



# MIT Open Access Articles

## *Towards the SocioScope: an Information System for the the Study of Social Dynamics through Digital Traces*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

<b>Citation</b>	Vaccari, Andrea et al. "Towards the SocioScope." Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM GIS '09. 52-61. ©2009 Association for Computing Machinery.
<b>As Published</b>	<a href="http://doi.acm.org/10.1145/1653771.1653782">http://doi.acm.org/10.1145/1653771.1653782</a>
<b>Publisher</b>	Association for Computing Machinery
<b>Version</b>	Author's final manuscript
<b>Citable link</b>	<a href="http://hdl.handle.net/1721.1/55366">http://hdl.handle.net/1721.1/55366</a>
<b>Terms of Use</b>	Attribution-Noncommercial-Share Alike 3.0 Unported
<b>Detailed Terms</b>	<a href="http://creativecommons.org/licenses/by-nc-sa/3.0/">http://creativecommons.org/licenses/by-nc-sa/3.0/</a>



**Andrea Vaccari**  
**Francesco Calabrese**  
**Bing Liu**  
**Carlo Ratti**

**Towards the SocioScope: an  
Information System for the  
Study of Social Dynamics  
through Digital Traces**

# Towards the SocioScope: an Information System for the Study of Social Dynamics through Digital Traces

Andrea Vaccari  
Massachusetts Institute of  
Technology  
77 Massachusetts Ave  
02139 Cambridge, MA, USA  
avaccari@mit.edu

Bing Liu  
University of Illinois at Chicago  
851 S. Morgan Street  
60607 Chicago, IL, USA  
liub@cs.uic.edu

Francesco Calabrese  
Massachusetts Institute of  
Technology  
77 Massachusetts Ave  
02139 Cambridge, MA, USA  
fcalabre@mit.edu

Carlo Ratti  
Massachusetts Institute of  
Technology  
77 Massachusetts Ave  
02139 Cambridge, MA, USA  
ratti@mit.edu

## Categories and Subject Descriptors

H.4.0 [Information Systems Applications]: General; H.2.8 [Database Management]: Database Applications—*Spatial Databases and GIS*

## General Terms

Design, Management, Security

## Keywords

Digital traces, social dynamics

## ABSTRACT

Over the past decade there has been an explosion in the deployment of pervasive systems like cell phone networks and content aggregators on the Internet that produce massive amounts of data as by-products of their interaction with users. This data is related to the actions and opinions of people and thereby to the overall dynamics of cities, how they function and evolve over time.

Here we propose the implementation of a new information system, the *SocioScope*, that streamlines the operations of data gathering and data storage and that supports the diverse needs of the research community, local authorities, enterprises, and citizens. Our *SocioScope* allows for the collection, integration, and study of different types of spatio-temporal data.

We present two use cases and an analysis of the requirements for the system. We introduce the multi-tier architecture that supports data gathering, storage, integration,

analysis, and visualization. We describe the technical implementation of each module of the system and of the security mechanism for data access control. We conclude with a discussion on the opportunities and challenges of the study of digital traces and with an brief illustration of future directions of work.

## 1. INTRODUCTION

Over the past decade there has been an explosion in the deployment of pervasive systems like cell phone networks and content aggregators on the Internet that produce massive amounts of data as by-products of their interaction with users. This data is related to the actions and opinions of people and thereby to the overall dynamics of cities, how they function and evolve over time. Electronic logs of cell phone calls and geotagged photographs are examples of *digital footprints* [13] that today allow researchers to better understand how people flow through urban space [14], and could ultimately help those who manage and live in urban areas to configure more liveable, sustainable, and efficient cities [29]. Moreover news and descriptions of events, as well as blog posts and online reviews of products and services are forms of *buzz* that can often be geocoded to build semantic indexes of different parts of a city [23, 2].

Indeed there already exist many projects that focus on the study of digital footprints [13, 25, 18] and the implementation of new mobile services [21, 3, 20] and interfaces [19, 26, 9]. These studies, however, present two major limitations. First, they tend to consider only one data set at a time, and therefore they fall short on the inherent limitations of the particular trace (e.g. spatial resolution or statistical representativeness). Second, they do not corroborate their results with other types of urban traces and with results from more traditional methodologies typical of transportation engineering and urban planning. To complement these projects and study the significance of digital traces from a multidisciplinary perspective, we developed WikiCity [5, 6], a system that facilitated the flow of real-time information between the city and its inhabitants.

WikiCity Rome, in particular, was a proof-of-concept ap-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM GIS '09, November 4-6, 2009, Seattle, WA, USA.

Copyright 2009 ACM ISBN 978-1-60558-649-6/09/11 ...\$10.00.

plication that demonstrated the potential for the participatory provisioning of urban information by local authorities, service providers, businesses, and citizens. Its goal was to promote the active use of digital traces for real-time location-based decision making in the city. It consisted of a map of Rome, Italy that provided real-time information about cell phone network activity in the form of a heat map of the aggregate number of calls, and details about the public events, the position of busses, and the news of *La Repubblica*, a major Italian newspaper. During its development, we faced challenges to collect and analyze data from different data providers (e.g. telecom operators and public authorities), with social and legal issues like privacy concerns and the elaboration of non disclosure agreements, and technological problems like the heterogeneity of the data, which came in different formats, resolutions, and scales, and the capacity to store and stream large amount of information.

We believe that these challenges constitute a barrier to the development and application of pervasive systems and digital traces in the context of the city. Here we propose the implementation of a new information system, the *SocioScope*, that streamlines the operations of data gathering and data storage and that supports the diverse needs of the research community, local authorities, enterprises, and citizens. Our *SocioScope* allows for the collection, integration, and study of different types of spatio-temporal data like cell phone network activity (e.g. number of calls and text messages), geotagged picture collections from content aggregators like Flickr.com, descriptions of public events from online repositories like Upcoming.org, posts and reviews from websites like Twitter and Yelp, and census statistics and land use information from government databases.

Our implementation includes a data integration system that provides a unified view of the data sets, thus hiding the differences in resolution and scale between them. It also includes a set of standard operations that can be applied to the integrated data sets: for example it is possible to calculate a *sanity map* of the cell phone network activity to compute the average number of phone calls generated in different areas of the city at 10 pm of a Saturday. The system also provides different methods to visualize the data and the results of the operations: for example it is possible to calculate a *flow map* of the movements of tourists in a city through the analysis of geotagged pictures, to visualize and navigate it with a web interface and then to export it into a KML file that can be opened with Google Earth.

