IQT QUARTERLY

VOL. 5 NO. 2 FALL 2013

MORE THAN MAPS

The Analytics Behind Geospatial Technologies

IQT IN-Q-TEL

IQT Quarterly is a publication of In-Q-Tel, Inc., the strategic investment firm that serves as a bridge between the U.S. Intelligence Community and venture-backed startup firms on the leading edge of technological innovation. *IQT Quarterly* advances the situational awareness component of the IQT mission, serving as a platform to debut, discuss, and debate issues of innovation in the areas of overlap between commercial potential and U.S. Intelligence Community needs. For comments or questions regarding IQT or this document, please visit www.iqt.org, write to iqtquarterly@iqt.org, or call 703-248-3000. The views expressed are those of the authors in their personal capacities and do not necessarily reflect the opinion of IQT, their employers, or the Government.

EDITORIAL

IQT Quarterly, published by In-Q-Tel, Inc.

Editor-in-Chief: Emma Resnick

Theme Editor: Todd Stavish

Contributing Editor: Brittany Carambio

Managing Editor: Lisa L. Bader

Design by Lomangino Studio LLC

Printed in the United States of America

IQT QUARTERLY

Identify. Adapt. Deliver.™

TABLE OF CONTENTS

GEOSPATIAL SOFTWARE IS EATING THE WORLD

By Todd Stavish

Marc Andreessen, one of the most influential figures in Silicon Valley, famously stated in a 2011 essay, "Software is eating the world." Andreessen's belief is that we are in the middle of a broad technological and economic shift in which software companies are poised to take over large swathes of the economy. Essentially, no matter what industry you are in, software companies of all shapes are emerging to offer solutions to every business challenge imaginable. While Andreessen's assertion was primarily referring to Internet companies, the trend is true regarding geospatial software as well. A new generation of mapping startups is radically transforming both geospatial processing capabilities and the economics of the industry.

These new startups are taking advantage of three dramatic and fundamental changes in the geospatial industry. The first change is the rise of open alternatives: open source, open standards, and open data. The second major change is the ability to perform analytics at scale by utilizing low-cost cloud computing and machine learning to revolutionize geospatial processing. The final change is the democratization of user interfaces to enable non-technical users to perform basic geospatial analysis on their own, without the aid of highly trained and often expensive geospatial professionals.

The pattern of open source alternatives commoditizing entrenched proprietary software vendors has already been established in operating systems and databases. A successful open source geographic information system (GIS) needs to contain a relational database with geospatial extensions so that it can support geographic objects, an application server to publish data from major spatial data sources, tiling to accelerate the delivery of data, and a user interface.

Open source software companies offer both technical and economic benefits to their customers. Open source business models are oriented towards long-term and low-cost subscriptions instead of large licensing events that the proprietary vendors prefer. Open source companies are often able to offer equivalent functionality at a lower cost than proprietary vendors because they are able to coordinate software development teams worldwide with minimal or no direct expenses. From a technical perspective, open source architectures are extensible and support Intelligence Community customizations. For example, security enhancements that are required to work with classified data can be built into the architecture.

Open standards, on the other hand, ensure interoperability. The Open Geospatial Consortium (OGC) is the leading international consortium that defines geospatial interface standards. These open standards empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications.

Open standards empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications.

The expansion of open data is perhaps one of the most exciting developments that has recently occurred for GIS analysts. The OpenStreetMap (OSM) project provides a crowd-sourced, free, editable map of the world. OSM was inspired by the success of Wikipedia and has grown to over one million users plotting location data worldwide. Proprietary map data tends to be sparse in developing nations because of the lack of commercial appeal to substantiate the cost. OSM, however, has very rich data in the developing world because of contributions from non-government organizations, map data donations, student participation, and hobbyist geographers.

The second major shift affecting the geospatial industry is the ability to inexpensively perform analytics at scale. There are a handful of geospatial startups that are using cloud computing to apply brute force computation and machine learning techniques to location data. The ability to process large amounts of location intelligence allows advertisers, for example, to understand consumer behavior and target mobile users with specialized ads. In addition, database companies are building geospatial extensions to perform geo-processing at scale on an organization's own datasets.

Democratization, the third major shift to the geospatial industry, has been driven by non-technical users' exposure to web mapping products like Google Maps. Web mapping has created an appetite for non-GIS analysts to perform their own basic map creation and analysis. There has been an emergence of startups with cloud-based data publishing tools that enable non-technical users to perform basic GIS analysis and generate their own maps with their own data. Technical users can also integrate maps into their mobile apps.

Historically within the Intelligence Community, geospatial data was considered a separate type of intelligence. This is no longer true. Geospatial intelligence is a part of all intelligence-gathering disciplines. For example, images now embed geo-tags, and collections have locations attached. Signals Intelligence requires IP address geo-location and triangulation of mobile devices. Open Source Intelligence analyzes social media with location data and utilizes crowd-sourced mapping. Location reveals what the data means, and in some forms of analysis, location may be the largest indication of intent.

__Todd Stavish__ is a member of the Technical Staff within IQT's Advanced Analytics and Infrastructure practice. Previously, as a Chief Systems Engineer with InfiniteGraph, Stavish was responsible for customer acquisition. Prior to that, with Socrata, he was responsible for technical account management as well as conducting business development and field marketing for federal programs. Stavish received a bachelor's degree in computer science from Saint Bonaventure University.

A LOOK INSIDE: MORE THAN MAPS

In recent years, consumers have seen a major uptick in the availability and usability of geolocation services and apps. Tracking a storm in proximity to a user's house, checking for well-reviewed establishments while travelling in an unknown area, and being able to monitor your child's location from afar are all well-worn consumer uses made possible by the availability of geodata. This edition of the *IQT Quarterly* discusses innovations in analytics, mobile, open data, and data management technologies that enable such services.

Norman Barker opens the issue with a discussion of Cloudant. Cloudant's technology is based on a clustered implementation of Apache CouchDB, a NoSQL database. Barker discusses the features of a multi-cluster Cloudant deployment and how this technology can store, index, and query massive quantities of spatial data.

Next, George Percivall of the Open Geospatial Consortium (OGC) discusses how this international non-profit organization develops consensus standards among its members that guide developers in the geospatial field. Percivall outlines some of these OGC standards as well as the OpenPOI (points of interest) Database Project, which contains more than 8 million names and point locations and links to other commercial, government, and academic POI databases.

In an excerpt from the journal *Cartography and Geographic Information Science,* researchers from George Mason University analyze how virtual communities that transcend national boundaries but coalesce around a state-specific issue are formed through online social interactions. Using geographical information from social media feeds, these communities can be identified, mapped, and analyzed to better understand globally-distributed communities.

Chris Herwig follows with an explanation of how MapBox has used cloud computing to advance the process of delivering satellite data to users, making live imagery available for immediate publication. In doing so, MapBox enables satellite data to be turned into actionable intelligence.

Graham Neray's article discusses how developments in data management technologies affect the usability of geospatial data. He explains why traditional relational databases are not equipped to maximize the potential of location-based data, and discusses how alternative open-source NoSQL databases, such as MongoDB, can provide rich, interactive analysis.

Evaluating the accuracy and completeness of Volunteered Geographic Information (VGI) — geospatial data that is contributed by average citizens — is examined in an excerpt from the *ISPRS International Journal of Geo-Information.* The article considers methods for evaluating the data and what the results mean for the ongoing use of VGI.

Finally, Kevin Merritt closes the issue with an article explaining how public data — specifically geospatial data — has been unlocked by advances in cloud infrastructure and mobile computing, allowing it to be accessed by everyday users. While previously only specially-trained GIS analysts could have made use of such data, with this evolution non-experts can create detailed maps and apply public information to civic issues, as in the use cases that Merritt describes.

As the broad consumer offerings of geolocation services expand, IQT will persist in identifying technologies that enable optimal use of geospatial data, while recognizing that maximizing geospatial analytics will require multi-faceted approaches. Q

A CHANGE OF COURSE: CLOUD SERVICES AND LOCATION-AWARE APPLICATIONS

By Norman Barker

Mission-critical applications are increasingly relying on geospatial data points. Applications such as mapping and guidance software and counterterrorism analytics need to retrieve data from a central storage repository to deliver a location-based response to the user. This is a requirement that federal agencies, along with other types of organizations, are grappling with now.

Companies such as Cloudant are developing geospatial database services using a variety of underlying database and hosting technologies to facilitate this demand for cloud database services that can underpin location-aware applications. Cloudant's architecture is based on BigCouch, which is a highly available, fault-tolerant, clustered implementation of Apache CouchDB, a NoSQL database that stores data as JSON documents (think of "documents" as "database records").

While Cloudant appears to the end user as one Apache CouchDB instance, it is in fact one or more BigCouch nodes in an elastic cluster, acting in concert to store and retrieve documents, index and serve views, and serve applications. Clusters behave according to concepts outlined in Amazon's Dynamo paper. Each BigCouch node can accept requests, data is placed on partitions based on a consistent hashing algorithm, and quorum protocols are applied for read and write operations.

A multi-cluster Cloudant deployment (Figure 1) has the following characteristics:

1. Partitioned to span multiple data centers, all of which accept reads and writes. This allows Cloudant's Domain Name System to route applications to the closest copy of data and fail over traffic in disaster scenarios.
2. HTTPS provides the security of industry-standard SSL encryption.
3. The Global Load Balancing Tier handles routing within the data center, but can also securely route traffic to other data centers in the event of a load balancing outage in one data center.
4. Each database node is its own completely independent machine, creating a highly available shared-nothing architecture.
5. When a database node accepts a request, it coordinates with the other nodes to save data or respond to a query.

Additionally, Cloudant provides callable functions such as full-text indexing and search. These functions act in concert to store and retrieve JSON, index and serve specific views of the stored data, and serve CouchDB-compatible applications. Solutions are then provided to customers — usually developers — as a database-as-a-service (DBaaS), either as a public cloud solution or as a hosted private cloud implementation in which data security, compliance, or other customer requirements make shared hosting impractical or unacceptable.

In all cases, DBaaS solutions should provide a scalable database that is distributed and fault tolerant, so that

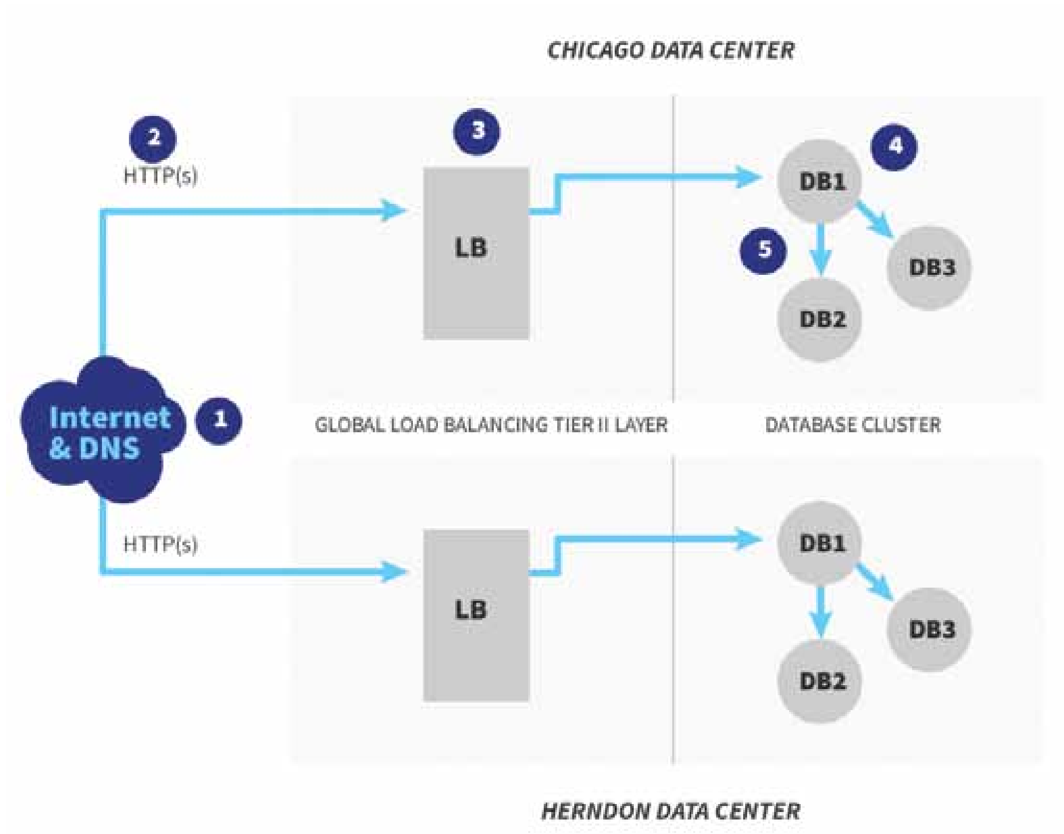


Figure 1 | Multi-cluster Cloudant deployment

if it fails, it gracefully recovers with no data loss to the customer or the end user.

