# GEOINTELLIGENCE: DATA MINING LOCATIONAL SOCIAL MEDIA CONTENT FOR PROFILING AND INFORMATION GATHERING

**Peter Hannay and Greg Baatard**
Secau – Security Research Centre, School of Computer and Security Science,
Edith Cowan University, Perth, Western Australia

p.hannay@ecu.edu.au, g.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. 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).

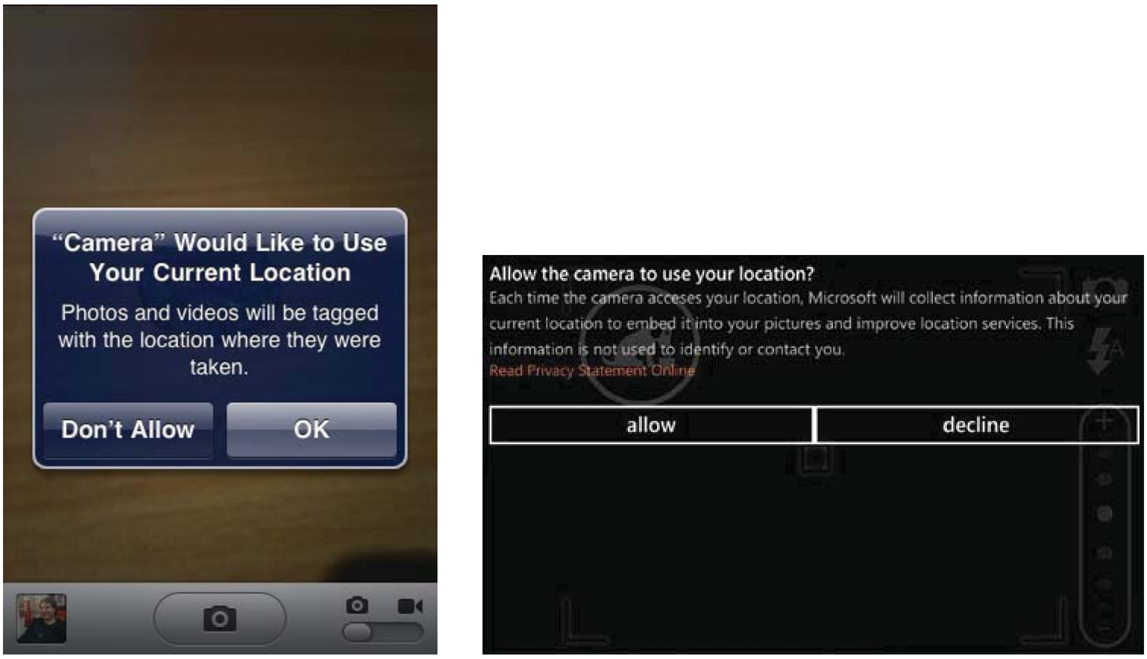
## HOW GEOTAGGING WORKS

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.

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 (Figures 1) (Valli and Hannay 2010).



"Camera" Would Like to Use Your Current Location

Photos and videos will be tagged with the location where they were taken.

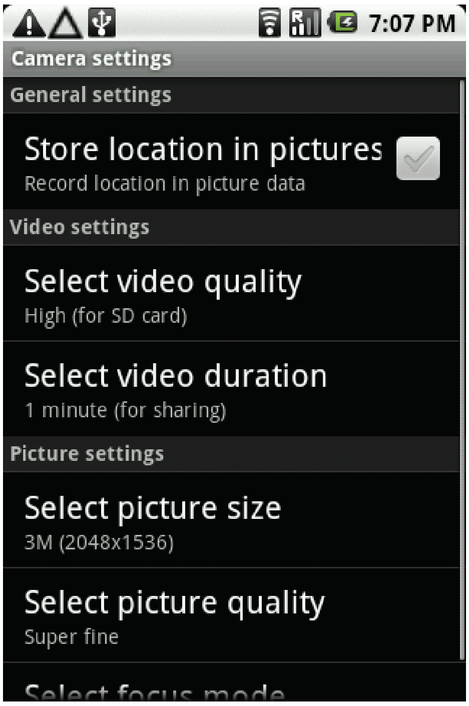
Allow the camera to use your location?

