MASTER IN COMPUTER SCIENCE

Une ontologie pour le profilage des sites de réseaux sociaux par rétro ingénierie

Crémer, Véronique

Award date:
2011

Awarding institution:
University of Namur

Facultés Universitaires Notre-Dame de la Paix Namur,
Faculté d'informatique
Année académique 2010-2011

Une ontologie pour le profilage des sites de réseaux sociaux par rétro ingénierie

Véronique Crémer

Mémoire présenté en vue de l'obtention du grade de master
en Sciences Informatiques

Résumé

Les sites de réseaux sociaux, il y a des centaines. Des millions de personnes y sont inscrites. Ils offrent la possibilité de créer un profil pour ensuite tisser des liens avec d'autres membres. Ils permettent également publier des messages, des articles, des photographies, etc. Tout ceci laisse des traces dans les bases de données de ces sites, ce qui n'est pas toujours sans conséquence. C'est justement cette connaissance que ce mémoire, dans un premier temps, tente d'évaluer via une représentation de la connaissance, c'est-à-dire, une ontologie.

Dans un deuxième temps, le mémoire envisage la problématique du risque de connexion de deux profils, ainsi que de ces conséquences en cas de fusion de deux sites de réseaux sociaux différents.

Abstract

There are hundreds of social network sites. Millions of people are registered. They offer the opportunity to create a profile and then create links with other members. They also allow to post messages, articles, photographies, etc.. All this leaves traces in the databases of these sites, which is not always without consequence. It is precisely this knowledge that this master thesis, firstly, is looking through a knowledge representation, ie, an ontology.

In a second step, the master thesis considers the problem of the risk of connecting two profiles, as well as the consequences in case of merging of two different social networking sites.

Je tiens à remercier toutes les personnes qui m'ont aidées lors de l'élaboration de ce mémoire.

D'abord le professeur Englebert, pour sa disponibilité et ses conseils avisés, ainsi que le professeur Petit et le professeur Claire Lobet pour leur aide.

Ensuite mes filles, Bérénice et Mélusine pour leur soutien et leur patience, ainsi que leur papa, pour son aide durant toutes ces années d'étude.

Table des matières

Chapitre 1

Introduction

Les sites de réseaux sociaux, appelés « social network sites » en anglais, offrent la possibilité à leurs utilisateurs de créer un profil, c'est-à-dire une identité virtuelle sur le net. Ils permettent également de tisser des liens avec d'autres membres via les listes d'amis. Une fois inscrit, les utilisateurs peuvent alors publier des messages, des articles, des photographies, etc.

Des sites de ce genre, il y en a des centaines. On en parle régulièrement dans les médias. Ils sont devenus très populaires. Des millions de personnes de par le monde se connectent fréquemment pour discuter avec des amis, publier du contenu, etc. Tout ceci laisse des traces: toute l'activité d'un membre, ainsi que ses informations de profil sont ainsi conservées. À cela, il faut également ajouter toutes les informations fournies par des tiers tels que amis ou sites partenaires. Beaucoup d'informations personnelles peuvent ainsi être rassemblées et stockées en un seul endroit.

Le fait que les sites aient la possibilité de stocker autant d'informations sur leurs membres n'est pas toujours sans conséquence. En effet, celles-ci peuvent être analysées ou revendues; ou encore, avec un impact beaucoup plus direct, elles peuvent se retrouver dans le domaine publique et servir soit à des fins malhonnêtes, soit à entacher la réputation d'une personne. Par conséquence, la connaissance qu'un site a la possibilité d'avoir à propos de ses membres peut avoir une très grande valeur. C'est justement cette connaissance que ce mémoire tente d'évaluer via une `représentation de la connaissance`, c'est-à-dire, une ontologie.

Pour ce faire, nous avons essayé de déterminer quelles sont les informations collectées. Le domaine étant fort vaste et absolument pas documenté,

nous avons dû restreindre la recherche aux informations données directement par l'utilisateur, ainsi que par son activité au sein même du site.

Une fois l'ontologie construite, nous avons examiné l'évolution de la connaissance en cas de fusion des données de deux sites de réseaux sociaux. En effet, ces derniers sont actuellement en pleine mouvance et les rachats ne sont pas rares. Un individu pourrait alors voir deux de ses profils connectés. Les risques de connexion ainsi que les conséquences sur la connaissance relative à l'individu ont donc été envisagés.

Dans le deuxième chapitre de ce document, nous avons tenu à préciser les termes `site de réseau social` , ceux-ci étant souvent employés de façons diverses. Le troisième chapitre envisage quelques enjeux liés aux sites de réseaux sociaux. Le chapitre quatre examine quelques cas pouvant amener un tiers à avoir accès à une partie ou à la totalité des données stockée sur le site. Les chapitres cinq et six définissent les notions d'ontologie et de raisonneur. Le septième chapitre consiste en la description de la méthode utilisée afin de répondre à la question posée par le mémoire. Le chapitre huit décrit les résultats que nous avons extrait de l'ontologie. Le chapitre neuf envisage quelques cas particuliers. Enfin, le dernier chapitre conclut ce travail.

Chapitre 2

Les réseaux sociaux

Depuis quelques années, les sites de réseaux sociaux - ou social network site - sont en pleine effervescence. Actuellement, il ne se passe pas une semaine, voire quelques jours sans qu'on en entende parler. Pour s'en rendre compte, il suffit de faire une recherche dans les actualités de la semaine de Google avec les termes « réseau social » pour voir plusieurs milliers de résultats s'afficher (voir figure 2.1).

Les personnalités importantes se doivent d'avoir leur profil dans un ou plusieurs sites. De Didier Reynders [1] à Sarah Palin en passant par Barack Obama, dont la campagne présidentielle était même centrée sur les réseaux sociaux [2], les politiciens ne peuvent ignorer ce moyen de séduire les électeurs.

Dans le cadre des protestations en Afrique du Nord et au Moyen-Orient en ce début de 2011 [3], les réseaux ont offert un vent de liberté en autorisant l'échange d'idées, d'informations, en permettant le rassemblement des opposants au gouvernement, et ce, malgré les tentatives de répression du pouvoir [4].

Un film retraçant, de manière romancée, la création du réseau social Facebook [5] est même sorti fin 2010.

Rares sont les personnes qui n'ont jamais entendu parler de Facebook ou

1. Article du Vlan Bruxelles `blogs.vlan.be/vlanbruxelles/reynders-bat-di-rupo-sur-facebook/`
2. Article du New York Time `www.nytimes.com/2008/07/07/technology/07hughes.html`
3. Wikipedia `fr.wikipedia.org/wiki/Protestations_dans_les_pays_arabes_de_2010-2011`
4. Voir, par exemple, l'article du Monde diplomatique `http://blog.mondediplo.net/2011-02-15-La-revolution-arabe-fille-de-l-Internet`
5. The social Network, film américain réalisé par David Fincher

FIGURE 2.1 – La recherche de « réseau social » dans les actualités de la semaine fournit plusieurs milliers de résultats

de Twitter. Tous en ont un idée plus ou moins précise. Nous allons tenter de définir de manière rigoureuse ce qu'est un site de réseaux social.

2.1 Introduction à la notion de sites de social network

Nous allons tenter d'apporter une définition à la notion de site de réseau social, en partant du plus basic, c'est-à-dire la définition du mot réseau, pour en arriver à la notion de site de réseau social.

2.1.1 Notion de réseau

Selon [13], le mot réseau est un mot ancien. Il vient du latin « retiolus » qui signifie petit filet. C'est un diminutif de « retis » (filet). Ce terme a notamment donné le mot « rets » [6], ainsi que réticulaire [7].

En vieux français « retiolus » devient « réseul » (XVIIe siècle) puis « rézeau » (fin XVIIe siècle) et enfin réseau. Il désigne un « tissu en forme de rets », destiné à la capture de poisson, d'oiseau, mais aussi utilisé pour retenir les cheveux (d'où résille [8]). On y trouve ainsi une notion de maillage et de capture.

Au Moyen âge, le mot réseau est également appliqué au système sanguin à cause de sa forme de filet; mais sans notion de circulation. En effet, à cette époque, on ignorait le rôle du cœur, ainsi que le principe du recyclage du sang. Pour cela, il faut attendre le début du XVIIe siècle.

Avec cette extension s'ajoute ainsi l'idée implicite de circulation le long des fibres du réseau. Cela devient encore plus explicite quand, au XIXe siècle, le terme est utilisé pour désigner l'ensemble des routes ou des voies de chemin de fer d'une région.

Toujours au XVIIIe siècle, l'abbé La Caille introduit la notion de « réseau » de triangle. Il s'agit de découper un espace par des triangles dans le but de tracer une carte. Il ajoute ainsi une notion géographique au terme.

Les réseaux ont également été repris par les mathématiques dans la théorie des graphes. Un réseau peut être vu comme un ensemble de nœuds reliés entre eux par des liens.

6. Filet pour prendre des oiseaux, des poissons - dictionnaire Larousse
7. Qui a la forme d'un filet, d'un réseau - dictionnaire Larousse
8. Filet pour envelopper les cheveux - dictionnaire Larousse

C'est vers le milieu du XIX siècle, que la sociologie s'empare du terme et commence à l'utiliser pour désigner un ensemble d'individus ainsi que les liens qui unissent ces mêmes individus.

Le terme est donc associé à l'idée de maillage fermé et de circulation le long de ces mailles.

2.1.2 Notion de réseau social

Un réseau social représente une structure sociale dynamique faite de nœuds (individus ou organisations) reliés entre eux par des canaux (des relations sociales). Ainsi, selon Lazega [9] [12],

> *Un réseau social est généralement défini comme un ensemble de relations d'un type spécifique (par exemple de collaboration, de soutien, de conseil, de contrôle ou d'influence) entre un ensemble d'acteurs.*

On retrouve dans cette définition la notion de maillage dans les relations, ainsi que la notion de circulation dans la transmission d'informations entre les individus le long du maillage.

L'analyse de ces réseaux est fondamentale dans plusieurs domaines tels que la sociologie, l'anthropologie, la géographie, la psychologie,... Toujours selon Lazega [12],

> *L'analyse des réseaux est une méthode de description et de modélisation de la structure relationnelle de cet ensemble.*

Parmi les grands noms de l'analyse, citons, par exemple, Jacob Lévy Moreno [10] (1892-1974) dont le test sociométrique permet de décrire la structure d'un groupe. Il consiste à demander à chaque personne qui, parmi les membres du groupe, elle souhaiterait ou ne souhaiterait pas avoir comme compagnon (ami, collègue, voisin,...). La résultante est un sociogramme (dessin) permettant de mesurer des relations affectives dans un groupe. Il met ainsi en évidence les leaders (les membres du groupe choisis par une majorité de membres), les isolés (les membres qui ne recueillent que de l'indifférence), les parias (membres systématiquement rejetés).

9. Professeur de sociologie économique à l'Université de Lille-1 et auteur de plusieurs livres sur le thème des réseaux sociaux

10. Médecin psychiatre, psychosociologue, théoricien et éducateur américain d'origine roumaine, il est l'un des pionniers de la psychothérapie de groupe.

L'analyse des réseaux sociaux puise abondamment dans les mathématiques avec la théorie des graphes. C'est Moreno qui a ouvert la voie avec son sociogramme. La théorie des graphes fournit non seulement une méthode de représentation graphique, mais également un ensemble de concepts, de classifications de propriétés,...

Plus d'informations peuvent être trouvées sur le site de l'INSNA (Network for Social Network Analysis, `http://www.insna.org/`), l'association professionnelle d'analyse de réseaux sociaux.

2.1.3 Notion de réseautage social

Le terme « réseautage » est un mot récent inventé afin de traduire le terme anglais « networking. » Il signifie « disposition en réseau, maillage ». C'est l'action correspondant au verbe « réseauter » défini par « se construire un réseau ».

En sociologie, « un réseautage social » a une signification différente de celui de « réseau social. » : un réseau social est un maillage entre des individus; un service de réseautage social est un outil facilitant la connexion des individus.

> *Le réseautage social (...) se rapporte aux moyens mis en œuvre pour relier les personnes entre elles [4]*

Une association d'anciens, une agence de rencontre,... permettent de tisser des liens ; une recommandation est aussi un moyen de tisser des liens. Ils offrent donc des services de réseautage.

Avec l'arrivée d'internet et de Web 2.0 - les internautes deviennent acteurs - la notion de réseautage a été ainsi étendue aux sites permettant de tisser un réseau social sur la toile.

2.1.4 Notion de site de réseau social - Social Network Site

En repartant de la définition des différents mots composant l'expression, un « site de réseau social » peut être défini comme un site internet permettant d'exprimer un ensemble de relations d'un type spécifique entre un ensemble d'acteurs.

De même, on peut définir un « site de réseautage social (Social networking site) » comme un site internet favorisant la constitution d'un réseau social.

Dans la littérature, les termes « site de réseau social (social network site) » et « site de réseautage social (social networking site) » sont souvent utilisés indifféremment. Il y a pourtant une différence. Le site de réseautage favorise l'extension ou la création d'un réseau social, quand le site de réseau social exprime simplement des relations entre des acteurs.

Boyd [11] et Ellison [12] proposent l'utilisation du terme « network (réseau) » au lieu de « networking (réseautage) ». Elles considèrent que le terme networking n'est pas adapté. Elles basent leur argumentation sur le fait que, excepté sur quelques sites bien spécifiques tel que LinkedIn, la plupart des utilisateurs ne cherchent pas à étendre leurs réseaux sociaux, mais à se connecter à des personnes déjà connues. La plupart des utilisateurs vont majoritairement se mettre en rapport avec une personne faisant déjà partie de leur réseau social. Il y a bien sûr des exceptions, tels des politiciens, des entreprises, etc. essayant de se faire connaître.

Le terme « site de réseau social » est un site sur lequel des utilisateurs sont connectés à d'autres utilisateurs, formant ainsi un réseau. Un « site de réseautage social » correspondra plus à un site permettant d'agrandir un réseau.

2.1.5 Notion de profil

Un réseau social sur le web est basé sur des liens tissés entre des identités virtuelles. Une identité virtuelle est une identité créée sur le Net. Elle peut être le reflet d'une identité réelle ou être fabriquée. C'est la face publique qu'un individu accepte de dévoiler sur les sites. Elle peut se traduire soit simplement par l'utilisation d'un pseudo, soit par une identité fidèle à l'identité réelle, soit par une identité totalement différente,... Le profil sur un site de réseau social est l'ensemble des informations stockées sur ce même site et relative à l'identité virtuelle.

La majorité des définitions d'un réseau social sur le Net inclue la notion de profil. Par exemple, sur le blog du Figaro, Laurent Suply [13] énonce, en

11. Chercheuse pour Microsoft Research et membre du Harvard Law School's Berkman Center for Internet and Society, elle est connue pour ses recherches sur les pratiques des jeunes dans le contexte des réseaux sociaux

12. Professeur-assistante à l'université de l'état du Michigan. Ses recherches explorent entre autre le thème des réseaux sociaux, de la représentation de soi sur les sites de rencontre,...

13. Journaliste au Figaro

parlant d'un site de réseau social :

> *Le terme désigne un site internet permettant à l'internaute de s'inscrire et d'y créer une carte d'identité virtuelle appelée le plus souvent « profil »(...)*

Un utilisateur crée un profil en entrant les informations requises : nom, adresse email, date de naissance, pseudo... Ensuite, le membre peut l'enrichir par des photographies, la liste de ses centres d'intérêts ou toutes autres informations qu'il souhaite associer à son profil.

L'étape suivante consiste à rechercher d'autres utilisateurs afin de créer des liens via le site. Les personnes choisies comme « amies » le seront sur base d'éléments de leur profil comme par exemple le nom, l'adresse email, l'école, l'intérêt pour un sujet donné,...

