

GeoIntelligence: Data Mining Locational Social Media Content for Profiling and Information Gathering

Peter Hannay¹ & Greg Baatard²
Edith Cowan University
¹p.hannay@ecu.edu.au
²g.baatard@ecu.edu.au

Abstract

The current social media landscape has resulted in a situation where people are encouraged to share a greater amount of information about their day-to-day lives than ever before. In this environment a large amount of personal data is disclosed in a public forum with little to no regard for the potential privacy impacts. This paper focuses on the presence of geographic data within images, metadata and individual postings. The GeoIntelligence project aims to aggregate this information to educate users on the possible implications of the utilisation of these services as well as providing service to law enforcement and business. This paper demonstrates the ability to profile users on an individual and group basis from data posted openly to social networking services.

Keywords

geotagging, geocoding, GPS, security, smartphone, Facebook, Twitter, Flickr, EXIF

INTRODUCTION

The widespread implementation of technologies such as GPS, A-GPS and other location based services in smartphones, cameras and portable music players has resulted in a plethora of location-aware devices. This functionality has been utilised by numerous developers in applications and services that go well beyond the domains that typical users expect to be location-aware (Junglas and Watson 2008). While most users understand that mapping and navigation programs require location-awareness, they may not realise that said capabilities are utilised in some games, social networking applications, camera functionality and numerous other applications and services (Wagner, Lopez et al. 2010).

The proliferation of locational data in publically available information has been encouraged largely by social media services such as Facebook, Google+, Twitter, Flickr and Foursquare. Such services allow users to share information in the form of text, images and videos. Although the availability of the information can be controlled, a significant proportion of users across these services make it public – either through choice, apathy or ignorance (Lindqvist, Cranshaw et al. 2011). A concept known as “geotagging” allows for the user’s locational data to be published as metadata alongside the information being shared. The locational data provides context to the information being shared, and can be used to “check in” to venues. This paper investigates the scope of locational data provided via such services – the feasibility of mining the data and the importance of raising awareness of the quantity of data available via these services (Freni, Vicente et al. 2010; Elwood and Leszczynski 2011; Lindqvist, Cranshaw et al. 2011).

HOW GEOTAGGING WORKS

The concept of geotagging originated in high-end digital cameras, with integrated or peripheral-based GPS receivers allowing them to encode locational data in image files created with the camera. The mid-2000s saw smartphones with integrated cameras and location-awareness features reach maturity, and the geotagging of images taken on such devices became commonplace. Modern social media services make use of locational data in images and also utilise browser and operating system-based techniques to geotag user-submitted content. Details of how geotagging is implemented in these systems are presented below.

Geotagging of Images

Locational data in images is typically stored as metadata in Exchangeable Image File Format (EXIF) – the standard format used to store metadata in image files taken using a digital camera. EXIF data is stored as a series of tags and associated values in the header of an image file. Common tags include the make and model of the camera and various camera settings such as exposure time and focal length. The EXIF data can also contain information that uniquely identifies the device that has taken the image (JEITA 2002). Tags for locational data include latitude, longitude and altitude. This information can be populated at image creation time when using a

location-aware device, or added manually at a later date. Adding locational EXIF data to an image post-creation typically involves using software to correlate timestamps of images with timestamps in the log of a GPS receiver carried with the camera while pictures are taken. EXIF data can be read and utilised by various applications, services and programming languages, an example of this is shown in Figure 1 below.



```

GPS Version ID      : 2.3.0.0
GPS Latitude Ref    : South
GPS Latitude        : 37 deg 48' 14.72"
GPS Longitude Ref   : East
GPS Longitude       : 144 deg 58' 21.76"
GPS Altitude Ref    : Below Sea Level
GPS Altitude        : 106 m
GPS Time Stamp      : 04:14:06
GPS Satellites      : 3
GPS Status          : Measurement Active
GPS Measure Mode    : 2-Dimensional
GPS Img Direction Ref : Magnetic North
GPS Img Direction   : 225
GPS Map Datum       : WGS-84
GPS Area Information : MELBOURNE MUSEUM
GPS Date Stamp      : 2012:04:11
    
```

Figure 1 – JPEG Image and corresponding GPS EXIF Data

In the case of JPEG files there are a number of tags used to indicate various pieces of data stored within. These tags are indicated by value of 0xFFXX, where XX is an identifier for the specific type of tag. The starting tag for image data is 0xFFD8 and the ending is 0xFFD9, anything outside of this is likely to be some form of meta data, as the information between these flags can include any value, it is important to exclude it prior to searching for other tags (Vermeulen 2005). The tag for EXIF data is 0xFFE1, followed by a value indicating the size of the EXIF record and followed by the EXIF data itself, an example is shown in Table 1 (Vermeulen 2005).