The Heart of the Application

When location-based services are added to the dynamic of the DBaaS, these solutions stop being simply repositories in the Cloud for datasets and become the heart of many applications. They provide not only the data that they deliver to the user, but to some extent the underlying logic (the location-aware context) that makes the application function and useful to the end user.

Location-based services are a method of delivering relevant data to a user based on his or her current position. Part of a service guarantee is delivering the requested data with minimal latency. A geospatial cloud that is providing services beyond a localized area typically spans multiple data centers. Routing users to the nearest data center is essential for providing a usable service. This can be achieved through access control lists or by using the IP address of the user to locate the request.

Intelligence analysts use geography as both a source and the content of the data they evaluate. Geospatial data is a key input in counterterrorism and weapons proliferation analysis, and many more intelligence disciplines. Using bounding polygons to monitor communications and their proximity to key locations around the world is critical to countering terrorist activity. Strategic interdictions of weapons proliferation rely on geospatial monitoring of materials-transfer around the world. In addition, the ability to track the spread of pandemic disease, the global transfer of narcotics, and the travel of high-value targets would be limited without geospatial data.

However, the integration of geospatial analytic capabilities into standard intelligence practices can be challenging. Cloudant's geospatial analytic capabilities eliminate the challenges of how to store, index, and query massive quantities of spatial and temporal data. Cloudant includes indexing capabilities based on standards like GeoJSON to support the storage of different coordinate systems (WGS84, UTM, NGR, etc.) and automates geo-hashing, proximity, bounding (box, polygon, radius, ellipse), intersection, and temporal queries. Building rich new geospatial applications on top of Cloudant's service allows analysts to stay ahead of adaptive threats and disruptive events.

Cloud Databases for Distributed Applications

Multi-version concurrency control (MVCC) and over-the-wire database replication are methods to get an eventually consistent system, while making sure critical applications are always available. Eventual consistency with online networks is typically measured in seconds, so that users receive relevant information quickly. An advantage of using MVCC and replication is that a system or device can be disconnected and reconnected to a network at any time and then synchronized with the rest of the database cluster.

This scalability feature for servers is also relevant for geospatial location-based services and data collection. An ad hoc network of devices can be formed and data replicated between the members. A user with a device can receive a filtered subset of the geospatial data from the larger system, go offline to work locally, then return later and synchronize the edited data.

Cloud data layers leverage a model of eventual consistency using an over-the-wire data replication protocol. This replication protocol can be used to filter and transfer data from servers to the edge and back again. It is this feature that can be used for location-based services and data collection to provide applications that are not supported by alternative geospatial services.

The Latest Standards

The Open Geospatial Consortium (OGC) is an international consortium developing interoperability standards for geospatial data. The organization has two specifications that are of interest: Web Feature Service (WFS) and the proposed GeoPackage specification. The latter defines a geospatial database format for importing data onto devices and other systems, and the WFS specification defines methods for querying and updating data between systems and devices. Traditionally, implementations of WFS include 'LockFeature' to ensure the serialization of the Geography Markup Language (GML) feature. However, extending the WFS response to indicate resource conflicts rather than just the identifiers for newly created or updated resources would extend the applicability of WFS to include the MVCC model and, hence, allow implementations to pick the goals of availability and partition tolerance that can be implemented at a lower total cost than a system that has consistency guarantees.

The proposed GeoPackage specification does not include an over-the-wire protocol to transfer data to mobile devices or between systems, but does target the disconnected or limited connected network environments. With the open replication protocol, it is possible to implement a model of eventual consistency using WFS and GeoPackage, if the updated results response includes indicating resource conflicts.

Eventual consistency has been proven to work in other sectors, and the application of this approach to location-based services and data collection is a model that fits geospatial in a domain that has typically embraced transactional updates. Frequently disconnected and ad hoc networks fit eventual consistency. It is exciting that, with minor specification changes, a MVCC model can be added to existing WFS clients and servers that will enable vendors to be interoperable and allow users to pick services that fit their requirements.

Current standards have laid strong groundwork for the establishment of working protocols, developer best practice, and compatibility of location-based services going forward, but there remains scope for improvement and expansion of current ratified standards. Doing so will address both the evolving nature of end user devices, as well as the demands of applications wishing to mine or harvest geospatial data. For now, the OGC continues to evaluate specifications, and listen to the input of members of the organization as to how to augment and ratify these as workable industry standards in the medium-term. With minor specification changes, vendors will be interoperable and users can pick services that fit their requirements.

Norman Barker *is the Director of Geospatial at Cloudant, an IQT portfolio company. He is a specialist in developing geospatial data discovery and dissemination products. Barker has been developing geospatial programs for more than 10 years and leads the development of distributed geospatial indexes for Cloudant. His primary interest is in how to use unstructured geospatial data. He is also an open source developer with contributions to MapServer, GDAL, and PostGIS.*

EMERGING STANDARDS: UNIVERSAL FILE FORMATS, POINTS OF INTEREST, AND THE SENSOR WEB

By George Percivall

The Open Geospatial Consortium (OGC) is an international industry consortium of 481 companies, government agencies, and universities participating in a consensus process to develop publicly available interface standards. OGC standards support interoperable solutions that "geo-enable" the web, wireless, and location-based services and mainstream IT. The standards empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications. Standards are developed in a unique consensus process supported by the OGC's industry, government, and academic members to enable geoprocessing technologies to interoperate, or "plug and play".

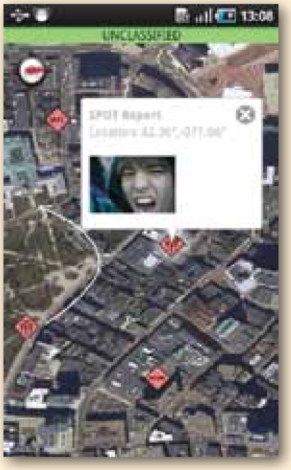
GeoPackage

GeoPackage is a universal file format for geodata. GeoPackage is open, standards-based, and application- and platform-independent. It is the modern alternative to formats like GeoTIFF, SDTS, and shapefile. At its core, GeoPackage is simply an SQLite database schema. Users need only know SQL to use GeoPackage on any desktop or mobile operating system on the market.

GeoPackage was carefully designed to facilitate widespread adoption and use of a single, simple file format by both commercial and open-source software applications — on enterprise production platforms as well as mobile handheld devices.

GREAT FOR MOBILE

Mobile device users who require geospatial information, including active maps in disconnected or limited network connectivity environments, can benefit in several ways. GeoPackage efficiently supports download of data, cached for offline use and eventual upload of any updates to the data made while offline. A common challenge for developers of mobile apps is the limited and uncertain storage available on devices and the likelihood that each location app will require its own potentially proprietary geospatial data store. These separate application-specific data stores may contain the same geospatial data, wasting the limited storage available and requiring custom coding for data translation, replication, and synchronization to enable different apps to share the same worldview. In addition, many existing geospatial data stores are platform-specific, which means that users with different platforms must translate data to share it. Only GeoPackage meets the requirements for a lightweight, standards-based, and vendor-neutral geodata format.



CROSS-PLATFORM SHARING

GeoPackage is built on SQLite and can therefore be used easily by a broad spectrum of software developers in a consistent way on every major mobile and desktop platform in the market. SQL-savvy programmers can easily add GeoPackage support to their apps.

GeoPackage was designed in an open manner by the geospatial developer community. The format is nearing adoption in the OGC as a standard and is open for public comment. Many companies have already implemented GeoPackage based on the draft specification.

The Sensor Web Standard

In today's world, most sensors have proprietary software interfaces defined by their manufacturers. New application programming interfaces are requested and developed for each type of sensor. This situation requires significant investment on the part of developers, with each new sensor or project involving multiple systems. It also requires investment on the part of the providers of sensors, gateways, and services where observations are used. Standardized interfaces for sensors in the Internet of Things (IoT) will permit the proliferation of new high-value services with lower development overhead and wider reach. This will also lower the cost for sensor and gateway providers.

The past decades have witnessed dramatic increases in the precision of location information (see Figure 1). Global Positioning System (GPS) was an incredible innovation in the 1990s. Prior to GPS, accurate location was determined by an intensive human survey process. Survey geo-location and GPS enabled many applications associated with region-centric and feature-centric information. Geographic Information Systems (GIS) applications enabled greater awareness and analytical capability using a feature-based modeling of environments. New consumer capabilities were established (e.g., car navigation systems with accuracy to several meters).

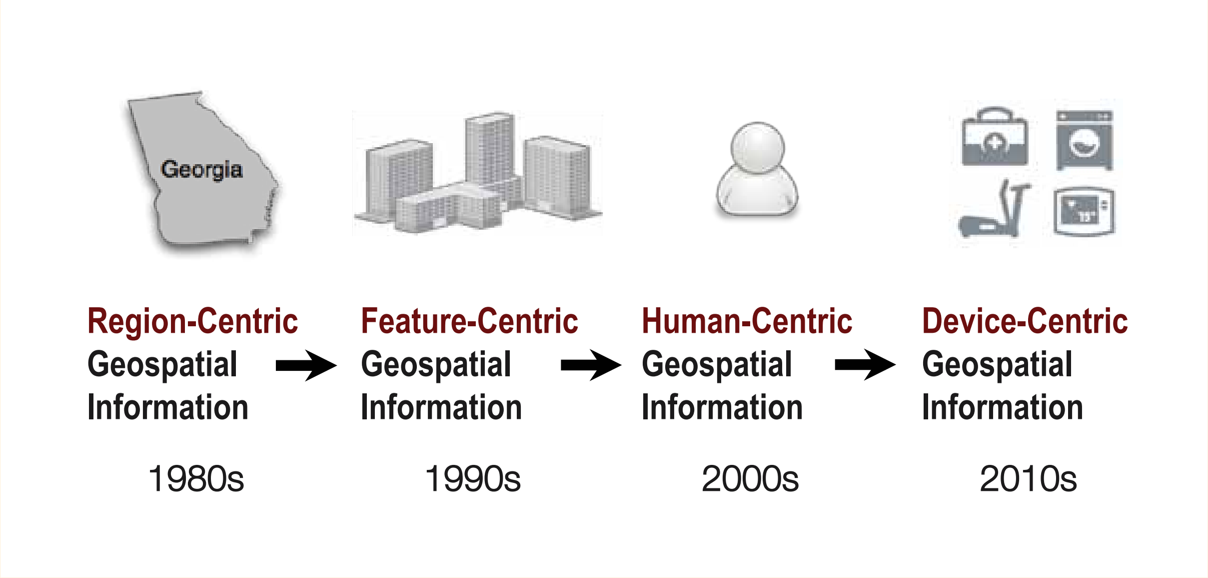
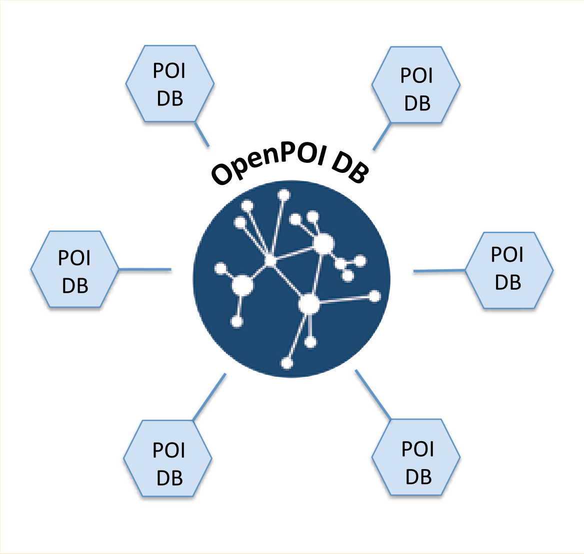


Figure 1 | Progression of geospatial information



Shortly after the turn of the millennium, smartphones combined GPS with other location positioning methods to increase the usability of location at a human-centric level. The most notable addition to the positioning method was the use of Wi-Fi signals fused with GPS to provide a quicker fix and higher accuracy through the use of multiple sensors. In order to achieve the anticipated benefits of IoT, location accuracy will again need to undergo a leap of innovation in positioning technology. Not only will the precision of the location need to be higher, but orientation will become more critical (which way am I looking?), as will the need to provide a widespread and common approach to identifying indoor location.