The system also support a security mechanism based on the traditional concepts of *roles* and *privileges*. The latter have been extended to support the definition of spatial or temporal constraints: a user can be limited not only in the type of data and operations that she can access, but also in the ranges and scales of the viewport and timeport of a query. For example, she might be limited to access only data in the New York City area, or only data at the scale of the borough; moreover she might be limited to access data that is at least one month old, or data that is aggregated at least at the scale of the day.

It should be noted that here we are proposing a general-purpose information system, neglecting a specific application in favor of an exploration of the broad domain of social dynamics. In the future, we will document the use of the system in specific applications like the monitoring of traffic conditions or emergency situations in the city. In the

remainder of this paper we present two use cases for the *SocioScope* information system. Based on them, we propose an analysis of the requirements for the system and we introduce the multi-tier architecture that supports data gathering, storage, integration, analysis, and visualization. Then, we describe the technical implementation of each module of the system and of the security mechanism for data access control and we conclude with a discussion on the opportunities and challenges of the study of digital traces and with an brief illustration of future directions of work.

## 2. RELATED WORKS

There exist many projects that attempt to extend geographical information systems within the context of urban environments. Yu et al. [30] recognized the need for a collaborative framework for location-based services for information searching. Their approach is based on global ontologies (e.g. WordNet or OntoLingua) and shared ontologies intended to exhibit the commonness and characteristics of the associated local sources of information according to geographical scales, domains, formats, or user profiles. While the framework implements a module to respond to users' queries by analyzing the query, locating relevant sources of information and merging the information together, it ignores aspects of data collection and structuring as well as those of data protection and management.

WIPER [21] is a system for emergency response that leverages the cellphone network data analysis to detect emergency situations, simulate their evolution, and provide courses of action in dealing with them. It is based on a three-layered architecture composed by a *Data Source and Measurement* layer that physically resides on the cellular network and handles collection, storage and preprocessing of the data, a *Detection, Simulation and Prediction* layer that resides on the WIPER network and processes the streaming data to detect anomalies and runs simulations for prediction and mitigation, and a *Decision Support* layer that provides a web-based application for the emergency response managers. The system provides a good architecture for collecting and handling cellular activity logs, and highlights one of the major issues in this process: the lack of control over the provider's network and therefore the inability to use real-time standard protocols in collecting the data.

A more thorough approach is proposed by Microsoft Research in two sister projects, MSR Sense [19] and SenseWeb [26]. MSR Sense and SenseWeb successfully overcome some of the major drawbacks of ad hoc applications. In particular, they ease the process of publishing streams of data as a useful service and they provide primitives to query sensors and aggregate data based on keywords or location. Nonetheless, while their primitives are sufficient to support simple geospatial applications, they are not flexible enough to support the needs of the research community. In particular, they fail to consider data ownership and privacy issues that can be solved by allowing data providers to control their data streams and specify the access level of different types/groups of users, and to define a mechanism to extend the spatio-temporal operations applicable to the data. These issues are however all addressed in our solution.

Few commercial solutions have been developed to collect and manipulate urban data. NAVITIME [3] is an example of urban information system that focuses on multi-modal mobility and provides wayfinding support that considers all

modes of transportation and encompasses the entire traveling activity. It is a mobile phone-based navigation service used by almost 2 million users in Japan that incorporates various modes of transportation, such as walking, driving, and riding trains, buses, taxis and airplanes, to provide a comprehensive routing system and to interactively guide users with maps, itineraries, voice prompts, vibration alerts, progress bars, and so on. The major drawback of NAVITIME is that it is closed-source and therefore it is not possible on one side to easily plug-in new data sources, and on the other side to customize or extend the services offered.

Other projects instead provide strong algorithms for the analysis of urban dynamics and would benefit from an integrated system to share data visualization techniques. For example, UrbanSim [8] is a sophisticated simulation systems that models the long-term impacts of transportation and land-use alternatives on a region's urban processes over several decades. It uses information related to the region's households, jobs, travel routes, and real estate to provide estimations of their future trends and to generate indicators like population and employment density that can be useful for stakeholders. UrbanSim is open source software and we believe that it would strongly benefit by having access to additional information like the cellular network usage of a specific region, but also that the results of its analysis could be functional to other projects.

The lack of interoperability in these systems is symptomatic of many challenges. First the data that needs to be collected and integrated is very diverse, ranging from geospatial to demographic data, from environmental measurements to cellular activity logs. Moreover data is in the hands of several players that often do not use standard protocol for collecting and storing it and do not have majors interest in sharing it. Finally, even if data can be collected and integrated, the amount of storage required would quickly become the primary technological limitation to address [4].

### 3. USE CASES AND REQUIREMENTS

To motivate our design choices, we envisioned how the system could be collaboratively used by different stakeholders. We present two scenarios: the first describes how researchers like social scientists, practitioners like urban planners, local authorities, and data providers like telecom operators may study digital traces to develop a common vocabulary and a shared set of tools and practices towards a grounded theory of *digital sociology* [29]; the second exemplifies the social implementation of the system by businesses and citizens to develop urban mash-ups and situational applications.

#### 3.1 Use Cases

In a first scenario, providing a standardized mechanism to collect and analyze aggregate logs of their cell phone network activity, SocioScope can be used by telecom operators to monitor and analyze statistical information about their network that are orthogonal to their current analyses. While call accounting systems focus on single users to analyze specific information used to bill the subscriber, SocioScope allows to study aggregate information of the overall patterns and behaviours of their customers. For example: Where do young people go at night or during the weekends? How do they flow from one place to adjacent areas of interest? What parts of the city are characterized by similar types of dwellers or activities?

Telecom operators may extend access to the data to social scientists that could understand the laws governing local and remote communication behaviors, and to urban planners that could assess land use and mobility flows. Analytical tools would be used by different actors to support such studies: these range from quantitative indicators of qualitative properties of the urban context to the visualization of hotspots and flows within and between cities. Analyses on aggregate logs of network activity, for example, could reveal overall patterns of behavior of people [16], and could complement reports on land use and census data [24].

