Projecte Fi de Carrera
Anàlisi de perfils d'usuari en xarxes socials

Ismael Florit Zacarias
Enginyeria en Informàtica

Cristina Pérez Solà
Consultora

Data de Lliurament
3 de Gener de 2014

Agraïments

A la meva companya que tant s'ha esforçat a donar-me ànims per aconseguir aquesta titulació i al meu avi que tant va creure en donar-me suport per entrar en aquest sector.

Resum introductori

Perquè la informàtica no només són videojocs, virus i lletres indesxifrables a la pantalla, sinó una ajuda per la societat moderna que ens capbussa en el futur a una velocitat mai vista.

Ara bé, cal tenir en compte que aquesta velocitat pot fer que perdem pel camí la percepció de la realitat i la saviesa de saber que aquesta eina que se'ns ha ofert s'ha de saber fer servir. No hem d'oblidar mai que tota la informació que circula per la xarxa se'ns podria tornar en contra si se'n fa un mal ús.

Per aquest motiu hem d'apendre a no confondre la realitat amb la ficció que ens ofereix el fet d'amagar-nos darrera d'una pantalla, que també es una altra realitat –encara que sigui intangible-.

Provarem doncs de mostrar que tota informació pot ser accessible d'una forma o altra i que disposar d'ella d'una forma organitzada por tornar-se en contra si cau en males mans.

Taula de continguts

1. Introducció

1.1. Introducció del problema a resoldre

La constant evolució de les tecnologies mòbils ens està portant a un canvi d'estil de vida; depenem cada cop més de dispositius cada cop més sofisticats que en permeten fer més coses en menys temps, planificar més tasques amb menys esforç i "sociabilitzar-nos" de forma immediata i remota.

Tota aquesta informació aportada mitjançant els dispositius mòbils, afegeix una característica extra a les xarxes socials, i és que la informació es comparteix de forma pràcticament instantània i en el moment que passa.

Hem de destacar que tota la informació que confiem als nostres aparells mòbils no resideix únicament als dispositius. Habitualment les dades viatgen d'una forma privada o pública a desenes de xarxes socials que queden alimentades per la interacció dels usuaris i posteriorment se'n podrien extreure estadístiques per a diferents usos com estudis de mercat i estudis de perfils d'usuari [1][2] o fraus [3] o assetjaments [4].

El problema que vol exposar aquest projecte és el fet de la vulnerabilitat de les dades personals dels usuaris d'aquests tipus de xarxes socials més directes que segueixen i emmagatzemen dades personals amb una precisió esfereïdora (coneixement de recorreguts d'anada a la feina, de tornada a casa, hores de realització d'esports, creació de patrons per les aficions o gastronomia, entre molts altres).

En termes absoluts, no hi ha cap problema que diferents xarxes socials coneguin totes aquestes dades ja que som nosaltres mateixos, els usuaris, qui compartim aquestes dades amb els nostres "coneguts" de dins d'aquestes xarxes socials. El problema esdevé quan deixem de controlar la visibilitat de les nostres dades que en ocasions és una tasca difícil de configurar bé per la quantitat d'opcions disponibles o fins i tot per una traducció "computeritzada" dels textos de les característiques (on podem deixar perdre algun detall i acabar mostrant dades públicament a la xarxa que no volíem fer).

Així doncs, en aquest projecte volem deixar en evidència que el fet de compartir moltes dades privades en les xarxes socials sense posar gaire interès al nivell de visibilitat del perfil, podria posar en perill la nostra integritat (física o moral). Veure vídeo [5].

1.2. Context global del projecte

El projecte es centra en l'estudi de l'obtenció de dades extretes d'internet sobre els usuaris que es converteixin en "objectiu". Totes les dades que s'analitzaran s'hauran obtingut de forma lícita amb la intenció futura de generar un informe sobre els usuaris (sense cap tipus de pirateig ni accés no autoritzat a les dades privades d'aquests), posant al descobert dades com hores de connexió, hores fora de casa (per feina o lleure) o fins i tot dades precises d'ubicació.

S'ha de tenir en compte que totes les dades trobades a les xarxes socials es relacionaran entre elles però que no sempre estaran completes i que per tant no

ens podem refiar dels resultats i no haurien de servir com a referent en cap altre objectiu que no sigui clarificar la perillositat d'exposar la gran quantitat de dades que un usuari pot arribar a exposar a internet sense saber-ho.

1.3. Objectius del Projecte Final de Carrera

L'objectiu d'aquest projecte final és la definició d'un procediment comprovat de com obtenir dades de les diferents xarxes socials i d'altres fonts d'internet (blogs, webs de referència i documents d'accés públic) per a desenvolupar un perfil o patró d'informació i extreure'n dades de la vida privada d'algun usuari.

Basant-nos en diferents dades clau d'un usuari com poden ser:

- Nom i cognoms de l'usuari
- Correu electrònic
- Nom d'usuari comú o àlies (equivalent en ocasions a l'adreça de correu)
- Altres
    - Id d'usuari extret del codi font d'un client web d'una xarxa social
    - Id d'usuari extret de les bases de dades (generalment no xifrades) dels clients mòbils d'aquestes mateixes xarxes socials.

En les diferents fases del projecte tractarem els avantatges o mancances a l'hora de fer servir cadascuna d'aquestes dades com a punt de partida per extreure informació de l'usuari objectiu de la recerca.

El llistat d'objectius a assolir durant el desenvolupament del projecte son:

1. Cerca i llistat de vulnerabilitats trobades a la xarxa per tal d'extreure dades de les diferents xarxes socials actuals; aquest pas fa referència a la cerca de bugs reconeguts o descoberts per part de les pàgines web que farem servir com a punt de referència per extreure'n dades. També es proposaran algunes comandes de cerca de buscadors (principalment Google) que ens permeten buscar determinats tipus d'arxiu en determinats dominis (que farem servir per buscar possibles arxius annexats en blogs o servidors de compartició d'arxius)
2. Mitjançant les dades clau que puguem obtenir d'alguns usuaris observats se n'intentarà extreure el màxim de dades personals possibles per fer-ne un inventari.
3. Definició d'una plantilla que pugui reflectir de forma clara cadascuna de les dades que s'han pogut extreure de l'usuari i procedimentar el com s'ha pogut arribar a cadascuna d'aquestes dades.
4. En base a les dades trobades i tal com s'hagi pogut procedimentar la seva cerca; cal automatitzar-ne (per totes les vies possibles) l'extracció de dades de la xarxa.

    a. Donat el caire del tipus de cerca (molt manual i basat en el com s'aniran torbant les dades), l'automatització contemplarà més aviat l'estructuració de les dades a extreure, es a dir, indicar una URL o arxiu d'internet i treballar-lo automàticament per fer-ne ús en "l'informe sobre l'usuari"

b. L'extracció de les dades haurà de realitzar-se periòdicament per tal d'obtenir la informació el més actualitzada possible, ja que ens basem en xarxes socials que actualitzen el seu contingut gairebé en temps real.

Es tractarà cadascun d'aquest objectius de forma extensa per exposar les necessitats i resultats de cadascuna de les fases amb l'objectiu de trobar els condicionants que fan que les dades exposades a les diferents xarxes socials, programaris per smartphone o altres (consultes extretes de cercadors, per exemple) siguin una font d'informació privada que tota mesclada tregui a la llum més informació privada de la que els usuaris s'imaginen.

N'extraurem llavors el següent objectiu final del projecte:

5. Com a objectiu final caldrà trobar la manera de protegir les dades personals, donant consells pel registre en xarxes socials, blogs i d'altres, de forma segura per a que no puguin donar peu a la relació de les dades personals i patir l'anomenat ciber-assetjament.

    a. La configuració de la privacitat de les dades ha de ser un bon començament per tal de trobar una solució als abusos que poden patir-se si se'ns coneixen dades privades.

1.4. Planificació inicial

Veiem ara una planificació detallada que pendrem com a punt de partida per tal d'anar desenvolupant el projecte i fer-la servir com a base de partida per cadascuna de les xarxes socials que es volen analitzar.

Planifiquem l'inici de recerca d'informació per al primer semestre del curs 2013-2014; l'anàlisi de les fonts d'informació comença el 7 d'octubre.

Del 7 al 13 d'octubre;

- Cerca de les víctimes amb més interacció a les xarxes socials exposades
- Accés i reconeixement a les dades públiques de la xarxa Facebook (donat un usuari)
    - Ens centrarem en trobar assistència a events i viatges, fotografies i llistat dels seus contactes.
- Accés i reconeixement a les dades publicades a Facebook de Runtastic, Nike+ o similar

Del 14 al 20 d'octubre

- Accés i reconeixement a les dades públiques exposades de la xarxa Twitter (dona t un usuari)
- Accés i reconeixement a les dades públiques exposades de la xarxa LinkedIn
    - Ens centrarem en la trajectòria professional de l'usuari (llocs de treball i cronològic) i els seus contactes més directes (que posteriorment intentarem creuar amb els usuaris d'altres xarxes)
- Accés i reconeixement a les dades publicades de Runtastic, Nike+ o similar

Del 21 al 27 d'octubre

- Extracció de dades de Flickr i Tumblr; en concret l'extracció de les dates de publicació i vincles a les fotografies publicades (si es possible).
    - Cal considerar que l'hora de publicació pot determinar en quin moment s'ha fet i si es tracta d'una càmera de fotos convencional o un telèfon mòbil amb connexió a internet (generalment s'annexen les fotos a través de l'ordinador quan es disposa de xarxa wifi)
    - Les fotografies no tractades poden contenir dades EXIF amb informació relativa a la fotografia (profunditat de camp, obertura, etc.), però també data i ubicació de realització

Del 28 d'octubre al 3 de novembre

- Extracció de les bases de dades i anàlisi de l'estructura del client mòbil per la xarxa Foursquare

Del 4 al 6 de novembre

- Extracció de dades i títols dels articles d'un blog personal o professional
    - Generalment escrits en base a plantilles, l'anàlisi ha de ser molt estructurat.

Del 7 al 9 de novembre

- Resum de les estructures de dades obtingudes per l'extracció de dades de les diferents xarxes socials.