Selon Boyd et Ellison, la notion de profil fait partie intégrante de la notion de site de réseau social:

> *We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site. [6]*

Selon les auteurs, un site de réseau social doit donc avoir 3 caractéristiques:

- Le maintien d'un profil
- La possibilité de créer des connexions avec d'autres utilisateurs
- La possibilité de voyager à travers ces connexions en cliquant sur un nom se trouvant parmi une liste d'amis.

Le dernier point est un ajout par rapport aux définitions précédentes. Il permet de se rapprocher de la définition d'un réseau dans le sens qu'il permet ainsi de « dessiner » la carte d'un réseau. Dans la quasi majorité des sites, la liste des amis d'un individu (virtuelle) est visible soit aux personnes autorisées à voir le profil, soit à tout le monde. Pour quelques exceptions, cette visibilité est paramétrable.

2.1.6 Réseau social

Pour la suite de ce document, nous considérerons qu'un site de réseau social est un site qui permet au moins

- de créer un profil: c'est la représentation d'une identité virtuelle nécessaire à la création de liens virtuels entre les membres d'un réseau.
- de rechercher des « amis » : indispensable afin de choisir les individus avec qui une relation sera souhaitée, et donc de créer des liens.
- de créer des liens avec d'autres profils: c'est le principe même du réseau social.

Ces caractéristiques sont présentes dans tous les sites de réseaux sociaux classiques. On peut également affirmer que les sites de blogs ou de diffusions tels que Skyblog ou Youtube sont des sites de réseaux sociaux parce qu'ils respectent les fonctionnalités décrites ci-dessus. Par contre, un site comme Wikipédia [14]ne peut pas être considéré comme un site de réseau social. En effet, même s'il permet la création de profils, il n'est pas possible de créer des liens entre les profils.

2.1.7 Les premiers sites

Les sites de réseaux sociaux sont les purs produits du Web 2.0. Avec le Web 1.0, internet se focalisait sur le contenu d'un site. L'introduction du Web 2.0 a déplacé le centre d'intérêt vers l'utilisateur en lui permettant de devenir acteur, d'interagir. On a vu ainsi l'émergence

- de sites d'expression tels que les blogs,
- de sites permettant le contact entre individus comme les services de chat, les listes de diffusion, etc.

Selon Wikipedia, le plus ancien site de réseautage social est Classmate.com, créé en 1995. Il permettait à des personnes d'une même école, ou d'une même entreprise,... de se rejoindre sur le net. Néanmoins, il ne fut pas possible de créer un profil ou de lister ses amis avant plusieurs années. Le premier site de réseau social, avec toutes les fonctionnalités inhérentes, était SixDegrees.com, lancé en 1997.

Tous les sites de réseaux sociaux ne sont pas nés « site de réseau social ». Certains ont commencé leur vie comme site de partage de fichiers, de messagerie, de chat, de liste de diffusion, de blog,... Les fonctionnalités typiques

14. Projet d'encyclopédie collective établie sur Internet, `http://fr.wikipedia.org`

des sites de réseaux n'ont été ajoutées que par la suite. Par exemple, Skyrock(Skyblog) était un site proposant aux utilisateurs de créer leur blog. Il n'est devenu un « site de réseau social » en intégrant des services tels que la création de profil ou l'ajout d'amis, qu'à partir de mai 2007. Myspace est un autre exemple. Au départ, il s'agissait d'un site de stockage de données, qui est devenu, à partir de 2004 un vrai site de réseau social.

2.2 Classification

Dans ce mémoire, nous allons essayer de d'évaluer la connaissance qu'un site de réseau social peut avoir de ses membres. En d'autres mots, nous allons tenter de déterminer quelles sont les informations pouvant être contenues dans les bases de données de ces sites. Pour ce faire, nous allons analyser par rétro ingénierie quelques sites caractéristiques. Afin de nous aider à choisir les plus représentatifs parmi la foison de sites existants, nous allons tenter de les classer selon différentes caractéristiques.

FIGURE 2.2 – Quelques sites de réseaux sociaux parmi d'autres

2.2.1 Classification selon l'accessibilité

Un site peut être accessible à

- tout le monde (public). Par exemple: Facebook, Myspace, Hi5, Flickr, Skyrock, Friendster,...

- à certaines personnes uniquement (privé, i.e. uniquement sur invitation). Par exemple: aSmallWorld, Faceparty, grono.net, Parano.be,...
- selon un critère spécifique. Par exemple:
  - ASUIsTalking: accessible uniquement aux personnes ayant une adresse email de l'université de l'Arizona
  - Biip: accessible uniquement aux personnes ayant un numéro de téléphone Norvégien
  - College Tonight: accessible uniquement aux personnes ayant une adresse email finissant avec .edu

Dans ce mémoire, nous ne travaillerons qu'avec des sites publiques. Il est en effet difficile d'accéder, et donc de déterminer les données collectées dans les sites dont l'accès est limité. De plus, les sites les plus populaires et les plus utilisés sont en grande majorité des sites publiques.

En ce qui concerne les sites de type privé, vu justement leur caractère privé, on peut supposer que, d'une part, les membres sont plus enclin à se dévoiler, et que d'autre part, les données personnelles stockées par le réseau sont mieux protégées [15].

2.2.2 Classification selon la divulgation de l'identité

Selon [9], on peut classer les sites de réseaux sociaux en fonction de ce qui est affiché aux internautes afin de leur permettre d'identifier un profil. Certains sites dévoilent le nom complet. Ce sont ceux qui encouragent l'utilisation de la véritable identité. Par exemple, extrait des conditions d'utilisation de Facebook:

> *Les utilisateurs de Facebook donnent leur vrai nom et de vraies informations les concernant.*
>
> *Vous ne fournirez pas de fausses informations personnelles.*
>
> *Vous mettrez vos coordonnées, exactes, à jour.*

En s'inscrivant à Facebook, les utilisateurs acceptent les conditions d'utilisation du site. Ils sont donc contractuellement obligés de s'identifier correctement. On peut alors supposer que la majorité des profils sont de vrais profils et que l'identité dévoilée par le profil est la véritable identité [16]. Le

15. Témoignage d'un membre de ASmallWorld `http://tfmc.blogs.com/the_flying_monkey_circus/2007/11/a-small-world-c.html`

16. Réseaux sociaux et sécurité, article de l'AWT, `http://www.awt.be/web/sec/index.aspx?page=sec,fr,foc,100,067`

but est de faciliter les recherches et les retrouvailles entre personnes ayant déjà des liens. Dans cette catégorie, outre Facebook, on peut citer, Copains d'avant, Twitter,...

D'autres sites cachent la véritable identité de l'utilisateur et se contentent d'afficher le prénom. Il s'agit, par exemple de sites tels que Friendster, delicious.com. Il offre une (faible) protection à l'identification d'un profil.

Enfin, certains, plus rares, tel que Match.com, découragent l'utilisation de la véritable identité. Ils tentent de faire en sorte que le lien entre la personne réelle et la personne virtuelle soit difficiles à « deviner ».

Néanmoins, quelque soit la politique adoptée, les sites encouragent leurs utilisateurs à associer au profil une photographie de soi identifiable. Or celle-ci pose de sérieux problèmes vis-à-vis de la vie privée, parce qu'elle peut servir à identifier la personne se cachant derrière un profil qu'il croit anonyme.

2.2.3 Classification selon le thème

Les sites de réseaux sociaux peuvent avoir un thème particulier, réunissant ainsi des personnes ayant un intérêt commun. Parmi ces différents types, citons:

- Les plateformes de partage. Le but de ces sites est la diffusion de fichiers tels que vidéos, photographies, musiques,... L'objectif est de rendre le document accessible. Exemple de site: Youtube, Dailymotion
- les réseaux ayant un thème spécifique. Le but est de réunir des personnes autour d'un centre d'intérêt commun. Par exemple:
  · La photographie: Piczo, Woophy
  · Les jeux: Avatars United, Gaia Online
  · La religion: Soul Harvest, MyChurch
  · ...
- Les réseaux professionnels. Le but de ces réseaux est de faciliter le contact professionnel. Par exemple, un recruteur pourra y effectuer une recherche afin de trouver le candidat pour un travail donné; une entreprise pourra y rechercher des partenaires, etc.

2.2.4 Classification selon le moteur

Sur des sites de rencontres tel match.com, le profil a un rôle important. C'est lui qui va donner envie - ou non - de contacter un internaute [15]. La

relation est basée principalement sur ce qu'on est. De même, sur les sites plus classiques tel Facebook, pour augmenter la chance d'avoir de nouveaux amis, il faut donner un maximum d'informations dans son profil. L'application du site propose généralement un questionnaire assez exhaustif qu'il suffit de remplir pour en dévoiler beaucoup sur soi. Le profil est le moteur sous-jacent à la dynamique du site.

Dans les weblogs, tel LiveJournal, le profil passe au second plan [15]. L'important, c'est le contenu publié, ce que la personne dit. Un ami, c'est une personne que l'on autorise, ou même que l'on souhaite voir lire nos écrits. Ce contenu est tout aussi problématique, sinon plus, que les sites de réseaux sociaux classiques. En effet, en plus des informations de profils, certains internautes n'hésitent pas à dévoiler des choses très personnelles : des rencontres amoureuses, des expériences de vie, des problèmes de couple, etc.

2.2.5 Classification selon la confidentialité

Les possibilités de gestion de la confidentialité sont très variables d'un site à l'autre.

Sur des sites tels que Facebook ou Flickr, il est possible de paramétrer la confidentialité pour un grand nombre de choses. Des éléments du compte tels que photographies ou informations personnelles peuvent être rendus inaccessibles à tout ou une partie des membres du réseau. Ce sont généralement les sites divulguant la véritable identité qui permettent le plus de limiter les accès, donnant ainsi un sentiment de confidentialité, de sécurité. Mais la plupart des utilisateurs de ces sites ne modifient pas les paramètres de confidentialité instaurés par défaut [15]. Le niveau de privacité de la majorité des comptes dépend donc de la manière dont le logiciel supportant le site de réseau social est codé. Et celui-ci n'est pas implémenté afin de servir l'utilisateur, mais bien l'intérêt du site de réseau social. Les paramètres par défaut sont ainsi souvent beaucoup trop permissifs [17].

Sur d'autres sites, généralement des sites de rencontre, tel que match.com, il n'est pas possible de limiter l'accès aux informations du profil. En effet, le

17. Article de Electronic Frontier Foundation (EFF), organisation non gouvernementale international fondée en 1990 aux États-Unis, dont l'objectif essentiel est de défendre la liberté d'expression sur Internet. `http://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly`

but de ces sites étant la rencontre, il faut en montrer un maximum. Plus un utilisateur montre d'informations de lui-même, plus il a des chances d'avoir de nouveaux« amis ».

2.2.6 Classification selon le financement

Un site de réseau social a besoin de rentrées financières pour survivre : les serveurs, la gestion des comptes, l'électricité et autres coûts, demandent des moyens financiers parfois conséquents. Ces rentrées peuvent avoir plusieurs origines, telles que:

- vente d'espaces publicitaires ciblés en fonction du publique tel Facebook, MySpace, Skyrock, Flickr...
- ventes de données relatives à leurs membres tel Facebook [18]
- vente d'espaces publicitaires ciblés en fonction du contenu tel Youtube, Flickr,...
- vente de produits, de services ou de fonctionnalités supplémentaires tels qu'impressions de photographies sur différents supports pour Flickr ou assurances pour animaux sur DogCity

Le financement derrière un site peut avoir une certaine importance dans le choix des données collectées. Par exemple, si le site vend des espaces publicitaires ciblés en fonction du publique, il sera probablement plus enclin collecter les informations nécessaires à l'analyse des intérêts ses membres.

18. Article de The Gardian, `http://www.guardian.co.uk/technology/2007/sep/13/guardianweeklytechnologysection.news1`

Chapitre 3

Enjeux

En se référant à la théorie du « six degrés de séparation »- en théorie, tout individu est relié à un autre individu par une chaîne social de maximum 5 maillons (i.e. personnes) [1] - le réseau social fournit une porte d'entrée vers le monde entier! Chacun a ses motivations pour s'inscrire: pour rester en contact avec certaines personnes, pour se retrouver entre personnes ayant une passion commune, partageant des goûts semblables, pour partager et diffuser des vidéos, des photographies, pour rencontrer de nouveaux « amis » comme sur les sites de rencontres, pour étendre son réseau d'affaire, pour trouver un travail, pour faire comme tout le monde, parce que c'est la mode...

Il y a beaucoup de bonnes raisons de s'inscrire sur ces sites. Et il est clair qu'avoir un réseau social étendu peut-être bénéfique : pour trouver un emploi, développer un réseau d'affaire, etc.

De plus, les sites de réseaux sociaux facilitent les rencontres entre personnes ayant des points communs, partageant la même passion, ou les mêmes goûts.

C'est également une source de conseils, d'aide ou de réconfort, tel le mini site de réseau social Stobacco [2] dont le but est d'aider les fumeurs à arrêter. Ou encore cet autre site « Mémoire des vies [3] » proposant un espace de souvenirs de personnes disparues et qui, par la même occasion permet de trouver conseils et réconforts.

Malheureusement, l'usage de ces sites peut également comporter des risques aux conséquences parfois très fâcheuses ; ceux-ci sont majoritaire-

1. Théorie émise par Frigyes Karinthy, reprise par Stanley Milgram
2. `www.stobacco.com`
3. `www.memoiredesvies.com`

ment liés aux informations personnelles laissées par les utilisateurs. Nous allons nous focaliser sur certains de ces enjeux.

3.1 La Réputation Numérique

La réputation numérique est l'image que l'on se fait d'un individu à partir de toutes les informations présentes sur le net et relatives son sujet. Ces informations peuvent avoir de multiples origines : un avis sur un site d'achat, un blog, un commentaire une discussion sur un forum, etc. (voir figure 3.1).

Beaucoup de personnes ont une « réputation en ligne ». Cette réputation est basée sur les traces - y inclus le ou les profils sur les différents sites de réseaux sociaux - laissées par ou au sujet d'un individu sur la toile. Ces traces peuvent être stockées pour de longue période, soit dans les bases de données d'un site de réseau social, soit dans le cache de Google [4], soit dans des archives web tel que la Wayback Machine [5].

La réputation numérique peut être mauvaise: une photo d'une soirée un peu trop arrosée, un(e) ex qui lynche son ancien(e) ami(e) via son blog, un mécontent lors d'une transaction commerciale,...

L'opinion sur un individu est susceptible se retrouver limitée à ces sites qui peuvent donner une image peu flatteuse de la personne ; image qui peut être liée à des faits relativement anciens. Lorsqu'un individu se fait « googeliser », toutes ces traces apparaissent.

Pour certains, c'est la seule source permettant à un quidam de se faire une opinion sur un tel. Il n'est donc pas étonnant de voir cette réputation ressurgir dans la vie réelle. Il existe quantité de cas où des demandeurs d'emplois se sont vu recalés suite à des photographies compromettantes sur le net; Comme ce candidat qui s'est vu montrer une photographie de ses fesses à un entretien d'embauche [6]. Le cas Yoan Demarq est un autre exemple typique : suite à une transaction commerciale qui c'est mal déroulée, un homme (alias Tranquillosse) a placé une vidéo sur Dailymotion dans laquelle il expliquait s'être fait arnaquer par un certain Yoan Demarq lors d'une vente

4. Article du Journal du Net, `http://www.journaldunet.com/ebusiness/le-net/actualite/la-cnil-veut-initier-les-reseaux-sociaux-au-respect-de-la-vie-privee.shtml`

5. Service fournit par L'Internet Archive (IA) qui permet aux utilisateurs de voir les versions archivées de pages Web, `http://www.archive.org/`

6. RTL Info, `http://www.rtlinfo.be/info/archive/211684/`

FIGURE 3.1 – Cartographie de l'identité numérique (source Flickr)
La réputation numérique est basée sur les traces laissées par les utilisateurs. Ici, elles sont classées en fonction de l'activité sur le net. Pour plus d'information, voir l'article associé : `http://www.fredcavazza.net/2006/10/22/qu-est-ce-que-l-identite-numerique/`

sur le site « Leboncoin ». Rapidement, Yoan Demarq est devenu, pour le net, un escroc.

