 'corrected' raster.

Tsetse distributions expand and contract with seasonal climate. They achieve a minimum distribution at the peak of the dry season; these regions of minimum tsetse distribution are termed 'reservoirs' (DeVisser et al., 2010). Of relevance to our case study, we can use these minimum distributions as opportunities to 'reset' the model predictions so as to reduce any errors of omission that may have resulted over the previous season from incorporating volunteered information. In doing so, we can ensure that TED model predictions are reliable estimates of the minimum distribution of tsetse.

To test the functionality of the VGIS, we will simulate the implementation of the system to test the evaluation of volunteered information and the integration of this information with the DeVisser's tsetse distribution model. These simulations will primarily explore the assessment of volunteered reports of tsetse presence, under three scenarios. The first illustrates the case a report fills in a gap in the predicted distribution of tsetse (Figure 5a). Presumably, this is a product of error in the estimation of

Figure 5. Users may volunteer reports of tsetse presence under a range of scenarios. (A) Illustrates the case where a report fills in a gap in a patch of tsetse, likely correcting an error in TED model predictions. (B) Illustrates the case where a report establishes connectivity between two isolated patches of tsetse. (C) Illustrates the case where a report of tsetse presence is spatially isolated from the predicted distribution of tsetse. In each case, a user is presented with a prediction of tsetse distribution from the TED model (Column 1). Users identify an error in the model, observing tsetse in an area where they are not predicted to occur, and submit a report (Column 2 - black box). The report is submitted for reliability assessment; if deemed reliable, TED model predictions are updated to reflect the new information (Column 3 – black box).



safety distribution. As stated previously, the TED model is designed to minimize errors of commission at the expense of added errors of omission. The assumption here is that the gap observed in the distribution is a product of this process. When a report is made, occurring within the bounds of this, the probability of that report being accurate is reasonably high. Therefore, our model should assign a high reliability score to that report.

The second case involves a report that connects two clusters of tsetse. In this case, we assume that patches of tsetse distribution should, for the most part, maintain connectivity in some

way. If a report is made that establishes connectivity between patches (Figure 5b), there's a high likelihood that this report is reliable. Therefore, our model should assign a score that reflects this likelihood.

Finally, we simulate the case in which a report is made which places tsetse in a region that is isolated from predicted patches of tsetse distribution (Figure 5c). Since we have no prior reason to believe tsetse occur in this region, based on model projections, there is a low likelihood that this report is true. Therefore, our model should assign a reliability score that emphasizes the extreme nature of this report. Through these simulations, we can identify the effective threshold for reliability.

CONCLUSION AND LIMITATIONS

Communications barriers present one of the most prominent barriers to disease surveillance programs. When communications between health care providers and regional health authorities break down under passive surveillance systems, there is a need to make attempts to directly collect disease incidence data directly. Yet, there are significant hurdles to implementing active surveillance programs (e.g., costs and logistics). By adopting concepts of crowdsourcing, public participation, and volunteered GIS, we can open the door for an intermediate solution for disease surveillance. Such an intermediate solution employs citizens to collect surveillance information, increasing the manpower available to collate the information. It may not then be necessary to dispatch health professionals to procure the data directly. Targeted campaigns can also be utilized to solicit participation on behalf of the public to assist in collecting surveillance data. Finally, our approach to assessing the credibility of volunteered information increases the utility and reliability of data obtained from these campaigns.

However, there are significant limitations to a full implementation of the VGI in our case study. The region in Kenya in which we are working is remote; there are significant challenges in terms of communications connectivity, reliable electricity, and necessary hardware; AMREF (American Medical and Research Foundation) and the African Conservation Centre (AAC) have made significant improvements to local infrastructure, but much more is required. Cost remains the most substantial hurdle for regional implementation. Our utilization of open source solutions mitigates, but does not eliminate this challenge. Absent assistance from international partners, the likelihood of full implementation of the disease surveillance system will certainly remain in the domain of our scientific and development collaborators.

Connors et al. (2011) draw attention to the potential value of incorporating additional sources of information (e.g., Twitter, Flickr), aside from direct volunteering through a VGIS, to allow for increased participation; however, in doing so we would be introducing new types of uncertainty to the models. Our current design attempts to limit error exclusively to those of omission (i.e., we have tried to ensure that TED model predictions are estimations of the minimum area tsetse are distributed). In this way, we have greater confidence over the areas TED predicts tsetse to occur. When users volunteer reports of tsetse occurrence, they do so by providing GPS coordinates of their location (this is done in the background through the iOS application). Incorporating Twitter feeds, geo-tagged Flickr photos, among others would on the one hand provide us with more information; however the cone of location uncertainty of that information is much greater and far less tractable. These sources of volunteered information represent important avenues for future development, particularly in the broader field of VGI; but at this time are beyond the scope of what we believe to be possible to include in our project.