In preparation for sensor web, OGC created a new standard working group (SWG). The goal of the Sensor Web SWG is to develop one or more standards based on existing IoT protocols while also leveraging the existing and proven OGC Sensor Web Enablement (SWE) family of standards.

SWE standards are the only ones that focus on the meaning of sensor measurements and on making the sensor observations useful to end user applications. SWE standards allow users to assess the fitness for use of observations and allow accurate processing to create derived information suitable to the user's needs.

In much the same way that HTML and HTTP standards enabled the exchange of any type of information on the web, the OGC SWE standards enable the discovery of sensors and corresponding observations, exchange, and processing of sensor observations, as well as the tasking of sensors and sensor systems.

The OGC's OpenPOI Database Project

OpenPOI is a very large, freely available database containing points of interest (POIs) and links to other commercial, governmental, and academic POI databases. Not counting the POIs in the linked databases, the OGC's database contains the names and point locations for over 8 million businesses and civic places.

The OGC is looking for developers and POI publishers to participate in creating the largest system of linked geospatial data on the web. OGC's OpenPOI database provides a reference implementation of a POI encoding standard moving toward adoption in OGC.

The POIs currently in the OpenPOI database have been collected from many sources, including GeoNames (which includes U.S. Geological Survey, U.S. National Geospatial-Intelligence Agency, CIA World Factbook, hotels.com, and over a hundred sources), Factual, OpenStreetMap, and the Harvard China Biographical Database. More POI datasets are coming online every month.

This unique new service offers application developers a single point of entry to all of these data sources. If enough people — data developers, software developers, and their users — find the OpenPOI database useful, this service could organize the world's information geographically. It is intended to be a widely used public resource governed by policies similar to other open source data initiatives, such as Wikipedia.

WHY CREATE YET ANOTHER POI DATABASE?

Most POI databases are meant to provide rich, detailed information about the places in their database. In contrast, the OpenPOI database is meant to provide a comprehensive registry of links to other POI databases that contain the actual information about the places. Any organization can add to it and anyone can use it.

The database can be queried by bounding box or by name via simple requests that implement the OGC Web Feature Service (WFS) Interface Standard. Each POI also has a concise, permanent, and unique RESTful URL.

If enough people — data developers, software developers, and their users — find the OpenPOI database useful, this service could organize the world's information geographically.

All queries return data in either XML or JSON format. Considering how ubiquitous the need for POI information is, it is surprising that international standardization efforts have been few. In many ways, one could consider POIs a fundamental requirement of any spatial data infrastructure. POIs are also important in the commercial sector in personal navigation and social networks. For example, several social media outlets have made location such an integral part of their data model that almost every activity a user engages in can be tagged with location, weaving places of interest seamlessly into the fabric of their social platform.

The OGC's vision is the realization of the full societal, economic, and scientific benefits of integrating electronic location resources into commercial and institutional processes worldwide. The OGC's mission is to make this vision a reality based on developing publicly available geospatial standards. OpenPOI will advance the POI standard and implementations.

The OGC's vision and mission, experience with legal and organizational issues involved in openness, and rich connections to the world's providers and users of geospatial data and technology put it in a unique position to launch and host a resource such as the OpenPOI database.

Attracting collaboration partners is proving to be an achievable goal for the not-for-profit OGC. Wikipedia is an exemplar of previous success in this model. If Wikipedia were not a neutral provider, and had not been seeded with a large amount of useful data free of onerous licensing terms, it probably would never have succeeded as it has. Because the OGC is recognized and respected in the information technology world and organized to perform outreach, it is positioned to build the OpenPOI database brand in the web and mobile market space as few other initiatives could do. Q

__George Percivall__ is an accomplished leader in geospatial information systems and standards. As Chief Engineer of the Open Geospatial Consortium (OGC), he is responsible for the OGC Interoperability Program and the OGC Compliance Program. Prior to joining OGC, Percivall held senior positions at NASA and the GST's Geospatial Interoperability Group. Previously, he led developments in Intelligent Transportation Systems with the U.S. Automated Highway Consortium and General Motors Systems Engineering. He holds a BS in Engineering Physics and an MS in Electrical Engineering from the University of Illinois at Urbana.

DEMARCATING NEW NATIONAL BOUNDARIES:

Mapping Virtual Communities Through Social Media Content

By Anthony Stefanidis, Amy Cotnoir, Arie Croitoru, Andrew Crooks, Jacek Radzikowski, and Matthew Rice

By substituting physical with virtual interaction, social media has introduced a novel avenue for community building, transcending established administrative boundaries to diffuse ideas and information across space. As the topics that unite these communities vary widely, covering the broad spectrum from trivial to substantial issues, individuals may participate in multiple virtual communities, exploring various dimensions of their identities.

Of particular interest for this article are the global virtual communities formed around issues of specific national interest, just like the global activist communities that emerged as a response to the Arab Spring events. During such events, a state's core community (namely, its own citizens within its boundaries) is expanded through the addition of satellite communities from abroad that are involved with this state's issues and become connected to its citizens. We refer to these connected virtual communities as the polycentric virtual equivalent of that state. These virtual polycentric communities can play an important role in shaping global opinion about, and response to, events affecting a specific state, ranging from natural disasters to political crises. Identifying, mapping, and analyzing these virtual communities is a novel challenge for our community. We present these communities relative to established conventions of statehood, address the harvesting of relevant geographical information from social media feeds, and discuss the challenge of visualizing such information.

Connected Virtual Communities as Novel Hybrid States

Social media enables the general public to communicate with their peers, sharing information with them instantly and constantly in an effortless and intuitive way. In doing so, social media enabled the formation of virtual communities that shared many of the characteristics of communities formed in the physical space. As these virtual communities are formed across state boundaries, their projection onto the physical space also spans many states. When the topic that brings together such a virtual community is a national issue of one state (e.g., a natural disaster affecting it), this is equivalent to expanding that reference state's citizenry to include the involved overseas communities.

Accordingly, a state's core community (namely, its own citizens within its boundaries) is expanded through the addition of satellite communities from abroad that are involved with this state's issues and become connected to its citizens. This process generates a polycentric state structure, where a state comprises not only its own citizens, within its established borders and under its jurisdiction, but also satellite communities from abroad, sharing an interest or active involvement with this state's affairs (Figure 1).

The traditional state and its borders are substituted by its virtual equivalent, in the form of connected communities of stakeholders, and the aggregate of their spatial footprints. The state itself clearly remains as the central unit in this new union, and satellite communities may vary in their membership size; the green community in Figure 1 has more members than the red community. The links among these communities may stem from various aspects of life, ranging from cultural and economical to political, thus creating multiple levels of such connections.

A Polycentric Syria

As a test case for this article we are using Syria, with its strong geopolitical importance given the prolonged civil war that started as part of the Arab Spring. Using a system prototype that we have developed to harvest such information, we collected Twitter feeds through keyword queries to Twitter's streaming API.

In order to demonstrate how the structure of a virtual polycentric Syria can be gleaned through the analysis of social media content, tweets that included mentions to Syria or its hashtag equivalent #Syria were collected over a period of one week (10 July through 17 July 2012). As per standard Twitter terms of use, these tweets are a random 1 percent worldwide sample of Twitter traffic on this topic. We harvested a total of over 700,000 tweets about Syria (for a rate of approximately 4,000 tweets per hour). From among these tweets, 325,612 (approximately 45 percent of the total) had some sort of geolocation information associated with them, ranging from state and city level down to precise coordinates.

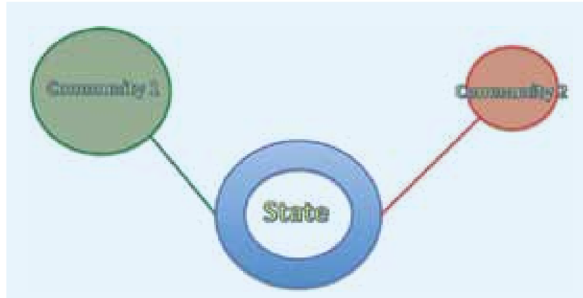


Figure 1 | A polycentric state structure, comprising the core nation-state and satellite communities with active involvement in this state's affairs

In order to identify the spatial distribution of these contributions, we generated a density surface of the origin locations of geolocated tweets in our data corpus. The map in Figure 2 is visualizing the global epicenters of the discussion about Syria in Twitterdom, as the hotspots from which these contributions originate. In addition to the major epicenter in Syria (and across its neighboring states), the secondary regions of concentration comprise the Gulf States (spanning Kuwait, Saudi Arabia, and the United Arab Emirates), Western Europe, and the United States (primarily the two coasts, and Southeast).