Each time the camera acceses your location, Microsoft will collect information about your current location to embed it into your pictures and improve location services. This information is not used to identify or contact you.

Read Privacy Statement Online

*Figure 1 – 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).



7:07 PM

Camera settings

General settings

Store location in pictures ☑
Record location in picture data

Video settings

Select video quality
High (for SD card)

Select video duration
1 minute (for sharing)

Picture settings

Select picture size
3M (2048x1536)

Select picture quality
Super fine

Select focus mode

*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 2011).

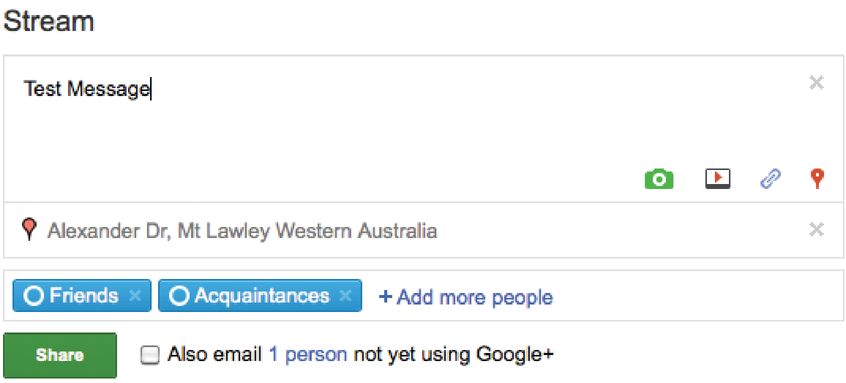
By default, Facebook’s privacy settings only make this information available to members of the user’s friend list. This, and the obfuscation of precise locational data via user-defined places, minimises the public disclosure of locational data via Facebook. Locational data is *not* included in images or video uploaded to Facebook.

**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).



Stream

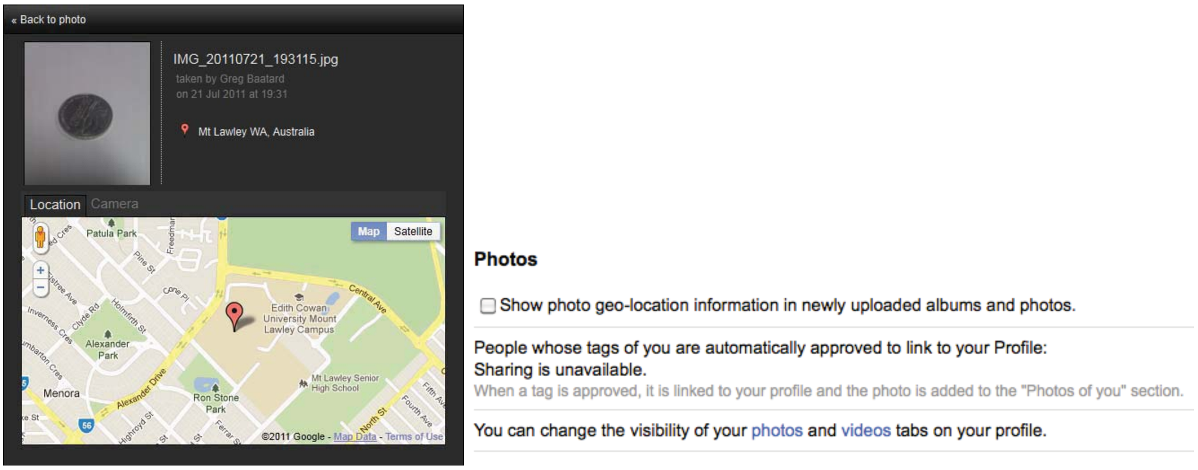
Test Message

Alexander Dr, Mt Lawley Western Australia

□ Also email 1 person not yet using Google+

*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).



IMG_20110721_193115.jpg
taken by Greg Baatard
on 21 Jul 2011 at 19:31

Mt Lawley WA, Australia

Location
Camera
Map
Satellite
Patula Park
Edith Cowan University Mount Lawley Campus
Alexander Park
Menora
Ron Stone Park
Mt Lawley Senior High School
Central Ave
Alexander Drive
©2011 Google - Map Data - Terms of Use

**Photos**

□ Show photo geo-location information in newly uploaded albums and photos.