Del 11 de novembre fins al 1 de desembre

- Integració dels parsejadors* d'informació per les diferents xarxes socials
- Repàs de les estructures de dades per tal de generar un format comú fàcilment que sigui fàcil de representar gràficament

Del 24 de novembre fins al 1 de desembre

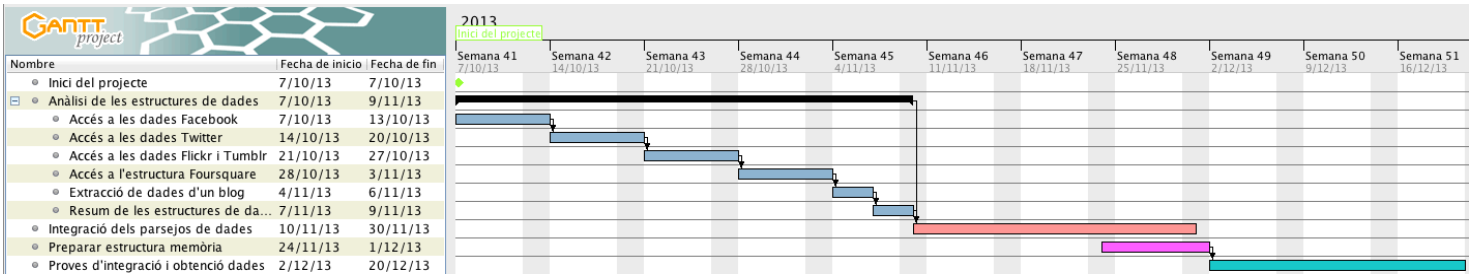
- Preparar l'estructura de la memòria per anar representant els resultats de l'extracció de dades de les diferents xarxes socials en plantilles formatejades que facilitin la relació de les dades extretes per un usuari.

Del 1 al 19 de desembre

- Proves d'integració dels sistemes desenvolupats
- Obtenció de dades actualitzades dels usuaris que s'han fet servir com a objectiu

Veiem a la figura 1.1 el diagrama de Gannt de la planificació inicial proposat pel desenvolupament del projecte.

Figura 1.1. Diagrama de Gannt de la planificació inicial



1.5. Capítols del desenvolupament de la memòria

En els següents capítols, veurem com s'ha anat desenvolupant el projecte per tal de comprovar com n'és de fàcil obtenir dades de forma lícita del usuaris de les xarxes socials més habituals (entre moltes).

- Capítol 2.1. Fase d'anàlisi i disseny d'extracció de dades

  Com a punt de partida, hem de comprovar fins a quin punt és viable extreure dades de les xarxes socials i a més, quines dades extreure.

  Hem de fixar-nos en quins mètodes ens poden oferir les xarxes socials per portar a terme el nostre estudi. Com veurem amb més profunditat, tindrem l'oportunitat de treballar sobre tres conceptes:

  - Accés controlat per integracions pròpies de la xarxa social (APIs)
  - Interpretació del propi codi HTML* que generen les xarxes socials per oferir les webs als seus usuaris
  - Obtenció de les dades emmagatzemades en els clients pròpis o de tercers de les xarxes socials.

- Capítol 2.2. Estratègies d'extracció de dades

  Un cop analitzades les possibles vies per l'extracció de les dades, hem de formalitzar un patró per cadascun dels mètodes proposats per l'obtenció de la informació.

  Tinguem en compte que cadascun dels mètodes van molt relacionats a l'estructura que tingui la xarxa social examinada, ja que cadascuna d'elles tria la seva pròpia integració de dades i al tractar-se generalment d'empreses independents, acostumen a fer servir els seus pròpis mètodes.

- Capítol 2.3. Observacions en l'evolució de la privacitat

Durant el desenvolupament del projecte, s'han anat produïnt canvis en alguna de les xarxes socials. Veurem la forma de comprovar que els desenvolupaments poden ser modificats i per tant s'haurà d'adaptar el producte final en el temps si els canvis son molt significatius.

En termes generals, si l'accés a les dades es fa mitjançant la pròpia interfície d'integració de les xarxes socials, els canvis haurien d'estar planificats i anunciats. Ara bé, si com a mètode d'extracció de dades es fa servir el parseig de les pàgines HTML o l'interceptació de la informació que emmagatzemen les aplicacions client, la cosa canvia.

Hi ha una part molt important de procés manual si l'extracció es fa per HTML o per aplicacions client. Però com vol comprovar aquest projecte, es possible fer-ho i relativament senzill.

- Capítol 2.4. Desenvolupament per cadascuna de les xarxes

  En aquest capítol veurem exemples i proves que s'han realitzat per extreure dades amb els mètodes determinats de cadascuna de les xarxes socials.

  Tinguem en compte la importància de poder accedir a les dades, ja que si no fós possible per cap dels mètodes proposats en el projecte, aquesta xarxa hauría de quedar exclosa de l'estudi.

  Per altra banda, hem de valorar que les dades extretes poden haver de necessitar d'una observació més "humana", en el sentit que es basarà en la interpretació de imatges o mapes.

- Capítol 2.5. Disseny i planificació per generar informes

  Un cop extretes les dades, s'han de proposar mètodes per guardar les dades. Identificar de quina forma s'han de poder tractar i plantejar mètodes per a la seva interpretació.

- Capítol 2.6. Descripció de la implementació de tests

  En aquest darrer capítol sobre el desenvolupament del projecte, veurem exemples de codi de les extraccions de dades obtingudes.

2. Desenvolupament

2.1. Fase d'anàlisi i disseny del sistema d'extracció de dades de les xarxes socials

En termes generals, les xarxes socials més famoses amb les quals ens trobem avui en dia tenen un accés a les dades dels usuaris relativament fàcil. Hem de considerar que aquestes xarxes socials tenen com a objectiu principal expandir-se en quantes més plataformes millor i per aquest mateix motiu tenen la intenció de facilitar a terceres parts (desevolupadors anònims o fins i tot empreses) la integració dels seus sistemes.

L'accés a les dades, via una interfície oberta anomenada API*, sempre sol venir controlat o restringit amb certes notacions de seguretat: la primera restricció de seguretat que ens trobarem a la totalitat de les xarxes és l'autentificació de l'aplicació que vol accedir les dades. Aquest control és molt important per la xarxa social, per a què cap màquina pugui accedir massivament sense control ni vigilància a les dades dels seus usuaris.

En ocasions, no serà possible una extracció via API (bé perquè no existeix o bé perquè el propi sistema no ho permet per tal de protegir certes dades privades dels usuaris).

En aquests casos en que l'extracció via una interfície no es possible, s'han buscat dos mètodes més que ens permetran d'accedir a la informació. Abans de res cal pensar que aquestes xarxes socials estan pensades per expandir-se com més millor, per tant accedir-hi no ha de ser difícil per altres canals com:

- Webs (accessibles via un usuari/perfil creat prèviament)
- Accés a les dades en memòria cau de les aplicacions clients (generalment, mòbils)

Observacions:

- *Per a l'accés a les dades via web, tinguem en conta que existeixen versions lleugeres d'aquestes, on el codi HTML serà més fàcilment parsejable.*
- *L'accés a les dades en memòria cau s'haurà de fer mitjançant algun terminal que tingui accés ROOT** i poder examinar totes les dades privades de l'aplicació.*

Per les dades que necessitem extreure per aplicacions que es basen sobre altres xarxes socials per compartir els seus continguts (Runtastic, Nike+ o fins i tot Foursquare), haurem d'establir patrons de publicació contra les altres xarxes socials. Per aquests casos, l'extracció de dades inicial serà semblant a com es fa l'extracció del contingut tal i com es fa en la xarxa social on es publica i posteriorment s'obtindrà la dada d'aquests contiguts específics.

Quan una aplicació externa publica contingut (sempre sota autorització de l'usuari i propietari d'una xarxa social), ho fa seguint unes plantilles de text que ens seran molt còmodes per trobar quan s'ha produït una d'aquestes publicacions i extreure'n:

- Data de publicació
- Data de l'esdeveniment (esport, menjar a un restaurant, visita a un museu, etc.).

- Vincle extern a la xarxa social on s'ha publicat l'esdeveniment on podrem veure més dades (recorregut de l'esport realitzat, informació sobre el restaurant, etc.)

En resum, la informació la tenim a l'abast de la mà, només hem de trobar la forma d'extreure-la i fer-la servir per obtenir patrons de comportament dels usuaris observats.

La mateixa dinàmica que fa que aquestes xarxes socials siguin tant conegudes i exteses, fa que la informació dels usuaris quedi exposada per a poder-ne realitzar qualsevol estudi.

2.2. Estratègies per l'extracció de dades de les xarxes socials

Ens trobarem que cadascuna de les xarxes socials té les seves pròpies implementacions en qualsevol dels mètodes que haguem de fer servir per extreure les dades, per tant haurem de tenir un procés molt "manual" per tal de desgranar cadascuna de les formes d'accés.

Pensem que ens enfrontem a xarxes socials de caire molt diferent encara que totes tinguin el mateix objectiu de compartir dades dels usuaris, i per tant, l'estructura de les dades estarà organitzada de forma diferent. Malgrat això, com que tenim un objectiu molt concret en l'extracció de dades, la informació final de la qual disposarem està limitada i podrem generalitzar els resultats.

Per cadascun dels mètodes que podrem fer servir per extreure dades, hem de tenir en compte que la constant evolució d'aquestes xarxes socials pot fer que el mètode triat sigui ineficient en un futur (bé per haver canviat l'estructura de dades o fins i tot per canvis en la política de privacitat de les dades publicades).

Veiem doncs els mètodes plantejats per a fer possible l'extracció de dades:

2.2.1. Parseig de dades obtingudes de les webs ofertes per les xarxes socials

L'evolució de les xarxes socials ens ha portat a tenir cada cop pàgines web més completes que permeten una forta interacció amb els coneguts o "amics" sense haver de disposar d'una aplicació d'escriptori o mòbil per interactuar-hi. Inclouen xats, animacions, visualització de galeries de fotos i vídeos, etc.

La complexitat d'aquestes webs està basada en Javascript o Flash. El codi HTML obtingut amb les webs d'aquest estil és doncs molt ofuscat si fem una simple crida GET mitjançant un programa client bàsic que no interpreti aquests dos llenguatges (o d'altres que compleixin el mateix objectiu de generar més codi HTML).

Ara bé, si mitjançant qualsevol navegador visitem la web de la xarxa social d'on volem extreure dades, ella mateixa s'encarregarà d'executar les actualizacions de les dades y per tant també la generació final del codi HTML que l'usuari veurà. Una vegada obtingut aquest HTML podem introduir-lo en un parsejador DOM* pensat per obtenir la informació que volguem. Amb aquest mètode, obtindrem la mateixa informació que un usuari pot veure a la web.

Per a ser més eficients i pràctics farem servir les versions mòbils de les webs d'on volguem extreure dades. Cal pensar que la inferior potència de càlcul dels dispositius mòbils, amb un accés a dades més lent y en general amb poc espai de pantalla per mostrar dades, la representació de les webs és més senzilla (tant visualment i per tant en codi HTML, com computacionalment –on l'HTML s'envia directament interpretat pel servidor més que interpretat en el client per Javascript o d'altres-).