In order to assess the structure of the polycentric virtual Syria, we clustered our data by state and identified the key international communities that are involved with the topic of Syria. These communities have been ranked by the absolute number of tweets contributed by their members during the period of study. While the absolute value is an important metric of participation, it has an obvious population bias: large nations are expected to contribute more feeds than small. In order to remove this bias, we introduced the concept of a participation index, defined as the number of tweets per thousand citizens for each of these countries. This allows us to recognize for example that while the United States is

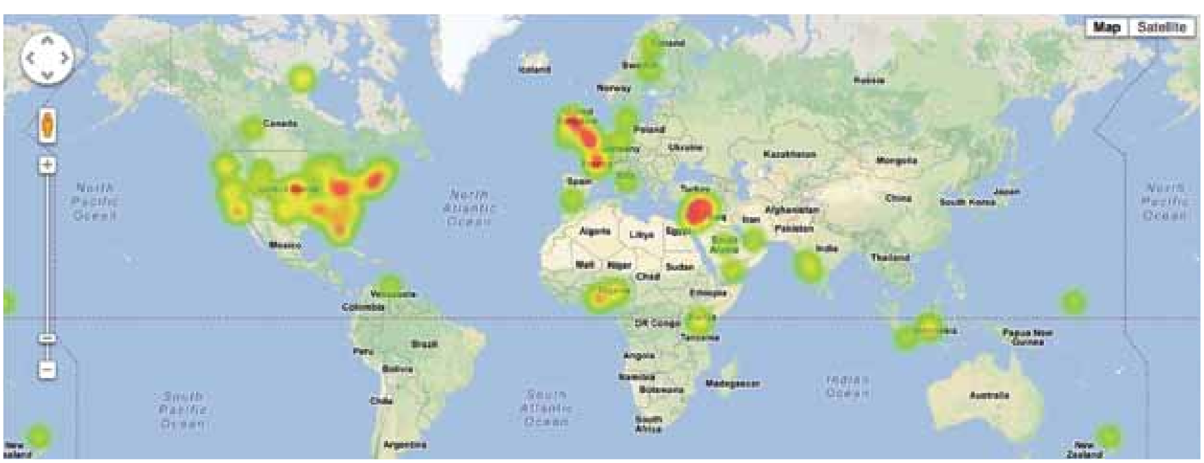


Figure 2 | A point density analysis revealing hotspots in geolocated tweets

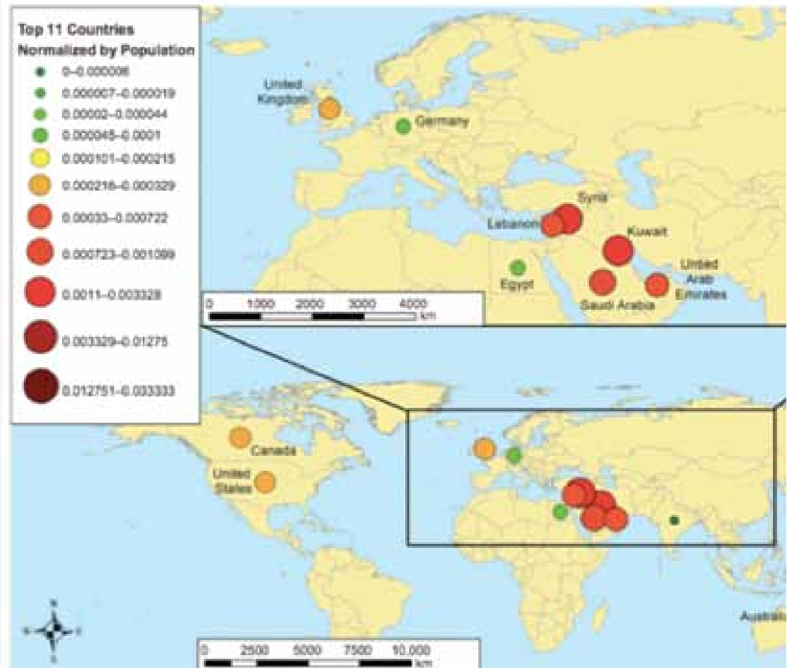


Figure 3 | The top 11 international communities participating in discussion of Syria on Twitter, plotted with their normalized participation index metrics

responsible for more Twitter traffic on the subject, Syria (as expected) and Kuwait (an interesting observation) are much more involved with the problem, with their participation indexes being tenfold that of the U.S. The normalized index serves as a more suitable metric (compared to absolute number of tweets) to express the level of involvement of various states on this particular topic. This information is plotted in Figure 3, which shows the spatial distribution of these key actors.

Outlook

The proliferation of social media has led to the emergence of a new type of geospatial information that defies the conventions of authoritative or volunteered geographic information, yet can be harvested to reveal unique and dynamic information about people and their activities. In this article we addressed the formation of virtual communities around national issues and the identification of polycentric virtual equivalents of states. This offers us a new perspective to geopolitical boundaries, showing how the world is structured and connected despite its political boundaries rather than because of them.

Dr. Anthony Stefanidis *is a Professor with the Department of Geography and Geoinformation Science at George Mason University, and the director of Mason's Center for Geospatial Intelligence. He holds a Ph.D. in Digital Photogrammetry from The Ohio State University. His areas of expertise include the analysis of digital imagery and video, spatiotemporal information modeling and analysis, sensor networks, and the harvesting of geospatial information from social media.*

Amy Cotnoir *is a Ph.D. student in Earth Systems and Geoinformation Science at George Mason University. Her area of expertise relates to creating and analyzing data in GIS and her research interests include social media, human geography, and international relations from a network perspective.*

Dr. Arie Croitoru *is an Assistant Professor with the Geography and Geoinformation Science Department at George Mason University, and a member of Mason's Center for Geospatial Intelligence. He holds a Ph.D. in Geoinformatics from Technion University in Israel. His research focuses on three interrelated research streams: geosocial analysis and location-based mass collaboration systems (LB-MCS), spatiotemporal data mining, and digital image processing. Currently, Croitoru's work focuses on developing novel algorithms for modeling, abstraction, and automated knowledge discovery in location-aware social media, and on the design and development of geosocial analysis platforms.*

Dr. Andrew Crooks *is an Assistant Professor with the Department of Computational Social Science and member of the Center for Social Complexity at George Mason University. He holds a Ph.D. from the University College London's (UCL) Center for Advanced Spatial Analysis. His area of expertise relates to working with GIS, creating and analyzing large spatial datasets, and finding patterns and insights from such data. Crooks is a recognized expert on the integration of GIS and agent-based modeling (ABM).*

Jacek Radzikowski *is a Senior Researcher with George Mason University's Center for Geospatial Intelligence. He holds degrees in Computer Science from Warsaw University of Technology, and in Computational Science from Mason. His areas of expertise are database management and development, software design, system implementation.*

Matthew Rice *is an Assistant Professor with the Geography and Geoinformation Science Department of George Mason University. He holds a Ph.D. in Geography from the University of California Santa Barbara. His research interests are geospatial information visualization and volunteered geographic information.*

EXCERPT FROM

Stefanidis, Anthony, Amy Cotnoir, et al. "Demarcating new boundaries: mapping virtual polycentric communities through social media content." *Cartography and Geographic Information Science.* 40.2 (2013). Print.

ENABLING GEOSPATIAL DATA DELIVERY

By Chris Herwig

The wealth of satellite imagery of the earth that is being gathered is growing each day. The types of professionals that use this information vary widely and include meteorologists, agriculturalists, geologists, conservationists, and educators. Despite the abundance of information available and the scope of fields that rely on satellite imagery, current methods for delivering this data to users have been outdated, until recently.

Traditionally, requesting satellite imagery through commercial imagery providers required a fax machine. The imagery request process involves multiple, often antiquated, communication hops to get access to pixel data coming from some of the most high-tech tools in outer space. Typically, a customer with the need to buy imagery navigates a field of imagery resellers and provides an "area of interest" and timeframe for when they would like to receive the imagery. The reseller faxes this information to a satellite imagery provider, who in turn responds with a quote. The quote is then delivered by the reseller to the customer that requested the imagery, who then decides whether or not to make the purchase. MapBox Satellite Live revolutionizes the imagery request process with direct, on-demand access that removes the unnecessary layers of communication between customers and providers.

Satellite Live is a cloud service that provides instant access to timely satellite imagery, with the goal of providing live imagery for publishing within hours of an event, anywhere in the world. The service provides any web-connected device the ability to request imagery directly from MapBox.com. Satellite Live uses an optimized processing workflow to ensure that imagery moves as quickly as possible from satellites to ground stations to MapBox.com.

Figure 1 | South America spanning Ecuador, Peru, and Brazil on a typical cloudy day

Pixels As Data

In addition to providing timely access, MapBox Satellite also allows users to moderately enhance the imagery. MapBox treats imagery as data and built algorithms to remove clouds and other atmospheric defects. Figure 1 shows a typical day in the Andes with a heavy cloud layer. The most common method of removing clouds is erasing the atmosphere by stitching together side-by-side images, taken months apart in different seasons. The stitching approach creates seams, producing a composite image with black zones of missing data. In contrast, Satellite Live selects one-tenth of an image's clearest pixels. It then averages all of the high quality pixels together and smoothes them out over the very clearest days. The result is the same region of the Andes, completely cloudless, with no distortion side effects (see Figure 2).

Live Pipeline

MapBox's scalable backend handles incoming imagery from Skybox, DigitalGlobe, RapidEye, and Astrium. Each of these satellite companies provides access to pre-event and post-event imagery of natural disasters and political unrest. New images are constantly added to the vendor-specific imagery endpoints. As new images appear, MapBox's processing pipeline downloads the source pixels, performs sensor-specific imagery processing algorithms, and converts the image into MBTiles (the format used to deliver maps online). This process is quick and seamless, and the end result is a robust imagery layer.

Turning Sensor Data into Actionable Information

MapBox lowers the barrier of entry for turning satellite data into actionable information and intelligence. The potential use cases for this kind of rapid imagery are incredible. Journalists can include the latest images available in their stories. Staffers briefing decision makers will be able to access the latest imagery, without needing a GIS background or special software.

In the iOS screenshot in Figure 3, the real-time imagery layers are easily composited on top of global satellite layers and the global street-level basemaps. Combining base-layers can easily be performed in an online user interface, without the need for a dedicated desktop tool.

Figure 2 | South America spanning Ecuador, Peru, and Brazil (cloudless)

MapBox maps provide cross-browser and cross-device compatibility and hosted maps are accessible from anywhere with an Internet connection.

The Larger Cloud Infrastructure

MapBox routinely uses cloud computing to process large geospatial jobs like rendering the street basemap for the entire world. In the context of rendering, the scalability challenges are the length of time to render and the storage space required for the rendering. Suppose you planned to render the world on multiple zoom levels. Even if you could ignore practical performance factors like PostGIS tuning, style optimization, and Mapnik performance, a rendering tool like TillMill, running on a single server, would take approximately 445 days. The space it would take to store all of the OpenStreetMaps tiles for two zoom levels is 54 terabytes. Using cloud computing, these rendering challenges can be avoided.

Another scalability benefit of cloud architecture is the ability to host hundreds of millions of map views per month. Map views have to be fast, displaying in a browser in the time it takes for other HTML content to load. In addition, the architecture must be resilient to massive spikes. MapBox's traffic load can spike up to

Figure 3 | Combining imagery layers enhances context and usefulness of satellite data

MapBox lowers the barrier of entry for turning satellite data into actionable information and intelligence.

500 percent during natural disasters like Hurricane Sandy. MapBox is entirely hosted in Amazon AWS. The availability of minutely metered server infrastructure together with smart management tools allows MapBox to provide guaranteed service to large media clients with millions of customers.

MapBox is powered by open source, largely by Node.js, Backbone.js, Puppet, Mapnik, and Jekyll. Cloud services are used in order to scale quickly with demand and avoid centralization. The company runs a cluster of EC2 instances as the primary application servers.

An Elastic Load Balancer divides traffic between the cluster of running servers and routes around clusters that become unresponsive.

MapBox serves maps from 30 globally distributed edge servers. These edge servers ensure that the maps are fast no matter where users are located. MapBox's georedundant infrastructure spans the entire globe for speed and reliability — even if there's a massive power outage or natural disaster, the company will be able to quickly and reliably provide the data necessary for analysis.

Chris Herwig *leads the Satellite team at MapBox, where he uses open source tools to acquire and process the source imagery that powers MapBox Satellite and other analytical products. Herwig is passionate about open data and open source. Prior to joining MapBox, he worked at the U.S. Department of State, where he was awarded the State Department's Franklin Award for his achievements in digital media for the Summit of the Americas. Herwig holds a Bachelor's degree in international area studies and human rights from Drexel University in Philadelphia, PA, where he attended as a Presidential Scholar.*

THE COMING WAVE OF LOCATION-BASED SERVICES

By Graham Neray

What if the President could view all military assets deployed in Afghanistan today in a single dashboard? What if he could visualize, manage, and analyze feeds from video, equipment sensors, and intelligence reports in real time?

What if the head of the TSA could track and monitor feeds — surveillance, security abnormalities, no-fly passengers — from every U.S. airport and leverage real-time geospatial analytics to prevent the next terrorist attack?