People whose tags of you are automatically approved to link to your Profile:
Sharing is unavailable.
When a tag is approved, it is linked to your profile and the photo is added to the "Photos of you" section.

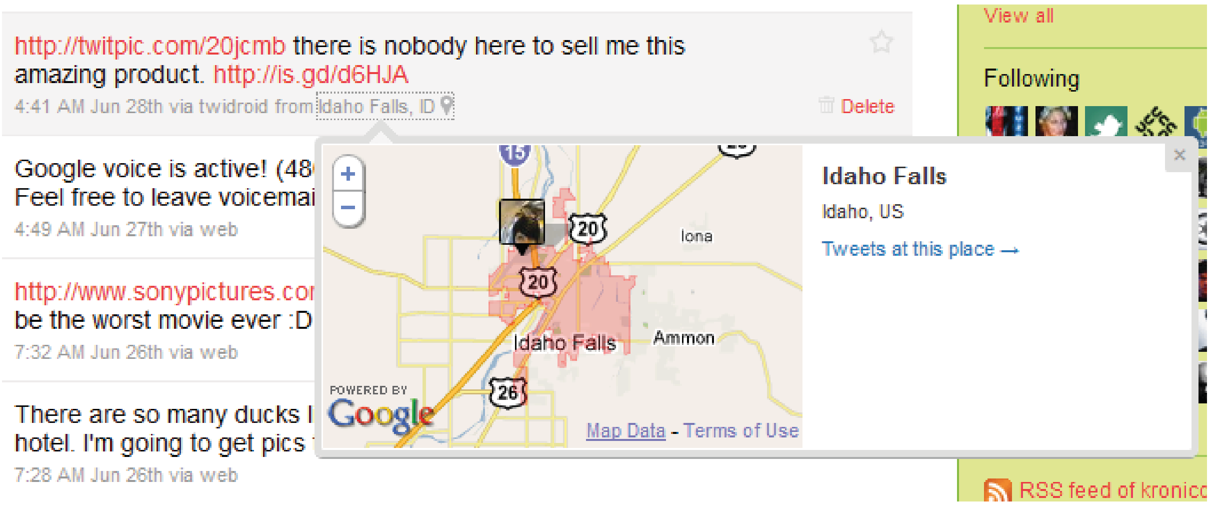
You can change the visibility of your photos and videos tabs on your profile.

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

Similar to Facebook, posts made in Google+ are *not* disclosed publically 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.



http://twitpic.com/20jcmb there is nobody here to sell me this amazing product. http://is.gd/d6HJA
4:41 AM Jun 28th via twidroid from Idaho Falls, ID

Google voice is active! (48
Feel free to leave voicemai
4:49 AM Jun 27th via web

http://www.sonypictures.co
be the worst movie ever :D
7:32 AM Jun 26th via web

There are so many ducks i
hotel. I'm going to get pics
7:28 AM Jun 26th via web

Idaho Falls
Idaho, US
Tweets at this place →

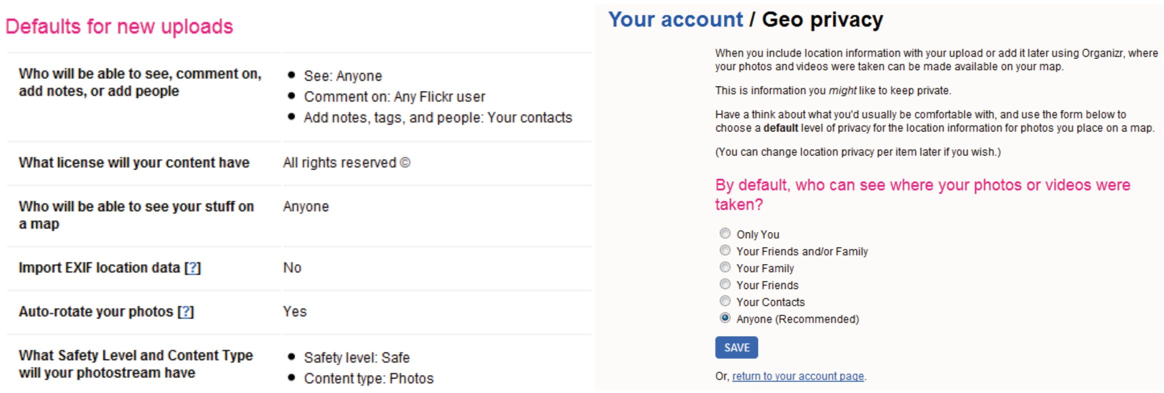
*Figure 6 – An example of a geotagged twitter posting via the web interface*