With real-time access to the results of these analyses, urban planners would be able to promptly detect and correct phenomena that reduce the livability and sustainability of the city: instead of planning year-long urban interventions and waiting months to evaluate their impact, they could shift to a more reactive configuration of the city. This could have important benefits for local municipalities in handling public events and responding to sudden emergencies: in particular, the analyses developed by the researchers could be used by decision makers without technical expertise to assess the criticality of a situation and to forecast how different solutions would work.

This scenario exemplifies how the collaboration between researchers, practitioners, local authorities and data providers could provide greater understanding of how the city function as a real-time control system and how it is stressed during public events and sudden emergencies. Two concrete examples of such application, both based on the SocioScope system, are the analysis of the New York City Waterfalls public exhibition of Summer 2008 (see [14, 15]) and the Obama | One People visualizations for the Inauguration Day of January 20, 2009 (see [22] and Figure 1). The former was a impact study of the NYC Waterfalls, a \$20 million public exhibition in New York aimed at attracting people towards the waterfront of the city, where we demonstrated how digital footprints have significant promise for conducting tourism studies and the optimization of tourism strategies, plans and marketing tools. The latter are two visualizations mobile phones call activity that characterized the inaugural crowd and answered the questions: Who was in Washington, D.C. for President Obama's Inauguration Day? When did they arrive, where did they go, and how long did they stay?

In a second scenario, learning from the success of social networks, user-generated content aggregators, mashups, and scripts on the Web, SocioScope can provide support for the development of situational applications, software that is designed for use by a specific social group, rather than for a generic set of users [28]. For example, a user could develop an automated jogging routes generator that takes into account air and noise pollution, and traffic patterns to provide suggestions of possible trips near the user. Moreover, another user could build on top of it and add support for weather forecasts to plan a weekly schedule in the country side that takes into account weather conditions. If critical mass was reached, we believe that the collaborative framework would become self-supporting, generating libraries of mashups hierarchically built one on top of the other that would provide useful decision support tools for different situations of urban living. Citizens could access such applications without having to know how the information is computed, being free to focus on the information provided for better informed decisions.

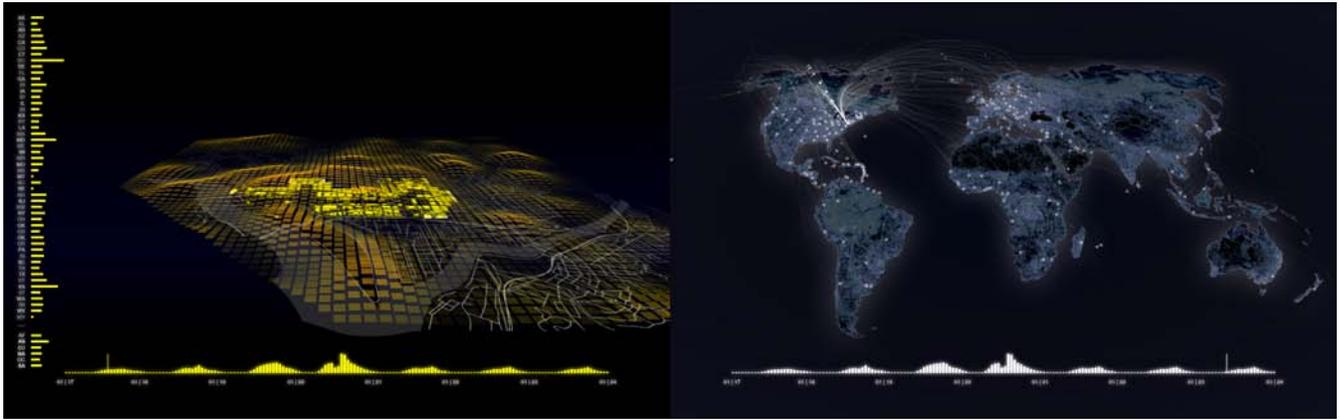


Figure 1: Obama | One people is an unprecedented analysis of Barack Obama’s Inauguration Day on January 20, 2009. The project features two visualizations of mobile phones call activity that characterize the inaugural crowd and answered the questions: Who was in Washington, D.C. for President Obama’s Inauguration Day? When did they arrive, where did they go, and how long did they stay?

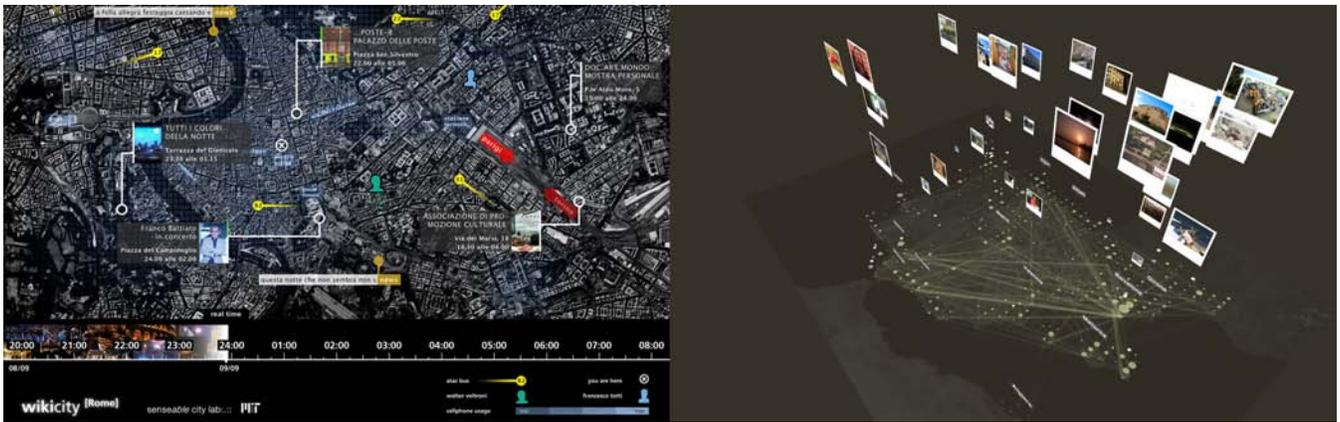


Figure 2: WikiCity Rome (on the left) is a map of Rome, Italy that provides real-time information about cell phone network activity in the form of a heat map of the aggregate number of calls, and details about the public events, the position of busses, and the news of La Repubblica, a major Italian newspaper. The World’s Eyes (on the right) illustrates the photos that people visiting Spain leave behind them as evidences of contemporary tourism in the country, uncovering the evolutions of the presence and flows of tourists.