Table 1 – A JPEG Header showing EXIF tag (green) and record size (blue).

ff d8	ff e1	81 9c	45 78	69 66	00 00	49 49	2a 00Exif..II*.
08 00	00 00	0f 00	0f 01	02 00	0a 00	00 00	c2 00
00 00	10 01	02 00	08 00	00 00	cc 00	00 00	12 01
03 00	01 00	00 00	01 00	00 00	1a 01	05 00	01 00
00 00	d4 00	00 00	1b 01	05 00	01 00	00 00	dc 00

While an increasing number of digital cameras are location-aware, smartphones are currently the most common locational-aware devices with camera capabilities – a trend which shows no sign of abating in the foreseeable future. The iPhone, Android and Windows Phone 7 platforms have support for geotagging of images included by default in the operating system. Geotagging is disabled by default on the iPhone and Windows Phone 7 devices, however the user will be prompted to enable the feature upon first launch of the camera application (Figure 2) (Valli and Hannay 2010).

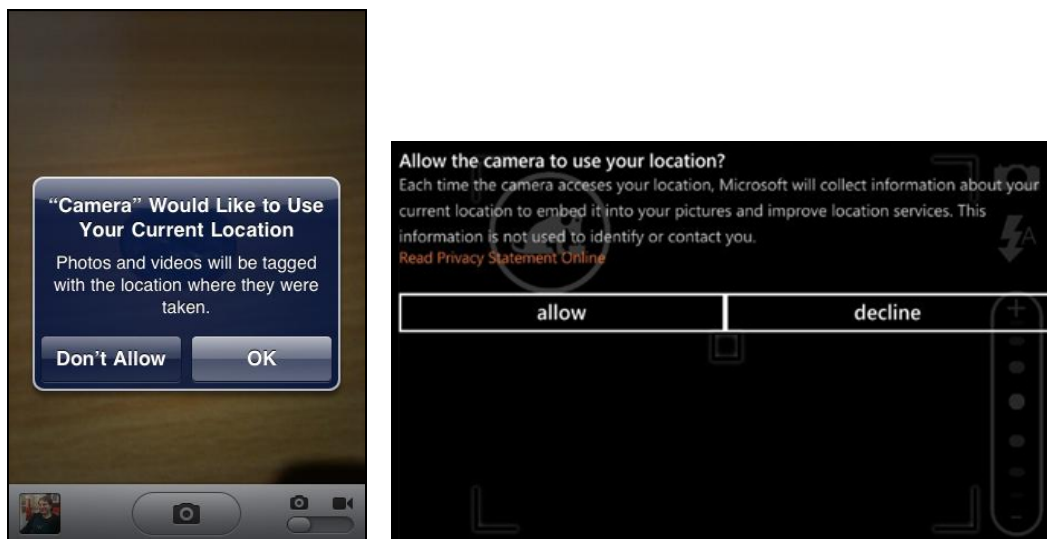


Figure 2 – iPhone (left) and Windows Phone 7 (right) camera location permission prompts

Geotagging is also disabled by default on Android devices, and must be manually enabled in the camera settings (Figure 3). Similar scenarios exist in other smartphone operating systems such as those of Palm and Blackberry devices. Although geotagging is disabled by default on most platforms, it is likely that users will enable the feature without considering its implications (Friedland and Sommer 2010).

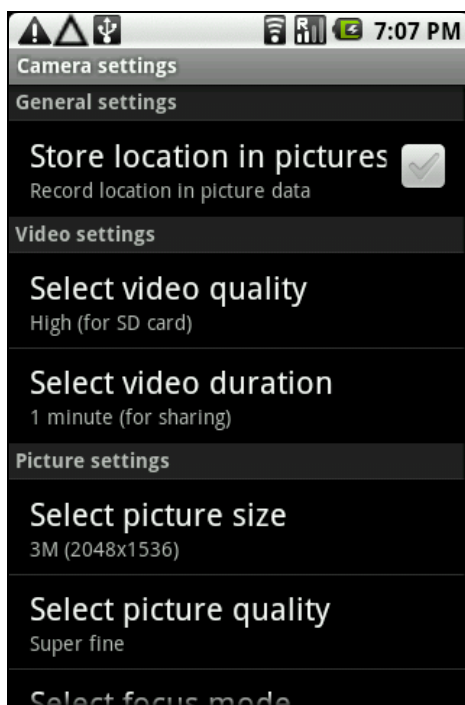


Figure 3 – Geotagging settings for Android

Geotagging in Facebook

Facebook is currently the dominant social networking platform, allowing users to share textual and multimedia content with others. In August of 2010, Facebook introduced a feature named “Places” that allows users to “check in” when using Facebook from a location-aware device such as a smartphone. When a user checks in, they are presented with a list of nearby user-defined places based on their current location. The user selects (or creates) the appropriate place and optionally describes what they are doing to complete the check in (Facebook 2012).