Usant qualsevol dels dos tipus de web (escriptori o mòbil) tractarem les dades de la mateixa manera, mitjançant un parsejador DOM, però haurem de tenir cura i verificar que les estructures son les mateixes o tenen el mateix nom, ja que moltes vegades, les dierents versions web poden estar desenvolupades per diferents equips de programadors o senzillament, ls diferents tipus de plataformes no permeten un llenguatge tant extens per assignar noms de variables i/o atributs.

2.2.2. Accés a les dades de les xarxes via APIs públiques

Moltes xarxes socials ofereixen un accés a les seves dades per què programadors anònims puguin desenvolupar altres clients y així ampliar la seva xarxa a molts més usuaris.

S'ofereixen una sèrie de biblioteques o consultes http (generalment REST*) que faciliten als programadors l'accés obtenint estructures fàcils d'integrar i també formes estàndard d'enviament de dades per alimentar les xarxes socials.

Actualment, qualsevol de les funcionalitats proposades per la xarxa ve securitzada per tal d'oferir seguretat als usuaris i haurem de registrar qualsevol aplicació client que volguem desenvolupar per extreure'n dades.

Això no suposa cap problema, ja que es pot descriure l'aplicació d'estudi d'aquest projecte com a un client més, però si que ens pot arribar a impedir fer consultes massives sobre un usuari o un grup d'usuaris o un grup de dades en concret si el servidor detecta moltes crides de forma massiva.

Cal afegir que l'accés via APIs públiques és més simple un cop s'ha aconseguit fer una aplicació base, ja que les xarxes socials tenen un conjunt de mètodes molt semblants entre ells i la informació vé molt millor estructurada que si optem pel parseig de dades HTML.

De totes maneres, la integració API és més laboriosa que un parseig HTML perquè cada xarxar social tindrà els seus mètodes i patrons de programació per accedir-hi.

2.2.3. Anàlisi de les dades extretes de les bases de dades d'aplicacions client

Com a tercer mètode per obtenir dades, podem fer servir les pròpies aplicacions mòbils o d'escriptori que ens ofereixen les xarxes socials.

Amb la senzillesa d'accedir a una base de dades local (prèviament actualitzada pel client oficial), podem obtenir una gran quantitat d'informació que ens ajudin en el procés d'extracció de dades d'un usuari.

Cal tenir present que els telèfons intel·ligents ofereixen una multitud d'aplicacions, però per no extendrens farem servir només les oficials. Hem de suposar llavors que les dades guardades en local seran molt limitades (si es que els programadors de la app han près consciència de la importància de la privacitat i només emmagatzemen la informació necessària).

Durant el desenvolupament del projecte només s'ha trobat amb una app que ha posat mesures de seguretat per protegir la integritat de les dades dels seus usuaris.

En altres casos, o no es guardava informació i es veia tota en línia o quedava exposada sense cap tipus de xifrat:

- Només l'app Foursquare ha protegit les dades (i només parcialment) amb un xifratge molt senzill; el xifratge Cèsar [6].

Aquest mètode àmpliament conegut només es basa en la rotació dels caràcters del text, on cadascuna de les lletres de la frase xifrada es desplaça un número definit de posicions en l'alfabet.

> Més que l'encriptació, es pot tractar com un mètode d'ocultació o codificació de les dades ja que té un xifrat molt fàcil de trencar.

Aquest procés d'extracció de dades serà molt manual, s'ha d'investigar l'estructura interna de l'aplicació mitjançant la base de dades i fins i tot així podríem no poder accedir a tota la informació si alguna data no resideix físicament al terminal. Com hem comentat, alguns clients basen la seva activitat en les dades en temps real i només si hi ha connexió a internet.

2.2.4. Estructures de dades comunes on abstraure les dades extretes de les xarxes socials

Després de tot, ens trobem que totes les xarxes socials treballades tenen un funcionament semblant i encapsulen les següents dades:

- Codi o nom d'usuari
- Avatar de l'usuari
- Data de publicació de les mencions
- Comentaris o textos escrits per l'usuari

Altres dades de l'estructura (opcionals):

- *Usuaris relacionats*
- *Dades geolocalitzades*
- *Comentaris extra o links relacionats amb la menció*

Per tant, malgrat el diferent caire de les xarxes socials i el seu objectiu final (bé sigui per publicar continguts personals i privats o professionals), tots treballen amb una estructura similar, organitzant la informació publicada en blocs d'informació que pot ser fàcilment classificable malgrat el contingut que tinguin.

Aquesta estructura similar en origen ens ajudarà a emmarcar la informació molt satisfactòriament per mesclar les dades extretes de les diferents xarxes socials i poder per tant crear una cronologia que ens ajudi a conèixer el perfil dels usuaris observats.

En els casos on la informació a extreure sigui més diferent a l'estructura bàsica i comú de totes les xarxes socials, caldrà posar especial èmfasi per no perdre aquell contingut que pot ser determinant per exposar més informació de l'usuari.

2.3. Observacions en l'evolució de la privacitat i seguretat de les dades de les xarxes socials

Durant l'elaboració del projecte que les diferents xarxes socials, veiem que s'estan evolucionant continuament els productes (en qualsevol de les seves plataformes) per tal d'oferir més seguretat o més funcionalitat als usuaris. Aquestes actualitzacions no impliquen que no es puguin extreure dades de les xarxes socials, però si que cal en ocasions autentificar l'aplicació client o canviar els mètodes d'accés per extreure les dades o trobar altres mètodes per extreure la nova informació.

2.3.1. Autentificació OAuth

L'autentificació contra les xarxes es realitza generalment amb una *api key* (amb un protocol privat) o mitjançant *OAuth key [7]*.

Bàsicament OAuth es tracta d'un protocol d'autentificació que permet de conectar aplicacions de tercers amb portals web o xarxes socials, etc. I té la finalitat principal d'aprovar l'accés a totes les dades privades sense que l'usuari hagi de donar les seves credecials reals.

Un cop l'usuari decideix que l'aplicació client d'un tercer ja no pot tenir accés a les dades, que tant podría ser de lectura com d'escriptura, pot revocar-ne el permís des del portal d'administració i així guanya en no tenir que canviar la seva contrasenya principal.

Llavors tenim que per les següents xarxes socials hem d'obtenir una autorització prèvia a l'obtenció de dades si es volgués realitzar una extracció de dades via les seves APIs públiques:

- Facebook
- Twitter
- Tumblr y Flickr
- LinkedIn

Recordem que aquest mètode d'auntentificació només es requerit si féssim un accés a les dades mitjançant la seva API pública, mai si realitzem un parseig de les dades HTML o si accedim a les dades privades d'altres aplicacions client que també facin servir aquesta xarxa social.

2.3.2. Actualitzacions del servei

L'accés a la API pública, un cop construit, no hauría de canviar. Ara bé, sempre hem de tenir present que es poden produir canvis i hem d'estar previnguts.

Per canvis en la API podem consultar les pròpies pàgines orientades als desenvolupadors de les xarxes socials:

- Facebook - https://developers.facebook.com/roadmap/
- Twitter - https://dev.twitter.com/blog/category/announcements

- Tumblr - http://developers.tumblr.com/tagged/changelog
- Flickr - http://www.flickr.com/services/developer/changelog/
- Foursquare - https://developer.foursquare.com/docs/changelog

L'accés a les dades via parseig HTML pot ser encara més canviant que una implementació via API. Un simple redisseny visual o canvi en el framework de la web parsejada o el pas de la web d'escriptori a la web mòbil ja pot fer que tot el procés de parseig quedi alterat.

És un mètode fàcil, ràpid d'implementar i modificar, però té aquest risc.

Com a èina de proves, els navegadors Google Chrome o el Firefox disposen d'uns modes de funcionament que ens poden ajudar a extreure dades dels usuaris sense haver de desenvolupar les aplicacions client. Ambdós navegadors poden indicar a la web que se'ls retorni en versió mòbil.

Com a afegit, disposen d'una consola pensada per a desenvolupadors que permet llençar sentències Javascript* contra la web visualitzada i per tant, poder accedir al model d'objectes DOM per fer els tests. Un cop testejades, es poden aprofitar els noms desitjats dels atributs o elemets HTML que ens proporcionaran la informació que volem extreure.

El tercer mètode d'accés a les dades, basat en consultes SQL* a les bases de dades dels clients mòbils, també pateix el risc que se'n canvii l'estructura tot i que no hauria de ser gaire habitual que desaparegui informació, tot al contrari.

Si que podem pendre constància de possibles canvis observant el registre de canvis de l'aplicació client si el seu desenvolupador els especifica.

2.3.3. Seguretat i privacitat de les dades

Ningú posa en dubte que els desenvolupadors de les xarxes socials posen especial èmfasi en la protecció de les dades dels seus usuaris, ja que de la seguretat i fiabilitat que ofereixin dependrà l'èxit o el fracàs de la plataforma.

Per altra banda, sempre hi haurà atacs contra aquestes xarxes socials per diferents motius (interesos creats, venjances, venda de dades, etc.) però no ens centrem en això per aquest projecte.

Les webs, com a plataforma global d'accés a les xarxes socials, estan molt ben securitzades, però les aplicacions mòbils emergents tenen buits en la protecció de les dades.

L'ofuscació* del codi font (que es podría obtenir decompilant les aplicacions mòbils) es una realitat per totes les aplicacions client oficials testejades en les xarxes socials examinades en aquest projecte, però no podem dir el mateix de les bases de dades que contenen informació privada de l'usuari.

Tot i que l'accés a la base de dades d'aquestes aplicacions client no és immediata i requereix d'una certa preparació del dispositiu, amb una mica de temps es pot aconseguir una còpia de les bases de dades.

> Primer de tot, s'ha d'obtenir accés ROOT del dispositiu per tal de poder accedir a la seva àrea privada on suposadament les apps poden deixar informació compromesa.
>
> Un cop tenim accés a l'àrea privada del dispositiu, tenim accés lliure a la seva estructura de dades i visibilitat directe de les bases de dades o altres arxius de preferences d'usuari.
>
> *Més informació a http://faqsandroid.com/root/*

Facebook, Twitter, Tumblr i Flickr han deixat les seves bases de dades obertes, sense cap tipus d'ofuscació o encriptació de les dades, el que suposa una falla de privacitat ja que algú que obtingués un dispositiu mòbil d'altri podría llegir les dades privades o converses sense haver d'accedir ni tant sols a l'aplicació, encara que l'usuari i propietari del telèfon hagués canviat la contrasenya d'accés a la xarxa social.