A similar model could be implemented by businesses to streamline logistics and operations, or to support a one-time task that affects a small group of employee. Applications that share and aggregate information among employees, like structured wikis, are all example of situational applications. As an example, a company could assemble a trouble call system that integrates its existing customer database, command and control system, and fleet management system with a GIS server to locate an affected customer based on a phone number; the system would then find the affected service, perform network tracing and analysis, identify the potential remedy and, finally, route the closest service technicians to the proper location with detailed instructions on the work to be performed [10]. Interestingly, a preliminary approach towards this direction is currently available to IBM’s employees [7]. The IBM Situational Applications Environment was built to help individuals and small teams create ad hoc composite applications to address their immediate business needs. In the same way, Socio-

Scope could be used to support situational applications that need access to spatio-temporal data and geospatial analyses in a controlled environment.

Two examples of situational applications for the citizens are WikiCity Rome (see [6] and Figure 2 on the left) and World’s Eyes (see [12] and Figure 2 on the right). The former is described in the Introduction and was the original project that motivated the development of SocioScope: it consisted of a map of Rome, Italy that provided real-time information about cell phone network activity in the form of a heat map of the aggregate number of calls, and details about the public events, the position of busses, and the news of La Repubblica, a major Italian newspaper. The latter is a set of visualizations that illustrate the photos that people visiting Spain leave behind them as evidences of contemporary tourism in the country, uncovering the evolutions of the presence and flows of tourists. The analysis and mapping of this data allows to understand the attractiveness of leisure cities and their points of interest.

## 3.2 Requirements

Based on the above scenarios, we have compiled a list of requirements for the system. The foremost requirement is the seamless gathering, storage, integration, and provision of spatio-temporal data collected from systems that are independent and unaware of each other. Such systems generate heterogeneous data sets that are stored on different platforms, use different representation formats, and follow different conventions. For these reasons, the gathering process requires a malleable mechanism in order to connect multiple data sources using different communication protocols and data formats, and to support both new and legacy sensing platforms and pervasive systems. Moreover, the storage process requires a flexible data persistence model that has to be space- and time-efficient and should still support data types that are characterized by different levels of resolution and covering different geographical scales from street-level, to intracity- and intercity-level. The collection-to-integration workflow should also take into account issues related to transfer bandwidth, handling of dense datasets, data freshness, data ownership, and propagation latency [1].

As discussed by Schoupe [27], monitoring and reporting processes require more cohesive and streamlined workflows, improved comparability and better accountability, and higher-quality data with respect to timeliness, completeness, accuracy, accessibility, consistency, integrity, neutrality, and robustness. These requirements elicit the need to complement the current panorama of disparate systems with heterogeneous information and ad-hoc searches and downloads, with an information systems that use open standards to support the dynamic management of heterogeneous networks of sensors, data flows, and interactive services on the web. This can be obtained through the use of open geospatial protocols and interfaces as defined by the Open Geospatial Consortium, the Global Spatial Data Infrastructure Association, and the Open Source Geospatial Foundation.

The system must also be able to provide useful information to users with or without special knowledge, to support scientific investigations as well as real operations, and to enable decision support for both single-actor and multi-actor actions. These functionalities should be provided by one common portal which adapts accordingly to the user's profile and scales the complexity of the interface according to her needs. The portal should also stimulate the participation and public sharing of data and knowledge by allowing users to easily implement a new data stream and to maintain ownership of it. Finally, the system should enforce authentication and authorization to control data access and avoid abuses and vandalism. It should also guarantee high standards of ethics and privacy, even if these two widely-discussed problems have yet to find commonly accepted satisfactory solution.

## 4. SYSTEM ARCHITECTURE

SocioScope supports a workflow that begins with the gathering of urban data, continues with the storage and integration of bits of data coming from different sources, and finishes with the analysis and visualization of urban information. This streamlined process is reflected in the architecture of the system represented in Figure 3 which is subdivided into a stack of five layers, each exposing a interface based, when available, on open standards. The five layers are:

**Data Gathering.** This layer provides standard tools for uploading data. This operation can be event-based, in which case data is transferred to the system either once or periodically, or continuous, in which case a data stream is permanently connected to the system. The transfer can be over FTP/SFTP/HTTP connection, UDP packets, web services, database transactions, and email-based uploads. It can also process different file formats, like GeoRSS, KML, CSV files, GIS application-specific formats like Shapefiles, and other non-standard-formats.

**Data Storage.** This layer stores the raw data into a geospatial DBMS. It also stores meta-data related to users' credentials, roles, and permissions. This task requires to define a schema model that is flexible enough to accommodate different data representations, resolutions, and scales, but also capable to efficiently store the data without too much overhead and to retrieve information within soft real-time constraints.

**Data Integration.** This layer provides a mechanism for the integration of data types, retrieved either from historical data sets or real-time data streams. As depicted in Figure 3, this tier receives data from the gathering and storage layers using open geospatial standards and database transactions and returns the data in a unified view that hides the differences in resolutions and scales.

**Data Analysis.** This layer provides standard operations and urban indicators that can be applied to the integrated data served by the integration layer. For example it is possible to calculate a *sanity map* of the cell phone network activity to compute the average number of phone calls generated in different areas of the city at 10 pm of a Saturday. The layer also provides a web service for the automated retrieval of the results of the analyses: this allows to extend the system with external analytical tools.

**Data Visualization.** This layer consists of different methods to visualize the data and the results of the operations. A web interface based on Flash allows to explore data and to identify interesting trends and correlations, while another web service allows to define spatio-temporal queries and to export the results of the analyses into a KML file that can be opened with Google Earth.