By Facebook’s current default privacy settings, this information is only shared with members of the user’s friend list. This, and the obfuscation of precise locational data via user-defined places, minimises the public disclosure of locational data via Facebook. Locational data is *not* included in images or video uploaded to Facebook, even if the source image contained locational EXIF data.

Geotagging in Google+

Google+ is a social networking platform in beta at the time of writing, offering similar features to Facebook. All posts made to Google+ can be geotagged, and this is achieved using the W3C’s Geolocation API:

The Geolocation API defines a high-level interface to location information associated only with the device hosting the implementation, such as latitude and longitude. The API itself is agnostic of the underlying location information sources. Common sources of location information include Global Positioning System (GPS) and location inferred from network signals such as IP address, RFID, WiFi and Bluetooth MAC addresses, and GSM/CDMA cell IDs, as well as user input.

(W3C 2010)

Hence, while a geotagged post from a desktop computer with no wireless capability may determine a coarse location via the user's IP address (this technique is prone to error), a geotagged post from a smartphone with GPS capabilities will likely result in much more accurate locational data (Figure 4).

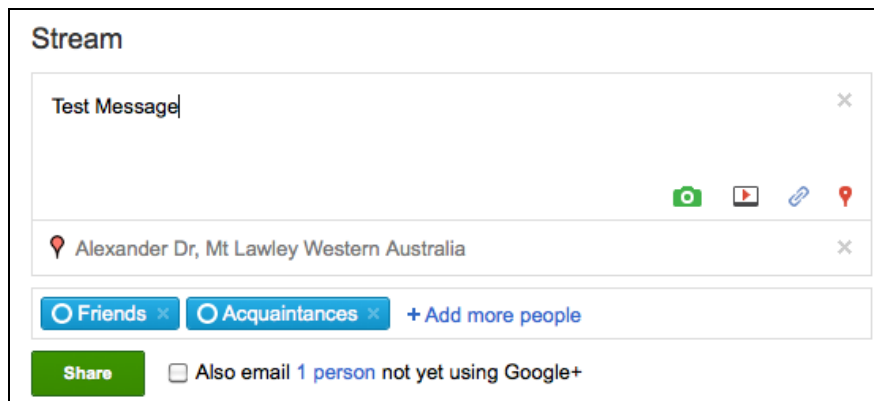


Figure 4 – An example of locational information being included in a Google+ post

Google+ also supports geotagged images via locational EXIF data, however this feature is disabled by default and must be manually enabled (Figure 5).

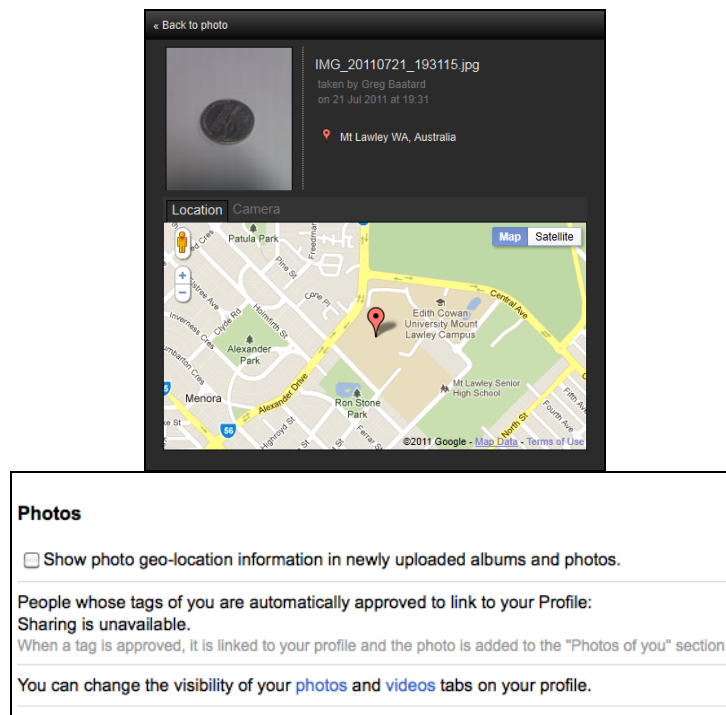


Figure 5 – Examples of geotagged data from Google+ (left) and permissions (right)

Similarly to Facebook, posts made in Google+ are *not* publically disclosed by default.