What if U.S. Intelligence could inspect and aggregate IP addresses, geospatial tags, and log files to identify international cyber criminals?

While the U.S. Government has been working with geospatial data for decades, a wave of innovation across wireless, sensor, and data management technologies is unearthing a slew of new applications that take advantage of this data. Declining mobile chipset costs have lowered smartphone prices, making them affordable for the masses. Sensor prices continue to drop, and they can now be embedded in anything from cars to smart watches. The growth of smartphone adoption, with over 60 percent penetration in the U.S., has created demand for innovative apps.[1] The ubiquity of high-speed connectivity over wireless networks makes it feasible to deliver bandwidth-intensive apps on a mobile device. Verizon's 4G LTE network, for instance, now covers 95 percent of the U.S. population.[2] Lastly, new data management technologies — alternatives to the relational database — that are more flexible, scalable, and can accommodate diverse data types in a single repository enable new applications on the backend.

Invasion of the Apps

Much of the innovation in this area is happening in the consumer market. Google Glass incorporates sensor and geospatial data in real time to provide augmented reality. Uber's mobile cab-hailing service is turning the taxi industry upside down. Nike's FuelBand tracks users' movements throughout the day — including running, walking, basketball, dancing, and other everyday activities — then calculates calories burned and visualizes progress in a dashboard.

The ability to incorporate location-based or geospatial data into mobile and desktop applications changes what's possible, not only for consumers, but for enterprises and government organizations as well.

The consumer market has, however, been the driving thrust behind location-based apps for several reasons. One reason is that these apps require users to transmit data over wireless networks, which in many cases are not secure enough for government entities. Perhaps more importantly, traditional relational databases — the bread and butter of government IT — are poorly-equipped to provide the rich, interactive experiences users have come to expect. This is why most location-based services are based on alternatives to relational databases, like MongoDB.

When it comes to location-based services, relational databases fall short in a number of ways.

Geospatial Data at Scale With the proliferation of smartphones and other mobile devices, sensors, and machine-generated data, the volume of information that is geocoded and suitable for location-based services is massive. Geospatial capabilities are mature in relational databases, but making these systems scale to large data volumes and numbers of users while maintaining performance at a reasonable cost has been challenging.

Data Diversity The modern location-based app must incorporate a variety of data types: unstructured data, semi-structured data, polymorphic data, and of course, geospatial data. The underlying database, therefore, must have a data model that is flexible enough to support a myriad of data types in a single data store. The traditional relational database, however, was built to support relatively static applications with highly structured data. It flattens data into two-dimensional tabular structures of rows and columns, enforced by a rigid schema, making it challenging or even impossible to support the requirements of leading-edge geospatial apps.

Continuous Global Availability The relational database of yesteryear lived in a single data center. If the database went down in the middle of the night, someone woke up the database administrator, who would fix the problem as soon as she could. By contrast, modern applications need to be available and replicated in real time across the world, and downtime is unacceptable.

How We Build and Run Apps Today

To address these shortcomings, a number of companies are foregoing the relational database in favor of MongoDB, an open-source NoSQL database.

Unlike relational databases, MongoDB is designed for how we build and run applications today. Instead of storing data in rows and columns as one would with a relational database, MongoDB stores a binary form of JSON documents (BSON). Relational databases have rigid schemas that must be defined up front. By contrast, documents in MongoDB have flexible, dynamic schemas that can vary across documents and can evolve easily over time. This model makes it possible to represent rich objects with nested documents and arrays in a way that is more natural and that mirrors the world it's intended to model. Further, the document model makes it easy to incorporate disparate types of data in a single data store and in real time, such as feeds from video, equipment sensors, and intelligence reports.

MongoDB also has native auto-sharding, which allows users to scale from single server deployments to large, complex multi-data center architectures. It also provides built-in replication with automated failover to enable high availability. Lastly, MongoDB has a rich set of geospatial capabilities, including GeoJSON support, geospatial indexes, and geospatial analytics using the Aggregation Framework with circle, line, rectangle, and polygon intersections.

Industry Adoption of MongoDB

Here are a few examples of how enterprises and government bodies are using MongoDB to transform how they use geospatial data:

City of Chicago The City of Chicago is using MongoDB to cut crime and improve municipal services by collecting and analyzing geospatial data in real time from over 30 different departments. For instance, in a given area, it might evaluate the number of 911 calls and complaints, broken lights, stolen garbage cans, liquor permits, and abandoned buildings, determining that an uptick in crime is likely (see Figure 1). It needs to marry structured and unstructured data at scale and to conduct in-place, online analysis. With legacy technologies, this would be challenging at best, infeasible at worst.

Top 5 Global Telephone Company A mobile operator is using MongoDB to monetize underused legacy data from wireless towers. Like many telephone companies, this operator collects data on the locations of its customers. Rational network investment hinges on knowing which cell sites require more capacity, or where more cell sites are needed. This operator uses MongoDB to power a new service through which a business can push location-specific offers to the company's subscribers in real time when they are in the vicinity of that business. This provides a new revenue stream for the telephone company and brings its legacy systems to life.

Top 5 Industrial Equipment Manufacturer A top industrial equipment manufacturer is using MongoDB to power a cloud-based analytics platform that ingests, stores, and analyzes readings (e.g., temperature, location) from its customers' equipment. It then presents the readings back to customers via a web interface — including visualization, key metrics, and time series

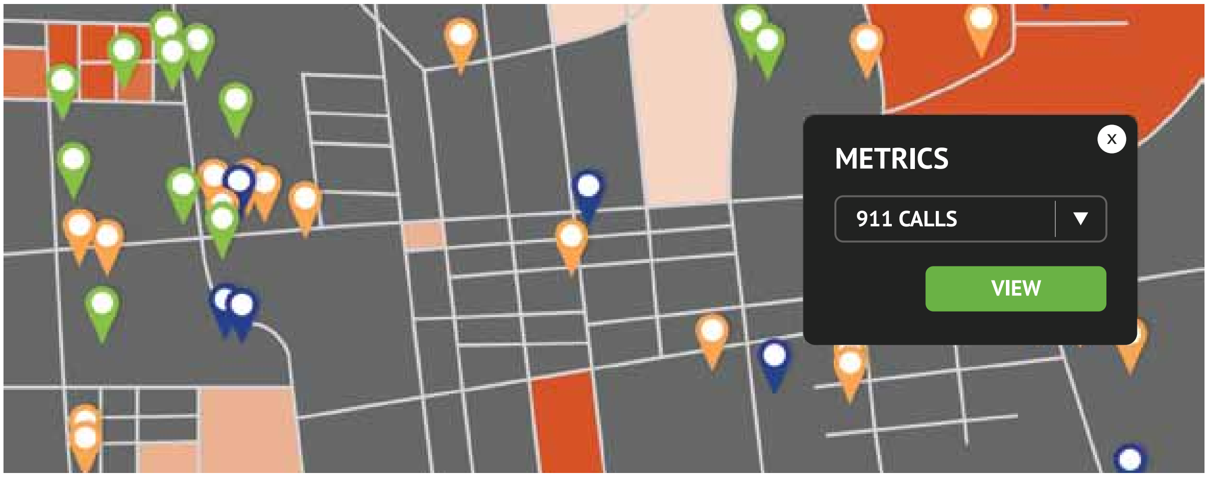


Figure 1 | The City of Chicago is using MongoDB to cut crime and improve municipal services by collecting and analyzing geospatial data in real-time from over 30 different departments.

analysis — to help them make better decisions about their businesses, such as where to provision equipment and how to increase facility efficiency. Additionally, the company uses the data to inform internal product development. The application stands out in an industry that has seen little innovation in the last half-century with the ability to drive new revenue streams.

Leveraging Location in the U.S. Government

Advances in geospatial capabilities and database scalability enable a number of unique applications for the defense and intelligence communities.

Military Asset Analytics The military owns and manages fleets of tanks, drones, Humvees, and other assets. While it may currently have the ability to see where these assets are at a given point, the military could expand its understanding of how the equipment is used, and furthermore, optimize its usage. MongoDB's scalable model and native geospatial capabilities enable the military to ingest and analyze sensor data to optimize vehicle and platoon deployment, fuel consumption, and other metrics, both in real time and retrospectively. MongoDB's flexible data model makes it possible to do all this in a single data store and to provide a 360-degree view of military assets.

Surveillance Data Aggregation Government agencies are exploring new ways to increase national security, including innovative surveillance data collection. This data is pouring in from a variety of sources and in massive volumes. MongoDB provides an adaptable and scalable platform for aggregating various surveillance feeds and making sense of the data through real-time analyses.

Crime Data Management and Analytics Legacy criminal record systems based on relational databases are often brittle and difficult to adapt to legal and regulatory changes. MongoDB makes it easy and cost-effective to adapt the technology as the law evolves and to incorporate information from disparate sources. Law enforcement organizations can deploy innovative, real-time analytics to identify offender-specific, geospatial, and other crime patterns quickly and effectively.

Government is increasingly faced with growing data volumes and shrinking budgets, new demands, and legacy infrastructure. Given the rise of innovative geospatial applications in the market and the advent of open source technologies, solutions are available that will allow organizations to make the most of their data. Q

***Graham Neray** is a Product Marketing Manager at 10gen, an IQT portfolio company. Neray works closely with customers, partners, and the open-source community to articulate how MongoDB is quickly becoming the world's most popular database. Prior to joining 10gen, he was a Senior Business Analyst at CSMG, a boutique management consulting firm specializing in the high-tech and telecom industries. He graduated from Brown University with a concentration in Political Science.*

REFERENCES

[1] "Mobile Majority: U.S. Smartphone Ownership Tops 60%." *Nielsen Newswire.* June 6, 2013.
[2] Fitchard, Kevin. "Verizon wraps up LTE rollout; plans all-VoIP phone launch for late 2014." *GigaOm.* June 27, 2013.

ASSESSING COMPLETENESS AND SPATIAL ERROR OF FEATURES IN VOLUNTEERED GEOGRAPHIC INFORMATION

By Steven P. Jackson, William Mullen, Peggy Agouris, Andrew Crooks, Arie Croitoru, and Anthony Stefanidis

Improvements in communications technology and information availability are having a significant impact on the field of geography as they enable the general public to produce geospatial products for mass consumption on the Internet. As technology continues to improve and the Internet becomes more accessible, the amount of geospatial data generated by citizens without formal geographic training is expected to rapidly increase.[1] Thus, Volunteered Geographic Information (VGI) is bringing the general public into the realm of map production functions traditionally reserved for official agencies.

The focus of this article is on the accuracy of crowdsourced VGI. It has been noted that public participation in geospatial mapping on the web has allowed citizen groups to map and provide local knowledge context that significantly advances the mapping project; however, others have noted that the characteristics of the information are less rigorous than traditional scientific data collection reporting. A step towards understanding the potential data quality of volunteered data would be to quantify key quality characteristics for geospatial data that can reasonably be expected to be included in contributed datasets, and then compare those characteristics against reference sources of data to quantify data quality.

Accuracy and Completeness Considerations for Volunteered Geographic Information

While data quality has been at the center of the research agenda since the definition of GIScience, geographers Michael Frank Goodchild and G.J. Hunter presented a discussion of the method for comparing two datasets whereby the tested source of data is compared to the reference source of data.[2,3] The reference dataset is assumed to represent ground truth while the test dataset is measured against the reference dataset. Comparisons between datasets are common within the literature; however, the methods in this paper are tailored for comparing point features.

Considering the rather ad hoc nature of VGI approaches, completeness is as important as accuracy when it comes to assessing the quality of the contributed information. Our research continues the trend in evaluating both completeness and accuracy, but extends the notion of completeness to the comparison of individual point features representing area features, and assesses accuracy generally.