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 2011). 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. The default, and recommended, settings are to include locational data and make all content publically available (Figure 7).



Defaults for new uploads

| | |
|---|---|
| **Who will be able to see, comment on, add notes, or add people** | • See: Anyone<br>• Comment on: Any Flickr user<br>• Add notes, tags, and people: Your contacts |
| **What license will your content have** | All rights reserved © |
| **Who will be able to see your stuff on a map** | Anyone |
| **Import EXIF location data [?]** | No |
| **Auto-rotate your photos [?]** | Yes |
| **What Safety Level and Content Type will your photostream have** | • Safety level: Safe<br>• Content type: Photos |

Your account / Geo privacy

When you include location information with your upload or add it later using Organizr, where your photos and videos were taken can be made available on your map.

This is information you *might* like to keep private.

Have a think about what you'd usually be comfortable with, and use the form below to choose a **default** level of privacy for the location information for photos you place on a map.

(You can change location privacy per item later if you wish.)

By default, who can see where your photos or videos were taken?

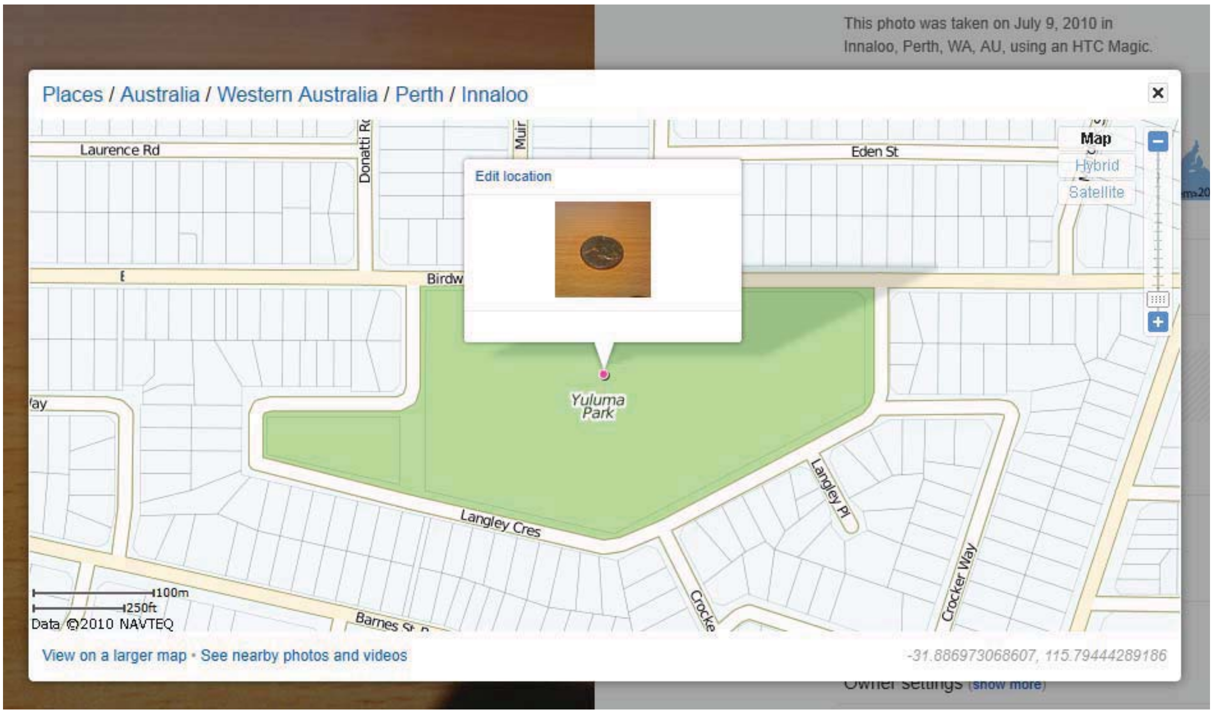
○ Only You
○ Your Friends and/or Family
○ Your Family
○ Your Friends
○ Your Contacts
● Anyone (Recommended)

SAVE

Or, return to your account page.