Les réseaux sociaux offrent la possibilité de gérer, en partie, son ou ses profils [17]. L'utilisateur a le choix d'y intégrer les informations qu'il souhaite et d'en cacher d'autres. Un profil est la face publique qu'un individu consent (de manière consciente ou non) à dévoiler sur le net. Il a donc un certain contrôle sur ce que les autres verront de lui et ainsi gérer, en partie, cette réputation. Une autre manière d'influencer sur sa réputation est d'utiliser son ou ses réseaux afin de créer soi-même des traces. Quant aux amis, comme dans la vie réelle, ils peuvent être de précieux appuis et fournir des recommandations quand cela s'avère utile.

Afin de gérer correctement son profil, il faut comprendre comment fonctionne la confidentialité sur le site. La manière dont celle-ci est implémentée est parfois incohérente, obscure , et surtout différente pour chaque site. Ainsi, des étudiants d'Oxford ont été surpris de constater que non seulement le réseau de l'université d'Oxford de Facebook est accessible à tout membre de l'université, y compris professeurs et surveillants, mais qu'en plus, leur profil est visible, par défaut, à tous membres appartenant au même réseau [7]. D'autres exemples de la manière dont Facebook et autres peuvent gâcher la vie de quelqu'un, voir [8]

3.2 Collecte d'informations personnelles

Les réseaux sociaux facilitent la collecte d'informations personnelles par des personnes pas toujours bienveillantes. Outre les données visibles directement via son profil, un utilisateur peut voir ses données divulguées de beaucoup de façons (voir section 4), et cela peut avoir des conséquences, comme, par exemple rendre crédible un faux profil, ou faciliter le piratage d'un compte de site web.

Le piratage d'un site tel que Yahoo, Msn, Twitter, etc. peut être rendu possible grâce aux questions secrètes. En effet, ces sites proposent aux utilisateurs de créer une question secrète lors de l'inscription. En cas d'oubli du mot de passe, il suffit alors de répondre à la question afin de pouvoir redéfi-

7. Article de The Guardian, `http://www.guardian.co.uk/media/2007/jul/17/digitalmedia.highereducation`

8. Article de The Independent, `http://www.uni-europa.org/unisite/Events/Webmasters/PDF08/Facebook1-fr.pdf`

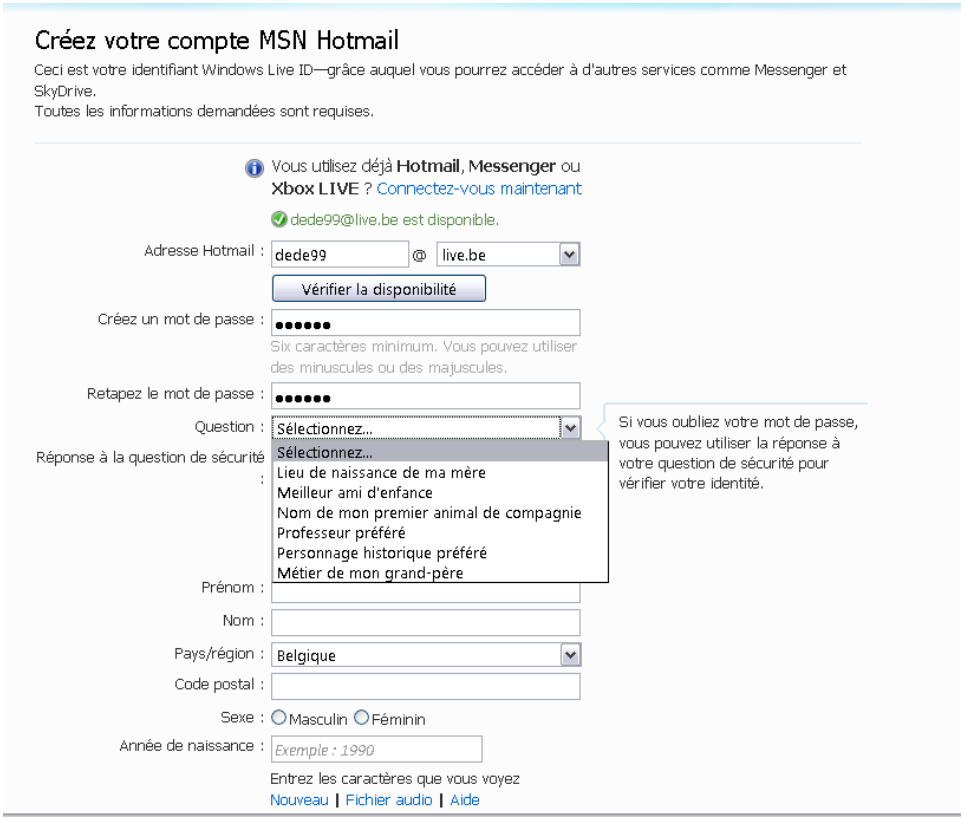

FIGURE 3.2 – En cas d'oubli du mot de passe, certains sites proposent une question secrète afin de pouvoir redéfinir le mot de passe du compte

nir le mot de passe (voir figure 3.2). Cela signifie donc que toutes personnes connaissant la réponse peut, d'une part changer le mot de passe sans l'autorisation du titulaire du compte, mais d'autre part, avoir accès à d'autres données potentiellement encore plus confidentielles. Ainsi, en utilisant cette technique, « Hacker Croll » [9] a accédé aux boites mail de certains employés Twitter, ce qui, de fil en aiguille, lui a permis, entre autre, d'accéder à l'interface d'administration du site Twitter, ainsi qu'a des documents confidentiels, des comptes bancaires, etc.

Les risques encourus ne se limitent pas uniquement aux comptes relatifs à des sites web. Des informations telles que nom, adresse, date de naissance, et le nom d'un animal familier, d'un parent, d'un frère ou d'une sœur sont

9. Article du journal Le Monde `www.lemonde.fr/idees/chronique/2010/05/12/hacker-croll-de-l-ingenierie-sociale-a-la-question-secrete_1350033_3232.html`

déjà suffisantes pour courir un risque de vol d'identité [23]. Les risques liés à l'usurpation d'identité sont nombreux. Un hacker peut ainsi:

- ouvrir une ligne téléphonique. Les factures seront bien évidement envoyées à la victime,
- faire des achats
- ouvrir un compte en banque et y effectuer des crédits au nom de la victime,
- « prouver » son identité en cas de déclaration de perte de carte d'identité. De cette manière, l'usurpateur pourra posséder une nouvelle carte toute neuve et authentique.
- ...

Afin de sensibiliser le public aux dangers liés au vol d'identité, la société Fellowes [10] a créé un site web `www.securisezvotreidentite.be`. Plus d'informations sur le vol d'identité peuvent être trouvées sur ce site.

3.3 Faux profil

N'importe qui peut ouvrir un compte sur un site de réseau social, sans avoir à y prouver son identité. Certaines personnalités ont ainsi été surprises d'apprendre qu'elles avaient un ou plusieurs profilsFacebook [11].

L'utilisation sur le net de l'identité d'une personne par une autre n'est pas réservée aux personnalités connues. On peut se retrouver dans une situation où des membres du réseau discutent ou font des échanges, en pensant communiquer avec un ami, alors qu'il s'agit d'un imposteur.

Le but de cette usurpation peut parfois être très malveillant. Elle peut avoir comme objectif, par exemple de:

- Faire endosser la responsabilité d'actes illégaux par un autre. Il y a quelque temps, un jeune homme de 19 ans a été placé en garde à vue par la police, accusé d'agressions sexuelles sur des femmes. Après enquête, il s'est avéré qu'un autre homme avait utilisé les photographies et autres informations laissées sur le site Skyblog par notre quidam; et ce dans le but de contacter des jeunes femmes et de leur donner rendez-vous. [12].

10. Fabricant néerlandais de destructeurs de documents
11. Article de Info du Net, `http://www.infos-du-net.com/actualite/12869-facebook-prison-maroc.html`
12. Article de La Voix du Nord, `http://www.lavoixdunord.fr/Region/actualite/`

- Se faire passer pour un amis afin de voler des informations personnelles dans le but voler l'identité d'un internaute. En effet, une fois le faussaire accepté comme ami, il a accès à toute une série d'informations, comme le nom de jeune fille de la mère, qui combinée et ajoutée à d'autres renseignements de valeur peuvent favoriser l'usurpation de l'identité.
- Se venger en donnant une mauvaise réputation. L'usurpateur va pouvoir donner une image peu flatteuse de sa victime, ce qui peut avoir des conséquences assez importantes dans la vie réelle.
- Cacher son âge. Un « cyber-prédateur » à la recherche d'enfants pourra plus facilement entrer en relation avec un autre enfant si celui-ci pense être en contact avec quelqu'un de son âge. Les sites de réseaux sociaux ont beau essayer d'instaurer des mesures, mais tant qu'il n'y aura pas de vérification réelle de l'âge, ça ne sera pas très efficace.
- S'amuser au dépend d'une « amie », comme par exemple, cette étudiante de 15 ans qui croyait avoir poussé au suicide son amoureux virtuel, qui , après enquête, c'est révélé être un faux profil inventé par de bonnes copines [13]

Les réseaux sociaux nécessitent donc une vigilance de tout instant afin d'écarter au maximum les risques d'usurpation d'identité, tant au niveau des informations personnelles laissées sur le net et permettant le vol d'identité, qu'au niveau d'individus créant de faux profils sur le net et dont les conséquences ne sont pas toujours anodines.

3.4 Profiling

Le profiling consiste à examiner - de manière automatique et via l'application de technique de datamining [14] - des données relatives à une personne sous le regard de résultats d'analyses statistiques afin de déduire - avec un certain taux d'erreur - de nouvelles connaissances sur cette personne, le but étant généralement l'aide à la décision.

Secteur_Region/2009/02/23/article_place-en-garde-a-vue-parce-qu-un-autre-s.shtml

13. Article de 7sur7 www.7sur7.be/7s7/fr/4134/Internet/article/detail/1232714/2011/03/08/Une-ado-de-15-ans-detruite-par-le-pire-canular-imaginable.dhtml

14. Le datamining est l'extraction de connaissances à partir de données - Définition de Wikipédia, fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es

Il s'agit donc d'utiliser les informations sur une personne afin de la classer dans une catégorie, cette catégorie étant associée à une probabilité de risque, ou de réaction face à un stimuli, etc.

Par exemple, le profiling est utilisé couramment dans le cadre des assurances. Les données personnelles d'un individu sont analysées afin de le placer dans une catégorie correspondant à un certain niveau de risque, et ce à la lumière de résultats d'analyses statistiques. De là, en fonction du niveau de risque associé, la société d'assurance va accepter ou non d'assurer une personne ou de décider d'appliquer un tarif plus ou moins élevé.

Cette technique est particulièrement intéressante en marketing: en fonction de la connaissance d'un individu, quel est le « risque » que celui-ci soit réceptif à une publicité pour des savons? Ou pour des DVD ? Et de là, choisir les annonces auxquelles l'internaute sera le plus réceptif...

Les réseaux sociaux peuvent employer les informations relatives à l'ensemble de leurs membres à des fins statistiques. Ils disposent ainsi de données statistiques permettant d'appliquer la technique de profiling à leurs membres. Ces derniers seront donc catégorisés en fonction des informations laissées sur le site.

Dans un article à la DH, la commissaire chargée des nouvelles technologies Viviane Reding [15]a rappelé que *les informations concernant une personne ne peuvent pas être utilisées sans son consentement préalable* .

Cela pose des problèmes au niveau de l'autonomie. L'autonomie, c'est [2] la *Faculté de se déterminer par soi-même, de choisir, d'agir librement.* C'est le droit de s'autodéterminer. Un individu a-t-il le droit de refuser d'être catégorisé ? Si l'utilisateur d'un service n'en a pas conscience, il est clair que son avis n'entre pas en ligne de compte. Et s'il en a conscience, le refus du profiling peut avoir comme conséquence le refus du service.

Par exemple, pour avoir accès à la majorité des sites de réseaux sociaux, l'internaute n'a d'autre choix que d'accepter que ses données soient soumises à des techniques de profilage afin de mieux cibler les publicités.

Pour continuer, les données déduites via le profiling automatique peuvent être totalement erronées. Plusieurs raisons peuvent intervenir :

- Qu'en est-il de la validité des catégories. Est-ce que les analyses statistiques ont été faites dans les règles de l'art? L'échantillonnage était-

15. `http://www.dhnet.be/infos/new-tech/article/257804/donnees-privees-et-publicite-sur-internet-bruxelles-sonne-l-alarme.html`

il correct? Les données sont-elles fiables? Les questions étaient-elles neutres? Le datamining correctement appliqué? On prétend que les populations immigrées sont plus délinquantes et qu'il faut par conséquent les contrôler particulièrement alors qu'en réalité le critère discriminant est celui du niveau socioculturel.

- La catégorie auquel l'individu devrait appartenir n'existe pas car il n'y a pas suffisamment de données dans l'échantillonnage que pour la faire apparaître.
- Le profiling n'est pas une science exacte. Il y a un certain pourcentage d'erreurs. Un individu peut donc être classé dans une mauvaise catégorie.
- Les catégories peuvent évoluer avec le temps. Par exemple, le risque d'un décès prématuré diminue avec les progrès de la médecine. Les catégories doivent donc être revues régulièrement.
- L'individu peut également évoluer avec le temps. L'assignation durable à une catégorie peut ainsi conduire à des erreurs

Sans des catégories robustes, le profilage ne donnera que des conclusions sans valeur. Le but du profilage étant généralement l'aide à la décision, s'il y a erreur de catégorie, la décision risque d'être inadéquate... Le bon sens d'un humain pourrait éventuellement rectifier les choses; mais dans les décisions automatiques il n'y a pas d'humain.

De plus, l'utilisateur a-t-il le droit de demander à changer de catégorie? S'il ignore avoir été catégorisé, la réponse est évidement non. Dans le cas contraire, le profilage repose sur un mécanisme qui est relativement nébuleux. Il est parfois difficile de remettre en question ce que l'on ne comprend pas. L'autonomie est donc mis à mal sans information claire de ce qui se passe.

Une fois classé dans une catégorie, il y a souvent décision. Quelle soit erronée ou non, cette décision peut également mettre à mal l'autonomie de la personne: elle peut se voir refuser l'accès à un service, ou ne se voir proposer que des choses qui ne l'intéressent pas. Il y a donc risque d'inégalité de traitement.

Sur le site de Facebook, par exemple, une liste de suggestions d'amis est proposée: [16].

La fonctionnalité de suggestions vous aide à retrouver des per-

16. `http://www.facebook.com/help/?faq=15325`

sonnes ou des Pages que vous connaissez sans doute. Facebook calcule les suggestions en fonction des réseaux auxquels vous appartenez, d'amis communs, des informations concernant votre formation et votre emploi, des contacts importés à l'aide de l'outil de recherche d'amis, et d'un certain nombre d'autres facteurs.

Les profils des amis proposés ne correspondent peut-être pas aux styles ou aux types de relations que le membre du réseau recherche.

Un dernier risque peut encore être cité: la stigmatisation d'une catégorie d'individu. Un tel ayant telles caractéristiques serait supposés avoir tel défaut...

3.5 Economique

Il existe plusieurs méthodes de financement d'un réseau social (voir 2.2.6). Les plus spécifiques aux réseaux sociaux sont d'une part, la vente d'espaces publicitaires, et d'autre part, la vente des données relatives à leurs membres.