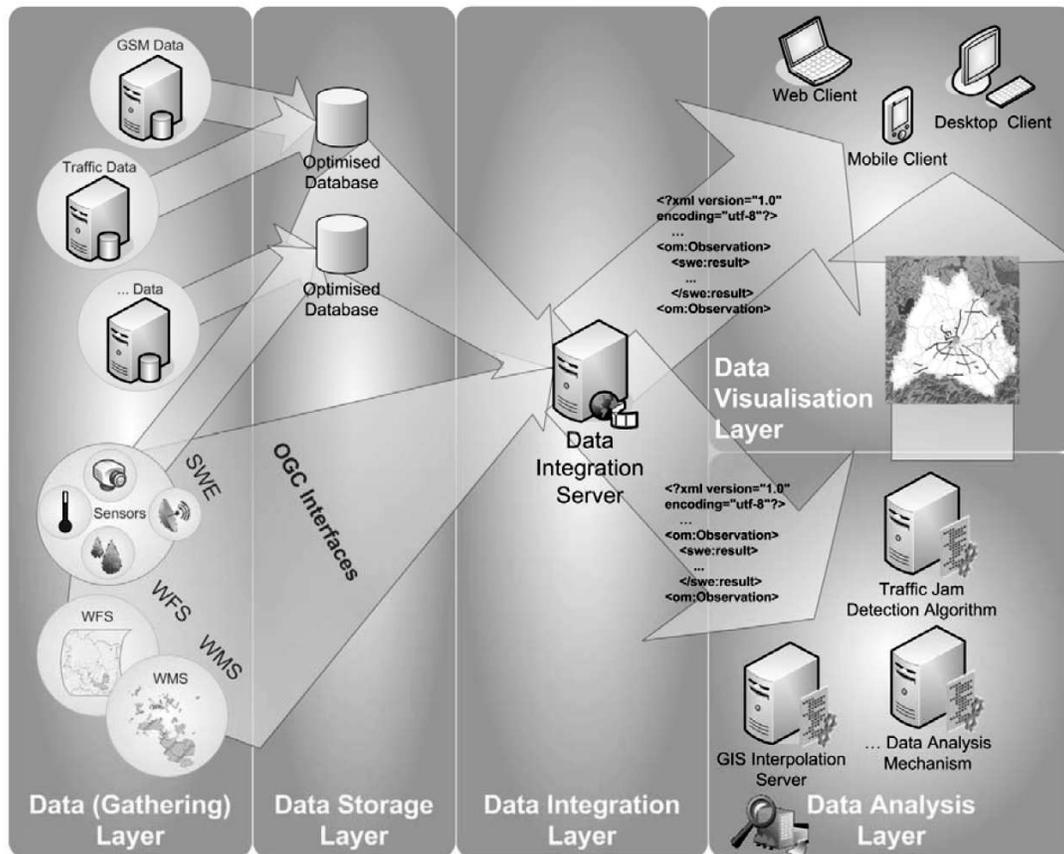
## 5. IMPLEMENTATION

SocioScope features a stable implementation of the data gathering, data storage, and data visualization layers, and a prototypical implementation of the integration and analysis layers. It also includes a complete implementation of security and logging cross-cutting concerns, and system-wide functions for management and monitoring.

### 5.1 Data Gathering and Persistence Layer

The Data gathering layer of the architecture was implemented for the following types of connections: FTP/SFTP/HTTP, email, UDP packets transfer and web services. A routine is associated to each data stream, and can be configured so as to retrieve data samples with a certain frequency or due to the occurrence of events (e.g. new e-mail arrived or UDP packet received). The retrieved data is then parsed from the specific data format, and converted to a Java object that can be manipulated by the Data persistence layer.

The Data storage layer uses the PostgreSQL DBMS with the PostGIS extension that natively support GIS data types



**Figure 3: The SocioScope system architecture is based on five layers that support the gathering of urban data, the storage and integration of bits of data coming from different sources, and the analysis and visualization of urban information.**

and operations. PostgreSQL also supports multiple schemas per database, a feature that is used to divide the data into spatial partitions, so that each partition is dedicated to a self-contained group of datasets that pertain the same geographical region. Moreover, an additional partition is dedicated to store metadata that is relevant for the system as a whole. The reasons that motivate this solution are two. First, using only one table per each data type would result in large tables and long query times, while partitioning tables in different schemas avoids the problem. Second, since the data stored is often confidential, partitioning in separate schemas allows to define users' groups that can access only the intended schemas.

The persistence model that we designed is aimed at providing a consistent organization of the data into a database such that different sets of co-relevant information are accessible and integrable by following a common organizational schema. The PUBLIC schema is used to store system-wide meta-data that is shared among schemas and users. It includes two relations used for storing users' data and credentials, two relations used by PostGIS for storing metadata information about geometric attributes, one relation that lists all the data providers and is used to track data ownership, and two relations used to store general-purpose information (e.g. location, borders, population) about world countries and global cities (see Figure 4). Spatial schemas are used to

store data supplied by the data providers in the context of a specific geographical region. For example, Figure 4 shows the tables storing data related to the projects of New York Talk Exchange [11] and of WikiCity Rome [6]. The former include relations to store the communication flows (phone traffic or Internet traffic) between New York and other cities, the end points of these communications, the protocols used, and the references to the original files that contained the information. The latter comprise the same relations to save locations and data files and also include relations to store the cellular network stations' characteristics and their activity statistics, and the movements of busses over time.

The schemas in Figure 4 exemplify many of the concepts of the organizational model described above. For example, the LOCATIONS table is used by every schema to define the *points of interest*, that is the places that are pivotal to the each spatial schema; while defined locally for each schema, the LOCATIONS table always uses the same set of attributes that can represent a geographic area as big as a country or as precise as a pair of latitude and longitude coordinates. Another example is the DATA\_FILES table that allow to keep track of the files that have been processed and loaded into the database, and to trace back the rows that were originated by a particular file, which is very useful in case a file was corrupted or contained errors in the data.