*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 2011).



This photo was taken on July 9, 2010 in Innaloo, Perth, WA, AU, using an HTC Magic.

Places / Australia / Western Australia / Perth / Innaloo

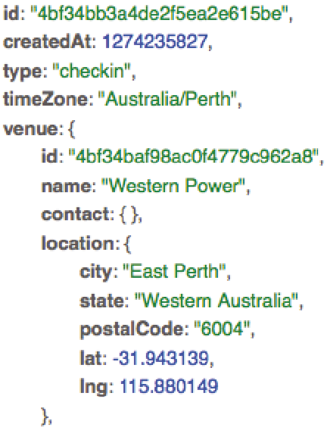
-31.886973068607, 115.79444289186

*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 check ins 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 2011).



```
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 situation. 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).

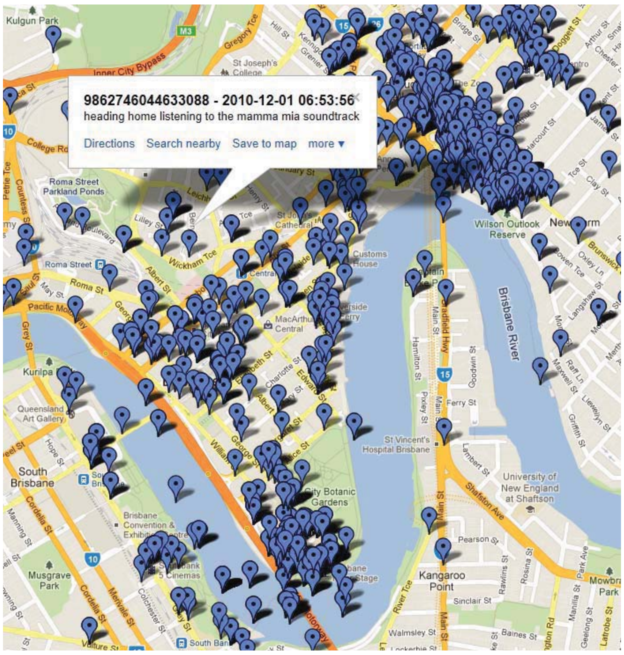
## INTRODUCING GEOINTELLIGENCE

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 or event within a certain timeframe of the same.

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 archived 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).



**9862746044633088 - 2010-12-01 06:53:56**
heading home listening to the mamma mia soundtrack

*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.

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 1.

*Table 1 – Potential uses of GeoIntelligence*

| Potential Use | Example of Use |
|---|---|
| Law Enforcement | Identify potential witnesses to an event by finding posts within a specific location and timeframe<br>Identify or locate suspects 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 a business<br>Determine common interests and frequent locations of clientele who post from within a business |
| Privacy Awareness and Education | Demonstrate the ability to deduce personal information such as home address 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, local shopping centre, hobbies and interests and their favourite bars and restaurants. With further analysis, it is possible to plot a user's daily routine and deduce where they are likely to be at a particular time.

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.

## 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.

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 July 1st, 2011, 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. (2011). "Graph API." Retrieved July 2nd, 2011, from https://developers.facebook.com/docs/reference/api/.

Flickr. (2011). "The App Garden - API Documentation." Retrieved July 1st, 2011, from http://www.flickr.com/services/api/.

FourSquare. (2011). "API Documentation." Retrieved July 4th, 2011, 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.

JEITA. (2002). "EXIF 2.2." 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 June 16th, 2011, 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.

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. (2011). "Streaming API." Retrieved July 2nd, 2011, from https://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.

W3C. (2010). "Geolocation API Specification Editor's Draft 10 February 2010." Retrieved June 13th, 2011, from http://dev.w3.org/geo/api/spec-source.html.

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.