Les données personnelles récoltées sur le site permettent aux publicitaires de mieux cibler leurs éventuels futurs clients. En effet, ces sites contiennent souvent beaucoup d'inscrits. Plus il y a d'individus inscrits dans le réseau, plus l'échantillonnage permettant des analyses statistiques est grand et plus le nombre d'individus touchés par la publicité est vaste. Et plus un individu donne des informations personnelles, plus il est aisé d'augmenter la connaissance que l'on a sur lui, et donc de le cibler. Un site de réseau social est un terrain idéal permettant l'application de la technique de profilage dans le but d'ajuster avec le plus de précision possible les publicités. Ces réseaux sociaux ont donc la possibilité d'offrir un espace publicitaire riche et unique permettant une publicité plus efficace que dans les médias traditionnels. Le fait qu'IPG [17] - société de marketing et de publicité - ait investi dans Facebook [18] (à hauteur de 0.5%) n'est pas innocent.

Un site de réseau social peut soit analyser lui-même les profils de ses membres afin de sélectionner les individus visés par une publicité spécifique,

17. InterPublic Group (IPG), `http://www.interpublic.com/`

18. `http://investors.interpublic.com/phoenix.zhtml?c=87867&p=irol-newsArticle&ID=877174`

comme par exemple Facebook [19], soit il peut déléguer cette analyse à un autre. C'est le cas pour les sites partenaires de Google AdSense [20] tel LiveJournal [21].

3.6 Liberté d'expression

Les réseaux sociaux permettent de s'exprimer en toute liberté. Chacun a le droit de créer des groupes de discussions relatifs à n'importe quel thème ou de faire connaître son opinion sur n'importe quel sujet. C'est, en partie, grâce à cette liberté que les protestations d'Afrique du Nord et du Moyen-Orient en ce début de 2011 [22] ont pu avoir lieu.

Mais qui dit liberté d'expression, dit également propos tendancieux, publications d'images irrévérencieuses ou de liens violant les droits d'auteurs (voir par exemple la figure 3.3), etc. Or, tout ce qui est publié sur ces sites et qui est visible par tout le monde est considéré comme publique (voir [7]). Les limites relatives à la liberté d'expression sont donc également applicables! Par conséquence, on ne peut pas publier une image sans le consentement de la ou des personnes se trouvant sur la photographie. On ne peut également pas calomnier ou diffamer quelqu'un, telle cette internaute renvoyée devant la justice française accusée d'avoir calomnié le Dr Delajoux [23]. De même, la partie visible d'un profil peut servir de preuve devant un tribunal, comme cet employé français licencié pour avoir critiqué sa hiérarchie [24].

Notons également que Facebook se donne le droit de diffuser comme bon lui semble toutes informations visibles par tout le monde. On trouve, en effet, dans le règlement relatif au respect de la vie privée, en parlant de ces dernières:

(...) elles peuvent aussi être indexées par des moteurs de re-

19. Sur Facebook, chaque membre peut créer une publicité et en désigner le publique cible
20. Programme permettant de diffuser des annonces ciblées `www.google.com/adsense`
21. `www.livejournal.com/myads/`
22. Wikipedia `fr.wikipedia.org/wiki/Protestations_dans_les_pays_arabes_de_2010-2011`
23. Dr delajoux avait opéré Johnny Hallyday en 2009, à Paris, Article de la RTBF, `http://www.rtbf.be/info/societe/detail_une-fan-de-johnny-hallyday-au-tribunal-pour-injures-sur-facebook?id=5663053`
24. Article de 7sur7, `http://www.7sur7.be/7s7/fr/4134/Internet/article/detail/1184847/2010/11/19/Denigrer-sa-hierarchie-sur-Facebook-peut-valoir-un-licenciement.dhtml`

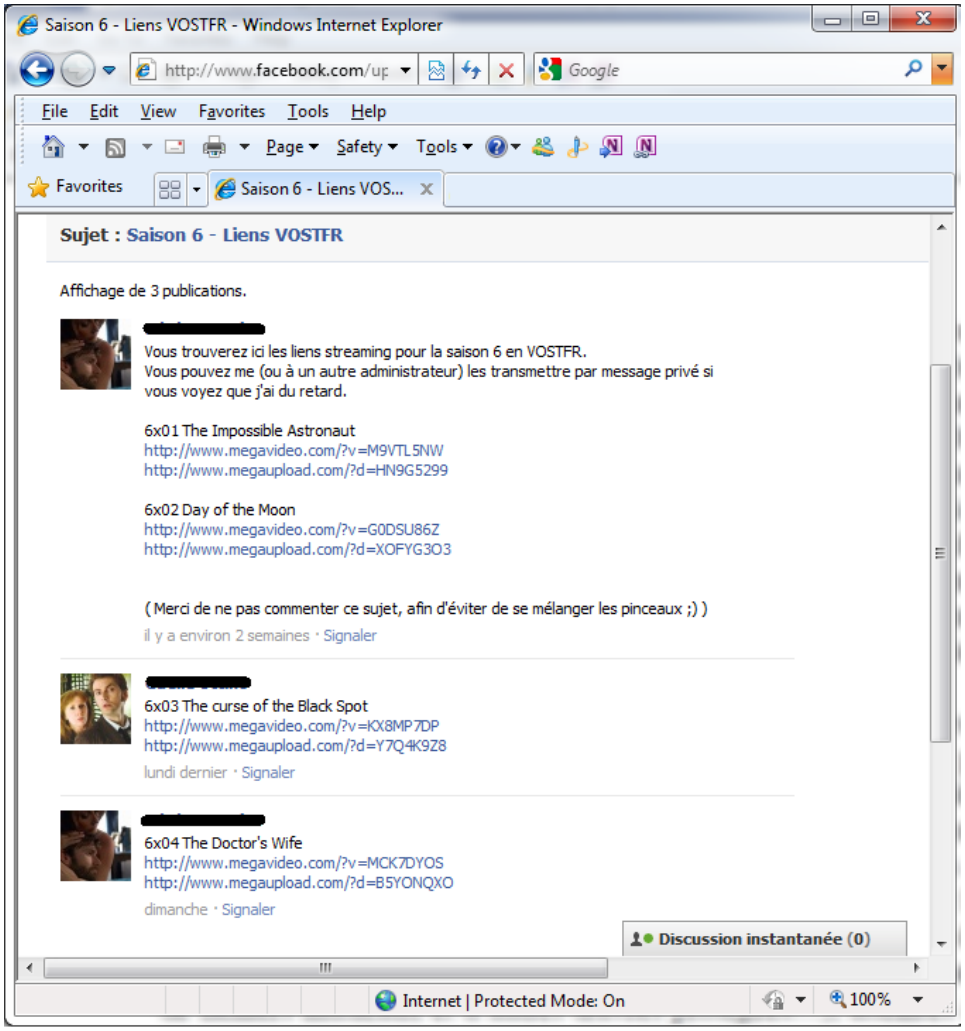

FIGURE 3.3 – Les réseaux sociaux facilitent la diffusion de films ou séries protégées par les droits d'auteurs

> *cherche tiers et être importées, exportées, diffusées et rediffusées par Facebook et par des tiers, sans restrictions de confidentialité.*

Facebook va même plus loin: il s'approprie tout contenu publié sur son site. On trouve, en effet, dans les conditions d'utilisation:

> *Pour le contenu protégé par les droits de propriété intellectuelle, comme les photos ou vidéos (« propriété intellectuelle »), vous nous donnez spécifiquement la permission suivante, conformément à vos paramètres de confidentialité et paramètres d'applications : vous nous accordez une licence non-exclusive, transférable, sous-licenciable, sans redevance et mondiale pour l'utilisation des contenus de propriété intellectuelle que vous publiez sur Facebook ou en relation à Facebook (« licence de propriété intellectuelle »).*

Il n'y a pas que sur Facebook que l'on retrouve ce genre de clause. par exemple, sur Flickr:

> *you grant Yahoo! the following worldwide, royalty-free and non-exclusive license(s),(...) the license to use, distribute, reproduce, modify, adapt, publicly perform and publicly display such Content on the Yahoo! Services (...)*

Il se peut qu'en Europe, cette clause soit reconnue comme abusive et donc annulée, mais ce n'est sûrement pas le cas dans tout les pays. Il est donc conseiller d'être très prudent et de réfléchir avant de publier tout contenu.

Chapitre 4

Divulgation de la vie privée

Beaucoup d'utilisateurs de réseaux sociaux postent tout et n'importe quoi sur leur site [23]. Toutes ces informations, quand elles se retrouvent dans de mauvaises mains, peuvent être collectées et utilisées par des tiers, et parfois à fin malhonnête. Des étrangers peuvent ainsi se retrouver à manipuler des données qui auraient dû rester privées.

Nous allons présenter quelques points pouvant être à l'origine de la collecte, et/ou de la divulgation d'informations personnelles et privées. Mais d'abord, nous allons préciser ce qu'est une donnée à caractère personnel.

La notion de données à caractère personnel est relative et n'est pas définie partout de la même manière. Selon la loi du 8 décembre 1992 relative à la protection de la vie privée à l'égard des traitements de données à caractère personnel:

> *on entend par « données à caractère personnel » toute information concernant une personne physique identifiée ou identifiable,...; est réputée identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale.*

Tous ne s'entendent pas sur la notion de données d'identification ainsi média6 [1] donne une autre définition,

> ***Non-Personally Identifiable Information***

1. Média6 est une société américaine spécialisée dans la collecte de données via les cookies à des fins de profiling `www.media6degrees.com/privacy/index.html`

We collect Non-Personally Identifiable Information ("Non-PII") from visitors to this Website. Non-PII is information that cannot by itself be used to identify a particular person or entity, and may include your IP host address, pages viewed, browser type, Internet browsing and usage habits, Internet Service Provider, domain name, the time/date of your visit to this Website, the referring URL and your computer's operating system.

Ainsi, selon Média6, une donnée qui ne peut, à elle seule, être utilisée à des fin d'identification n'est pas une donnée d'identification. Or dans la définition précédente, toute donnée pouvant, de manière directe ou indirecte conduire à l'identification est une donnée d'identification. L'adresse IP, par exemple, est une donnée d'identification dans le sens qu'elle permet d'identifier, de manière unique, un set de données. Une adresse IP permet également de localiser le pays d'un terminal qui se connecte à internet. En effet, les adresses IP sont gérées par l'IANA [2] qui documentent toutes ses attributions. Il est même possible d'affiner la localisation en connaissant la politique du fournisseur d'accès pour attribuer ses adresses. La localisation peut aussi être complétée par le traçage. Il s'agit de retracer le chemin qu'un paquet de données suit à travers le réseau pour aller d'une machine à une autre [3]. Enfin, via le fournisseur d'accès internet, il est tout à fait possible d'identifier la personne ayant utilisé, à un moment précis, l'adresse IP spécifiée. [4]. Ainsi, dernièrement, en France, un pédophile a été identifié grâce à son adresse IP [5]. Certains ne considèrent donc pas comme données personnelles des données du type numéro d'identification d'un terminal, ou d'un software...

Il y a atteinte à la vie privée lorsque des informations personnelles de l'utilisateur sont collectées ou manipulées sans le consentement de la personne. Nous allons décrire quelques procédés permettant ou facilitant cette atteinte à la vie privée.

2. Internet Assigned Numbers Authority, organisation d'ont le but est de gérer l'attribution des adresses IP `http://www.iana.org/`

3. Exemple de site permettant de tracer: VisualIPtrace `http://www.visualiptrace.com/demo.html`

4. Protection des données à caractère personnel en ligne: la question des adresses IP par European Data Protection Supervisor, `http://www.edps.europa.eu/EDPSWEB/webdav/shared/Documents/EDPS/Publications/Speeches/2009/09-04-15_adresses_IP_FR.pdf`

5. Article du journal Le point, `http://www.lepoint.fr/actualites-societe/2009-04-17/pedophilie-premiere-arrestation-des-cybergendarmes/920/0/335803`

4.1 Paramétrage des paramètres de confidentialité

La plupart des données relatives à un profil sont des informations à caractère personnel, et ne devrait donc pas être visible par tout un chacun. Certains sites, tel match.com, ne permettent pas à l'utilisateur de limiter l'accès aux éléments de son profil. Tout est publique. D'autres, par contre, fournissent des outils permettant de déterminer qui a accès à certaines informations.

Mais ces outils ne sont pas toujours suffisamment utilisés. Selon [9], la majorité des membres ne change pas les paramètres par défaut, qui sont souvent trop permissifs [6]. Et ce, pour plusieurs raisons: soit les internautes ignorent que cette possibilité existe; soit ils ne prennent pas la peine de les changer ; soit ça leur semble beaucoup trop compliqué. Les possibilités, ainsi que la manière de les configurer changent complètement d'un site à l'autre. Sans compter qu'un site peut changer complètement sa politique de vie privée, souvent sans aucun préavis, ce qui désoriente encore plus l'utilisateur. L'exemple de Flickr est assez parlant: tout les documents tels que « condition d'utilisation », « confidentialité »,... ne sont actuellement plus disponibles en français.

C'est donc l'implémentation du logiciel supportant le réseau social qui détermine les paramètres de confidentialité de la majorité des membres d'un réseau. La manière dont est codée l'application a donc une grande importance sur ce qui est effectivement divulgué.

Le but des réseau sociaux est différent de celui des internautes [7]. Les utilisateurs sont invités à révéler un maximum de choses. Voici le genre de message que l'on peut trouver dans un site tel que Flickr:

> *Les visiteurs de Flickr ont accès à votre page de profil. Vous pouvez empêcher une partie des utilisateurs ou Flickr d'avoir accès à vos informations.*
>
> *Ceci étant, un profil détaillé est bien plus utile pour tout le monde. Ne vous en faites pas trop.*

Si l'individu laisse son profil visible par n'importe qui, il est lui-même responsable de la possible perte de contrôle de ses données personnelles.

6. Article de Electronic Frontier Foundation (EFF), `http://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly`

7. Article de 01Net, `http://www.01net.com/editorial/509845/facebook-rend-publiques-certaines-donnees-privees-de-ses-membres/`

Beaucoup d'utilisateurs souhaitent donner le plus d'informations possible, dans le but, par exemple, de se faire le plus d'amis possible. Dans [15], on voit que

- 93.8% des utilisateurs révèlent leur sexe.
- 83.3% des utilisateurs révèlent la ville dans laquelle ils habitent.
- 87.1% des utilisateurs révèlent l'école supérieur dans laquelle ils ont étudié.
- 45.1% des utilisateurs révèlent leur adresse.
- 59.8% des utilisateurs remplissent et montrent le champ « à propos de moi »
- 67.8% des utilisateurs révèlent leur compte d'adresse de messagerie instantanée
- 83.8% des utilisateurs révèlent leur date de naissance.
- 92.3% des utilisateurs révèlent leur adresse email.
- 78.5% des utilisateurs révèlent le statut de leur relation amoureuse.

Certains utilisateurs développent une sorte de dépendance aux sites de réseaux sociaux [26]. Ils ne communiquent plus aux autres que via ces sites. Ils se soucient alors encore nettement moins des informations postées en ligne, ainsi que des risques encourus[15].

4.2 Le piratage

Un paramétrage correct de la confidentialité - c'est-à-dire un paramétrage reflétant le désir de l'utilisateur - ne garantit pas que cette confidentialité soit respectée. En effet, sur internet, rien n'est véritablement privé. Aucun site n'est à l'abri d'attaques de hackers qui trouveraient un chemin vers les données confidentielles. Dernièrement, un pirate proposait à la vente des millions de données personnelles issues de compte Facebook [8]

4.3 La négligence d'un ami

La divulgation d'informations privées peut-être due à un tiers. Il est en effet possible de marquer des photographies ou des vidéos en indiquant les personnes se trouvant sur le média. L'internaute peut avoir été suffisamment

8. Article du journal Le Monde, `micro.lemondeinformatique.fr/actualites/lire-1-5-millions-d-identites-facebook-volees-en-vente-3126.html`

prudent que pour ne pas dévoiler des photographies compromettantes, mais un « ami »pourrait l'avoir fait à sa place. Sachant que n'importe qui peut afficher l'identité des personnages sur la photographie, l'individu peut se retrouver dans une situation fâcheuse... Il est possible de limiter l'annotation des photographies à certaines personnes, mais encore une fois, l'option n'est pas connue de tous...