Critical to the success of VGIS for disease surveillance is adequate public participation. Too few reporters can make it difficult to assess credibility and limits the conclusions that can be drawn from the information collected; however if incentives for participation are carefully

considered, there can be a drive for individuals to accurately and reliably contribute to the system. Integration of volunteered information for disease surveillance, especially in low-income countries, can be used as an alternative to the high costs of active surveillance programs, which are often implemented in rural areas to learn more about disease prevalence. The prototype for a VGIS outlined in this study demonstrates how technology and participatory science can advance passive disease programs to improve public health in needed parts of the world.

ACKNOWLEDMENTS

This research was supported by the National Institutes of Health, Office of the Director, Roadmap Initiative, and NIGMS Award No. RGM084704A.

REFERENCES

Altmann, J., Alberts, S. C., Altmann, S. A., & Roy, S. B. (2002). Dramatic change in local climate patterns in the Amboseli basin, Kenya. *African Journal of Ecology*, *40*(3), 248–251. doi:10.1046/j.1365-2028.2002.00366.x.

Baird, T., Leslie, P., & McCabe, J. (2009). The effect of wildlife conservation on local perceptions of risk and behavioral response. *Human Ecology*, *37*(4), 463–474. doi:10.1007/s10745-009-9264-z.

Balram, S., & Dragićević, S. (2006). *Collaborative geographic information systems*. Hershey, PA: Idea Group Publishing. doi:10.4018/978-1-59140-845-1.

Batchelor, N., Atkinson, P., Gething, P., Picozzi, K., Fevre, E., Kakembo, A., & Welburn, S. (2009). Spatial predictions of Rhodesian human African trypanosomiasis (Sleeping Sickness) prevalence in Kaberamaido and Dokolo, Two Newly Affected Districts of Uganda. *PLoS Neglected Tropical Diseases*, *3*(12), e563. doi:10.1371/journal.pntd.0000563 PMID:20016846.

Bauer, B., Kabore, I., Liebisch, A., Meyer, F., & Petrich-Bauer, J. (1992). Simultaneous control of ticks and tsetse flies in Satiri, Burkina Faso, by the use of flumethrin pour on for cattle. *Tropical Medicine and Parasitology*, *43*(1), 41–46. PMID:1598507.

Bishr, M., & Kuhn, W. (2007). Geospatial information bottom-up: A matter of trust and semantics. In S. I. Fabrikant, & M. Wachowicz (Eds.), *Lecture notes in geoinformation and cartography* (pp. 365–387). Berlin, Germany: Springer Berlin - Heidelberg. doi:10.1007/978-3-540-72385-1_22.

Bishr, M., & Mantelas, L. (2008). A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal*, *72*(3), 229–237. doi:10.1007/s10708-008-9182-4.

Boroushaki, S., & Malczewski, J. (2010a). Measuring consensus for collaborative decision-making: A GIS-based approach. *Computers, Environment and Urban Systems*, *34*(4), 322–332. doi:10.1016/j.compenvurbsys.2010.02.006.

Boroushaki, S., & Malczewski, J. (2010b). Using the fuzzy majority approach for GIS-based multicriteria group decision-making. *Computers & Geosciences*, *36*(3), 302–312. doi:10.1016/j.cageo.2009.05.011.

Conati, C. (2004). How to evaluate models of user affect? In E. André, L. Dybkjær, W. Minker, & P. Heisterkamp (Eds.), *Affective dialogue systems* (Vol. 3068, pp. 288–300). Berlin, Germany: Springer Berlin - Heidelberg. doi:10.1007/978-3-540-24842-2_30.

Connors, J. P., Lei, S., & Kelly, M. (2011). Citizen science in the age of neogeography: Utilizing volunteered geographic information for environmental monitoring. *Annals of the Association of American Geographers. Association of American Geographers*. doi: doi:10.1080/000-45608.2011.627058.