Geotagging in Twitter

Twitter is a “micro-blogging” service that allows users to make short text-based posts. Posts can be restricted so that only the users who “follow” the poster can view them; however the default settings make all posts public. In November of 2009, Twitter added the ability to include locational data in posts. This data can be viewed by anyone who is able to access the post, as shown in Figure 6.

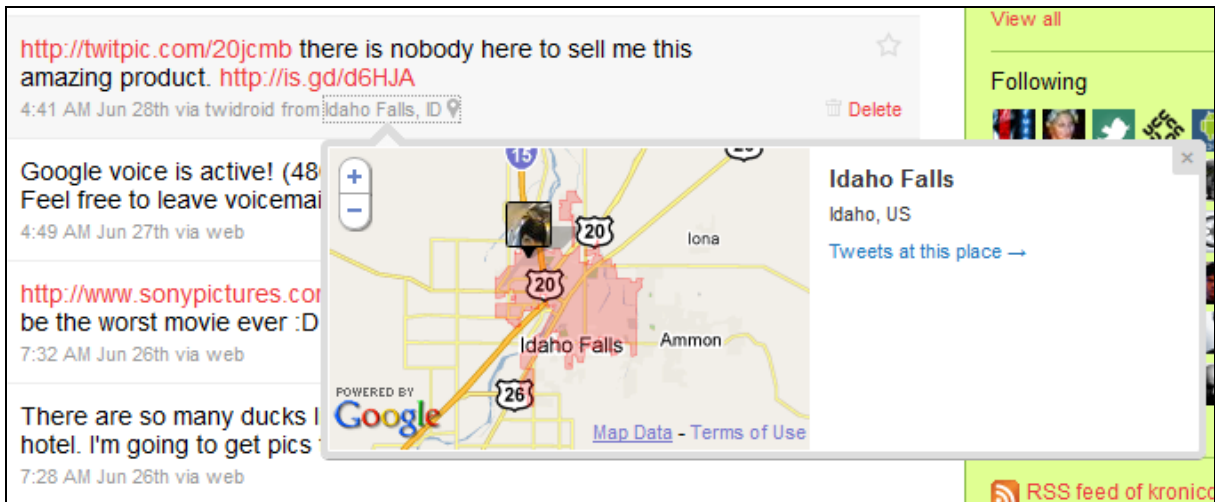


Figure 6 – An example of a geotagged twitter posting via the web interface

Location-inclusive Twitter posts include either a “Place ID” referring to a human-recognisable location, or the user’s current latitude and longitude at the time of posting (Twitter 2012). While posts made via the Twitter website utilise the user’s IP address to determine a coarse location, there are numerous Twitter client applications available for all smartphone platforms that use GPS and A-GPS to provide precise location data.

The default Twitter account settings do not enable the geotagging of posts, and most client applications on smartphones maintain this. As well as enabling the feature for all posts, individual posts can be geotagged.

Geotagging in Flickr

Flickr is a popular example of one of many photo-sharing websites, allowing users to upload images which can be viewed either by a restricted group of people or made publically available. Flickr uses EXIF data in uploaded images to provide locational information. By default, locational EXIF data is not imported. However, the recommended setting once the feature is enabled is to make the data publically available (Figure 7).