4.4 Liste de contacts et groupe

Les données de type relationnel (listes de contacts et/ou appartenances à un groupe) permettent de déduire de nouvelles informations sur les individus en se basant sur le principe « dis-moi qui sont tes amis, je te dirais qui tu es ». Ces analyses sont basées sur des données sociologiques, et permettent de déduire de nouvelles connaissances avec un certain taux de probabilité. Par exemple, un utilisateur qui n'a que des amis de plus de 50 ans n'est probablement pas un adolescent. L'article du Boston globe [9] sur le projet Gaydar [10] est une illustration de ce propos.

Il faut également savoir que la liste d'amis, ainsi que l'appartenance à un groupe est visible par tous. En effet, dans la majorité des sites de réseaux sociaux, il n'est pas possible de le cacher.

La fiabilité des résultats peut varier énormément en fonction de la nouvelle information que l'on cherche à déduire, de la méthode utilisée, ainsi que du site de réseau social. Cela va d'une performance relativement médiocre à une précision surprenante. Plus d'informations sur la fiabilité de ce genre d'analyses peut être trouvées dans le document [28].

4.5 Le phishing

L'intéressé peut avoir été contacté et mis en confiance dans le but de l'inciter à en révéler encore d'informations : des données personnelles peuvent être dévoilées au cours d'une discussion, une page personnelle peut servir à rediriger vers un site dangereux - style phishing -, un lien vers une vidéo peut s'avérer être un cheval de Troie, etc.

9. `www.boston.com/bostonglobe/ideas/articles/2009/09/20/project_gaydar_an_mit_experiment_raises_new_questions_about_online_privacy/`

10. Projet de deux étudiants du MIT qui cherche à déterminer si un individu est homosexuel ou pas en analysant le profil de la personne, ainsi que le profil des amis

4.6 Le site de réseau social

Le site de réseau social collecte et conserve toute une série d'informations sur un individu, et ce pour un laps de temps indéfini. Les données relatives à l'utilisateur peuvent être conservées, même après la suppression du compte. La plupart de ces données relèvent d'un caractère personnel, elles doivent être protégées. Il est de la responsabilité de l'organisme collecteur de prendre soin des données récoltées. Mais Facebook se dédouane de tout problème :

> *Nous ne pouvons donc en aucun cas garantir que le contenu que vous publiez sur ce site ne sera pas vu par des personnes non autorisées. Nous ne sommes en aucun cas responsables du non-respect des paramètres de confidentialité ou des mesures de sécurité en vigueur sur ce site.*

Il existe effectivement des cas de « fuites ». Il y a quelques années, des étudiants du MIT [11] avaient pu télécharger facilement les profils de certains membres de Facebook afin de faire du datamining . Il semble que cette brèche ait été comblée, mais rien ne dit que d'autre failles n'existent pas...

De plus, il faut savoir que tout les employés de Facebook ont accès à tous les profils, qu'ils soient privés ou publiques [12]....

Enfin, les données sont souvent transférées dans des pays tiers où les lois de protection de la vie privée ne sont pas toujours les mêmes:

> *En utilisant Facebook, vous acceptez que vos données personnelles soient transférées et traitées aux États-Unis.*

4.7 Photographies de profil

Les utilisateurs ont souvent l'habitude d'associer une photographie à leur profil. Selon[9], 61% des images de profil ont une qualité suffisante pour servir directement à l'identification et 80% contiennent des informations utiles à l'identification.

11. MIT Computer Science and Artificial Inteligence Laboratory (CSAIL), `http://groups.csail.mit.edu/mac/classes/6.805/student-papers/fall05-papers/facebook.pdf`

12. Article de Reference.be, `http://www.references.be/carriere/les-employ%C3%A9s-de-facebook-peuvent-se-connecter-%C3%A0-votre-profil`

Il existe actuellement des outils très performants permettant de faire le lien entre deux photographies [13], et donc de mettre en relation les informations associées à chaque image. Ainsi, des profils d'étudiants de la Carnegie Mellon University ont pu être identifiés en utilisant les portraits nominatifs des élèves, collectés sur site de l'université [9]. De la même manière, il est possible de connecter deux profils distincts grâce à leur photographie de profil. Si un internaute a un profil sur Facebook - utilisant la véritable identité - et un profil sur match.com - normalement anonyme -, et qu'il utilise des photographies ayant une qualité suffisante, le profil de match.com pourra être re-identifié. Dans ce cas, les informations du profil match.com - visibles par tous - seront associées à la personne.

Ce genre d'identification permet de mettre un nom sur un profil anonyme. Il permet également d'augmenter la connaissance sur un individu, et ce , sans le besoin, ni le consentement, ni même la connaissance de l'individu.

Outre cette connexions entre deux profils, une photographie peut donner beaucoup d'informations telles l'aspect physique d'une personne - comme par exemple la couleur des yeux ou des cheveux - , l'origine ethnique, l'appareil photographique utilisé, les habitudes vestimentaires, etc.

4.8 Re-identification après anonymisation

Outre pour des fins publicitaire, les sites de réseaux sociaux utilisent les données collectées à des fins statistiques.

> *Facebook peut utiliser les données de votre profil sans vous identifier en tant qu'individu vis-à-vis des tiers. Ces données nous permettent notamment d'estimer le nombre de gens au sein de votre réseau qui aiment tel morceau de musique ou tel film, ou encore en vue de personnaliser les publicités et promotions que nous vous proposons sur Facebook.*

Les lois réglementant l'utilisation et la collecte de données à des fins statistiques requièrent l'anonymisation des données: selon le principe 3.3 de la recommandation N°R(97) 18,

> *Les données à caractère personnel collectées et traitées à des fins statistiques doivent être rendues anonymes dès qu'elles ne sont*

13. On peut citer, par exemple, Picassa, `picasa.google.com/features-nametags.html`

plus nécessaires sous une forme identifiable.

Des données anonymisées sont ainsi fournies à des sociétés de marketing, à des développeurs d'applications, à des centres de recherche sur le datamining,...

Dans la pratique, il convient de déterminer quand des données peuvent être considérées comme anonymes, ainsi que la manière de les anonymiser. Selon cette même recommandation,

> *L'anonymisation consiste à supprimer les données d'identification afin que les données individuelles ne puissent être attribuées nommément aux diverses personnes concernées*

Dans la majorité des cas, cela consiste en la séparation des données identifiables dès qu'elles ne sont plus nécessaires. Néanmoins,

> *Une personne physique n'est pas considérée comme 'identifiable' si cette identification nécessite des délais coûteux et des activités déraisonnables. Lorsqu'une personne physique n'est pas identifiable, les données sont dites anonymes*

Des données dites anonymes pourraient ainsi ne plus l'être dans quelques années suite à l'évolution des technologies. Et cette technologie est déjà en train de rattraper les sites de réseaux sociaux. En effet, une équipe de l'université du Texas a réussi à créer un algorithme de ré-identification, basé sur les listes de contacts-amis, en faisant des recoupements sur les données publiées par différents réseaux sociaux[16]. Ils sont ainsi parvenu à ré-identifier 30% des utilisateurs ayant un compte sur Twitter et Flickr parmi un set de données de Twitter, et ce avec un taux d'erreur de 12%.

Une autre façon de désanonymiser des données est de les comparer à des données démographiques[9]. Par exemple, aux États-Unis, la majorité de la population peut-être identifiée en utilisant conjointement le code postal, la date de naissance et le genre. Or 45.8% des utilisateurs de l'étude de[9] dévoilent ces informations.

4.9 Application basée sur l'API

La plupart des sites de réseaux sociaux mettent à disposition des internautes une API [14] permettant de créer des applications que les membres pourront ensuite utiliser. Ces applications peuvent être de toutes sortes. Cela peut-être des jeux, des applications permettant d'envoyer des fleurs virtuelles, etc.

Sur Facebook, pour avoir accès à une de ces applications, l'utilisateur doit d'abord accepter que le logiciel accède à tout ce qui est contenu dans son profil. En refusant, il se voit interdit l'accès à cette application. En acceptant, l'utilisateur permet également à l'application de voir le profil (privé ou non) de ses amis. Il y a moyen d'empêcher l'accès de son profil par des applications installées sur la page d'amis via une option relativement cachée (Privacy -> Applications -> Other Applications [15]).

Si pour certaines applications, les données du profil peuvent être utiles, pour la majorité d'entre elles, l'accès à ces informations est loin d'être pertinent.

Ces applications tournent sur le serveur du développeur, donc sans véritable contrôle de la part du site de réseau social. La BBC a ainsi pu collecter toute une série de données privées [16].

4.10 Web beacon

Souvent, la collecte des données - personnelles ou non - se fait sans le consentement, ou même la connaissance de l'individu, par exemple via l'utilisation des « web beacon» ; ces derniers sont également appelés web bugs ou encore pixels espions. Un web beacon est une petite image transparente - afin d'être invisible - généralement de la taille d'un pixel – de manière à limiter le temps de chargement. Lorsqu'un utilisateur ouvre une page contenant un web beacon, le browser envoie une requête à un service dans le but de charger l'image. Dans cette requête sont incluses des informations sur la page ouverte (via un identifiant collé à l'image) ainsi que la source de la

14. Application Programming Interface : Bibliothèques de fonctions destinées à être utilisées par les programmeurs dans leurs applications

15. Article de The Washington Post, `http://www.washingtonpost.com/wp-dyn/content/article/2008/06/11/AR2008061103759_pf.html`

16. Article de BBC News, `http://news.bbc.co.uk/2/hi/programmes/click_online/7375772.stm`

demande. Le service sollicite alors l'enregistrement d'un cookie. Lorsque le même utilisateur ira sur une autre page contenant également un beacon appartenant au même service, le cookie sera alors utilisé pour associer les deux pages web. La navigation de l'utilisateur sera analysée tant que le cookie sera là.

Facebook a créé récemment une polémique à cause de son beacon : si monsieur X achète des billets d'avion sur un site partenaire, l'info est envoyée à Facebook qui en avertit les « amis » [17] en l'affichant sur le « mur ». Depuis le début, le beacon est optionnel; mais à l'origine l'utilisateur devait désactiver cette option. A l'heure actuelle, l'utilisateur doit volontairement l'activer. Néanmoins, que l'information soit ou non divulguée aux « amis », les données continuent d'arriver jusqu'à Facebook, et ce, que vous ayez autorisé le beacon ou non. Facebook n'est pas le seul à utiliser les web beacon. Par exemple, Flickr via Yahoo l'utilise également.

17. Article du journal Le Monde, `http://pisani.blog.lemonde.fr/2007/11/30/facebook-fait-volteface-ecoute-les-usagers`

Chapitre 5

Ontologie, OWL et Web sémantique

Sur le web, énormément d'informations sont disponibles. Beaucoup sont redondantes, obsolètes, incomplètes, voir incorrectes. Celles-ci sont lisibles et compréhensibles par les humains. On voudrait qu'elles soient également interprétables par une machine, afin de pouvoir les rechercher, les partager, les utiliser dans différentes applications par différents acteurs. A cette fin, il est nécessaire de pouvoir représenter cette connaissance dans un langage commun. Une ontologie est une représentation de la connaissance relative à un domaine. Afin de développer un langage commun exprimant une ontologie, le consortium W3C a investi dans le projet web sémantique `www.w3.org/2001/sw/SW-FAQ`. Cette analyse a conduit à la création d'un langage de modélisation de données distribuées sur le web : RDF, étendu par la suite à RDFS et OWL.

5.1 Ontologie

L'ontologie est la « Partie de la philosophie qui a pour objet l'étude des propriétés les plus générales de l'être, telles que l'existence, la possibilité, la durée, le devenir » [2].

Par extension, dans le contexte de l'informatique, une ontologie est une représentation de la connaissance relative à un domaine. Elle permet la réutilisation et le partage de cette connaissance par l'utilisation d'un vocabulaire commun. C'est largement utilisé dans des domaines tels que la gestion des

connaissances [1] ou l'intelligence artificielle.

Une ontologie inclut donc:

- La définitions de concepts
- L'arrangement des concepts en hiérarchie
- La définition de propriétés (attributs de concepts ou relations entre deux concepts)

Selon [8], les ontologistes distinguent deux types d'ontologies: les ontologies légères et les ontologie lourdes:

> *Lightweight ontologies include concepts, concept taxonomies, relationships between concepts, and properties that describe concepts. Heavyweight ontologies add axioms and constraints to lightweight ontologies. Axioms and constraints clarify the intended meaning of the terms gathered on the ontology*

La différence entre une ontologie lourde et une ontologie légère, est la notion d'axiome qui permet de raisonner sur une ontologie via des mécanismes d'inférence. Ils permettent également de préciser la sémantique des termes utiliser.

Un schéma entité-relation peut être considéré comme une ontologie légère car il est fait de concepts, de propriétés relatives aux concepts et de relations entre les concepts. Pour être considéré comme une ontologie lourde, ce schéma devrait pouvoir inclure des axiomes.

Nous reviendrons sur ces notions ultérieurement.

5.2 RDF

RDF - Ressource description Framework - est basé sur l'utilisation de triplets de la forme : (:sujet, prédicat (ou propriété), objet). Dans le vocabulaire RDF, sont fournis une propriété : type (relation entre un sujet et son type), ainsi qu'un « objet » : property (type général d'une propriété)

Exemple : Shakespeare est un écrivain

- Sujet : Shakespeare
- Prédicat : type
- Objet : écrivain

1. L'ensemble des initiatives, des méthodes et des techniques permettant de percevoir, d'identifier, d'analyser, d'organiser, de mémoriser, et de partager des connaissances entre les membres des organisations - Wikipédia

Sous une autre forme :

- :Shakespeare rdf:type écrivain

La signification d'un mot est souvent dépendante de son contexte. De plus, selon les sources, les définitions peuvent varier, être nuancées, ou carrément être contradictoires. En effet, tout le monde peut dire n'importe quoi sur n'importe quel sujet. Ceci est particulièrement vrai sur le web. Par exemple, le mot « homme » peut aussi bien être synonyme de « humain », que « humain de sexe masculin » ou que « humain de sexe masculin parvenu à l'âge de virilité ». Pour d'autres, le terme correspond à un ensemble de caractéristiques qu'un « animal » doit avoir afin d'être reconnu comme homme, en vue de résoudre la question : « A partir de quel moment les ancêtres des hommes peuvent-ils être qualifiés d'homme? ». Dans le contexte militaire, ce sera un soldat. On pourrait ajouter encore beaucoup d'autres définitions, dont certaines seraient probablement farfelues et erronées...

Il n'est pas question ici de choisir la « bonne » définition. Il n'y a d'ailleurs généralement pas de bonne définition mais juste des définitions qui dépendent de l'usage que l'on en fait. La modélisation doit donc pouvoir les accepter toutes. Afin d'intégrer cette notion dans RDF, on utilise un espace de nommage. Il s'agit d'un préfixe faisant référence à la source définissant le terme utilisé. Il se place devant les « : ». Par exemple, « rdf:type » renvoie à définition de « type » dans l'espace de nommage « rdf ».