Craglia, M. (2007). Volunteered geographic information and spatial data infrastructures: When do parallel lines converge. In *Proceedings of the VGI Specialist Meeting*, Santa Barbara, CA. Retrieved from http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Craglia_paper.pdf

Crosetto, M. (2001). Uncertainty and sensitivity analysis: Tools for GIS-based model implementation. *International Journal of Geographical Information Science*, *15*(5), 415–437. doi:10.1080/13658810110053125.

DeVisser, M., Messina, J. P., Moore, N., & Lusch, D. (2010). A dynamic species distribution model of Glossina subgenus Morsitans: The identification of tsetse reservoirs and refugia. *Ecosphere*, *1*(1), 1–21. doi:10.1890/ES10-00006.1.

Elmasri, R., & Navathe, S. (2008). *Fundamentals of database systems* (Vol. 2). Upper Saddle River, NJ: Pearson Education.

Elmes, G., Dougherty, M., Challig, H., Karigomba, W., McCusker, B., & Weiner, D. (2005). Local knowledge doesn't grow on trees: Community-integrated geographic information systems and rural community self-definition. In P. Fisher (Ed.), *Developments in spatial data handling* (pp. 29–39). Berlin/Heidelberg, Germany: Springer-Verlag. doi:10.1007/3-540-26772-7_3.

Elwood, S. (2006). Negotiating knowledge production: The everyday inclusions, exclusions, and contradictions of participatory GIS research. *The Professional Geographer*, *58*(2), 197–208. doi:10.1111/j.1467-9272.2006.00526.x.

Elwood, S. (2010). Geographic information science: emerging research on the societal implications of the geospatial web. *Progress in Human Geography*, *34*(3), 349–357. doi:10.1177/0309132509340711.

Fischer, G., Shah, M., Tubiello, F., & Velhuizen, H. (2005). Socio-economic and climate change impacts on agriculture: An integrated assessment, 1990-2080. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1463). doi:10.1098/rstb.2005.1744 PMID:16433094.

Flanagin, A., & Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3), 137–148. doi:10.1007/s10708-008-9188-y.

Goodchild, M. F. (2007a). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. doi:10.1007/s10708-007-9111-y.

Goodchild, M. F. (2007b). Citizens as voluntary sensors: Spatial data infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, *2*, 24–32.

Goodchild, M. F. (2010). Crowdsourcing geographic information for disaster response: A research frontier. *International Journal of Digital Earth*, *3*(3), 231–241. doi:10.1080/17538941003759255.

House, R., Javidan, M., & Dorfman, P. (2001). Project GLOBE: An introduction. *Applied Psychology*, *50*(4), 489–505. doi:10.1111/1464-0597.00070.

Langley, S. (2010). *Developing a participatory GIS network for the mapping of African trypanosomiasis in Kenya*. Unpublished Interview Data.

Langley, S. A., & Messina, J. P. (2011). Embracing the open-source movement for managing spatial data: A case study of African trypanosomiasis in Kenya. *Journal of Map & Geography Libraries*, *7*(1), 87–113. doi:10.1080/15420353.2011.534693 PMID:21686072.

Maitima, J. M. (2012). *The variabilities in plant phenological activities in response to rainfall and temperature variabiltiy in Nguruman*. Unpublished Manuscript.

McCall, M., & Minang, P. (2005). Assessing participatory GIS for community-based natural resource management: claiming community forests in Cameroon. *The Geographical Journal*, *171*(4), 340–356. doi:10.1111/j.1475-4959.2005.00173.x.

McKnight, K. P., Messina, J. P., Shortridge, A. M., Burns, M. D., & Pigozzi, B. W. (2011). Using volunteered geographic information to assess the spatial distribution of West Nile Virus in Detroit, Michigan. *International Journal of Applied Geospatial Research*, *2*(3), 72–85. doi:10.4018/jagr.2011070105.

Metcalf, S., & Paich, M. (2005). Spatial dynamics of social network evolution. In *Proceedings of the 23rd International Conference of the System Dynamics Society.*

Moore, N., Alagarswamy, G., Pijanowski, B., Thornton, P., Lofgren, B., & Olson, J. et al. (2012). East African food security as influenced by future climate change and land use change at local to regional scales. *Climatic Change*, *110*(3-4), 823–844. doi:10.1007/s10584-011-0116-7.

Moore, N., Messina, J. P. (2010). A landscape and climate data logistic model of tsetse distribution in Kenya. *PLoS ONE*, *5*(7), e11809. doi:10.1371/journal.pone.0011809 PMID:20676406.