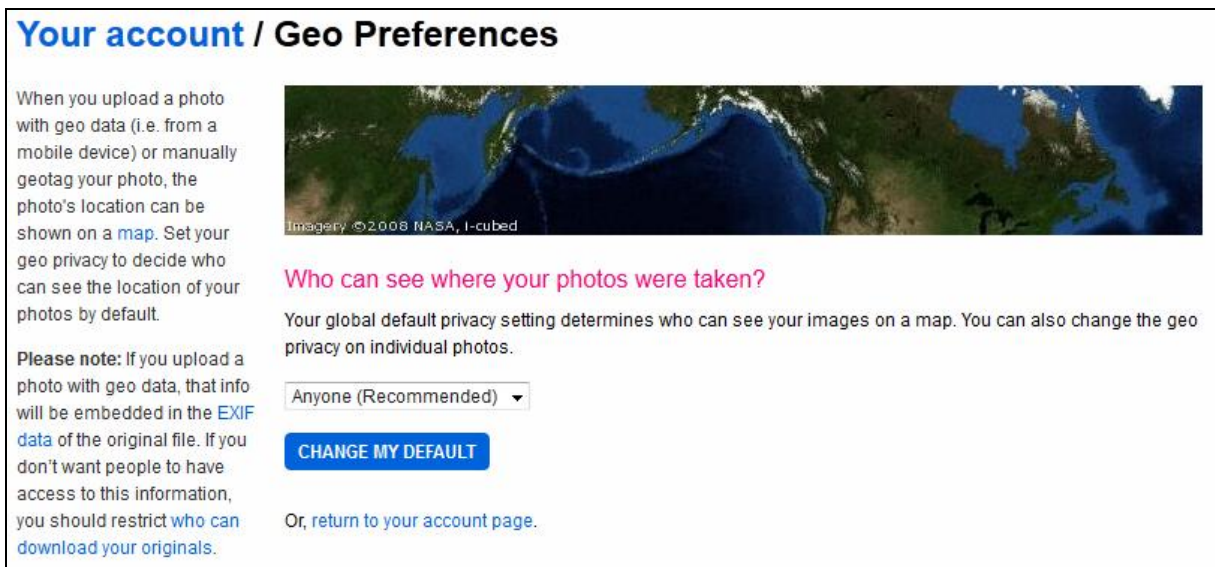


Figure 7 – Locational permissions settings for Flickr

Users can control the presence and availability of locational data on a per-picture basis. When viewing an image on the Flickr website, locational data (and other EXIF data) can be viewed and the location can be plotted on a map (Figure 8) (Flickr 2012).

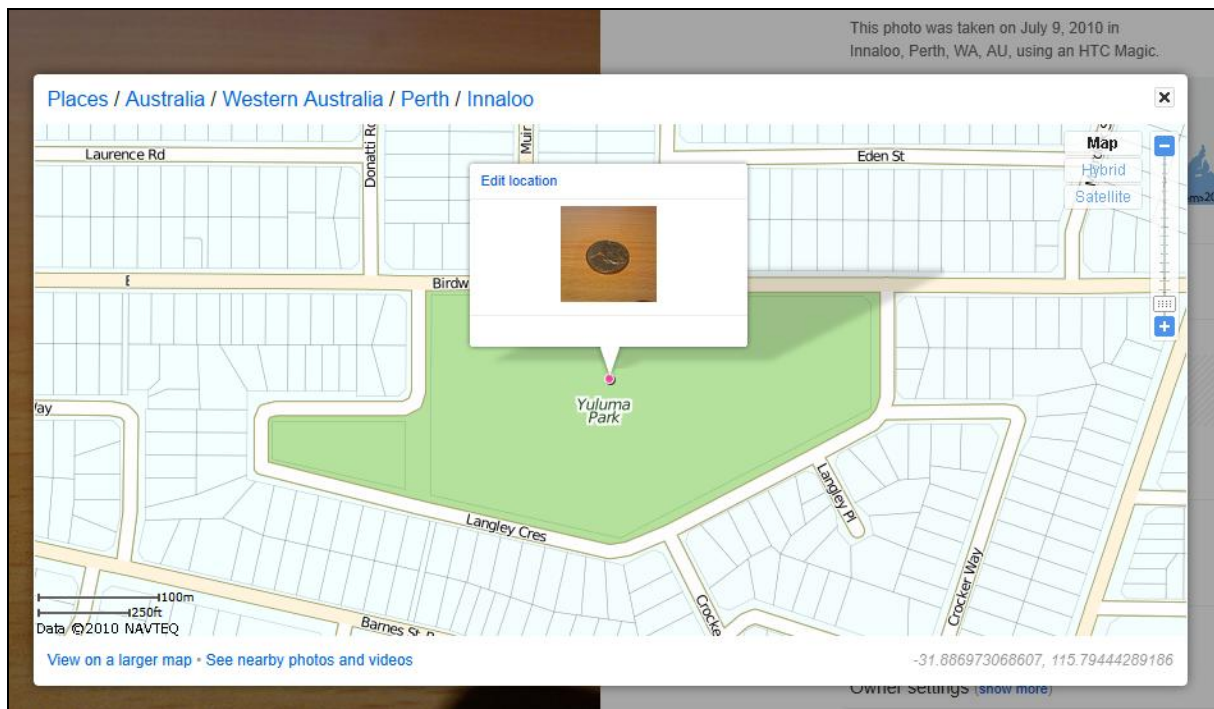


Figure 8 – Example of a geotagged image in Flickr as seen via the web interface

Geotagging in Foursquare

Foursquare is a location-based service available on mobile devices such as smartphones. Similar to (but pre-dating) Facebook’s “Places” feature, it allows users to check in at their current location in order to announce their presence and share short textual information.

While check ins are to user-defined places, the data submitted to Foursquare includes the user’s latitude and longitude. The visibility of a check in can be limited to the user’s friend list, but are public by default. The latitude and longitude submitted in a check in is also publically available via the Foursquare API (Figure 9) (Foursquare 2012).

```
id: "4bf34bb3a4de2f5ea2e615be",
createdAt: 1274235827,
type: "checkin",
timeZone: "Australia/Perth",
venue: {
  id: "4bf34baf98ac0f4779c962a8",
  name: "Western Power",
  contact: {},
  location: {
    city: "East Perth",
    state: "Western Australia",
    postalCode: "6004",
    lat: -31.943139,
    lng: 115.880149
  }
},
```

Figure 9 – Output from the Foursquare API showing latitude and longitude.