Autre exemple, si « lit:aCréé » renvoie vers une définition du terme « aCréé » dans le domaine de la littérature et signifie « est l'auteur de », l'affirmation : « Shakespeare est l'auteur de Hamlet » s'écrira :

```
:Shakespeare lit:aCréé Hamlet
lit:aCréé rdf:type rdf:Property
```

Le triplet - sujet, prédicat, objet - semble parfois trop limité pour exprimer ce qui doit l'être. Par exemple, comment traduire : « Shakespeare a écrit Hamlet en 1604 »? La technique est relativement aisée : il suffit en effet de considérer les prédicats qui nous intéressent en faisant en sorte qu'il ait tous un même sujet abstrait.

Notre exemple avec Shakespeare peut donc s'écrire :

```
q:n1  bio:author lit:Shakespeare;
      bio:title Hamlet;
      bio:publicationDate 1601
```

C'est une première approche de la réification. On transforme un sujet en « chose » de propriété.

Les triplets peuvent être représentés (sérialisés) de diverses manières. La sérialisation est l'écriture des triplets sous une forme textuelle. Il existe plusieurs formes possibles. Une des plus courantes est l'expression des triplets sous format XML (RDF/XML). La standardisation de ce format se trouve sur le site de W3C : `http://www.w3.org/TR/rdf-syntax-grammar/`. Voici un exemple de retranscription des triplets en XML :

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:lit="http://www.litvoc.com/rdf/">

<rdf:Description rdf:about="Shakespeare">
       <rdf:type>écrivain</rdf:type>
      <lit:aEcrit>Hamlet</lit:aEcrit >
</rdf:Description>

</rdf:RDF>
```

Avec :

- xmlns:xml name space - permet d'associer un espace de nommage (rdf, par exemple) avec sa définition
- rdf:about - ce dont on parle.

Le principe du web, c'est que les informations sont distribuées sur la toile. Pour rassembler ces données en RDF, c'est relativement aisé et cela revient à prendre tout les triplets. Le problème rencontré alors est de faire correspondre les sujets entre eux. Par exemple, au niveau machine, sans table de correspondance, « William Shakespeare » et « Shakespeare William » pourraient être considéré comme deux personnes différentes. Il nous faut donc un identifiant unique et universel qui serait utilisable partout sur le web et qui représenterait une même ressource; une ressource étant définie comme quelque chose ayant une identité. L'IETF - Internet Engineering Task Force - a définit l'URI - Uniform Resource Identifier - comme une suite de caractères devant respecter une syntaxe bien précise. La norme est

décrite dans le document `http://tools.ietf.org/html/rfc3986`. Les URI peuvent être de type « locator », « name » ou les deux :

- URL - Uniform Resource Locator - est un URI qui identifie la ressource par son moyen d'accès. Exemple : la syntaxe de RDF est identifiée par `http://www.w3.org/1999/02/22-rdf-syntax-ns#`
- URN - Uniform Resource Name - est un URI qui identifie la ressource par un nom. Exemple : « ISBN-13: 978-0192834164 » pointe, de manière unique vers une édition de Hamlet publiée par Oxford Paperbacks le 2 avril 1998.

Il est évident que trouver un URI pour chaque ressource est une gageure. Revenons à notre exemple : le prédicat (aEcrit) n'est pas défini de manière très précise. On voudrait plus d'informations sur cette propriété. RDFS - RDF Schema Language - permet de résoudre - en partie - ce problème.

5.3 RDFS

RDFS utilise RDF afin de décrire les relations entre les éléments. Pour ce faire, il définit l'objet Class et propose quelques prédicats : SubClassOf, Domaine, Range, ... ainsi que des règles d'inférences simples entre propriétés et classes permettant de raisonner sur les données.

Une classe est définie comme un ensemble de ressource. Par exemple, la classe chat regroupe tout les individus de type chat. Si Minou est un chat, alors il sera considéré comme une instance de la classe Chat.

La notion de SubClassOf lie deux classes et est définie par : si la classe A est une SubClassOf de la classe B, alors tout membre de la classe A est membre de la classe B.

```
Exemple :
   Minou :instanceOf :Chat
   :Chat rdfs:subClassOf :Animal
D'où on peut tirer une inférence :
   Minou :instanceOf :Animal
```

Le domaine relatif à une propriété lie la propriété à une classe tel que si un élément a cette propriété, alors il fait partie de la classe correspondant au

domaine. Par exemple, si le domaine de la propriété mariDe est définit par la classe correspondant à l'ensemble des hommes mariés, alors on pourra déduire que tout individu qui sera le mari de quelqu'un sera un homme marié. Il ne s'agit pas d'une contrainte, mais d'une inférence. Si on écrit qu'une femme est le mari de X, le modèle n'invalidera pas cette assertion. Il se contentera de déduire que cette femme est un homme marié.

Le range relatif à une propriété lie la propriété à une classe tel que si un élément correspond à la valeur de cette propriété, alors cet élément fait partie de la classe correspondant au range. Par exemple, si le range de la propriété mariDe est définit par la classe correspondant à l'ensemble des femmes mariées, alors on pourra déduire que tout individu qui sera le résultat de la propriété mariDe sera une femme mariée. Il s'agit également d'une inférence. Si on écrit que X est le mari d'un homme, comme pour le domaine, le modèle n'invalidera pas cette assertion. Il se contentera de déduire que cet homme est une femme mariée.

Enfin, la notion de subPropertyOf lie deux propriétés et est définie par: si la propA est une subPropertyOf de propB, et si X propA Y , alors X propB Y.

```
Exemple :
   :Shakespeare bio:mariDe :Anne Hathaway
   bio:mariDe rdf:type rdf:Property
   bio:mariDe rdfs:subPropertyOf bio:connait
D'où on peut tirer une inférence :
   :Shakespeare bio:connait:Anne Hathaway
```

RDF est fournit avec un « RDF query engine »utilisant le langage de requête SPARQL, similaire au SQL. Il permet ainsi de faire de la recherchée se basant sur l'union des triplets de base (i.e. fournis) et des triplets résultants des inférences. RDF(S) permet de représenter les données de manière consistance. RDFS fournit des primitives élémentaires pouvant servir à la création d'une ontologie simple.

5.4 OWL

OWL - Web Ontology Language - est un langage qui étend RDF et fournit des inférences plus complexes, permettant une meilleure expressivité

des relations et la création d'ontologies plus sophistiquées.

Deux types de propriété peuvent être distinguées : les « DataProperty » et les « ObjectProperty ».

Une DataProperty décrit une relation entre une classe et une valeur d'un certain type. Le type peut être un littéral RDF ou un type simple [2]. Par exemple, la propriété âge est une DataProperty entre un individu et les entiers: « :Anne :âge :25 »

Les ObjectProperty décrivent une relations entre deux classes. Par exemple, la propriété mariDe est une ObjectProperty entre la classe Homme et la classe Femme

Parmi les propriétés introduites par OWL, citons:

- inverseOf: si la propriété propA inverseOf de la propriété propB et si X propA Y, alors Y propB X
- Functional: si X prop Y et X prop Z, alors Y=Z
- EquivalentClasses: si la classe X EquivalentClasses Y et si a est un élément de la X, alors a est un élément de Y et réciproquement
- InverseFunctionalObjectProperty: si X prop Z et Y prop Z, alors Y=X
- etc.

Un autre apport de OWL est la notion de restriction (limitation). LA classe OWL:restriction est définie par la description de ses membres; en d'autres mots par les propriétés que ses membres doivent respecter pour faire partie de cette classe. Si une classe satisfaisait la condition, alors elle est un élément de cette classe. La combinaison des notions de restriction et de subclass permet d'ajouter des contraintes sur les classes. Par exemple, si on considère la propriété :eat, les végétariens peuvent être définis comme des éléments de la classe Individu mangeant uniquement des légumes.

5.5 RDFa

RDFa - RDF dans des attributs de HTML - est une recommandation de W3C et comme son nom l'indique, est une extension de RDF. C'est une syntaxe qui permet de structurer les données d'une page web. Via l'utilisation de tag, les informations d'une page web pourront être transformées en RDF. Les spécifications peuvent être trouvées sur le site de la W3C: `http://www.w3.org/TR/xhtml-rdfa-primer`

2. `http://www.w3.org/TR/2004/REC-owl-guide-20040210/#Datatypes1`

Exemple :