2.4. Mètodes d'extracció per cadascuna de les xarxes socials proposades al projecte

2.4.1. Facebook

Darrerament Facebook s'ha convertit en una de les xarxes socials més expandides i resulta particular conèixer algú que no disposi d'una conta d'usuari, tot i que només la faci servir en comptades ocasions i no per posar dades personals (com fotografies, reports de viatges, etc.) sinó com a plataforma per a difondre imatges i vídeos extrets d'internet.

L'API (o SDK com l'anomenen al seu racó per a desenvolupadors) ofereix principalment accés a les dades per a les següents plataformes de forma oficial:

- SDK per iOS
- SDK per Android
- SDK per Javascript
- SDK per PHP

Ells mateixos ofereixen un llistat de links per accedir als seus continguts fent servir altres llenguatges de programació i utilitzant llibreríes de tercers (les SDKs oficials que ofereixen son OpenSource i per tant replicables a altres llenguatges).

Així mateix existeixen molts altres projectes de codi obert que ens ofereixen la possibilitat d'accedir a les dades (sempre de forma autenticada contra Facebook) però amb més facilitat d'obtenció d'informació. Hem de tenir en conta que aquests projectes poden quedar desactualitzats en qualsevol moment després d'una actualització de la xarxa social i per tant podria no ser aconsellable si volem desenvolupar una app final que es pugui fer servir de forma prolongada.

Exemples de llibreries de tercers

- https://code.google.com/p/facebook-java-sdk/
- http://restfb.com/

Facebook ofereix als usuaris una versió mòbil de la seva web que permet fer una consulta relativament fàcil al seu contingut, molt més eficient que en la seva versió d'escriptori ja que el contingut es genera en servidor. Per tant, serà possible accedir-hi de forma senzilla.

També disposa d'una aplicació client mòbil i encara que la informació de la qual disposa s'actualitza només quan s'accedeix a l'aplicació en primer plà, certes dades queden residents al dispositiu repartides en diferents bases de dades (llistat de contactes emmagatzemat en JSON* –guardat en aquest format per a emular el servidor en cas de treballar offline-, missatges de chat, urls de fotografies publicades, entre d'altres).

El que més impacta de Facebook, per l'envergadura que té, és que la informació de la seva base de dades no està encriptada.

2.4.2. Twitter

Com a altra gran xarxa social actual, hem de considerar que la seguretat a la que estan subjectes les dades personals dels usuaris es molt gran.

Primer de tot, per a desenvolupar un client per a Twitter, hem d'obtenir una clau proporcionada per ells mateixos. Sense aquesta no serà possible accedir via API a les dades.

Twitter ofereix un sistema REST molt senzill que permet extreure multitud d'informació, ara bé, son ells mateixos qui ofereixen un llistat de llibreries de tercers que faciliten per diferents llenguatges de programació l'accés a les dades.

Exemples de llibreries de tercers

- http://twitter4j.org/en/index.html

Per la seva senzillesa, és molt fàcil extreure dades parsejant el perfil d'un usuari de Twitter tant en la versió d'escriptori com en el mòbil. Amb el nom dels objectes DOM trobat, l'accés es immediat.

Per altra banda, l'extracció de les dades via la base de dades de la seva aplicació client també és viable, ja que tenim accés a una taula de missatges on es veuen les mencions fetes per l'usuari, missatges relacionats, l'usuari que origen de la menció si no ha estat el propi usuari que l'haençatat, etc.

2.4.3. Foursquare

Les darreres actualitzacions d'aquesta xarxa de compartició de restaurants han fet incrementar molt la seguretat a la qual estan subjectes les dades.

Donat que ens trobem davant d'una xarxa que es centra en els dispositius mòbils (en concret iOS i Android), hi ha una dificultat afegida per accedir a les dades automàticament ja que caldria una app per examinar les dades locals dels terminals i a més ser "amic" de la víctima (no es permet configurar de cap de les maneres una visibilitat pública dels llocs que l'usuari visita).

Aquesta xarxa funciona en base als "checkins*" que un usuari fa en indrets que visita i donar-ne la seva opinió. Ara bé, la seva API només permet extreure informació dels checkins fets pel propi usuari (mai buscar el llistat de checkins fets per un altre, només rebre notificacions quan aquest ha fet un checkin https://developer.foursquare.com/start/realtime).

Així llavors se'ns farà complicat desgranar els patrons de comportament d'un usuari.

Però si s'aconsegueix ser "amic" de la víctima, rebrem així que es donin els checkins als restaurants que visita i quederan emmagatzemats a la base de dades local del nostre dispositiu mòbil i en podrem fer una posterior extracció mitjançant consultes SQL (fins i tot, en podrem obtenir –si ho ha especificat- la conta associada de Facebook o Twitter també per al seu posterior anàlisi).

2.4.4. Publicacions de recorreguts d'esports (Runtastic, Nike+)

La forta integració que tenen aquests sistemes de tracking amb les xarxes socials fa possible que compartir dades dels recorreguts que es realitzen practicant qualsevol esport sigui molt intuïtiva.

Per la mateixa raó que és molt fàcil publicar quins recorreguts i en quins horaris es fan, cal pendre consciència de la facilitat d'extreure un patró que existeix.

Generalment, aquestes plataformes d'esports publiquen un link públic a les xarxes socials (com Facebook, Twitter, etc.) que un cop obtingut pels mètodes comentats per les altres xarxes socials, podrem accedir sense cap tipus d'autentificació.

Obtindrem un mapa amb el recorregut i altres detalls que ens poden ajudar en el nostre estudi de comportament.

2.4.5. Publicacions d'imatges (Tumblr o Flickr)

L'objectiu inicial de tractar les xarxes socials de publicació de fotos més que veure quins llocs ha visitat un usuari o quins son els seus gustos en referència a les fotografies que fa, es vol centrar en les característiques tècniques que se'n poden extreure de les fotografies publicades.

Mitjançant l'examinat de les propietats EXIF d'una imatge podríem arribar a saber si:

- les fotografies estan tractades (generalment això esborra dades EXIF)
- marca i model de la càmera o telèfon mòbil
- dades pertinents a fotografia (obertura, profunditat, etc.)
- dades gps i temporals de la fotografia
- i darrerament, extreure la petjada digital de la càmera que ens permetria trobar per internet altres imatges fetes amb el mateix dispositiu.

  Podem trobar pàgines que mitjançant aquesta petjada digital, rastregen internet per trobar més imatges. Tot i que generalment aquest tipus de pàgines es fan servir per localitzar càmeres extraviades, podríem valorar fer-ne aquest nou ús.

  - http://www.stolencamerafinder.com/
  - https://www.lenstag.com/

Mitjançant l'aplicació mòbil de Flickr i l'accés a la seva base de dades podem obtenir el llistat de fotografies realitzades pel client, amb l'informació extra afegida (dades gps, conta associada de Facebook, Twitter o Foursquare, entre d'altres).

2.4.6. LinkedIn

En aquesta xarxa social professional ens tornem a trobar com a requisit una autorització per accedir a les dades de l'usuari final.

Tot dependrà de les dades que en volguem extreure. Si només necessitem fer un llistat de llocs de treball i estudis realitzats serà més que suficient accedir al seu perfil públic i fer un parseig HTML de les dades i obtenir el seu avatar ja que la

informació pública extreta amb aquest mètode a LinkedIn és molt extensa (si s'ha configurat com a pública).

Si ens plantegem quin tipus de perfil volem exposar en aquesta xarxa social, no hem de perdre de vista que és un mur on col·locar el currículum, per tant, cal suposar que la gran majoria d'usuaris configurarà una visibilitat molt alta del seu perfil per poder ser localitzat fàcilment (contrariament al que podría passar en altres xarxes socials).

Finalment, també veiem que podem extreure molta informació de l'aplicació mòbil d'aquesta xarxa social. Per cadascun dels nostres contactes, podem extreure noms, càrrec , url del seu avatar, missatges enviats i rebuts.

Pel perfil de l'usuari observat, a banda dels missatges enviats i rebuts, també podem fer un llistat d'estudis i de feines i càrrecs.

2.5. Disseny i planificació del software final per generar un informe sobre un usuari

Sota el pretext de trobar els usuaris observats que interaccionin en les diferents xarxes socials plantejades durant el projecte, la prova és extreure una eina on introduïts els perfils de les diferents xarxes socials se'n pugui obtenir un patró de comportament.

Per cada xarxa social s'extraurà, a banda d'un accés a la informació, les següents estadístiques:

- Mitjana de mencions a les diferents xarxes socials (per dia, setmana, etc)
- Finestra horaria on es fan generalment les mencions
- Interacció d'altres usuaris en les mencions

Se'n podran extreure altres dades, però ja vindran donades per una interpretació en funció de l'observació de les dades que es reflexin en l'informe resultant per cada usuari.

Cal tenir en compte que dades com:

- Hores d'estada a casa
- Hores de sortida (anar a treballar, dinar, sopar, vacances)
- Aficions o hobbies

, hauran de ser interpretades per una observació detallada i personal ja que cap software no pot valorar aspectes personals, textos escrits o fotografies exposades a les xarxes socials (a no ser que estiguin correctament categoritzades de forma semàntica).

2.5.1. Emmagatzemament de les estructures de dades resultants

Per desar l'estructura de dades i emmagatzemar informació relativa a diferents xarxes socials s'obtarà per un document XML* on poguem anar ampliant els nodes d'informació i per tant fer tant extensible con sigui necessari l'espai d'informació que es vagi obtenent (sempre pensant que no podrem extreure la mateixa informació dels diferents usuaris que es poguessin anar observant, ja que no tots ells tenen els mateixos tipus de perfil, amb la mateixa configuració de visibilitat configurada i fins i tot poden no tenir perfil en una determinada xarxa social).

XML base que anirem extenent en funció de les dades extretes:

```
<?xml version="1.0" encoding="UTF-8"?>
<perfil>
	<xarxa name="" url="">
	<nick>...</nick>
	<avatar>...</avatar>
	<mencions>
		<mencio>...</mencio>
	</mencions>
	<contactes>
		<contacte>...</contacte>
	</contactes>
	</xarxa>
</perfil>
```

I vist aquest XML, podríem en ja extreure l'objecte Java (que faríem servir per una implementació de test per obtenir dades dels usuaris).

Aquest objecte Java anomenat `Perfil.class`, engloba totes les dades que podem extreure d'un usuari en una xarxa social. Veure ANNEX 1.

Per acabar amb l'emmagatzemament de les dades d'un perfil, fent servir l'interface `Marshaller` de Java, obtindrem l'estructura XML de la menció. Veure ANNEX 2

2.5.2. Extracció de dades

Per tal de generar un patró d'extracció de dades el més generalista possible i no caure en un desenvolupament específic per cada xarxa social, totes les dades possibles s'han obtingut de les versions mòbils de les pàgines web oficials que ofereix cadascuna de les plataformes i d'aquesta manera aprofitar la base i l'estructura de l'aplicació de proves.