Quantifying completeness and accuracy of VGI will allow users of the data to better understand the data's utility. As VGI is gaining popularity, it leads to the generation of large volumes of geospatial data that can potentially complement and enhance traditional "authoritative" data sources. To enable tapping into this potential, we need a better understanding of the quality of VGI contributions, in particular their accuracy and completeness. This is even more important now, as VGI data collection is increasingly involving volunteers with little or no geographic training, who are producing geographical data. Consequently, there is a need to further study the quality characteristics of VGI.

A Comparison of Data Sources

We aim to extend the current state of knowledge on the topic of VGI quality by focusing on completeness and accuracy of point features within VGI data. Our analysis compared two VGI test datasets against a reference dataset and analyzed their differences. We also examined the particularities of a hybrid variant of VGI, and assessed its impact on the overall accuracy of the VGI product.

Figure 1 | School count by data source

Source	School Count
Oak Ridge National Laboratory (ORNL)	402
OpenStreetMap (OSM)	406*
OpenStreetMap Collaborative Project (OSMCP)	412

* Includes 48 historical school locations.

The reference data is based upon information from the Department of Education's lists of public and private schools. On behalf of the Federal government, Oak Ridge National Laboratory (ORNL) was asked to geospatially improve the location accuracy of the Department of Education data using repeatable methods. In addition to this reference dataset, two test datasets are also used in this case study.

The first VGI test dataset comprises school locations from the point of interest layer of OpenStreetMap (OSM). We also use a second test dataset, which is a product of the U.S. Geological Survey (USGS) OpenStreetMap Collaborative Project (OSMCP) Second Phase. OSMCP represents a hybrid variant of VGI in that it introduces limited oversight to the VGI process: the data are collected through VGI processes, peer-edited by volunteers, and provided to the volunteers by a government agency.

Figure 1 presents the total count of schools within the study area for each data source. While the total number of schools is reasonably close across the three sources, it is important to note ORNL and OSMCP data represent only active schools while OSM data includes approximately 12 percent historic schools which are likely no longer in existence and would, therefore, not match schools in either ORNL or OSMCP.

The results shown in Table 1 imply that these datasets are similar. However, a deeper assessment of the schools showed that only 281 schools are common to all three datasets, illustrating that simple feature count may not adequately evaluate spatial accuracy or completeness. It is clear that a systematic comparison of the datasets is required in order to evaluate the quality of the VGI test data in comparison to the reference data.

Figure 2 | Summary of record counts for data matching

Method	ORNL-OSMCP	ORNL-OSM
Intersection	357	287
Reference Complement Test	44	114
Test Complement Reference	63	99
Union	464	500
Reference Count	401	401
Test Count	412	406

Automated matching of the datasets was carried out using four different methods across each of the name and address fields from the dataset attributes. Our automated methods are incapable of dealing with some cases; therefore, the user must manually examine the unmatched records which remain after the automated processes to determine if any potential matches were missed.

The final process in the analysis includes computation of the intersection, union, and complement. The *intersection* dataset includes all records that appear in both datasets. The *reference complement test* dataset includes those records that are in the reference dataset, but not in the test dataset. The *test complement reference* dataset includes those records that are in the test dataset, but not in the reference dataset. These values can then be used in the computation of the accuracy and completeness.

Results of the Matching Methods

Examining the results of the automatic and manual matching methods, Figure 2 provides a summary of the counts that were obtained from each of the intersection, complement, and union calculations. The union and complement counts for the OSM comparison are higher, as would be expected, because of the lower matching rates caused by the absence of the address information; however, as is the case for comparing record counts, simply comparing the counts between these comparisons is insufficient when trying to understand the meaning of the results.

ASSESSING COMPLETENESS

Assessing completeness of the contributed data provides an understanding of the reliability of the reported results and allows assessment of the usefulness of contributed data as a potential data source for use by mapping agencies and researchers. The comparison of ORNL and OSMCP data showed that a total of 89 percent of the records were matched. Of that, 82 percent were matched automatically, while the remaining 7 percent were matched manually, indicating that the automated matching algorithm is successful.

The match rate for the ORNL and OSM data comparison was considerably lower, with only 71 percent of the total records matched. However, the manual match rate was over twice as high at 15 percent, with the automated match rate falling to 56 percent. Due to the relatively poor performance for completeness, the utility of OSM data as an alternative mapping source is questionable in the study area; however, the OSMCP data, which included approximately 9 out of every 10 schools, represented a significant improvement over the unconstrained OSM results of just over 7 out of every 10 schools.

Figure 3 | Spatial error for matched schools

	SPATIAL ERROR (m)					
Datasets	Count	Minimum	Maximum	Mean	St. Deviation	Median
ORNL-OSMCP	357	2	487	47	50	33
ORNL-OSM	287	2	1,848	190	314	43

As VGI is evolving, both in terms of participation and scope, a better understanding of its quality, the parameters that affect it, and the practices used to produce it will help enhance the utility of its products for geospatial analysis.

Analyses revealed that the OSM and OSMCP efforts captured schools that were not in the ORNL data. 28 percent of the OSM schools remained unmatched at the end of the analyses while 11 percent of the OSMCP records remained unmatched.

ACCURACY

While the completeness measure is assessed by comparing the matched and unmatched records, the accuracy examines only those records which had matches as identified previously in the intersection. It is expected that the overall accuracy of the OSMCP data will be high considering the quality control procedures that were part of the project; however, it is important to quantify the accuracy of the data since this is a fundamental element of geospatial analysis.

The spatial error is evaluated for each match and the results are located in Figure 3. While the minimum error was two meters in both comparisons, the maximum error for the OSM data was approximately four times that of the maximum error for the OSMCP. Both the mean and the standard deviation were higher for the OSM data with the latter indicating that the error within the OSM data varies more than in OSMCP. In addition, the median of the OSM data is lower than the mean, indicating that the data is skewed.

Because the error was so different between the two datasets, an additional effort was undertaken to examine the nature of the error distribution. The percent of matched schools within 150 m (cumulative) for both OSMCP and OSM can be shown to be 96 percent. These results indicate that either dataset would be equally capable of getting the user to the school property.

One additional assessment was undertaken in order to evaluate the accuracy of OSM *versus* OSMCP. In this final evaluation, the spatial error for matches from both datasets was compared to each other to determine which one is closer more often. Of the 281 matched schools, OSMCP schools were closer 58 percent of the time; however, OSMCP also has the largest difference (224 m).

Conclusions

Our observations indicate that the added rigor appears to improve both the completeness and accuracy as compared to the OSM data. The analysis of completeness showed that the OSMCP data captured close to 90 percent of the records in the reference ORNL database, while the OSM data captured approximately 70 percent of these records. The lower completeness result observed within the OSM data can be attributed to two factors: the OSM data does not include address information, and the collection methods employed for the OSM data do not include the formal quality control processes implemented within the collection methods for OSMCP. Lastly, 70 more OSMCP schools (357) matched the reference dataset than did OSM schools (287). Similar trends were identified with respect to positional accuracy, which reflects the spatial error between the locations of the two datasets, with OSMCP data appearing to be more accurate than OSM.

As VGI is evolving, both in terms of participation and scope, a better understanding of its quality, the parameters that affect it, and the practices used to produce it will help enhance the utility of its products for geospatial analysis. Based on this initial work, several areas of future work need to be explored. Unconstrained (and untrained) contributors do not always share a common understanding of the definition of what a feature is or where it should be located and the effect of the vagueness on data quality is not understood. The result of the vagueness can be degradation in the utility of VGI for decision making; however, these effects have not been studied. Lastly, there is a need to improve methods for evaluation of data that is currently labeled as "authoritative" because, as we have shown in this research, these datasets are not without error. Q

Steven P. Jackson *is a Ph.D. student in Geography and Geoinformation Science at George Mason University. His specialty is geographic information systems science.*

William Mullen *is a Ph.D. student in Geography and Geoinformation Science at George Mason University. His expertise is in geographic information systems.*

Dr. Peggy Agouris *is a Professor with the Department of Geography & Geoinformation Science at George Mason University, and the director of Mason's Center for Earth Observing and Space Research. She is also currently the interim Dean of Mason's College of Science. She holds a Ph.D. in Digital Photogrammetry from The Ohio State University. Her areas of expertise include digital image analysis for geospatial information extraction and spatiotemporal information modeling and analysis.*

Dr. Andrew Crooks *is an Assistant Professor with the Department of Computational Social Science and member of the Center for Social Complexity at George Mason University. He holds a Ph.D. from the University College London's (UCL) Center for Advanced Spatial Analysis. His area of expertise relates to working with GIS, creating and analyzing large spatial datasets, and finding patterns and insights from such data. Crooks is a recognized expert on the integration of GIS and agent-based modeling (ABM).*

Dr. Arie Croitoru *is an Assistant Professor with the Geography and Geoinformation Science Department at George Mason University, and a member of Mason's Center for Geospatial Intelligence. He holds a Ph.D. in Geoinformatics from Technion University (Israel). His research focuses on three interrelated research streams: geosocial analysis and location-based mass collaboration systems (LB-MCS), spatiotemporal data mining, and digital image processing. Currently, Croitoru's work focuses on developing novel algorithms for modeling, abstraction, and automated knowledge discovery in location-aware social media, and on the design and development of geosocial analysis platforms.*

Anthony Stefanidis *is a Professor with the Department of Geography and Geoinformation Science at George Mason University, and the director of Mason's Center for Geospatial Intelligence. He holds a Ph.D. in Digital Photogrammetry from The Ohio State University. His areas of expertise include the analysis of digital imagery and video, spatiotemporal information modeling and analysis, sensor networks, and the harvesting of geospatial information from social media.*

EXCERPT FROM

Jackson, Steven, William Mullen, et al. "Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information." *ISPRS International Journal of Geo-Information.* 2 (2013): 507-530. Print.

REFERENCES

[1] Mooney, P.; Corcoran, P.; Winstanley, A. Towards Quality Metrics for OpenStreetMap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 514–517.

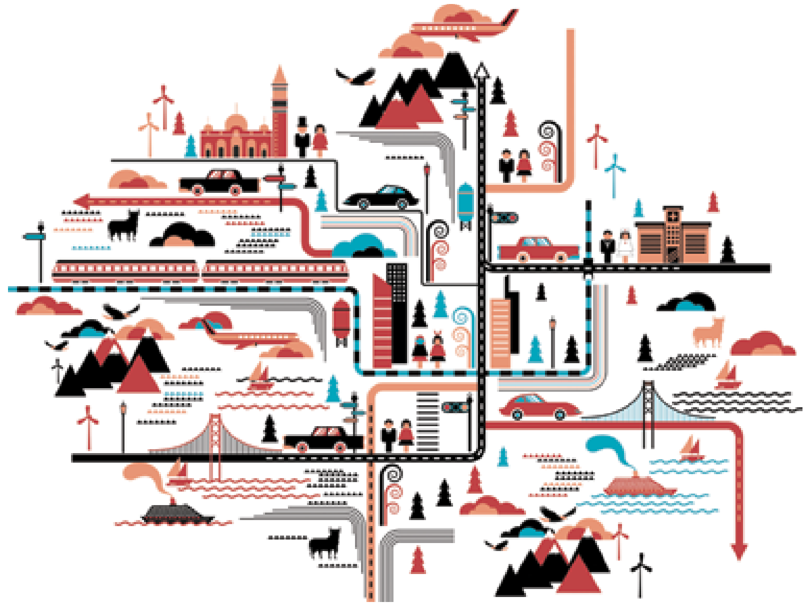
[2] Goodchild, M.F. Geographical information science. *Int. J. Geogr. Inf. Syst.* 1992, 6, 31–45.

[3] Goodchild, M.F.; Hunter, G.J. A simple positional accuracy measure for linear features. *Int. J. Geogr. Inf. Sci.* 1997, 11, 299–306.