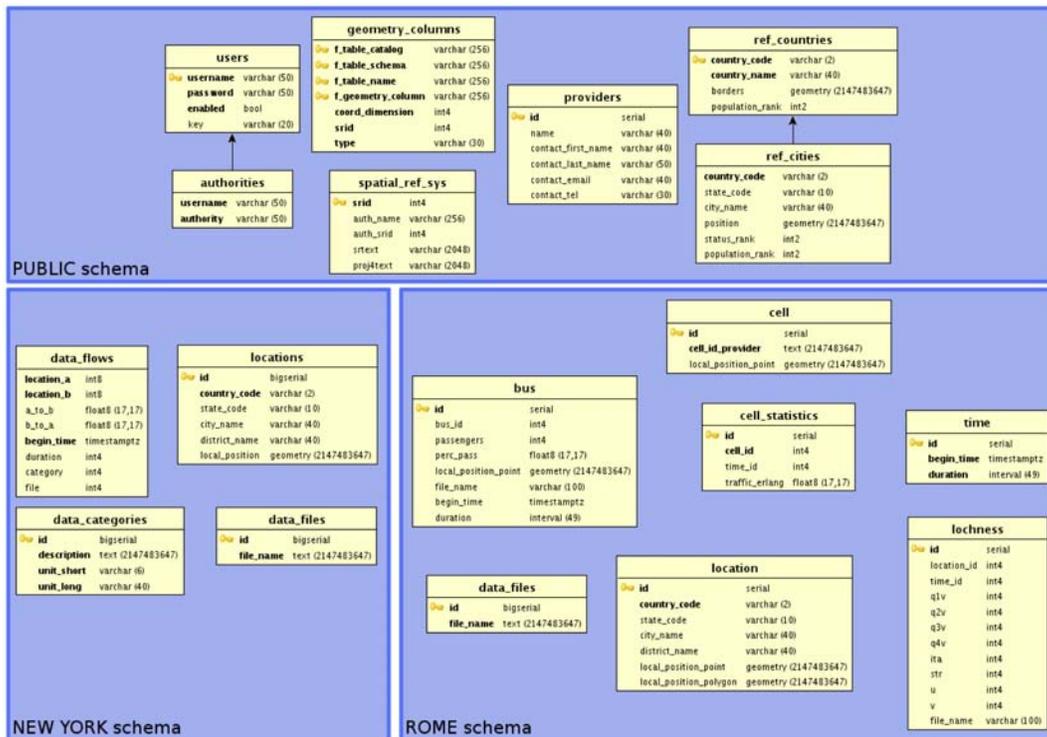


Figure 4: Illustration of the persistence model used in SocioScope. A public schema is used to store system metadata that is shared among schemas and users. The spatial schemas store diverse data like communication flows (phone traffic or Internet traffic) between cities, aggregate cellular activity logs (erlang traffic, number of calls, number of location updates), and the movements of busses over time.

## 5.2 Data Integration and Analysis Layers

The data integration and data analysis layers consist of a mechanism to project the raw data from the data storage layer to a spatial representation based on topological networks, and of a web service capable of processing spatio-temporal queries on the integrated data to return KML or CSV files that can be imported into Google Earth, Matlab, or Excel for further analyses or visual representations.

The integration mechanism is aimed at hiding the differences in resolutions and scales between the different data sets. While temporal integration is simple, requiring only to interpolate the data available in a given time range and to generate, if necessary, intermediate estimations, spatial integration is more complex and requires the definition of probes, or areas of interest, which are considered nodes of a topological network. Each raw data sets is then translated in a time series that can refer to either the nodes of the network or to the links that connect them. For example, cell phone calls can be associated to the nodes, and cell phone handovers can be associated to the links; at the same time, geotagged pictures are associated to the nodes, and the traces of a same photographer tacking picture in different areas of the city are associated to the links.

Different networks are allowed on the same geographical area, enabling to study the digital traces at different levels of resolutions. For example, one set could be the network of boroughs in New York City, and another the network of blocks. The web service makes use of this network representations to apply standard operations like a *sanity map*

of the cell phone network activity to compute the average number of phone calls generated in each probe at 10 pm of a Saturday, or a *flow map* of the movements of tourists between probes. The layer currently supports a wide range of analyses related to aggregate cellular activity logs, including spatio-temporal means and extraction of features from the sequences of data, and can integrate road networks and GeoRSS-based streams of news and events.

## 5.3 Data Visualization Layer

The data visualization layer provides two methods to visualize the data and the results of the operations: a web application based on Flash that allows to explore different data sets by navigating, panning and zooming on a map, and by adjusting the time range over which to follow the trends of different areas of interest of the city, and a web interface that allows to define spatio-temporal queries that can be issued to the web service in the analysis layer, and to export the results of the analyses into a KML file. Following the open philosophy that we adopted for the rest of the system, we use open technologies for interoperability whenever possible: for this reason, this layer makes use of the Google Maps API and the Yahoo! User Interface Library and leverages the interoperability with software like Google Earth, Matlab and Excel.

Figure 5 shows screenshots of the two web applications currently offered by SocioScope. On the left, users can select a viewport and time range over which to retrieve the data; if many data points are available in the selected viewport and

time range, the system automatically infer the best area and time resolution to use to aggregate the results. Users can also select the datasets and data analyses to apply, and the output template. The options available to users are limited by their roles: SocioScope first retrieves the viewport and time ranges over which users have read privileges, then based on their selection detects which datasets are available to them, and finally, based on the data type and other privileges, it detects which data processing and output templates users can use.

On the right, users can identify trends and anomalies that characterize a probe, compare similarities and differences between two or more probes, and study flows from one probe and adjacent areas of interest. A peculiarity of this application is the resemblance to the Google Finance interface: this design choice is motivated by the idea of providing analytical tools that are similar to those used to study the behaviour and performance of stocks in financial markets. Our approach likens a probe to a stock in a financial market: the activity of a probe is equivalent to the price of a stock, and therefore the same tools used to analyse and visualize stock markets can be used to analyse and visualize the activity of probes.

These applications address the needs highlighted in the use case presented in Section 3: with it, data providers, researchers, and planners can quickly define their own integration-analysis-visualization workflows. The ultimate goal is to create an infrastructure that allows users to publish open source scripts for different operations of integration, analysis, and visualization together with a documentation of their functionalities and usage. This would promote the collaborative creation of a public library of scripts that can be extended or reused to develop new workflows.

## 5.4 Security Model

The security model that we implemented is based on the concepts of *users* and *roles*. Roles define the type of account that a user holds, and the associated *privileges* represent the data she can access and the actions she can perform. The current SocioScope implementation supports three different roles:

**Administrators.** Administrators have full permissions over the system’s administrative functions. They are able to list users, modify their roles and permission, and analyze a log of their activities on the system. They are also able to control if the system’s components are working correctly, if the data streams are being received, and if the database and the web service are up and running. Finally they are able to restart the system if a module is not working properly and to lock down the system from external access.