Foursquare offers the ability to link a user’s Foursquare and Twitter accounts – automatically posting details of a check in to Twitter. Many users make use of this feature, substantially increasing the number of geotagged Twitter posts.

While there are numerous other services and websites that make use of geotagged content, the examples discussed above present an appropriate overview of the current environment. Although some services such as Facebook and Google+ do not share geotagged information by default, other services do. With each of the discussed services boasting over ten million users, there is a large amount of publically available geotagged social media content (Sorathia and Joshi 2009).

INTRODUCING GEOINTELLIGENCE

The concept of “GeoIntelligence” is simple: Mining publically available geotagged social media content. By aggregating the content and its associated locational data, a plethora of information can be deduced. The data can be analysed from numerous perspectives including time, location, user and keyword. Combinations of these are also possible, for example identifying users who were within 50km of a particular incident within two hours of its occurrence.

Obtaining geotagged social media content in a mineable format not particularly onerous thanks to the APIs available for developers to integrate with social media services. This allows code to be written which retrieves content – including locational data – from the services, such as posts to Twitter and check ins to Foursquare. Retrieving this data is standard functionality in the API, and does not breach the terms of service.

Social media service APIs present two main techniques for data mining. One technique is to write code which makes queries to the service(s), attempting to find the desired information by searching the content stored by the service(s). Although this technique is resource-light, it is problematic as limits are often imposed upon such queries by the API – often restricting the search criteria, frequency of searches and number of results. The second technique involves the ongoing retrieval of all geotagged content from the service(s) and archiving of the data in a local database. The database can then be queried without limitations. Obviously, such an approach is very resource-heavy, requiring substantial amounts of bandwidth, storage and processing power. Resource usage can be minimised if a relevant subset of the content can be identified, for example only retrieving and storing content from a certain region or timeframe. Despite the resource usage, the second approach is favourable as it allows for limitless querying and makes it easier to combine data from different social media services.

As a proof of concept, the authors created a web-based tool which autonomously collects all publically available geotagged Twitter posts. The details of each post are stored in a database and include the textual content of the post, latitude and longitude, post time, poster’s username and provided real name and the name of the client application used to make the post. The tool is written in the PHP scripting language and stores its content in a MySQL database. A web-based interface allows users to see recently collected posts, view each post on a map, view the poster’s Twitter profile and view all geotagged posts by a user on a map (Figure 10).

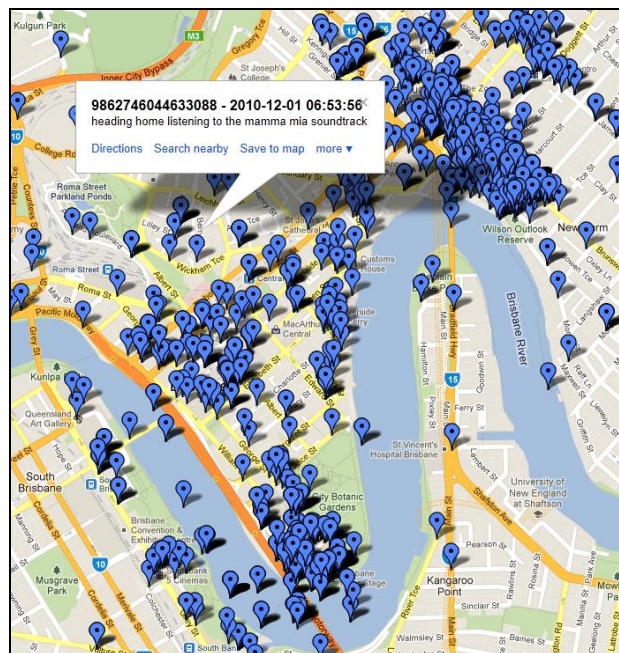


Figure 10 – A small fraction of the geotagged Twitter posts made by a particularly prolific user

Searches can be performed to find posts by keyword, username and location, with all matching posts being displayed on a map. All data displayed on maps is in Keyhole Markup Language (KML) format, an XML-based format designed to express points of interest on maps (KML 2.2 SWG Group 2010). Google Maps is used to display the KML data, which is generated by the web-application code.

While the tool in its current form represents a fairly basic implementation of GeoIntelligence, it has been able to demonstrate the concept's potential.

POTENTIAL USES OF GEOINTELLIGENCE

The authors have identified three main potential uses of GeoIntelligence, presented in Table 2.