DEMOCRATIZING MAPS:

Putting the Power of Maps in the Hands of Business Analysts

By Kevin Merritt



Rising citizen demand for access to public data has prompted governments worldwide to make more information available on the web in standardized, machine readable formats. In the process, many government leaders have discovered a strong interest among constituents for creating dynamic, interactive maps from publicly available data. Until recently, producing digital maps required technicians with specialized training, extensive labor hours, and expensive hardware and software. Due to advances in cloud and mobile technologies and the expansion of open APIs, mainstream users can now create beautifully detailed online maps in minutes from a variety of computing devices.

Technology innovation and political and cultural change are more closely linked than ever before. The broad acceptance of open data, which calls for broad access to and robust sharing of non-personal public information, underscores this reality. Mirroring the trajectory of business and consumer technology, the innovations forged in this open data movement are designed to promote ease-of-use, widespread adoption, and enthusiastic reuse by developers.

Geospatial Analysis: The Next Frontier in Open Data

The next step forward in open government initiatives involves the use of geospatial analysis to quickly reveal patterns and trends, identify geographic relationships among data points, and enable predictive modeling for better-informed civic planning. In fact, many of the agencies that have led the way in launching searchable, user-friendly data sites have already seen a spike in demand for mapping capabilities and location-based service applications.

Until recently, valuable geospatial data was mostly locked up in inaccessible, proprietary formats. The use of mapping and analysis tools was the exclusive purview of specially-trained GIS analysts, who relied on cumbersome and costly systems. Advances in cloud-based infrastructure and mobile computing technology, in conjunction with an ever-expanding array of open-API mapping services, are transforming the way that spatial information is generated and used in everyday contexts. These changes have far-reaching implications for business and government organizations, as well as members of the defense and intelligence communities.

Key Trends and Opportunities in the Market for Geospatial Data

Several technology trends are converging to revolutionize the use of geospatial data. These include the rise of big data, the growth of location-based information and analytics, the continuous consumerization of IT, and the maturity of cloud infrastructure and development platforms.

BIG DATA

As the digital universe gets larger, the variety of the data produced continues to expand. This creates an enormous challenge for organizations across all industries. Data savvy professionals are embracing this challenge by utilizing geocoding information to provide richer downstream analysis capabilities. This includes identifying spatial relationships and visualizing patterns in the data they are collecting.

LOCATION-AWARE DEVICES

A major cause for the surge in unstructured data is the proliferation of millions of location-aware smartphones, tablets, and other mobile devices. The ubiquitous use of cloud services by these devices and the connections of these devices to each other continue to contribute vast amounts of geospatial data to the digital universe. As the quality of this data improves over time, businesses and public sector agencies can use this information to boost productivity, solve complex resource distribution problems, and plan for contingencies with greater confidence.

MOBILE

Demand for lightweight mobile apps that incorporate web mapping capabilities will continue to increase. Geospatial solution providers will need to address the challenge of providing rich, interactive mapping functionality that can merge multi-source content on-the-fly while ensuring a high-quality, consistent user experience across devices. To keep training costs down and productivity high, organizations will require that these mapping-enabled applications work seamlessly with existing workflows and provide an intuitive interface that resembles everyday consumer applications.

CLOUD ADOPTION

The maturation of the cloud delivery model makes it easier and more cost-effective for organizations to deliver applications that use geospatial data. Spatial problems typically involve large volumes of data. This dovetails with the elastic storage and compute power that cloud networks provide. Furthermore, in these times of fiscal belt-tightening, cloud-based environments can offer massive infrastructure cost savings. Government organizations can leverage cloud technology to deliver sophisticated online mapping services to constituents faster and at a lower cost, while shifting budget resources to other high-priority projects.

To align with these trends, modern mapping applications will increasingly need to:

- Include point-and-click capabilities for creating a variety of maps using point, line, and polygon-based data, all without requiring specialized GIS training;
- Reflect the most up-to-date, multi-source information and be instantly refreshable through connections to sensors, enterprise databases, social media feeds, and other diverse data sources;
- Provide simple tools for layering different datasets on top of one another to enhance context and accelerate insights, even when source data is stored in different formats;
- Deliver a high-fidelity mobile map experience that can be accessed on the go and taken anywhere.

Helping Citizens Stay Connected with Online Maps

Using Socrata, mainstream users can produce data-driven online interactive maps, without the need for cumbersome hardware or GIS expertise. Webmasters, communication professionals, and online content creators can drag and drop existing location-based or geospatial files into the Socrata cloud to instantly generate highly detailed digital maps (Figure 1).

This flexible solution automatically converts text-based location fields (such as street addresses) into latitude and longitude coordinates that can be displayed on a map. The platform can also convert polygon-based data to render maps with sophisticated geospatial information describing shapes in the physical world.

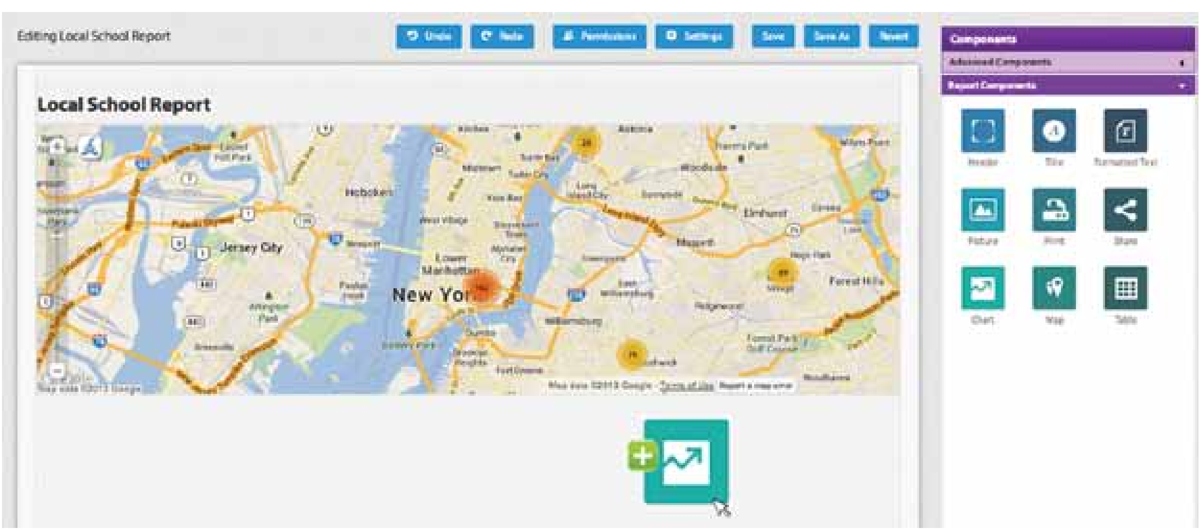


Figure 1 | Easy-to-use map building tools, like Socrata DataSlate, allow mainstream users to create data-driven, interactive maps

These shapes could represent waterways, transit lines, or road closures, to name a few examples.

Because the solution can be used as a plug-and-play cloud service with any Socrata Open Data Portal™, government agencies can incorporate these and other kinds of maps as part of a curated browsing experience to enhance engagement with web visitors. They can also take advantage of tools like Socrata DataSlate and mobile SDKs to optimize the digital mapping experience for mobile devices.

Public sector organizations have increasingly taken advantage of online mapping functionality to meet a variety of needs — from assisting with emergency management to facilitating simplified access to government services.

ONLINE MAPS AID RESPONSE TO HURRICANE SANDY

In planning for and responding to Hurricane Sandy, the New York City government leveraged its Socrata Open Data Portal infrastructure to quickly publish and share updated hurricane evacuation zone maps (Figure 2). Administration officials were able to continuously refresh these maps based on up-to-the minute flooding projections. By making this data available on its public data site, the city enabled developers to create additional emergency maps and applications. These online tools made it easier for residents to locate shelters, food distribution centers, and other resources, and to share this vital information with people in their networks. Mayor Bloomberg's office estimates that through the broad availability of these interactive maps, city service providers were able to reach 10 times as many people as they otherwise would have.

Figure 2 | New York City used the Socrata platform to create an interactive, mobile-optimized map of evacuation zones to provide vital information to residents following Hurricane Sandy

A THREE-YEAR PROBLEM SOLVED IN TWO HOURS

For several years, the Marine Board for the state of Oregon searched for a way to create an easy-to-update online map of resources for Oregon boaters. The Marine Board investigated the market for software mapping systems, but none of the solutions met its requirements for ease-of-use or affordability. After deploying the Socrata Open Data Portal, staff members were able to quickly and easily visualize datasets on a map. Within the first two hours of working with the solution, the Marine Board's Public Information Officer was able to create an online map showcasing the location of a multitude of marine resources, including marinas, boat launches, and more (Figure 3).

Geospatial Data for the Tech-Savvy Citizen

In this era of greater data and technology access, public entities are empowering citizens with the tools to not only research civic issues but to find solutions to them. A new generation of user-friendly mapping technology gives cities the opportunity to engage with their citizenry as co-creators, leveraging the skills and expertise they offer.

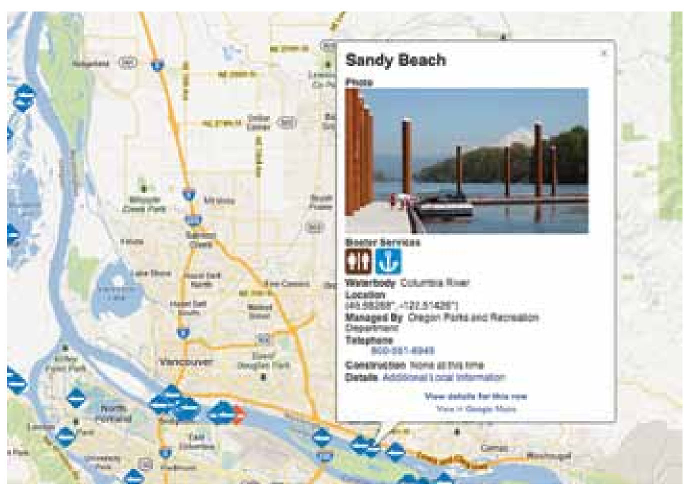
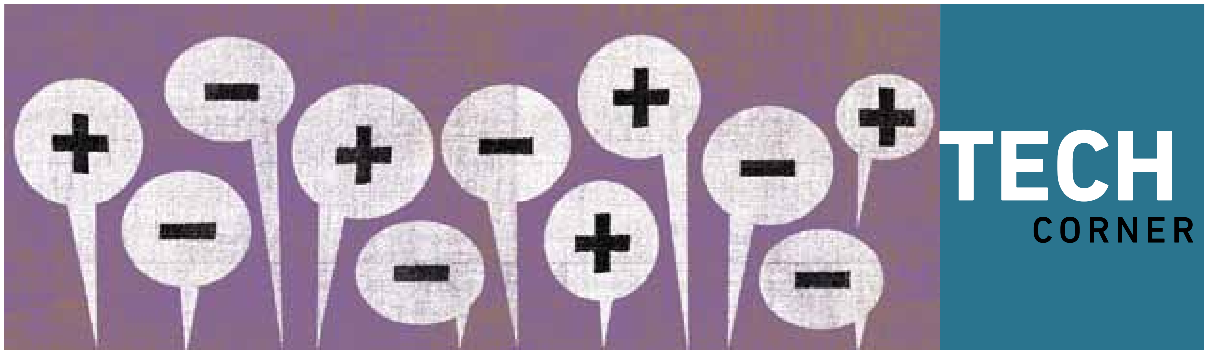


Figure 3 | The Marine Board for the state of Oregon created an online map to provide boaters with important resource and safety information

Kevin Merritt *is the founder and CEO of IQT portfolio company Socrata, a cloud-based, open data platform provider. Focused exclusively on the government sector, Merritt believes that breaking down data silos and unlocking the vast troves of data held by the government has the potential to spark tremendous economic growth and improve the programs and missions of agencies and the lives of everyday citizens. Merritt works out of Socrata headquarters in Seattle, Washington.*



To supplement the ***IQT Quarterly***'s focus on technology trends, *Tech Corner* provides a practitioner's point of view of a current challenge in the field and insight into an effective response.