Així doncs, s'ha desenvolupat una base que permeti parsejar el contingut HTML de les webs de perfils amb informació per extreure de forma comuna informació de les diferents xarxes socials.

Passos a realitzar per a l'extracció de les dades:

1. Obtenció del codi HTML de la pàgina de perfil de l'usuari observat (per una xarxa social en concret)

   Si l'accés a la pàgina HTML s'ha de fer de forma autentificada, caldrà primer guardar una versió de la web (versió mòbil) de forma local através del navegador de què disposem. Recordem que navegadors com Firefox o Google Chrome permeten visitar webs com si es tractés d'un navegador mòbil i així poder obtenir la versió reduida de HTML.

2. Parseig XML de la web obtinguda.

   HTML, com a llenguatge de marques*, es pot tractar com un arxiu XML. Llavors podem extreure'n el contingut desitjat fent servir un parsejador que carregarà l'objecte en memòria per a ser tractat posteriorment.

3. Extracció del Model d'Objectes del Document (DOM) pels nodes que volem obtenir:

   - Codi o nom d'usuari
   - Avatar de l'usuari (url de l'imatge)
   - Data de publicació de les mencions
   - Comentaris o textos escrits de les mencions

   Altres nodes DOM (opcionals):

   - *Usuaris relacionats*
   - *Dades geolocalitzades*
   - *Comentaris extra o links relacionats amb cada menció*

2.5.3. Autentificació en les extraccions de dades

Com he comentat, els usuaris de les xarxes socials poden configurar les seves pàgines de perfil com a públiques o privades només concedint accés a altres usuaris "amics".

Per a les dades de les pàgines públiques no hi hauria cap problema a l'hora d'accedir-hi mitjançant una consulta http i prenent-ne el contingut HTML. Ara bé, per les pàgines privades s'ha hagut de desenvolupar el mecanisme de d'autentificació (mitjançant usuari i contrasenya de l'"amic" de l'usuari observat) i conservació de cookies per tal de mantenir la sessió al llarg de l'extracció de dades.

Tot aquest mecanisme s'ajuda de les capçaleres HTTP utilitzades en les consultes GET o POST (segons la xarxa social ho requerís) i que per tant, han seguit un estàndard que ens ha facilitat la tasca de mantenir la sessió com si d'un navegador web es tractés [10].

2.5.4. Gestió d'errors

El parseig de dades realitzat contra els documents HTML de les versions mòbils de les webs de les xarxes es molt senzill. Només caldrà tenir en compte els següents aspectes:

- Si no tenim accés a la pàgina pública, no passa res, només que aquella part d'informe de l'usuari que podriem obtenir, quedarà buida i no haurem pogut recaptar tanta informació.

- Si no tenim dades útils per a l'extracció de l'usuari observat, només haurem de controlar que per cada element DOM que intentem extreure del document HTML, no es produeixi cap excepció per intentar accedir a un element no existent.

Donada la senzillesa d'un parseig DOM, els errors que es podem produir son mínims. Per altra banda, hem de controlar:

- Accés de lectura o escriptura dels documents HTML i XML.

Controlats aquests tres informes, el procés de generació del XML final no hauría de tenir cap problema més

2.5.5. Obtenció de dades de terceres fonts

Algunes de les dades que ens agradaría obtenir dels usuaris observats només es poden extreure de les aplicacions client (o del mateix contingut, si ens fixem en les fotografies).

La integració d'aquestes dades, malgrat haver-se de tractar mitjançant consultes SQL o extracció de dades EXIF , no han de suposar cap problema per afegir la seva informació als perfils XML generats, precisament per la seva facilitat d'extensió.

2.6. Descripció de la implementació i tests

2.6.1. Resultats de l'implementació d'extracció de de dades i avaluació dels resultats obtinguts en els tests.

L'avaluació de resultats d'aquest projecte pretén plasmar la facilitat amb la qual es poden obtenir i s'han obtingut dades publicades lliurament pels usuaris, però que a la vegada poden posar en compromís la seva privacitat.

La recopilació de les dades obtingudes de les diferents xarxes socials i els patrons de comportament del usuaris observats han vingut marcats pels següents factors:

- L'ús de les xarxes socials (no tots els usuaris tenen un perfil a totes les xarxes tractades en aquest projecte)

- El nivell d'interacció dels usuaris en les diferents xarxes socials (podem identificar dos tipus d'usuaris; els que utilitzen aquestes plataformes com a mitjà per a compartir contingut provinent d'internet –articles, contingut multimèdia, etc. - o com a mitjà provinent de la seva experiència –fotografies, vídeos o vivències personals-)

Tot i la facilitat per trobar les dades dels usuaris observats, cal dir que l'abast d'aquest projecte queda limitat a perfils per permetin l'accés a la seva informació; perfils amb visibilitat completament pública o usuaris que siguin "amics" amb l'usuari des del qual se'n llençaran les proves d'extracció de dades.

Així doncs, no farem servir cap tècnica il·lícita per obtenir les dades desitjades que ens permetran traçar un perfil dels usuaris observats (no farem servir ni procediments per trencar contrasenyes ni tampoc es desenvoluparan mètodes per trobar i explotar les vulnerabilitats de les xarxes socials).

L'extracció de les dades sempre vindrà lligada doncs al contingut disponible. Per les proves realitzades, els usuaris observats han publicat contingut de forma habitual.

2.6.2. Joc de proves i tests realitzats

Mitjançant un joc de proves, es vol donar una visió realista de l'objectiu principal del projecte; extreure dades de diferents xarxes socials per poder-ne fer posteriorment un tractament que ens doni un perfil o patró de comportament d'un usuari observat.

A continuació es detallen i es dona el resultat de les proves atòmiques realitzades per a completar una extracció de dades.

Farem servir la llibreria GWT per tal d'accedir al contingut HTML com si es tractés de Javascript però usant Java. Podem trobar aquesta llibreria a

https://code.google.com/p/google-web-toolkit/

i ens donarà les bases per accedir als objectes DOM.

Obtenir el codi HTML de la versió mòbil de la web d'una xarxa social

Per cadascuna de les xarxes socials examinades, hem trobat la seva equivalència mòbil que ens permetrà realitzar un parseig HTML de forma més senzilla:

- http://facebook.com → http://m.facebook.com
- http://twitter.com → http://mobile.twitter.com
- http://flickr.com → http://m.flickr.com
- http://es.linkedin.com → http://touch.www.linkedin.com

Exemple d'extracció per a una pàgina de perfil facebook (extracte HTML on es mostra una sola menció feta per l'usuari observat). Veure Annex 3.

Obtenir mitjançant DOM el llistat d'elements d'informació desitjats

Exemple de comandes DOM per l'extracció de dades del document HTML anterior (veure annex):

- Nom
  - `document.getElementsByClassName('tlName');`
- Avatar
  - `document.getElementsByClassName('profilePic);`
- Data de publicació
  - `document.getElementsByClassName('mfss);`

Amb aquestes extraccions obtindrem les següents dades de la menció:

- Nom
  - *Ismael Florit Zacarias*
- Avatar
  - https://fbcdn-photos-b-a.akamaihd.net/hphotos-ak-frc3/c0.40.960.639/s720x720/1459163_10202750354144847_1635441731_n.jpg
- Data de publicació
  - *El 13 de diciembre a la(s) 21:51*

Guardarem aquest joc de dades en un objecte Java que posteriorment farem servir per convertir a XML (Veure Annex 2).

Generació de l'estructura de dades XML amb la informació extreta

I finalment, fent servir la interfície `Marshaller` de Java, obtindrem l'estructura XML de la menció (Veure Annex 2).

```
JAXBContext context = JAXBContext.newInstance(Perfil.class);
Marshaller m = context.createMarshaller();
m.setProperty(Marshaller.JAXB_FORMATTED_OUTPUT, true);
Perfil object = new Perfil();
object.setName("Ismael Florit Zacarias");
```

```
object.setAvatar("http://server.com/file.ext");
object.setDate(new Date());
m.marshal(object, new FileOutputStream("perfil.xml"));
```

Recordem que no totes les dades que volíem obtenir es basaven en comunitats web; teniem altres objectius que requerien de consultes a bases de dades de les aplicacions mòbils client i també volíem extreure dades de les fotografies preses pels usuaris observats.

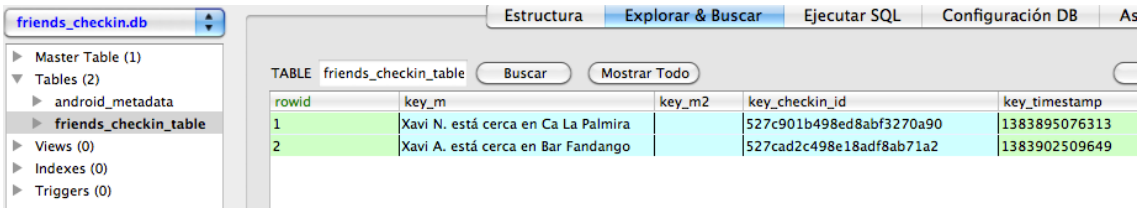
Obtenció de dades en base a consultes SQLite (per les apps mòbils)

Per l'obtenció de la base de dades d'un client mòbil, necessitarem un telèfon amb accés ROOT.

Un cop obtinguda la base de dades, només caldrà poder realitzar consultes SQLite amb qualsevol mànager (d'escriptori o mòbil) per poder veure les dades desitjades.

Com que aquesta base de dades treballa d'una forma relacional, és molt fàcil extreure un XML per les dades que hi conté; bàsicament cada fila de les taules requerides serà un element dins del XML i les seves columnes les podrem tractar com sub-elements XML o atributs (segons convingui)

Vegem una figura amb les dades consultades contra la app Foursquare que ens permet veure quins llocs ha visitat (per nom) un usuari observat –*veure figura-.*



En aquest punt, aconseguides aquestes dades a la base de dades guardada en local, hem de valorar com fem servir aquesta informació. Sabem que l'usuari de nom X ha fet una visita a cert indret, però no tenim manera de saber amb seguretat que hi ha estat, ja que no existeix confirmació real i tampoc tenim accés al mapa que es mostra en l'aplicació.

En aquests casos haurem de completar un XML de forma manual per tal de poder afegir al perfil general del nostre usuari les visites. Si que podrem per exemple saber el títol de la menció i l'hora

```
<?xml version="1.0" encoding="UTF-8"?>
<perfil>
<xarxa name="foursquare" url="">
<nick>COLUMNA 'firstname'</nick>
<avatar> COLUMNA 'firstname' + COLUMNA 'photoSuffix'</avatar>
<mencions>
      <mencio>
            <text>COLUMNA 'key_m'</text>
            <data>COLUMNA 'key_timestamp'</data>
```

```
      </mencio>
</mencions>
</xarxa>
</perfil>
```

Ens detindrem sobre la figura anterior per aportar una informació addicional a la xarxa social que ha generat aquesta taula de visites d'exemple.