Table 2 – Potential uses of GeoIntelligence

Potential Use	Example of Use
Law Enforcement	Identify potential witnesses to an incident by finding posts within a specific location and timeframe Identify or locate persons of interest by examining posts by certain users and determining a profile / daily routine
Business Intelligence and Marketing	Read posts and determine demographics of clientele who post from within or nearby a business Determine common interests and frequent locations of clientele who post from within or nearby a business
Privacy Awareness and Education	Demonstrate the ability to deduce personal information such as home address, school and daily routine in order to raise awareness of the privacy implications of geotagging

The usage examples provided are but a few of the potential applications of GeoIntelligence. If a user regularly makes publically accessible geotagged posts, it can be a trivial task to identify details such as their home address, workplace, school, public transport route, hobbies and interests and their favourite bars and restaurants. With further analysis and an appropriate amount of data, it is possible to plot a user's daily routine and deduce where they are likely to be at a particular time. Law enforcement agencies could use this information to identify persons of interest such as potential witnesses or suspects of an incident. While this is a theoretically positive usage of the data, it is easy to envisage how it could be misused or considered an invasion of privacy by some.

Marketing agencies or business could use this geotagged information for commercial gain – identifying customer demographics, interests and other locations their customers frequent. This could likely be achieved without the need for overly specific profiles of individuals. Such usage of the data is relatively benign, and comparable than the techniques used for targeted advertising on the Internet.

There exists the potential for such information to be used for malicious or nefarious purposes, for example stalking, harassment or burglary. The potential for misuse of geotagged social media content was highlighted by the website "pleaserobme.com", which utilised geotagged Foursquare and Twitter posts to advertise when people were not at home (Borsboom, Amstel et al. 2010; Freni, Vicente et al. 2010). In order to mitigate the potential for misuse of geotagged social media content it is important that public awareness is raised regarding the privacy and security implications of sharing such data. To this end, it is hoped that GeoIntelligence will be employed as a tool to raise awareness of this issue.

A number of similar applications, both web and smartphone-based, have been released over the past few years. These include Stalqer, an iPhone application which uses locational data from Facebook to track a user's friends (Hathaway 2009; Slattery 2009) and I Can Stalk U, a web-based awareness tool which utilised locational EXIF data in images posted to TwitPic – a service allowing photos and videos to be shared via Twitter (Jackson and Pesce 2010). More recently, an iPhone dating application named Girls Around Me, which uses publically available data from Foursquare to display the location, photographs and details of nearby women has caused controversy (i-Free Innovations 2012; Magid 2012; Wood 2012). Even the United States Army has released an article outlining the privacy and security risks of geotagging (Rodewig 2012).

LIMITATIONS & FUTURE DEVELOPMENTS

GeoIntelligence is of course not without limitations. It relies upon individuals making publically available geotagged posts. While this is becoming increasingly common as social media and location-aware smartphones become prevalent, it cannot be seen as representative sample of a population – even amongst the demographics where such activities are widespread. Furthermore, many users only geotag some of their posts to social media services, resulting in a data set which could be considered incomplete or inconsistent. GeoIntelligence can therefore not be used to determine information that requires a representative population sample or complete data set to be of value.

The future development of GeoIntelligence tools will be focused on the implementation and optimisation of analytical functions. Current plans for these functions include the ability to identify common daily routes and locations, subsequently highlighting deviations from these as items of interest. Additionally the ability to identify relationships between individuals based on locational proximity in a context sensitive manner, for example sharing locational proximity at a residential address may have more weighting than a workplace or restaurant (Crandall, Backstrom et al. 2010). The inclusion of additional data sources will assist in achieving the aforementioned goals, with reverse geo-coding databases and additional sources of social media information the key targets for this integration.

CONCLUSION

GeoIntelligence provides a demonstration of the potential value of publically accessible geotagged content from social media services. This potential exists within but is not limited to law enforcement, business intelligence, marketing and awareness. It should be noted however that GeoIntelligence highlights a serious issue in terms of the quantity of locational data being published publicly by individuals, as well as the potential for misuse of that data. As discussed, GeoIntelligence is not alone in a push to raise awareness of the potential risks of geotagging.

The increase in pervasiveness of locational-aware smartphones, cameras, tablets and laptops has caused a rapid increase in the amount of data observable via GeoIntelligence. As such the value and potential implications of misuse of this data have both seen dramatic escalations. It is important that users evaluate their use of these technologies and become aware of the amount of data they are making available. It is hoped that future developments in the area of GeoIntelligence is of value to society as a whole, but also helps to educate users of social media services as to the implications of sharing locational information.