BEYOND GIS: THE FUTURE OF GEOSPATIAL IS SPATIAL IT

A technology overview from IQT portfolio company Boundless

Government agencies are gathering more data and making more demands of that data in geospatial applications than ever before. These demands are increasingly being met by open source solutions — especially when that data has a geospatial component. Historically, the United States has funded proprietary vendors to grow their software offerings to meet new challenges like these. In many cases, these companies have become entrenched, slow to react to market needs, and often too expensive to deploy at scale. For these and many other reasons, agencies in the defense and intelligence communities are turning to open source software — backed by commercial maintenance and support — to replace proprietary offerings.

Government Adoption of Open Source

The rigid architectures and license costs of proprietary systems can be a burden to many agencies. As data needs increase (or worse, as time passes), so do these costs, making it difficult to budget effectively. Open source alternatives are free of software license costs and offer agility for organizations seeking to engineer their own services and solutions without the limitations of proprietary software. For those agencies that find the do-it-yourself possibilities of open source daunting, commercial open source companies like Boundless and 10gen provide enterprise-grade, expertly supported versions of open source software.

Boundless customers, including numerous defense and intelligence agencies, NOAA, Department of State, and the Federal Communications Commission (FCC), are finding superior value in commercially-supported open source software. When the FCC launched the Congressionally-mandated and open source-backed National Broadband Map (broadband.gov), they knew the site was going to receive heavy traffic. On its initial launch the site saw 500,000 visitors, with a peak load of 9,000 requests per second — more than five times the expected load.[1] As time went on, hits increased, prompting new mapping servers to be pressed into service to handle the traffic and ensure that the site stayed online. The ability to keep the site operational can be attributed to a modular, open source, web-centric approach and the support of Boundless, Michael Byrne, the FCC's Geospatial Information Officer, has stated that there was no way the agency could have afforded that level of proprietary hardware and software; the financial constraints would simply have limited the project's scalability. The FCC not only gained more robust and reliable infrastructure with open source, but was also able to access Boundless' support in the process.

The Technology Behind the OpenGeo Suite

For years open source software has offered superior scalability and reliability to proprietary software, and now open source also boasts more feature parity than ever, as companies work to match — and often exceed — their proprietary counterparts. Additionally, most open source technologies are built "from the ground up" using modern web-based technology, taking advantage of the web in ways proprietary systems with desktop legacies do not. For truly agile and innovative breakthroughs, organizations are discovering that it is wise to look beyond the old guard, proprietary software vendors.

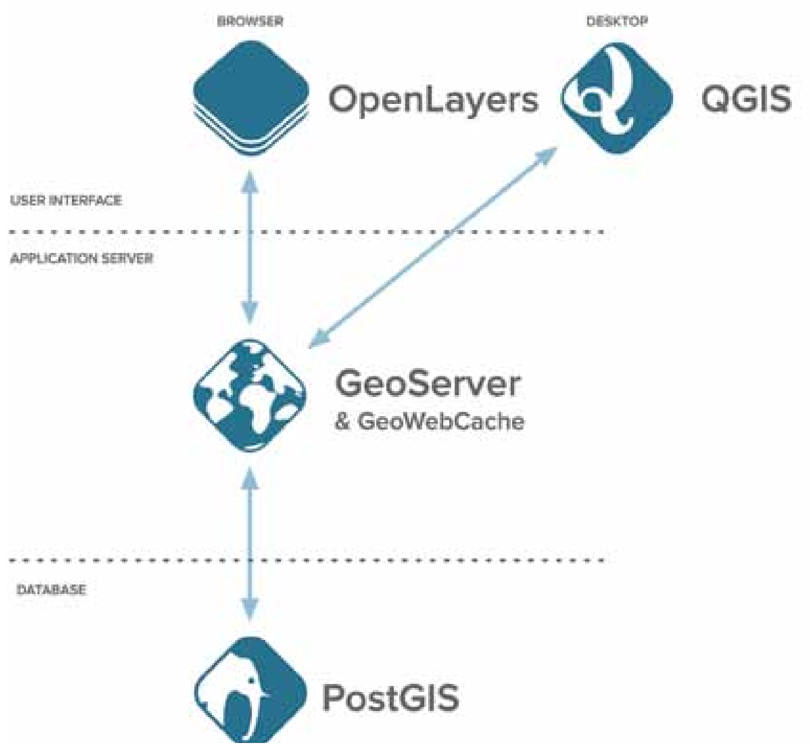
SPATIAL IT: GIS AFTER THE WEB

An effective web architecture is both complete and modular, and this is reflected in the architecture of Boundless' flagship product, OpenGeo Suite. Web applications typically contain a database or other data storage system at their architectures, some application logic in the middle, and a user interface layer at the top. The database and application layers interact via SQL over a network protocol (specific to the database vendor but usually abstracted away), and the application and interface layers interact via encoded documents transferred over Hypertext Transfer Protocol (HTTP).

The OpenGeo Suite makes use of a set of open source components that each fulfills a particular function. The foundation of the Boundless "stack" is a database (PostGIS) or file-based storage system. There are application servers in the middle tier (GeoServer and GeoWebCache), and a user interface layer on the top (OpenLayers). The database and application servers interact via SQL (with Open Geospatial Consortium-standard spatial extensions). The application servers and user interface layers interact via standard web encodings (XML, JSON, images) over an HTTP transport.

In many ways, the OpenGeo Suite web architecture is a "GIS" (geographic information system), but it is one that subordinates the "G" to the "IS". In classical GIS architectures, there is little differentiation from general-purpose corporate information systems. The Boundless web mapping architecture is distinguished from a standard application architecture, not in the arrangement or classification of the parts, but in what the parts do:

- The PostGIS database can answer spatial queries as well as standard attribute queries.
- The GeoServer map, or the feature server, can provide standardized web access to underlying GIS data sources.



- The GeoWebCache tile server can intelligently store and serve map tiles using standard web protocols for requests and responses.
- The OpenLayers map component can consume maps from multiple sources and provide tools for data editing and capture.
- QGIS is an open source GIS desktop tool with features, stability, and ease of use that rival proprietary alternatives.

A key feature of the OpenGeo Suite is interoperability. By employing OGC standards, any component of the architecture can be replaced with alternatives and vice versa. Component interoperability provides organizations a way to increase their return on investment by utilizing existing software infrastructure. For example, PostGIS can be interchanged with Oracle Spatial, SQL Server Spatial, DB2 Spatial, or ArcSDE. For web map access, GeoServer can be substituted with MapServer, ArcGIS Server, MapGuide, or any other WMS-capable map renderer. For web feature access, GeoServer can interchange with Ionic Red Spider, CubeWerx or any other fully featured Web Feature Server (WFS). Finally, for user map components, OpenLayers can substitute for Google Maps, Bing Maps, and other components.

For truly agile and innovative breakthroughs, organizations are discovering that it is wise to look beyond the old guard, proprietary software vendors.

MAPMETER: MANAGING SPATIAL IT IN PRODUCTION

Managing geospatial web services is not an easy task. In fact, many systems administrators are unable to measure their services performance metrics. For instance, how many requests services are getting, or if additional servers need to be brought online. Boundless is alleviating these pains with Mapmeter, an administration and management tool for analyzing GeoServer systems and enabling organizations to monitor the health of production deployments, optimize applications during development, and diagnose critical issues. With additional management data, administrators can better — and more cost effectively — make decisions about their geospatial deployments.

GEOGIT: MANAGING GEOSPATIAL DATA

With the proliferation of geospatial data, agencies across the defense and intelligence communities are drowning in derivative data and poor data management practices. GeoGit aims to solve this problem. It draws inspiration from the distributed version control system Git, but adapts its core concepts to handle the particular versioning issues of geospatial data. Users of GeoGit will be able to import raw geospatial data (currently from Shapefiles, PostGIS, or SpatiaLite) into repositories where every change to the data is tracked, can be viewed in a history, reverted to older versions, branched into sandboxed areas, merged back in, and pushed to remote repositories.

QGIS: A POWERFUL ALTERNATIVE FOR DESKTOP ANALYSIS

For decades, desktop GIS solutions have been the tool of choice for creators and analysts of geospatial information. While this traditional GIS concept is limited, forcing all spatial data through a specialized workflow, there are still many geospatial professionals that need a tool for accessing data or for running analytic processes that are too cumbersome for the web. To assist these professionals, Boundless supports QGIS, an open source desktop GIS tool with feature parity, stability, and ease of use that is beginning to rival proprietary desktop solutions. It offers a truly open alternative that lowers total cost of ownership, has no license fees, and runs on all common desktop operating systems.

THE OPEN SOURCE BENEFIT

Geospatial applications built with open source components and assembled into modern web architectures are becoming more and more common in the defense and intelligence communities. Given the robust functionality, flexible deployment, and lowered total cost of ownership of open source, this comes as no surprise. To support members of these communities in the adoption of supported open source solutions, companies like Boundless offer a wide variety of services. Unlike other software development models, open source can create a "virtuous cycle" where individual players may fund development of a feature that will benefit many others, furthering functionality and adoption. Q

***Boundless**, formerly known as OpenGeo, is an IQT portfolio company that provides open source geospatial software. To learn more, visit www.boundlessgeo.com.*

REFERENCES

[1] OpenGeo. *FCC Utilizes Open Source Tools to Create the National Broadband Map.* N.p., 28 Mar. 2011. Web. http://opengeo.org/publications/fcc/

IN THE NEWS

The *IQT Quarterly* examines trends and advances in technology. IQT has made a number of investments in analytic and geospatial tools, and several companies in the IQT portfolio are garnering attention for their unique solutions.

Palantir Technologies

IQT portfolio company Palantir develops software that provides visualization of data from large and disparate sources. The software has recently received national attention for its role in helping U.S. Immigration and Customs Enforcement (ICE) combat human trafficking, and is currently used for case management and investigative work. Palantir is used by a number of other state and local law enforcement agencies, including the New York City Mayor's Office of Special Enforcement. Palantir is headquartered in Palo Alto, California. IQT initially invested in the company in September 2005. **www.palantir.com**

Platfora

Platfora is an IQT portfolio company that has developed business intelligence applications that allow users to interact with large datasets in Apache Hadoop. The platform allows business users to access and query large amounts of information more efficiently, and at faster speeds than before. As a rising star in the business intelligence market, Platfora has garnered significant attention for its approach, including profiles in GigaOM, *The Register,* and the *Wall Street Journal.* Platfora is headquartered in San Mateo, California, and joined the IQT portfolio in September 2011. **www.platfora.com**

Cloudera

IQT portfolio company Cloudera provides a suite of tools designed to make it easier for organizations to process and interact with large datasets. The rapidly growing company has been making headlines as it develops partnerships, including a new engagement with big data analytics provider Appfluent. The relationship is aimed at improving data migration to Hadoop, the platform that sits at the heart of Cloudera's technology. The company is headquartered in Palo Alto, California. IQT initially invested in Cloudera in December 2010. **www.cloudera.com**

TerraGo

TerraGo is a developer of popular geospatial integration and collaboration solutions, based on its GeoPDF technology. The company recently ventured into place-based data discovery and visualization through its acquisition of Geosemble in 2012. Its new launch, TerraGo Vision, which integrates this place-based technology into a more comprehensive intelligence platform, has garnered attention from media outlets including InfoTech News and NewsEdge. TerraGo is headquartered in Atlanta, Georgia and has been part of the IQT portfolio since August 2006. **www.terragotech.com**

IQT®

IN·Q·TEL