Foursquare com a xarxa social per emmagatzemar, informar i compartir llocs d'interès de qualsevol caire, ens dóna la possibilitat de fer un seguiment acurat de les visites, moviments i horaris d'un usuari observat (prèviament assignat com a amic).

A banda de tota la informació que pugui ser extreta d'aquesta xarxa, ens hem trobat que és la única que es protegeix xifrant parcialment les dades dels usuaris (malgrat sigui amb un xifrat molt rudimentari). Com veiem a la següent figura:

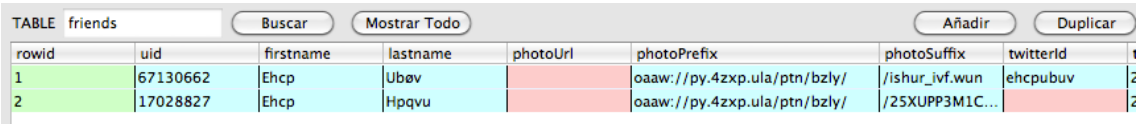


TABLE friends Buscar Mostrar Todo Añadir Duplicar

rowid	uid	firstname	lastname	photoUrl	photoPrefix	photoSuffix	twitterId
1	67130662	Ehcp	Ubøv		oaaw://py.4zxp.ula/ptn/bzly/	/ishur_ivf.wun	ehcpubuv
2	17028827	Ehcp	Hpqvu		oaaw://py.4zxp.ula/ptn/bzly/	/25XUPP3M1C...	

Inicialment, no ens hauríem donat compta del xifrat que fa servir aquest, però el detall del protocol de les URLs delata l'encriptació fàcilment:

oaaw://py.4zxp.ula/ptn/bzly/ → (ROT-19) **http**://ir.4sqi.net/img/user/

Òbviament, si executem la URL resultant, no obtindrem cap resposta correcta del servidor (probablement, hi hagi d'haver un login per accedir a la pàgina), però ja ens dóna pistes per poder esbrinar el nom d'usuari i relacionat amb els checkins, saber quin usuari ha fet què.

Obs: podrem desxifrar fàcilment la web http://superpatanegra.com/texto/index.php

En resum, és tota una irresponsabilitat deixar dades tant perilloses com saber exactament els entorns per on es mou un usuari a l'abast de qualsevol qui pugui obtenir una base de dades amb aquesta informació.

Obtenció de dades EXIF de les fotografies obtingudes

L'obtenció de les dades EXIF ha de ser molt curosa, perquè malgrat es basa en un estàndard, cada màquina de fotos o telèfon amb capacitat per realitzar-les, pot introduir informació en els camps que no desitgem i passar per alt el seu contingut. Veure Annex 4.

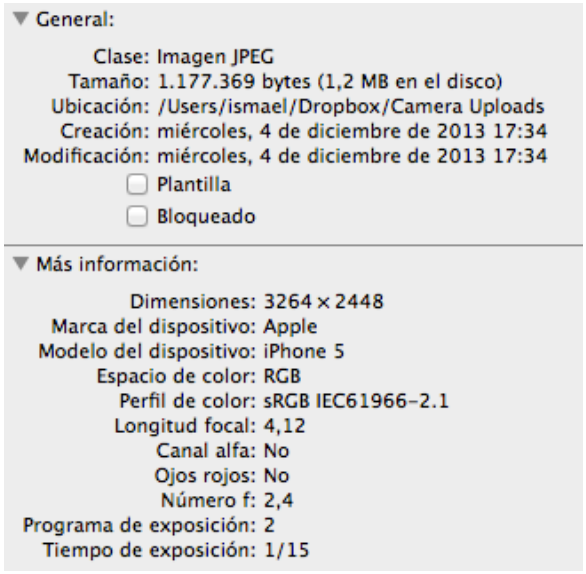
Per poder fer un seguiment genèric, pendrem dades com

- Data de presa de la fotografia

- Dispositiu (empremta digital, marca i model) *si està informat*
- Posició GPS *si està informada*

Observació: per facilitar l'extracció d'aquestes dades es fa servir una llibreria opensource (per Java, https://drewnoakes.com/code/exif/)

Exemple d'informació EXIF extreta pel sistema operatiu:



3. Conclusions i treball futur

Arribats a aquest punt del projecte, fem una valoració global del projecte i feina presentada.

L'objectiu base d'aquest projecte de final de carrera ha estat trobar un procediment i plantar les bases per tal de simplificar i organitzar la informació extreta de les diferents xarxes socials tracades i fer notar la importància de tenir configurada correctament la privacitat de les dades que s'exposen tant de forma pública o de forma privada als usuaris "amic" dins de les xarxes socials.

Tenir ben organitzada la informació extreta de les xarxes, ens permet tenir una plantilla de dades que tractar de la forma que més convingui.

Fem un repàs dels diferents punts assolits durant el desenvolupament d'aquesta memòria:

1. Avaluar el grau de dificultat per extreure dades de les diferents xarxes socials tractades en aquest projecte.

   Facebook, Twitter, Foursquare, LinkedIn, Flickr i Tumblr han estat avaluades amb l'objectiu d'analitzar la informació que podía ser extreta. A més, s'han aprofitat publicacions de Nike+, per mostrar com através d'altres xarxes aquest sistema de publicació de recorreguts esportius també ofereix informació per traçar perfils.

   S'ha observat que donada la filosofia de compartir informació amb els coneguts, no es poden amagar les dades que es mostren en aquestes xarxes socials. Sigui quin sigui el mètode d'extracció, la mateixa informació que els coneguts poden veure, pot ser extreta de forma automàtica i feta servir per guardar-se d'una forma estandaritzada per a posteriors tractaments.

   De la mateixa manera i de forma paral·lela, s'ha evaluat i comparat la informació extreta de les aplicacions client oficials de les diferents xarxes per demostrar que no només podem extreure informació d'una plataforma. Totes les plataformes disponibles han d'oferir la mateixa informació (webs, clients d'escriptori, clients mòbils, etc.)

2. L'extracció de les dades ha estat determinant per validar que obtenir informació de les diferents xarxes socials és possible. S'han desenvolupat petits programes per veure com a partir d'una estructura de dades se'n pot extreure informació per ser emmagatzemada de forma ordenada.

   Tot i tenir mecanismes més elegants per a l'extracció de dades, s'ha obtat per realitzar un parseig de les pàgines web de les diferents xarxes en contra de la implementació de n clients específics per cadascuna de les plataformes analitzades en aquest projecte. D'aquesta manera, sense haver d'implementar un nou client cada vegada que es vulgui examinar contingut d'una nova xarxa social, es podran fer servir les bases creades per parsejar codi HTML, agilitzant l'extracció dels nous perfils a les noves xarxes.

3. L'emmagatzemament de les dades extretes de les diferents xarxes s'ha plantejat de dues maneres.

   En primer terme, com que no només tenim una forma d'extreure contingut, s'ha buscat un mètode per poder guardar dades fent servir l'estàndard XML que és molt fàcil de generar i accedir des de moltes diferents plataformes i llenguatges de programació.

   En el nostre cas, hauria estat molt fàcil generar els XML resultats de cada extracció des de Java (fet servir per accedir a les APIs públiques), o també des de Java però generat através d'examinar els objectes DOM del codi HTML inicial, o finalment generat per consultes SQL si és que la informació extreta provenia de les bases de dades omplertes pels clients mòbils de les xarxes tractades.

   El segon punt que ens porta a emmagatzemar les extraccions en documents XML es basa també per la facilitat d'obtenció de les dades, també des de diferents plataformes i diferents programes d'anàlisi de dades.

   Així doncs, i en resum, XML ha estat una bona forma de generar contingut des de fonts diferents (automatitzades per parseig d'objectes DOM o manual en el cas de les consultes SQL de les bases de dades locals) i que pugui se portable entre diferents aplicatius en posteriors avaluacions de les dades provinents dels perfils dels usuaris observats.

Fent un repàs general del projecte, em sento en la posició de sentir-me positiu per la feina feta, on he pogut comprovar que totes les dades publicades a intenet poden tenir el seu punt favorable però també el seu punt negatiu.

El fet que sigui tant senzill obtenir un volum de dades significatiu sobre una persona i que posteriorment aquestes dades puguin arribar a ser fetes servir per tal de generar un perfil sobre aquesta persona, em qüestiona la fiabilitat real d'aquestes xarxes que malgrat posen de manifest la privacitat i seguretat de la informació que guarden, no deixa de ser important la cura que haurien de tenir els seus usuaris a l'hora de publicar continguts, dades personal i a més, triar amb quins "amics" comparteixen informació.

No és el mateix compartir informació amb un familiar directe que amb algú que s'ha conegut en un event professional o per altre banda amb persones conegudes en un ambient festiu. Cada grup d'"amics" potser hauria de tenir restriccions en la visibilitat dels continguts, que moltes vegades no es configura o bé per falta d'interès o bé per desconeixença de les possibilitats.

Passat el desenvolupament del projecte, veig que el resultat de totes les proves i la principal motivació inicial d'extreure dades de les diferents xarxes socials, passa per una millor educació dels usuaris ja que no es per falta de seguretat en l'obtenció de les dades ni errors en el disseny ni la implementació de les diferents aplicacions client on tenim un buit de seguretat.

Acabo aquestes conclusions amb el comentari per tots els informàtics conegut

"El pitjor virus informàtic es troba entre la cadira i el teclat"

Treball futur

Veiem ara algunes possibles evolucions de les estructures de dades generades per aquest projecte.

3.1. Integració de noves xarxes socials i ampliació de les dades extretes en les xarxes proposades

Per desenvolupar l'idea de concepte del projecte, s'han escollit algunes de les xarxes socials més populars actualment. De tota manera, se'n poden escollir multitud d'altres que actualment també estan de moda o comencen a sorgir.

Podem parlar d'altres tipus de xarxes socials com:

- Fòrums, on també en base a un nom o àlies d'usuari podem extreure els temps i horaris de participació
- Blogs que permetin comentaris i per tant extreure hores d'interacció
- Pàgines de cites, també en base a un nom o àlies o associat a un correu electrònic o al compte d'usuari d'una altra xarxa social més coneguda
- Aplicacions al núvol o webs que basen la seva activitat en consumir i compartir continguts (com Flipboard, http://barrapunto.com/ , http://stackoverflow.com, etc.)