**Data providers.** Data provider have limited control over the data streams that they are providing, and no control over other data streams. They can use the web interface to setup data streams, to verify that the data is being transferred correctly, and to pause/stop/restart the stream as they prefer. In the future, the initial setup phase where the data provider and the lab work together to set up the data exchange will be completely automatized.

**Normal users.** Normal users can only access the data, and have no administrative permissions. This role is automatically assigned during the registration process. Normal users are allowed to access the data and perform actions according to the roles that they hold. Different roles can allow

different data types to be accessed, and different operations and quotas (i.e. daily and weekly limits on the amount of data that can be downloaded) to avoid scraping and abuse. Roles have been extended to support the definition of spatial or temporal constraints: a user can be limited not only in the type of data and operations that she can access, but also in the ranges and scales of the viewport and timeport of a query. For example, she might be limited to access only data in the New York City area, or only data at the scale of the borough; moreover she might be limited to access data that is at least one month old, or data that is aggregated at least at the scale of the day.

The security model is divided in two processes: authentication and authorization. SocioScope supports two methods for authenticating users: a standard username and password form login enables users to authenticate against the web-based management application, while a public key can be used to authenticate against the web service presented above. When the identity of the principal has been established by the authentication process, the principal’s granted authorities are loaded and matched against the roles that are required to access the services in integration and analysis layers.

## 5.5 Management and Monitoring

Administrative functions for the management and monitoring of the system are currently being developed. The management services available to system administrators include functions to list all registered users or only those currently logged in, to create, delete, enable, and disable user accounts, to update their properties and credentials, and to define roles and privileges of users and users’ groups. The management services for data providers instead include functions to create, update, and delete data streams which are defined by a name, a informal description of the content of the stream, and a formal description of the data format. Data providers can also temporarily stop and restart the streams and define which users and users’ groups have read access to them.

Monitoring functions for system administrators allow to list when each user logged in and out, which queries she performed, and their time of execution. They also allow to view the current status of the system: the uptime, the active processes, the available space on disk, and the size and number of rows of each database table. The monitoring functions for data providers allow to view the status of the data streams, the bandwidth used, the time of the last data provision, and a list of users who queried the data. In the future, we will implement functionalities to define alerts on the data streams using a Event Processing Language (EPL) that can be used by the providers to be notified of important events related to the data itself (e.g. a sudden increase in the cellular network activity) or how it is access through the system (e.g. a user attempting to access the data over his/her daily limit).

## 6. DISCUSSION

SocioScope modular system architecture allows the iterative development of new domain objects and service methods that can be added to the system in a plug-and-play fashion using the Java Spring Framework and its inversion of control container. This approach eases the integration of new features and allows to focus the development efforts were

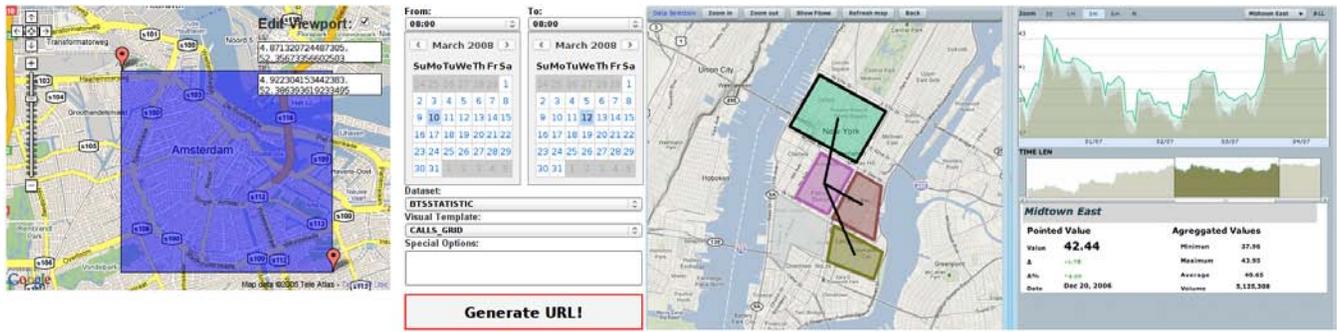


Figure 5: Web interfaces of the data visualization layer: a web interface based on Flash that allows to explore different data sets by navigating, panning and zooming on a map, and by adjusting the time range over which to follow the trends of different areas of interest of the city, and a web application to define spatio-temporal queries that can be issued to the web service in the analysis layer and to export the results of the analyses into a KML file.

they are most needed, implementing small components that can be easily reused. By leveraging the benefits of aspect-oriented programming for security and logging functions, the cross-cutting concerns are separated from the logic of the system and encapsulated into advices that can be easily re-configured as required.

The data persistence layer allows to successfully aggregate and provide access to the urban data managed in SocioScope using a multi-schema PostgreSQL database extended with PostGIS. The persistence model provides a consistent organization of the data that simplifies the access to the datasets and their analyses. By partitioning the datasets into different schemas, the model allows to handle the ever-growing amount of data gathered by SocioScope and to control the access to the data itself and extend the model to accommodate new information when required.

The security model is based on roles and associated privileges. These include the traditional concepts of ownership and readability and add discrimination along the following spatio-temporal dimensions: type of data, geographical area and resolution, time range and frequency, geospatial integration and processing operations, and representation models. Moreover the monitoring system stores details users' operations, which can be used to control their activity and limit the access to the data.

SocioScope is today used to gather, store, and analyse urban data related to a wide range of pervasive systems in New York, Washington D.C., London, Rome, Amsterdam, Copenhagen, and Graz. Currently the system is used by a selected group of researchers and data providers to test its functioning and understand what other features should be added to the system before opening it up to the public at large.