REFERENCES

- Borsboom, B., B. v. Amstel, et al. (2010). "Please Rob Me." Retrieved April 10th, 2012, from <http://pleaserobme.com/>.
- Crandall, D. J., L. Backstrom, et al. (2010). "Inferring social ties from geographic coincidences." Proceedings of the National Academy of Sciences **107**(52): 22436-22441.
- Elwood, S. and A. Leszczynski (2011). "Privacy, reconsidered: New representations, data practices, and the geoweb." Geoforum **42**(1): 6-15.
- Facebook. (2012). "Graph API." Retrieved April 10th, 2012, from <https://developers.facebook.com/docs/reference/api/>.
- Flickr. (2012). "The App Garden - API Documentation." Retrieved April 10th, 2012, from <http://www.flickr.com/services/api/>.
- Foursquare. (2012). "API Documentation." Retrieved April 10th, 2012, from <https://developer.foursquare.com/>.
- Freni, D., C. R. Vicente, et al. (2010). Preserving location and absence privacy in geo-social networks. Proceedings of the 19th ACM international conference on Information and knowledge management. Toronto, ON, Canada, ACM: 309-318.
- Friedland, G. and R. Sommer (2010). Cybercasing the joint: on the privacy implications of geo-tagging. Proceedings of the 5th USENIX conference on Hot topics in security. Washinton, DC, USENIX Association: 1-8.

- Hathaway, J. (2009). "Stalqer iPhone app finds people, whether they sign up or not." Switched. Retrieved April 10th, 2012, from <http://downloadsquad.switched.com/2009/10/28/stalqer-iphone-app-finds-people-whether-they-sign-up-or-not/>.
- i-Free Innovations. (2012). "Girls Around Me." Retrieved April 10th, 2012, from <http://girlsaround.me/>.
- Jackson, B. and L. Pesce. (2010). "I Can Stalk U." Retrieved April 10th, 2012, from <http://www.icanstalku.com/>.
- JEITA. (2002). "EXIF 2.2." Retrieved April 10th, 2012, from <http://www.exif.org/Exif2-2.PDF>.
- Junglas, I. A. and R. T. Watson (2008). "Location-based services." Commun. ACM **51**(3): 65-69.
- KML 2.2 SWG Group. (2010). "OGC KML." Retrieved April 10th, 2012, from <http://www.opengeospatial.org/standards/kml/>.
- Lindqvist, J., J. Cranshaw, et al. (2011). I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. Proceedings of the 2011 annual conference on Human factors in computing systems. Vancouver, BC, Canada, ACM: 2409-2418.
- Magid, L. (2012). "Girls Around Me App Is a Reminder To Be Aware What You Share." Forbes. Retrieved April 10th, 2012, from <http://www.forbes.com/sites/larrymagid/2012/04/09/girls-around-me-app-is-a-reminder-to-be-aware-what-you-share/>.
- Rodewig, C. (2012). "Geotagging poses security risks." United States Army. Retrieved April 10th, 2012, from http://www.army.mil/article/75165/Geotagging_poses_security_risks/.
- Slattery, B. (2009). "Stalqer iPhone App Redefines Creepy." PCWorld. Retrieved April 10th, 2012, from http://www.pcworld.com/article/174567/stalqer_iphone_app_redefines_creepy.html.
- Sorathia, K. and A. Joshi (2009). My World – Social Networking through Mobile Computing and Context Aware Application. Intelligent Interactive Assistance and Mobile Multimedia Computing. D. Tavangarian, T. Kirste, D. Timmermann, U. Lucke and D. Versick, Springer Berlin Heidelberg. **53**: 179-188.
- Twitter. (2012). "Streaming API." Retrieved April 10th, 2011, from <http://dev.twitter.com/docs/streaming-api>.
- Valli, C. and P. Hannay (2010). Geotagging Where Cyberspace Comes to Your Place. Proceedings of the 2010 International Conference on Security & Management. Las Vegas, CSREA Press: 627-632.
- Vermeulen, G. (2005, April 1st). "Description of Exif file format." Retrieved February 12th, 2012, from <http://gvsoft.homedns.org/exif/exif-explanation.html>.
- W3C. (2010). "Geolocation API Specification." Retrieved April 10th, 2012, from <http://www.w3.org/TR/geolocation-API/>.
- Wagner, D., M. Lopez, et al. (2010). Hide and seek: location sharing practices with social media. Proceedings of the 12th international conference on Human computer interaction with mobile devices and services. Lisbon, Portugal, ACM: 55-58.
- Wood, M. (2012). "Girls Around Me and the end of Internet innocence." CNet News. Retrieved April 10th, 2012, from http://news.cnet.com/8301-31322_3-57408165-256/girls-around-me-and-the-end-of-internet-innocence/.