Morris, D. L., Western, D., & Maitumo, D. (2009). Pastoralist's livestock and settlements influence game bird diversity and abundance in a savanna ecosystem of southern Kenya. *African Journal of Ecology*, *47*(1), 48–55. doi:10.1111/j.1365-2028.2007.00914.x.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth sciences. *Science (New York, NY)*, *263*(5147), 641–646. doi:10.1126/science.263.5147.641 PMID:17747657.

Rosenberg, T. (2011, Apr 28). Crowdsourcing a better world. *NYTimes.com*. Retrieved from http://opinionator.blogs.nytimes.com/2011/03/28/crowdsourcing-a-better-world/?pagemode=print

Sieber, R. (2006). Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers. Association of American Geographers*, *96*(3), 491–507. doi:10.1111/j.1467-8306.2006.00702.x.

Snyder, J. D., & Merson, M. H. (1982). The magnitude of the global problem of acute diarrhoeal disease: a review of active surveillance data. *Bulletin of the World Health Organization*, *60*(4), 605–613. PMID:6982783.

Stonebraker, M., & Kemnitz, G. (1991). The POSTGRES next generation database management system. *Communications of the ACM*, *34*(10), 78–92. doi:10.1145/125223.125262.

Stonebraker, M., & Rowe, L. (1986). The design of POSTGRES. In *Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data* (pp. 340-355).

Sui, D. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, *32*(1), 1–5. doi:10.1016/j.compenvurbsys.2007.12.001.

Tarimo-Nesbitt, R., Golder, T., Dransfield, R., Chaudhury, M., & Brightwell, R. (1999). Trypanosome infection rate in cattle at Nguruman, Kenya. *Veterinary Parasitology*, *81*(2), 107–117. doi:10.1016/S0304-4017(98)00194-0 PMID:10030753.

Terblanche, J. S., Clusella-Trullas, S., Deere, J. A., & Chown, S. L. (2008). Thermal tolerance in a south-east African population of the tsetse fly Glossina pallidipes (Diptera, Glossinidae): Implications for forecasting climate change impacts. *Journal of Insect Physiology*, *54*(1), 114–127. doi:10.1016/j.jinsphys.2007.08.007 PMID:17889900.

Thacker, S. B., Choi, K., & Brachman, P. S. (1983). The surveillance of infectious diseases. *Journal of the American Medical Association*, *249*(9), 1181–1185. doi:10.1001/jama.1983.03330330059036 PMID:6823080.

Turner, A. (2006). *Introduction to neogeography*. Sebastapol, CA: O'Reilly Media, Inc..

Waller, R. (1990). Tsetse fly in western Narok, Kenya. *Journal of African History*, *31*(1), 81–101. doi:10.1017/S0021853700024798.

Watson, H. J., Wixom, B. H., & Goodhue, D. L. (2004). Data warehousing: The 3M experience. In H. R. Nemati, & C. D. Barko (Eds.), *Organizational data mining: leveraging enterprise data resources for optimal performance* (p. 202). Hershey, PA: Idea Group Inc. doi:10.4018/978-1-59140-134-6.ch014.

WHO. (2005, August 22-26). Control of human African trypanosomiasis: A strategy for the African region. In *Proceedings of the 55th Session Regional Comittee for Africa*.

Yaukey, P. H. (2010). Citizen science and bird–distribution data: An opportunity for geographical research. *Geographical Review*, *100*(2), 263–273.

ENDNOTES

1 Michigan State University IRB# x10-1134
2 http://mapserver.org

Shaun A. Langley received his bachelor's degree in Wildlife Ecology from the University of Wisconsin - Madison, USA; master's degree in Entomology from Michigan State University, USA. He is working towards his PhD in Geography at Michigan State University, USA, with a focus on Volunteered GIS.

Joseph P. Messina is a professor in the Department of Geography at Michigan State University. He holds appointments with the Center for Global Change and Earth Observations and AgBio-Research. He received his PhD in Geography from the University of North Carolina at Chapel Hill in 2001. His research focus concerns disease ecology, geospatial analysis, complex systems, and earth observation. Current funded research includes the disease ecology of Malaria (NSF) and Sleeping Sickness (NIH) in Kenya and complex, integrated models of climate and urbanization (NASA) in China. He teaches courses in medical geography, remote sensing, and GIS.

Sue C. Grady is a medical/health geographer in the Geography Department at Michigan State University. She has a PhD in Earth and Environmental Sciences from the City University of New York and a MPH degree in International Public Health from Tulane University.