3.2. Interfície gràfica

La interfície gràfica no ha sigut l'objectiu de desenvolupament d'aquest projecte. Una futura possibilitat es pot basar en representar les dades extretes en una aplicació (tant web com d'escriptori com mòbil) per poder veure de forma ràpida totes les dades d'un usuari observat i "conèixer" el seu estil de vida.

Ja que s'ha decidit emmagatzemar les dades en documents XML, no hauria de ser difícil integrar-les en qualsevol GUI*, sigui la plataforma que sigui i amb total independència del llenguatge de programació que es faci servir.

3.3. Extracció massiva i mineria de dades

Un cop tenim feta la base que ens permetrà desgranar informació d'un usuari des de les seves xarxes socials, no seria gens descabellada l'idea de poder obtenir informació de forma massiva per fer-la servir per centenars de casos [8]:

- Estudis de mercat i tendències
- Actes delictius contra la propietat privada o intel·lectual de les persones o empreses
- Actes delictius contra la integritat física de les persones
- Fraus i espionatge

4. Glossari

API: Interfície que especifica com diferents elements d'un programari han d'interactuar

Checkin: Procés d'inscripció a una activitat, ubicació, etc.

DOM: Model d'objectes d'un document HTML

EXIF: Especificació per a arxius d'imatge on emmagatzemar metadades relacionades amb aquesta

GUI: Interfície gràfica d'usuari. Relatiu al conjunt d'eines que permeten interactuar de forma més intuitiva amb la informació que ofereix un ordinador

HTML: Llenguatge de marcat que permet estructurar informació i relacionar els documents en forma d'enllaços

Java: Llenguatge de programació

Javascript: Llenguatge de programació

JSON: Estàndard obert basat en text dissenyat per a intercanvi de dades llegible per humans

Llenguatge de marques: Combinació de dades i etiquetes que contenen informació adicional sobre l'estructura del text o la seva presentació

Ofuscació de codi font: Procés de manipulació del codi font per a què s'executi amb el mateix resultat, però sigui més difícil de llegir a ulls d'un programador

Parsejador: Interpretador de codi en base a unes regles

REST: Arquitectura de programari basada en http per a la transmissió d'informació

ROOT: Compte especial d'usuari dins d'un computador que permet realitzar absolutament totes les accions disponibles en el sistema sense restricció

SQL: Llenguatge de comunicació amb bases de dades relacionals

XML: Llenguatge de marcat extensible pensat per a definir estructures de dades

5. Bibliografia

[1] ¿Cómo se gestiona la publicidad en las redes sociales y como se gestiona?
http://www.prnoticias.com/index.php/marketing/1100-entrevistas-prmarketing/20118412-icomo-se-gestiona-la-publicidad-en-las-redes-sociales-y-cuanto-cuesta

[2] Así son los perfiles i tipos de usuarios...
http://www.puromarketing.com/16/16554/perfiles-tipos-usuarios-invaden-medios-redes-sociales.html#

[3] Fraudes y bulos en las redes sociales
http://www.europapress.es/sociedad/sucesos-00649/noticia-policia-alerta-nuevos-fraudes-bulos-redes-sociales-motivo-dia-inocentes-20131228095107.html

[4] El 15% dels adolescents s'ha sentit assetjat
http://punttic.cat/enquesta_facua_menors_xarxes_socials

[5] Vidente usa Facebook
http://www.youtube.com/watch?v=qMefWXCjJtY

[6] Xifratge Cèsar
http://ca.wikipedia.org/wiki/Xifratge_de_C%C3%A8sar

[7] OAuth
http://es.wikipedia.org/wiki/OAuth

[8] Mineria de dades
http://ca.wikipedia.org/wiki/Mineria_de_dades

[9] Facebook costing 16-32s jobs in tough economic climate
http://ondeviceresearch.com/blog/facebook-costing-16-34s-jobs-in-tough-economic-climate?utm_source=On+Device+Research+newsletter&utm_campaign=060ae12252-YPCC_Index_May_20135_22_2013&utm_medium=email&utm_term=0_9d4e76bb4b-060ae12252-337096201#sthash.F2xfgFzu.4xgDGYMm.dpbs

[10] Maintain session
http://www.journaldev.com/1907/java-servlet-session-management-tutorial-with-examples-of-cookies-httpsession-and-url-rewriting

4. Annexos

4.1. Annex 1

```
package com.florit.uoc.pfc;

import java.util.ArrayList;
import java.util.Date;

public class Perfil{

  public String nomXarxa;
  public String urlXarxa;

  public int id;
  public String nick;

  /** url de l'imatge */
  public String avatar;

  public ArrayList<Mencio> mencions;
  public ArrayList<Contacte> contactes;

}

/**
 * Comentari escrit per l'usuari observat a la xarxa
 * social
 *
 */
class Mencio{

  public int id;

  public String text;
  public Date data;

  /** Usuaris anomenats a la menció */
  public ArrayList<Contacte> mencionats;

}

/**
 * Contacte associal a un tercer perfil "amic"
 * de l'usuari observat.
 *
 */
class Contacte extends Perfil{

}
```

4.2. Annex 2

```
JAXBContext context = JAXBContext.newInstance(Mencio.class);
Marshaller m = context.createMarshaller();
m.setProperty(Marshaller.JAXB_FORMATTED_OUTPUT, true);
Mencio object = new Mencio();
object.setName("Ismael Florit Zacarias");
object.setAvatar("http://server.com/file.ext");
object.setDate(new Date());
m.marshal(object, new FileOutputStream("mencio.xml"));
```

i resutat

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<perfil>
    <id>0</id>
    <avatar>http://server.com/file.ext</avatar>
    <nick>Ismael Florit Zacarias</nick>
</perfil>
```

4.3. Annex 3

Part del document HTML (versió mòbil) d'una menció al perfil de Facebook d'un usuari

```
<div class="tlUnit accelerateContainer hasSecondaryActions
accelerate acw" id="u_6_12" data-sigil="tlUnit marea"><div
class="tlUnitActor tlUnitTop"><div class="tlUnitActorWrap"><div
class="mhideout" style="display:table-cell;" data-sigil="m-hide-
outside"><div class="tlUnitActorProfilePic"><i class="img
profilePic profpic" style="background:#d8dce6
url(&quot;https://fbcdn-profile-a.akamaihd.net/hprofile-ak-
prn2/s34x34/276354_1379589842_1583814775_q.jpg&quot;) no-repeat
center;background-size:100% 100%;-webkit-background-size:100%
100%;width:34px;height:34px;" aria-hidden="true" aria-
label="Ismael Florit Zacarias" role="img"></i></div></div><div
class="tlUnitHeader"><span class="tlName tlActor">Ismael Florit
Zacarias</span> <div class="tlHeaderMetadata"><span
class="mfss"><abbr data-
store="{&quot;time&quot;:1386967894,&quot;short&quot;:false}"
data-sigil="timestamp" data-store-id="65">El 13 de diciembre a
la(s) 21:51</abbr><b><span role="separator" aria-hidden="true">
· </span></b><i class="feedAudienceIcon img sp_u0asc7
sx_dbdfdf"></i></span></div></div></div></div><div
class="tlAboveUnit"><span class="tlActorText">"Cena"
lista</span></div><div class="tlUnitContent"><div
class="mhideout" data-sigil="m-hide-outside"><a
class="darkTouch"
href="/photo.php?fbid=10202750354144847&amp;id=1379589842&amp;se
t=a.2011131962827.2118979.1379589842&amp;source=46&amp;refid=17"
><i class="img tlPhoto tlPhotoAttachment" data-
store="{&quot;mode&quot;:&quot;scale-photo-
attachment&quot;,&quot;aspectRatio&quot;:1.5,&quot;padding&quot;
:10,&quot;widthRatio&quot;:1,&quot;outerPadding&quot;:null}"
id="u_6_13" style="background-image: url(https://fbcdn-photos-b-
a.akamaihd.net/hphotos-ak-
frc3/c0.40.960.639/s720x720/1459163_10202750354144847_1635441731
```

```
_n.jpg); background-color: rgb(255, 255, 255); background-size:
100%; -webkit-background-size: 100%; width: 1244px; height:
829px; background-position: 50% 50%; background-repeat: no-
repeat no-repeat;" aria-label="Una foto de Ismael Florit
Zacarias." role="img" data-sigil="orientation-resizable
touchable" data-store-id="57"></i></a></div></div><div
class="tlBelowUnit"><div class="feedbackInlineWrap newUFI
async_like async_composer inlineShare" id="u_6_14"><div
class="slimLike" id="counts_feedback_inline_10202750354144847"
data-sigil="feed-ufi-trigger"><a
href="/photo.php?fbid=10202750354144847&amp;id=1379589842&amp;se
t=a.2011131962827.2118979.1379589842&amp;source=46&amp;refid=17"
><span class="like_def">2 Me gusta<span role="separator"
class="sep" aria-hidden="true">·</span></span><span
class="like_opt">3 Me gusta<span role="separator" class="sep"
aria-hidden="true">·</span></span><span class="cmt_def">1
comentario</span></a></div><div class="ufiBorder
ufiContainer"><div class="inlineCommentLike _51rb _51rc"
id="b_feedback_inline_10202750354144847"><div
class="ufiActions"><div class="button _40pa equalWidth"><a
class="touchable like_def"
href="/a/timeline/feedback/unit/?ut=69&amp;wstart=0&amp;wend=138
8563199&amp;hash=-
4611215054248343282&amp;impressionid=f03bb776&amp;action=like&am
p;profileID=1379589842&amp;nodeID=u_6_14&amp;shareID=10202750354
144847&amp;gfid=AQDAL15byyowRgfp&amp;refid=17" role="button"
aria-label="Me gusta" data-method="post" data-ajaxify-
class="async_like" data-toggle-class="like" data-sigil="like
touchable ajaxify toggleable"><i class="centerAligned img
sp_u0asc7 sx_2f133e"></i><strong>Me gusta</strong></a><a
class="like_opt touchable"
href="/a/timeline/feedback/unit/?ut=69&amp;wstart=0&amp;wend=138
8563199&amp;hash=-
4611215054248343282&amp;impressionid=f03bb776&amp;action=unlike&
amp;profileID=1379589842&amp;nodeID=u_6_14&amp;shareID=102027503
54144847&amp;gfid=AQDHHn4HDQjNQa7M&amp;refid=17" data-
method="post" data-ajaxify-class="async_like" data-toggle-
class="like" data-sigil="like touchable ajaxify toggleable"><i
class="centerAligned img sp_u0asc7 sx_0259b0"></i><strong
class="selected">Me gusta</strong></a></div><div class="_5lb3
button equalWidth"><a class="touchable"
href="/photo.php?fbid=10202750354144847&amp;id=1379589842&amp;se
t=a.2011131962827.2118979.1379589842&amp;source=46&amp;refid=17"
onclick="" role="button" aria-label="Comentar" data-
sigil="ufiCommentLink feed-ufi-trigger feed-ufi-trigger
touchable"><i class="centerAligned img sp_u0asc7
sx_f6d8ff"></i><strong>Comentar</strong></a></div><div
class="button _4-r9 equalWidth"><a class="touchable" data-
store="{&quot;share_id&quot;:&quot;10202750354144847&quot;,&quot
;behavior&quot;:&quot;custom&quot;}"
href="/sharer.php?sid=10202750354144847&amp;refid=17"
role="button" aria-label="Compartir" data-sigil="share-popup
touchable"><i class="centerAligned img sp_u0asc7
sx_c4d06f"></i><strong>Compartir</strong></a></div></div></div><
/div></div></div><div data-
store="{&quot;target&quot;:&quot;u_6_12&quot;,&quot;editOptions&
quot;:{&quot;hideBtn&quot;:{&quot;uri&quot;:&quot;\/a\/timeline\
/editunit?ut=69&amp;hash=-
4611215054248343282&amp;impressionid=f03bb776&amp;nodeID=u_6_12&
amp;timelineContext=\u00257B\u002522profile_id\u002522\u00253A13
79589842\u00252C\u002522start\u002522\u00253A0\u00252C\u002522en
d\u002522\u00253A1388563199\u00252C\u002522query_type\u002522\u0
```