## 7. CONCLUSIONS

SocioScope is a urban information system specifically developed to support the needs of researchers, practitioners, local authorities, and enterprises interested in studying through digital traces. SocioScope streamlines the collection of spatio-temporal urban data, controls access to the confidential data based on roles and associated privileges, and provides tools to query and visualize the information. The system has already been used successfully in the analysis of the impact of

the Waterfalls public exhibition in New York City [16, 17] and to develop the Obama | One People visualizations [22]. We contend that SocioScope can facilitate and foster the research activity in digital sociology by offering tools that can be used to connect data providers with researchers and practitioners and by defining a common set of methodologies to collect, store, analyze, and visualize urban data [29].

Future work will focus both on the front-end and the back-end systems. For the former, we intend to extend the web application to support complex queries that require integration and processing of different data types and to offer a set of shared visualizations that allow to explore queries results directly from the web browser. For the latter, we plan to implement a transaction registry and a data exchange protocol to overcome the overhead of XML-based data files and to complement the current workflow, which best fits the retrieval and analysis of historical and soft real-time data, with technologies for Event Stream Processing (ESP) and Complex Event Processing (CEP) that could support the real-time monitoring of urban features (e.g. to detect traffic jams, car accidents, and sudden emergencies.)

## 8. ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their comments and feedback that helped to improve the manuscript. The authors would also like to thank Joost Bonsen, who wisely coined the term *SocioScope*.

## 9. REFERENCES

- [1] T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich. Mobiscopes for human spaces. *IEEE Pervasive Computing Magazine*, 2007.
- [2] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA, 2004. ACM.
- [3] M. Arikawa, S. Konomi, and K. Ohnishi. NAVITIME: Supporting pedestrian navigation in the real world. *IEEE Pervasive Computing Magazine*, 2007.

- [4] M. Balazinska, A. Deshpande, M.J. Franklin, P.B. Gibbons, J. Gray, S. Nath, M. Hansen, M. Liebhold, A. Szalay, and V. Tao. Data management in the worldwide sensor web. *IEEE Pervasive Computing Magazine*, 2007.
- [5] F. Calabrese, K. Kloeckl, and C. Ratti. *Handbook of Research on Urban Informatics*, chapter WikiCity: Real-Time Location-Sensitive Tools for the City. IGI Global, 2008.
- [6] F. Calabrese, K. Kloeckl, and C. Ratti. *Handbook of Research on Urban Informatics*, chapter WikiCity: real-time location-sensitive tools for the city. IGI Global, 2008.
- [7] L. Cherbakov, A.J.F. Bravery, and A. Pandya. Soa meets situational applications, <http://www.ibm.com/developerworks/webservices/library/ws-soa-situational1/>.
- [8] J. Davis, P. Lin, A. Borning, B. Friedman, P.H. Kahn, and P.A. Waddell. Simulations for urban planning: Designing for human values. *IEEE Computer Magazine*, 39(9):66, 2006.
- [9] S.B. Eisenman, N.D. Lane, E. Miluzzo, R.A. Peterson, G.S. Ahn, and A.T. Campbell. Metrosense project - people-centric sensing at scale. *ACM Conference on Embedded Networked Sensor Systems*, 2006.
- [10] ESRI. ArcGIS - engineered for interoperability. *Technical Report*, 2006.
- [11] New York Talk Exchange. <http://senseable.mit.edu/nyte/>.
- [12] World's Eyes. <http://senseable.mit.edu/worldseyes/>.
- [13] F. Girardin, F. Calabrese, F. DalFiore, C. Ratti, and J. Blat. Uncovering the presence and movements of people from user-generated content. *IEEE Pervasive Computing Magazine*, 2008.
- [14] F. Girardin, A. Vaccari, A. Gerber, and C. Ratti. Quantifying urban attractiveness from the distribution and density of digital footprints. *Journal of Spatial Data Infrastructure Research*, 4:175–200, 2009.
- [15] F. Girardin, A. Vaccari, A. Gerber, and C. Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. *11th International Conference on Computers in Urban Planning and Urban Management*, 2009.
- [16] F. Girardin, A. Vaccari, A. Gerber, and C. Ratti. Quantifying urban attractiveness from the distribution and density of digital footprints. *Journal of Spatial Data Infrastructure Research*, forthcoming.
- [17] F. Girardin, A. Vaccari, A. Gerber, and C. Ratti. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. *11th International Conference on Computers in Urban Planning and Urban Management*, forthcoming.
- [18] M.C. González, C.A. Hidalgo, and A.L. Barabási. Understanding individual human mobility patterns. *Nature Magazine*, 2008.
- [19] J. Liu and F. Zhao. Towards semantic services for sensor-rich information systems. *2nd IEEE/CreateNet International Workshop on Broadband Advanced Sensor Networks*, 2005.
- [20] R.N. Murty, A. Gosain, M. Tierney, A. Brody, A. Fahad, J. Bers, and M. Welsh. Citysense - an urban-scale wireless networking testbed. *IEEE Conference on Technologies for Homeland Security*, 2007.
- [21] A. Pawling, T. Schoenharl, P. Yan, and G. Madey. Wiper: An emergency response system. *6th International Conference on Information Systems for Crisis Response and Management*, 2008.
- [22] Obama | One People. <http://senseable.mit.edu/obama/>.
- [23] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, New York, NY, USA, 2007. ACM.
- [24] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: analyzing cities using the space-time structure of the mobile phone network. *Environment and Planning B*, 2008.
- [25] Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census - explorations in urban data collection. *IEEE Pervasive Computing Magazine*, 2007.
- [26] A. Santanche, S. Nath, J. Liu, B. Priyantha, and F. Zhao. Senseweb - browsing the physical world in real time. *Microsoft Research*, 2006.
- [27] M. Schouppe. Towards a single information space for the environment in europe. *Expert Consultation Workshop: ICTs and Climate Change*, 2008.
- [28] C. Shirky. Situated software, [http://www.shirky.com/writings/situated\\_software.html](http://www.shirky.com/writings/situated_software.html).
- [29] A. Vaccari, L. Liu, A. Biderman, C. Ratti, F. Pereira, J. Oliveirinha, and A. Gerber. A holistic framework for the study of urban traces and the profiling of urban processes and dynamics. *12th International IEEE Conference on Intelligent Transportation Systems*, 2009.
- [30] Shijun Yu, Marie-Aude Aufaure, Nadine Cullot, and Stefano Spaccapietra. Location-based spatial modelling using ontology. *6th AGILE Conference*, 2003.