```
0253A36\u00257D&amp;act=2&amp;gfid=AQDmo_sGE5rQ80AB&quot;},&quot
;editPostBtn&quot;:{&quot;uri&quot;:&quot;\/edit\/post\/dialog\/
?cid=1379589842\u00253A306061129499414\u00253A69\u00253A0\u00253
A1388563199\u00253A-
4611215054248343282&amp;ct=1&amp;nodeID=u_6_12&amp;redir=\u00252
Fismael.florit&quot;,&quot;rel&quot;:&quot;dialog&quot;},&quot;e
ditPrivacyBtn&quot;:{&quot;uri&quot;:&quot;\/privacy\/selector\/
dialog\/?cid=1379589842\u00253A306061129499414\u00253A69\u00253A
0\u00253A1388563199\u00253A-
4611215054248343282&amp;ct=5&amp;nodeID=u_6_12&amp;auto=1&amp;gf
id=AQCMCvOpObo9NuP4&quot;,&quot;rel&quot;:&quot;dialog&quot;},&q
uot;deleteBtn&quot;:{&quot;uri&quot;:&quot;\/a\/timeline\/editun
it?ut=69&amp;hash=-
4611215054248343282&amp;impressionid=f03bb776&amp;nodeID=u_6_12&
amp;timelineContext=\u00257B\u002522profile_id\u002522\u00253A13
79589842\u00252C\u002522start\u002522\u00253A0\u00252C\u002522en
d\u002522\u00253A1388563199\u00252C\u002522query_type\u002522\u0
0253A36\u00257D&amp;sa&amp;act=remove_content&amp;gfid=AQAFhHl3D
ubOTm5n&quot;}}}" data-sigil="m-unit-popup-opener"><a class="sec
_3u4 editButton" href="#" role="button" aria-haspopup="true"
data-sigil="touchable unit-popup-causal"><i class="img sp_u0asc7
sx_3ec5c0" data-sigil="unit-popup-context"><u>Más
opciones</u></i></a></div></div>
```

Extracció de la informació d'una menció a Facebook via Java

```
package com.florit.uoc.pfc;

import java.io.File;
import java.io.IOException;
import java.util.Collection;
import java.util.HashMap;
import java.util.Map;

import javax.xml.parsers.DocumentBuilder;
import javax.xml.parsers.DocumentBuilderFactory;
import javax.xml.parsers.ParserConfigurationException;

import org.apache.commons.collections.MapUtils;
import org.w3c.dom.Document;
import org.w3c.dom.NamedNodeMap;
import org.w3c.dom.Node;
import org.w3c.dom.NodeList;
import org.xml.sax.SAXException;

public class ExtraccioMencio{

  public static void main(String[] args) throws ParserConfigurationException,
SAXException, IOException{

    File arxiu = new File("/Users/ismael/Documents/workspace/PFC/mencio.xml");
    DocumentBuilderFactory dbFactory = DocumentBuilderFactory.newInstance();
    DocumentBuilder dBuilder = dbFactory.newDocumentBuilder();
    Document doc = dBuilder.parse(arxiu);

    //optional, but recommended
    //read this - http://stackoverflow.com/questions/13786607/normalization-
in-dom-parsing-with-java-how-does-it-work
    doc.getDocumentElement().normalize();


    HashMap<String, String> dadesMencioFacebook = new
```

```
HashMap<String,String>();
    dadesMencioFacebook.put("nom", "tlName");
    dadesMencioFacebook.put("avatar", "profilePic");
    dadesMencioFacebook.put("data", "mfss");

//    HashMap<String, String> dadesMencioTwitter = new
HashMap<String,String>();
//    dadesMencioTwitter.put("nom", "screen-name");
//    dadesMencioTwitter.put("avatar", "avatar");
//    dadesMencioTwitter.put("data", "metadata");
//
//    HashMap<String, String> dadesMencioLinkedIn = new
HashMap<String,String>();
//    dadesMencioLinkedIn.put("nom", "title-v2");
//    dadesMencioLinkedIn.put("avatar", "profile-photo");
//    // corresponent a la data d'una feina (que tractariem com una mencio)
//    dadesMencioLinkedIn.put("data", "cell-subtitle-v2");


    Collection<String> dades=dadesMencioFacebook.keySet();
    for(String d:dades){
      String dAttr = dadesMencioFacebook.get(d);

      NodeList nodes=doc.getChildNodes();
      for(int i=0;i<nodes.getLength();i++){
        Node node=nodes.item(i);
        System.out.println(d+": "+getClassValue(dAttr, node));
      }

    }
  }



  private static String getClassValue(String name, Node n){
    if(n==null)
      return null;

    if(n.getAttributes()==null || n.getAttributes().getNamedItem(name)==null){
      if(n.getAttributes()!=null){
        NamedNodeMap attrs=n.getAttributes();
        for(int i=0;i<attrs.getLength();i++){
            if(attrs.item(i).getNodeName().equals("class")){
              if(attrs.item(i).getNodeValue().contains(name)){

if(n.getTextContent()==null||n.getTextContent().trim().length()==0){
                  for(int j=0;i<attrs.getLength();j++){
                    if(attrs.item(j).getNodeValue().contains("http")){
                      return attrs.item(j).getTextContent();
                    }
                  }
                }else
                  return n.getTextContent();
              }
            }
        }
      }

      NodeList subN=n.getChildNodes();
      for(int i=0;i<subN.getLength();i++){
        String res=getClassValue(name,subN.item(i));
        if(res==null){
          continue;
        }else{
          return res;
        }
      }
    }
```

```
    return null;
  }
}
```

i resultant

```
data: El 13 de diciembre a la(s) 21:51 ·

avatar: background:#d8dce6 url("https://fbcdn-profile-
a.akamaihd.net/hprofile-ak-
prn2/s34x34/276354_1379589842_1583814775_q.jpg") no-repeat
center;background-size:100% 100%;-webkit-background-size:100%
100%;width:34px;height:34px;

nom: Ismael Florit Zacarias
```

4.4. Annex 4

Extracció de les dades EXIF d'una fotografía.

```
package com.florit.uoc.pfc;

import java.io.File;
import java.io.IOException;
import java.util.Collection;

import com.drew.imaging.ImageMetadataReader;
import com.drew.imaging.ImageProcessingException;
import com.drew.metadata.Directory;
import com.drew.metadata.Metadata;
import com.drew.metadata.Tag;

public class ExtraccioEXIF{

  public static void main(String[] args) throws ImageProcessingException,
IOException{

    Metadata metadata = ImageMetadataReader.readMetadata(new
File("/Users/ismael/Dropbox/Camera Uploads/2013-12-04 17.34.23.jpg/"));
    Iterable<Directory> metadataDirs=metadata.getDirectories();
    for(Directory dir:metadataDirs){
      Collection<Tag> tags=dir.getTags();
      for(Tag tag:tags){
        System.out.println(tag.getTagName()+": "+tag.getDescription());
      }
    }
  }

}
```

i obtenim la sèrie de dades (molta més de la que podriem necessitar per un estudi)

```
Compression Type: Baseline
Data Precision: 8 bits
Image Height: 2448 pixels
Image Width: 3264 pixels
Number of Components: 3
Component 1: Y component: Quantization table 0, Sampling factors 2 horiz/2 vert
Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Exposure Time: 1/15 sec
F-Number: F2,4
Exposure Program: Program normal
ISO Speed Ratings: 3200
Exif Version: 2.21
Date/Time Original: 2013:12:04 17:34:23
Date/Time Digitized: 2013:12:04 17:34:23
Components Configuration: YCbCr
Shutter Speed Value: 1/15 sec
Aperture Value: F2,4
Brightness Value: -7889/1739
Metering Mode: Spot
Flash: Flash did not fire, auto
Focal Length: 4,12 mm
Sub-Sec Time Original: 083
Sub-Sec Time Digitized: 083
FlashPix Version: 1.00
Color Space: sRGB
Exif Image Width: 3264 pixels
Exif Image Height: 2448 pixels
Sensing Method: One-chip color area sensor
Scene Type: Directly photographed image
Exposure Mode: Auto exposure
White Balance Mode: Auto white balance
```

```
Focal Length 35: 66mm
Scene Capture Type: Standard
Lens Specification: 103/25 103/25 12/5 12/5
Lens Make: Apple
Lens Model: iPhone 5 back camera 4.12mm f/2.4
Make: Apple
Model: iPhone 5
Orientation: Top, left side (Horizontal / normal)
X Resolution: 72 dots per inch
Y Resolution: 72 dots per inch
Resolution Unit: Inch
Software: 7.0.4
Date/Time: 2013:12:04 17:34:23
YCbCr Positioning: Center of pixel array
GPS Latitude Ref: N
GPS Latitude: 47.0° 33.0' 21.86999999999955"
GPS Longitude Ref: E
GPS Longitude: 7.0° 35.0' 44.45999999999742"
GPS Altitude Ref: Sea level
GPS Altitude: 248 metres
GPS Time-Stamp: 16:34:22 UTC
Thumbnail Compression: JPEG (old-style)
X Resolution: 72 dots per inch
Y Resolution: 72 dots per inch
Resolution Unit: Inch
Thumbnail Offset: 1190 bytes
Thumbnail Length: 6119 bytes
```