```
<p xmlns:dc="http://purl.org/dc/elements/1.1/"
     about="http://www.example.com/books/wikinomics">
     Dans son dernier livre
     <em property="dc:title">Wikinomics</em>,
     <span property="dc:author">Don Tapscott</span>
     explique les profonds changements technologiques,
    démographiques et économiques.
     Ce livre a été publié en
     <span property="dc:date" content="2006-10-01">octobre 2006</span>.
 </p>
```

5.6 Vocabulaire proposé par Facebook

Facebook veut être présent partout sur le web. À cette fin, il souhaite faciliter l'entrée et la sortie de flux d'informations relatif à ses membres via l'utilisation du web sémantique.

Pour ce faire, mi-avril 2010, à la conférence F8 [3] Facebook a introduit de nouvelles fonctionnalités dont les widgets (ou social plugin) et l'Open Graph Protocol.

5.6.1 Open Graph Protocol

Le but de l'Open Graph Protocol est de mieux connaître le contenu des pages visitées. Il s'agit donc d'un langage de description du contenu.

Open Graph Protocol est un protocole basé sur RFDa. Il définit des tags à ajouter dans une page web, permettant ainsi de décrire le contenu de la page visitée. Facebook pourra donc analyser de manière plus précise la nature de la navigation de ses membres sur des pages externes au réseau social.

Les termes de bases correspondent à la description d'un document: titre, type, description, etc.

3. http://apps.facebook.com/feightlive/

Terme	Description
og:title	titre de l'objet
og:type	type de l'objet (film, ville,etc.)
og:image	url d'une image représentant l'objet
og:url	url de l'objet, utilisé comme identificateur dans le graphe
og:description	description de l'objet

Les termes correspondant à la localisation permettent de situer l'élément dont il est question dans la page web. Par exemple, une page sur le Colisée indiquera comme tag de localisation: Rome, Italie.

Terme	Description
og:latitude	Par exemple « 37.416343 »
og:longitude	e.g., "-122.153013 »
og:street-address	Par exemple « 1601 S California Ave »
og:locality	Par exemple « Palo Alto »
og:region	Par exemple « CA »
og:postal-code	Par exemple « 94304 »
og:country-name	Par exemple « USA »

Si le document est relatif à une personne qui peut être contactée, les informations de contacts peuvent également être spécifiées

Terme	Description
og:email	Par exemple « me@example.com »
og:phone_number	Par exemple « mbox+1-650-123-4567 »
og:fax_number	Par exemple « +1-415-123-4567 »

La description complète du vocabulaire peut être trouvée sur le site `http://opengraphprotocol.org`.

5.6.2 Les widgets

Facebook fournit une série de widgets permettant d'utiliser les fonctionnalités de Facebook, tel « J'aime » - ou « Like »en anglais - (voir figure 5.1) sur des pages externes au site, via l'ajout de plugin (voir figure 5.2). Un plugin Facebook est une extension que l'on ajoute à une page web afin de proposer des fonctionnalités supplémentaires. Des données sont ainsi transférées de la page web concernée vers Facebook: tel utilisateur a indiqué

FIGURE 5.1 – La fonctionnalité `J'aime` de Facebook permet de recommander du contenu

qu'il aimait telle vidéo. D'autres informations sont également transférées de Facebook vers la page web: « voici la liste d'amis qui ont également aimé cette page ». Afin de mieux cerner ce qui est « aimé », Facebook encourage l'utilisation des tags définis par l'Open Graph Protocol. Facebook sera ainsi informé que vous avez aimé une vidéo relative à tel sujet.

5.6.3 Implication

Facebook devient omniprésent sur web. En une semaine, plus de 50.000 pages ont adopté un ou plusieurs plugins [4].

Un site web tire avantage à implémenter ces widgets. En effet, d'une part, il ajoute des fonctionnalités « sociales » très facilement. Il offre ainsi la possibilité au visiteur de la page de noter du contenu, ou de le commenter. D'autre part, Facebook touche des millions d'internautes. C'est un excellent moyen de faire de la publicité gratuite, via « Vos amis ont aimés... »

En contre partie, le widget transfère toutes les informations relatives à la navigation sur le site. L'utilisation de l'Open Graph protocol, ainsi que des widgets proposés par Facebook risquent donc d'augmenter de manière significative la quantité d'informations que le site possède à propos d'un individu. En outre, il est plus que probable que Facebook ait spécifié des règles d'inférences. Ils peuvent donc raisonner de manière automatique sur les données ainsi reçues. Facebook est de plus en plus en mesure de mieux cartographier les centres d'intérêt relatifs à un membre du réseau.

Il n'est pas sûr que les sites implémentant ces widgets, ainsi que les utilisateurs de ces widgets soient conscients du transfert vers Facebook de toutes ces informations. Tout se fait de manière transparente. Ce n'est pas très loyal vis-vis des internautes. Le principe de base est le même que le web beacon déjà utilisé, mais il provoque beaucoup moins de réactions de la part du public.

4. `www.allfacebook.com/2010/04/50000-websites-add-facebooks-like-button-and-social-plugins-in-first-week/`

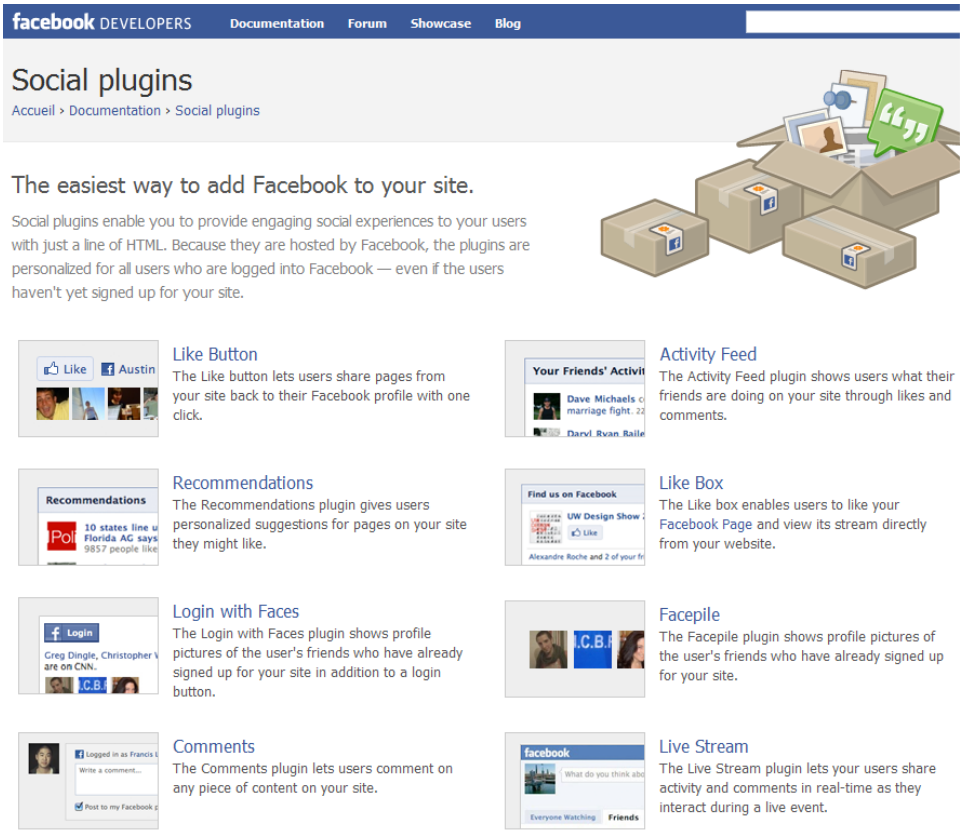

FIGURE 5.2 – Widgets proposés par Facebook
(source:http://developers.facebook.com/docs/plugins/)

Il devient de plus en plus difficile d'éviter ce réseau social. Chaque page visitée utilisant les fonctionnalités Facebook rappelle à l'utilisateur l'existence de Facebook. Quant aux sites extérieurs à Facebook, soit ils utilisent les widgets, soit ils risquent d'être ignorés de la vaste communauté des membres. Facebook étend ainsi sa présence sur le web.

Chapitre 6

Description logique et raisonneur

Dans notre travail, nous souhaitons raisonner sur l'ontologie relative aux sites de réseaux sociaux. Pour cela, nous allons utiliser les langages de descriptions logiques. En effet, pour décrire un domaine en utilisant les descriptions logiques, il faut d'abord définir les concepts principaux, puis les relations entre les concepts et/ou les individus. Il s'agit donc d'une démarche ontologique.

Une des caractéristiques des descriptions logiques, c'est la capacité de raisonner. Le raisonnement permet de résoudre des problèmes d'inférence et ainsi permet de mettre en évidence des connaissances implicites à partir de connaissances explicites. Nous allons donc dans un premier temps introduire les notions de description logique. Ensuite, nous parlerons des raisonneurs.

6.1 Description logique

Dans un article de Franz Baader [1] et de Werner Nutt [2][14], les logiques de description sont définies comme une famille de langages qui permettent de représenter la connaissance d'un domaine avec une sémantique formelle basée sur la logique. Les éléments fondamentaux d'une description logique sont:

- les concepts: catégories générales d'individus

1. Chercheur en informatique travaillant pour la DFKI (German Research Institute of Artificial Intelligence) ainsi que la TUD (Dresden University of Technology)
2. Professeur à la Free University of Bozen-Bolzano

- les relations entre les concepts et leurs propriétés
- les individus: éléments réels appartenant au domaine.

Dans une logique de description, la représentation de la connaissances est répartie en deux niveaux: la TBox et la ABox.

La TBox (Terminological Box) définit une terminologie, un vocabulaire. C'est la modélisation d'un domaine en termes de concepts, de propriétés et de rôles. Un rôle représente une relation entre deux individus. La TBox contient de même les concepts résultants de combinaisons de concepts et de rôles. Cela permet, par exemple de donner un nom à une construction complexe. Les descriptions élémentaires sont appelées concepts ou rôles atomiques.

La TBox contient également les axiomes terminologiques. Ils permettent de spécifier des liens entre les différents concepts et rôles. Ils sont souvent de la forme

$$C \sqsubseteq D \ (R \sqsubseteq S)$$
$$\text{or}$$
$$C \equiv D \ (R \equiv S)$$

ou C, D sont des concepts (R, S sont des rôles).

La ABox (Assertional Box) contient les assertions relatives à des individus du domaine, les assertions étant exprimées à partir du vocabulaire défini dans la TBox. Plusieurs ABox peuvent être associées à une même TBox, chaque ABox correspondant à un ensemble d'individus exprimés via les concepts et les rôles définis dans la TBox.

On donne une sémantique à une description via une fonction d'interprétation: un concept correspond à un ensemble d'individus, un rôle à une relation binaire entre des éléments de l'ensemble et un individu à un des éléments de cet ensemble.

Si Δ^I est un ensemble non vide et si $.^I$ est une fonction telle que
- à chaque individu a du domaine, on assigne $a^I \in \Delta^I$
- à chaque concept C, on assigne $C^I \subseteq \Delta^I$
- à chaque rôle R on assigne $R^I \subseteq \Delta^I X \Delta^I$

alors $\left\langle \Delta^I, .^I \right\rangle$ est une interprétation.

Une interprétation I satisfait
- l'axiome terminologique $C \sqsubseteq D$ si et seulement si $C^I \subseteq D^I$.

- l'axiome terminologique $C \equiv D$ si et seulement si $C^I = D^I$.

Une interprétation est un modèle si et seulement si l'interprétation satisfait tous les axiomes de la TBox.

Les langages de description logique se différencient par les constructeurs fournis afin d'élaborer des concepts et/ou des rôles composés. Le langage AL [3] a été introduit par [Schmidt-Schauß and Smolka, 1991] et constitue un langage minimal. La fonction d'interprétation a également été étendue au constructeur offert par AL :

- Le concept atomique A
- Le concept universel : $\top$ interprété par $\top^I = \Delta^I$
- Le concept bottom : $\bot$ interprété par $\bot^I = \emptyset$
- La négation atomique : $\neg A$ interprété par $\neg A^I = \Delta^I / A^I$
- L'intersection de concepts : $C \cap D$ interprété par $(C \cap D)^I = C^I \cap D^I$
- La restriction de valeur : $\forall R.C$ interprété par $\left\{a \in \Delta^I | \forall b : (a,b) \in R^I \rightarrow b \in C^I\right\}$
- La quantification existentielle limitée : $\exists R.\top$ interprété par $\left\{a \in \Delta^I | \exists b : (a,b) \in R^I\right\}$

Par exemple, $\neg Femme$ décrit les individus qui ne sont pas femmes, $Personne \cap Femme$ décrit les personnes qui sont des femmes, $Personne \cap \forall aDesEnfants.Fille$ décrit les personnes dont tous les enfants sont des filles et $Personne \cap \exists aDesEnfants.\top$ décrit les personnes ayant des enfants.

Différentes extensions existent. Par exemple, le langage ALU autorise tout les constructeurs d'AL, plus le constructeur U qui correspond à l'union $C \cup D$. Autres exemples, OWL-lite et OWL-DL, deux sous langages d'OWL.

OWL-lite est la version la plus simple et équivaut au langage $SHIF$ avec

- U : union $C \cup D$, interprété par $(C \cup D)^I = C^I \cup D^I$
- ε : quantificateur existentiel typé $\exists R.C$ interprété par $\left\{a \in \Delta^I | \exists b \in C^I : (a,b) \in R^I\right\}$
- R^+ : transitivité des rôles
- $S = ALU\varepsilon$ augmentée de R^+,
- H = hiérarchisation des rôles $R_1 \subseteq R_2$ interprété par $R_1^I \subseteq R_2^I$
- I = rôle inverse, interprété par $\left\{(b,a) \in \Delta^I X \Delta^I | (a,b) \in R^I\right\}$
- F : contrainte de cardinalité et admettant deux constructeurs:$1R$ et $2R$ qui correspondent respectivement à $\left\{a \in \Delta^I | \left|\{b \in \Delta^I | (a,b) \in R^I\}\right| = 1\right\}$ et $\left\{a \in \Delta^I | \left|\{b \in \Delta^I | (a,b) \in R^I\}\right| >= 2\right\}$

,

3. Attributive Language

La version OWD-DL est équivalente au langage $SHOIN$; donc, par rapport à OWL-lite, deux constructeurs ont été ajoutés:

- O permet la description d'un concept par l'énumération des individus qui le composent et est interprété par $\{a_1, a_2, \ldots, a_n\}^I = \left\{a_1^I, a_2^I, \ldots, a_n^I\right\}$
- N est une extension de la contrainte de cardinalité F. Elle autorise trois constructeurs:
  - $\left\{a \in \Delta^I \middle| \left|\{b \in \Delta^I | (a,b) \in R^I\}\right| = n\right\}$
  - $\left\{a \in \Delta^I \middle| \left|\{b \in \Delta^I | (a,b) \in R^I\}\right| >= n\right\}$
  - $\left\{a \in \Delta^I \middle| \left|\{b \in \Delta^I | (a,b) \in R^I\}\right| <= n\right\}$

6.2 Raisonner sur une ontologie

Les principales tâches d'un raisonneur sont:

- de déterminer si une description est satisfaisable, une description étant satisfaisable s'il n'y a pas de contradiction dans le modèle.
- de déterminer si une description est la subsomption d'une autre, c'est-à-dire de déterminer si la description est plus générale qu'une autre
- de vérifier que les assertions de la ABox sont consistantes. Une ABox est consistante s'il existe un modèle, c'est-à-dire s'il existe une interprétation satisfaisant tous les axiomes de la TBox.
- de s'assurer que les assertions de la ABox permettent de préciser à quel(s) concept(s) est attaché un individu.
- de lister les individus instanciant un concept particulier.

Ces propriétés permettent à l'utilisateur d'un raisonneur de poser des questions relatives au domaine modélisé. En charge, alors, au raisonneur de répondre à la question en utilisant différents algorithmes.

Plus un langage accepte de constructeurs, plus il est expressif et permet de définir les différents concepts du domaine de manière précise. Mais plus une modélisation est expressive, plus elle devient complexe et il devient alors difficile de raisonner dessus. Un problème de raisonnement est dit décidable si une machine de Turing peut le résoudre en un nombre fini d'étapes. Lorsque l'expressivité est trop grande, le langage peut devenir non décidable. Il s'agit alors de faire un compromis entre l'expressivité et la décidabilité. Le langage OWL correspond en fait à trois sous-langages: OWL-lite, OWL-DL et OWL-full. Les langage OWL-lite et OWL-DL sont décidables. OWL-full est la version d'OWL la plus expressive. Elle est non décidable. Elle permet,

par exemple, d'utiliser des classes en tant qu'individu.

Il existe différentes implémentations de raisonneur utilisant différents algorithmes et supportant différents constructeurs. Nous discuterons du choix du raisonneur dans le chapitre suivant, car il sera fonction de l'expressivité de l'ontologie.

6.3 SWRL

Le langage OWL-DL est assez restrictif. Afin d'augmenter son expressivité, le langage SWRL - Semantic Web Rule Language - a été introduit. Une règle SWRL contient un antécédent et un conséquent. Elle exprime que si l'antécédent est vrai, alors la conséquence doit l'être également. Cela permet une très grande expressivité. Par exemple, dans OWL, on peut définir le concept d'oncle par:

```
intersectionOf(SubClassOf(Homme), estfrereDe(Pere)).
```

Mais on ne sait pas définir la relation estOncleDe. Avec SWRL c'est possible:

$$aEnfant(?x, ?y) \text{ ^ } estfrereDe(?z, ?x) -> estOncleDe(?z, ?y)$$

Mais cette très grande expressivité a un prix: SWRL n'est pas décidable. Il existe une restriction de SWRL décidable qui travaille exclusivement sur les instances.

Chapitre 7

Problème, solution et méthode

7.1 Le Problème

Le problème qui se pose ici est de déterminer la connaissance que les sites de réseaux sociaux peuvent accumuler à propos de leurs membres. Toutes ces informations sont collectées, stockées, utilisées, parfois vendues par les sites. Cette connaissance est critique et peut avoir des impacts dans la vie réelle.

Au départ, il y a les renseignements donnés par l'utilisateur, via le profil. Ceux-ci sont complétés par l'activité du membre sur le site: publication de posts, affiliation à un groupe, liste des amis, etc. Ensuite, il y a les informations glanées par le site à l'extérieur du réseau. La récolte peut avoir lieu via des web beacons, des cookies, etc. Elle peut également se faire sur les pages personnelles, les blogs, etc. C'est ce que nous appelons la connaissance explicite.

Ensuite, basée sur toutes ces informations, vient la manipulation des données qui permet de déduire de nouvelles connaissances: identification via recoupements avec des éléments démographiques, association avec un profil d'un autre réseau social, etc.

L'utilisation des techniques de profilage (section 3.4) permet encore d'accroître de manière significative toute cette connaissance, en catégorisant les individus et en donnant des probabilités de réactions à certains stimuli. Les réseaux pourraient donc prédire le comportement de certains individus face

à différentes situations, et ainsi tenter de les manipuler.

Nous appelons « connaissance implicite » toute connaissance déduite à partir des connaissances explicites.

Toutes ces informations critiques peuvent ainsi se retrouver dans les bases de données d'un site de réseau social. Outre l'utilisation potentielle par le site, ces données sont également susceptibles de se retrouver un jour dans de mauvaises mains. En effet, aucun site n'est à l'abri, par exemple, d'attaques de hackers (voir chapitre 4), dont les conséquences peuvent être fâcheuses (voir, par exemple, section 3.1). Les enjeux sont donc importants.

7.2 Solution et méthode

Le but de ce mémoire est de créer une ontologie modélisant la connaissance que les sites de réseaux sociaux peuvent accumuler à propos de leurs membres. Cette connaissance ne peut être qu'hypothétique, car il est difficile de savoir exactement ce que les sites collectent réellement. Ensuite, cette ontologie a été utilisée afin d'essayer d'évaluer les conséquences en cas de connexion de deux profils lors d'une fusion de deux sites.

7.2.1 L'ontologie

Dans un premier temps, nous avons essayé de modéliser la connaissance directe qui peut être glanée sur le site. En d'autres mots, nous avons essayé de déterminer quelles sont les données qui peuvent être collectées et stockées par les sites de réseaux sociaux, à l'intérieur même des sites.

N'ayant pas accès aux serveurs ni aux bases de données des réseaux, la connaissance n'a pu être extraite que par rétro-ingénierie, c'est-à-dire de manière externe, via la navigation sur ces sites.

L'ontologie concerne la connaissance à propos d'un individu. D'où, toutes les informations pouvant augmenter cette connaissance et qui sont fournies par le profil, ainsi par l'activité de l'individu au sein des sites ont été considérées. Nous avons listé les informations pertinentes de la manière la plus exhaustive possible. Il est à noté qu'il n'est pas toujours évident de déterminer si l'information est pertinente. Par exemple, la couleur de fond de la page sur le site de réseau social peut sembler sans importance, mais cela indique déjà que l'individu aime cette couleur. De plus, certaines informations

collectées le sont de manière transparente pour l'individu, tel l'adresse IP, le navigateur web utilisé, etc.

Selon le type de réseau social, les données collectées peuvent être de natures différentes. LinkedIn sera informé sur le passé et le présent professionnel d'un individu; Flickr contiendra des informations relatives à la passion de la photographie, etc. C'est pourquoi nous avons choisi quelques sites représentatifs. A cette fin, nous avons utilisé les classifications proposées précédemment à la section 2.2.

Six sites ont été retenus:

- Facebook - `www.facebook.com`
- Flickr de Yahoo- `www.flickr.com`
- LindedIn - `www.linkedin.com`
- Youtube de Google- `www.youtube.com`
- LiveJournal - `www.livejournal.com`
- match.com - `www.match.com`

Ces sites ont été choisis, d'une part parce qu'ils font parties des plus utilisés, et d'autre part, parce qu'ils permettent de couvrir les différentes catégories de nos classifications:

- Vis à vis de l'accessibilité, nous avons choisi de ne considérer que des sites publics.
- Vis-à-vis de la divulgation de l'identité :
  - divulgation complète du nom et prénom (Facebook)
  - utilisation d'un pseudo (match.com)
- Vis-à-vis du thème :
  - les réseaux ayant un thème spécifique (Flickr)
  - les réseaux professionnels (Lindedln)
  - les plateformes de partage (youtube)
- Vis-à-vis du moteur :
  - les réseaux « classiques » (Facebook)
  - les weblogs (LiveJournal)
- Vis-à-vis de la confidentialité :
  - les réseaux offrant la possibilité de paramétrer la visibilité de pratiquement tout (Facebook)
  - les réseaux sans aucune confidentialité (match.com)
- Vis-à-vis du financement :
  - Facebook vend principalement des espaces publicitaires personnali-

sés, ainsi que des données - anonymisées - de ses membres

- Flickr vend essentiellement des produits ou des services. La publicité, quant à elle, existe sur le site, mais est relativement discrète.
- Youtube propose des espaces publicitaires basés sur le contenu
- Match vend des services, ainsi que des espaces publicitaires basé sur le contenu [1] [2]
- LiveJournal vend des produit et des services, ainsi que espaces publicitaires basés sur le publiques ainsi que sur le contenu,
- LinkedIn vend des services, ainsi que des espaces publicitaires basés sur le publique.

Notons que nous n'avons considéré que les versions non-payantes des comptes utilisateurs.

Il s'agit d'une sélection arbitraire. Un autre choix aurait tout à fait été possible, pouvant donner éventuellement des résultats différents.

En outre, les logiciels supportant les réseaux sont en perpétuelle évolution. En effet, des mises à jours sont régulièrement proposées: telle fonctionnalité est supprimée, telle autre proposée,... Le modèle présenté correspondra donc à une « photographie » d'une situation à un moment précis, en l'occurrence, mars 2010.

Une fois les informations recueillies sur ces sites représentatifs, nous avons, pour chaque site, modélisé cette connaissance via un schéma ERA [3] dans DB-main [4]

Les modèles ont été construits en appliquant la même méthodologie que pour la construction d'un schéma conceptuel de base de données. En d'autres mots, détermination des concepts, arrangement de ces concepts selon des liens hiérarchiques, détermination des attributs et des relations entre les concepts.

Ensuite, nous avons fusionné ces modèles afin d'obtenir l'ontologie recherchée. De manière à garder une correspondance entre l'ontologie et les sites de réseaux sociaux d'origines, nous avons indiqué, pour chaque élément, via une méta-propriété (i.e. propriété d'une propriété), le ou les réseaux sociaux collectant l'information. Cette méta-propriété - appelée « Reseau

1. Politique sur la vie privée de match: `http://fr.match.com/misc/privacy_v.php`
2. Publicité sur Meetic (qui a fusionné avec Match Europe en fevrier 2009), `http://www.meetic.fr/misc/advertising_v.php`
3. Schéma Entité Relations Attributs
4. Outil de modélisation et d'architecture de données, `http://www.db-main.eu`

social » dans le schéma - a été définie pour les entités, les attributs et les relations. Elle est multivaluée, une même information pouvant, en effet, être collectée par différents sites. Par exemple, les émissions télévisées préférées sont collectées à la fois par Facebook et par Youtube.

La méthode pour construire le schéma global a été la suivante :

1. Construction d'un schéma relatif à Facebook. Ce schéma est le plus détaillé et le plus complet possible.
2. Construction d'un schéma plus léger pour Flickr. Ce schéma contient principalement les concepts relatifs à Flickr.
3. Construction d'un nouveau schéma intégrant le schéma de Facebook et les concepts de Flickr.
4. Achèvement du nouveau schéma en le complétant avec les éléments de Flickr non-encore intégrés
5. Faire de même avec les schémas de LiveJournal, Linkedin, Youtube et Match.com

Avant et après chaque étape, le schéma a été revu afin d'en vérifier la cohérence, de déterminer si des règles de hiérarchies peuvent s'appliquer, s'il n'y a pas de redondance, etc.

Les schémas propres à Flickr, LiveJournal, etc. utilisés pour la construction du modèle global n'ont pas besoin d'être exhaustifs car ils ne servent qu'à la construction du modèle global. Mais ils doivent être suffisamment complets afin de permettre la mise en évidence des concepts. Les schémas complets de Flickr, LiveJournal, etc. seront extraits du modèle global via la méta-propriété « Reseau social ».

L'intégration des différents schémas s'est faite de manière incrémentale car les schémas peuvent être relativement complexes, incluant des concepts souvent semblables, mais avec des nuances parfois importantes. Facebook a été choisi comme schéma de départ parce qu'assez généraliste et très riche en concepts.

Pour chaque modèle à intégrer, la recette de[10] a été appliquée : identifier les correspondances, résoudre les éventuels conflits et fusionner les composants du schéma à intégrer avec le schéma global.

Par exemple, il y a correspondance entre la notion de petites annonces de LiveJournal et d'offres d'emplois de LinkedIn. Dans le schéma global,

le concept d'offres d'emplois a ainsi été renommé `annonces d'emplois` et exprimé comme un sous-type du concept de petites annonces.

L'outil ne permettant pas la définition d'axiomes, l'ontologie obtenue sera donc légère.

7.2.2 Annotation

Afin de facilité la manipulation du modèle, les différents types d'information rencontrés dans l'ontologie ont été annotées également via l'ajout de méta-propriétés.

Nous avons essayé de trier les informations en les classant dans différentes catégories. Cela permet de préciser le type d'informations récoltées. Nous avons considéré les catégories suivantes:

- contact: toute information permettant de contacter l'utilisateur en dehors du site. Exemple: numéro de téléphone, adresse, etc.
- religieux: toute information pouvant, d'une manière ou d'une autre, donner des indications sur la religion pratiquée par l'individu. Par exemple, le prénom: il y a en effet une probabilité assez grande qu'un individu s'appelant Mohammed soit musulman
- politique: toute information pouvant donner des indications sur les opinions politiques
- intérêt: toute information relative aux goûts de l'utilisateur, à ce que qu'il aime ou n'aime pas
- physique: toute information pouvant donner des indications sur le physique d'une individu. Par exemple, origine ethnique, couleur des yeux, etc.
- identification: toute information pouvant servir à l'identification de l'individu parmi un set de données. Par exemple, le nom d'utilisateur, la page web personnelle, etc.
- dépend du contenu: le contenu doit être analysé afin de connaître le type d'information pouvant être extrait. Les informations de cette catégorie peuvent donner des indications sur la religion, la politique, la personnalité, etc.; Par exemple, la publication d'une photo d'un meeting politique avec un utilisateur marqué sur cette photo renseigne sur ses options politiques.
- géographique: toute information permettant de localiser un individu à un moment donné. Par exemple, la position géographique et la date

de prise de vue d'une photographie.
- sexualité: toute information relative à la sexualité de l'individu
- relationnel: toute information permettant de déterminer les personnes en relation avec l'individu, ainsi que le type de relation entretenue. Par exemple, les amis, les invitations, etc.
- profession/étude: toute information relative aux études et à la ou les professions exercées par l'individu
- personnalité: toute information relative à la personnalité, au caractère de l'individu.
- financier: toute information donnant des indications sur les revenus financiers de l'individu, ainsi que sur son niveau de vie

Une information pouvant appartenir à plusieurs catégories, la méta-propriété est multivaluée.

La répartition dans les catégories a été réalisée en fonction du « bon sens ». Mais il est probable que l'on puisse en déduire encre plus que ce qui a été spécifié dans le schéma. Les statisticiens, via des corrélations, pourraient sans aucun doutes compléter cette méta-propriété de manière beaucoup plus rigoureuse.

7.2.3 Mise en évidence de l'information

L'outil DB-main permet de marquer des éléments d'un schéma. Il y a cinq marquages différents possibles: « Mark1 » à « Mark5 ». Le logiciel propose également la génération de vues. Une vue est un schéma qui dérive d'un autre schéma - qui peut également être une vue - et qui contient les éléments ayant le marquage choisi.

Par exemple, on peut choisir de marquer tout les éléments venant de Facebook avec la « Mark1 », les éléments de Flickr avec la « Mark2 », etc. On peut ensuite générer une vue que l'on appellera viewfacebook et qui contiendra tout les éléments du schéma principal ayant la « Mark1 ». Par après, dans la vue viewfacebook, on peut marquer avec la « Mark1 » les éléments venant de Flickr pour générer une autre vue contenant tout les éléments de la vue viewfacebook ayant la « Mark1 » etc. Il y a une infinité de vues possibles.

Parallèlement, l'outil permet de changer la couleur des différents éléments d'un schéma ou d'une vue. Il s'agit d'une autre manière de mettre en évidence des informations.

Un plugin utilisant les capacités de marquage et/ou de mise en couleur des éléments des schémas de DB-Main a été implémenté. Un plugin DB-main est un programme développé en java (*.class) [5] utilisant l'API java JIDBM [6] permettant d'accéder, en lecture et/ou en écriture, aux référentiels de DB-Main.

Ce pluging facilite la manipulation de l'ontologie. Il permet également de mettre en évidence visuellement les points de correspondance entre les informations collectées par différents sites de réseaux sociaux. Par exemple, il est facile de voir toutes les informations collectées à la fois par Facebook et par Flickr.

7.2.4 Conversion du schéma ERA en OWL

Le schéma ERA utilisé pour représenter l'ontologie est trop limité. Il ne permet pas d'exprimer que des ontologies légères. Exprimer l'ontologie via OWL permet de compléter le schéma via, par exemple, l'application de mécanismes d'inférences et l'utilisation des raisonneurs. Le schéma a donc été converti.

La conversion du schéma ERA en OWL peut se faire de manière automatique. Pour ce faire, nous avons créé le modèle OWL sous la forme d'un fichier au format XML; fichier qui a été ensuite importé dans l'éditeur OWL choisi.

Pour chaque entité du schéma ERA, de manière logique, une « Classe » ayant le même nom a été crée. Les liens de hiérarchies ont été transposés via la propriété « SubClassOf ».

En ce qui concerne les méta-propriétés relatives aux entités, nous les avons transposées via les « AnnotationProperty ». Notons que la méta-propriété relative au réseau social a été transposée différemment. En effet, cette méta-propriété est nécessaire afin d'évaluer les risques et les conséquences d'une fusion via un raisonneur; Or, les AnnotationProperty ne sont pas supportées par les raisonneurs car ils ne sont pas OWL-DL. Pour chaque site de réseau social, une classe définie par les instances de concepts collectés par le site a été créée. Cela correspond donc a six classes: Facebook, Youtube, etc. Le fait qu'un type d'information n'est pas collecté par un site a

5. Pour question de compatibilité ascendante, DB-Main autorise toujours l'exécution de plug-in développé en Voyager 2(*.oxo)

6. Java Interface for DB-Main

été exprimé en spécifiant que le concept et la classe du site sont disjoints.

Pour chaque relation, un ObjectProperty a été crée. Dans notre modèle ERA, il n'y a pas de relation n-aire. Chaque relation est donc liée à deux rôles. La classe correspondant à l'entité ayant le rôle listé en premier dans le schéma DB-main a été considérée comme le domaine de cette propriété. L'entité correspondant au rôle listé en deuxième correspond au range. Les cardinalités d'un rôle correspondant au domaine ont été exprimées selon les règles suivantes:

- $[1-1]$ a été traduit via ObjectSomeValuesFrom et FunctionalObjectProperty
- $[1-N]$ a été traduit via ObjectSomeValuesFrom
- $[0-1]$ a été traduit via FunctionalObjectProperty
- $[0-N]$ n'étant pas contraignante, elle ne nécessite pas d'être transposée.

Les cardinalités d'un rôle correspondant au range ont été exprimées selon les règles suivantes:

- $[1-1]$ a été traduit via InverseFunctionalObjectProperty, ainsi que ObjectSomeValuesFrom appliqué à la propriété inverse
- $[0-1]$ a été traduit via InverseFunctionalObjectProperty
- $[1-N]$ a été traduit via ObjectSomeValuesFrom appliqué à la propriété inverse
- $[0-N]$ n'étant pas contraignante, elle ne nécessite pas d'être transposée.

Notons que notre schéma ne contient aucun rôle avec une cardinalité de type $[N-N]$.

Les méta-propriétés s'appliquant également aux relations, elles ont été transposées de manière similaire. Les méta-propriétés autres que « réseau social » ont été exprimées via les AnnotationProperty. Quant à la méta-propriété « réseau social » le fait qu'une propriété est non collectée par un réseau social a été en limitant le domaine et le range au complément de la classe relative au site.

Pour chaque attribut, une ObjectProperty et une classe ont été crées. Le domaine de cette propriété est la classe correspondant à l'entité parente. Le range est la classe relative à l'attribut. Par exemple, l'attribut « nom » de la classe Personne a été traduit par une classe Nom et un ObjectProperty entre la classe Personne et la classe Nom. Une cardinalité du type:

- $[1-1]$ a été traduite via FunctionalObjectProperty et ObjectSomeValuesFrom
- $[0-1]$ a été traduite via FunctionalObjectProperty
- $[1-N]$ a été traduite via ObjectSomeValuesFrom
- $[0-N]$ n'étant pas contraignante, elle ne nécessite pas d'être transposée.

Notons que notre schéma ne contient pas d'attribut avec une cardinalité de type $[N-N]$.

L'utilisation d'ObjectProperty pour les attributs a été préférée à la notion de DataProperty parce qu'un attribut est un concept, et qu'il peut s'exprimer avec des littéraux différents. Par exemple, le nom d'une personne peut s'écrire J.Smith ou John Smith ou Mr. J. Smtih, etc.

Les méta-propriétés relatives aux attributs ont été transposées de la même manière que pour les relations.

De manière générale, les constructeurs ont été choisis afin d'avoir une ontologie exprimée dans un langage de description logique décidable; en outre, cette ontologie doit être capable de répondre à des questions du type « un tel set d'information est-il collecté par tel site de réseau social? ».

7.2.5 Complétion du modèle OWL

L'ontologie traduite en OWL offre la possibilité d'exprimer la connaissance de manière plus précise, à condition de compléter le schéma.

L'expression de l'ontologie sous forme d'un schéma ERA ne permet pas de mettre en évidence des liens d'héritage entre les différentes relations et les différents attributs. La relation « est de la même famille que » peut être déclinée dans OWL comme une super-propriété du même nom, avec des sous-propriétés « est le frère de », « est la mère de », etc. De même, il est possible de créer, via OWL, une super-propriété: « connait » ancêtre de toutes relations du type: « est amis », « est de la même famille », etc.

ERA ne permet également pas l'écriture d'inférences. Or, on voudrait exprimer le fait, qu'à partir des données récoltées sur un individu, il est possible de déduire de nouvelles connaissances, en utilisant différentes techniques, comme par exemple, le profiling (voir section 3.4 ou 4.4). Notons qu'il n'est pas toujours nécessaire de recourir à des techniques compliquées afin d'augmenter la connaissance. Voici, par exemple, une inférence pouvant être déduite à partir d'un simple raisonnement:

```
Si
   X est la mère de Y
   X est la mère de Z,
Alors
   Y est le (demi-)frère de Z, et réciproquement.
```

Ce genre d'inférences peut parfaitement être exprimé dans le modèle OWL à l'aide de SWRL. Néanmoins, les inférences de ce genres ne seront ajoutées au modèle qu'avec parcimonie, car SWRL est non-décidable. De plus, il est illusoire de vouloir lister de manière exhaustive toutes les inférences pouvant être établies. Nous n'en établirons donc que quelques unes, si nécessaire, afin d'illustrer nos propos.

Notons que nous n'établirons pas d'inférences basée sur le principe « dis moi qui sont tes amis et je te dirai qui tu es » (voir section 4.4). En effet, cela nécessiterait des données sociologiques que nous n'avons pas et qui sont parfois difficiles à deviner. De plus, la fiabilité des résultats est très variable.

Nous ne tiendrons également pas compte d'informations dépendant uniquement du contenu. Cela demande, en effet, des outils spécifiques d'analyse de photographies, de textes... - outils que nous n'avons pas - afin d'en extraire la connaissance.

Enfin, le schéma ERA a dû être affaibli afin de fusionner les concepts des différents sites. Par exemple, si sur Facebook, la date de naissance est obligatoire alors que sur match.com, elle ne l'est pas, le modèle fusionné n'aura pas d'autre choix que de spécifier une cardinalité [0,1]. Par contre, dans le modèle OWL, cela ne pause aucun problème. On peut parfaitement exprimer que tout individu qui est instance de Facebook a des valeurs pour la propriété « date de naissance ».

Le langage OWL complété de SWRL permet donc à la fois de définir des liens d'héritage entre des propriétés, d'établir des axiomes permettant d'inférer et de préciser les cardinalités en fonction du site de réseau social. Le modèle OWL, résultat de la conversion automatique décrite dans la section précédente, nécessite donc d'être revue et complétée manuellement. Notons qu'il ne sera jamais complet car il est impossible d'y intégrer toutes les inférences possibles.

7.2.6 Choix de l'éditeur OWL et du raisonneur

Comme éditeur OWL, notre choix c'est porté sur Protégé [7], car c'est un outils open source et parmi les plus utilisés, donc stable, avec une grande communauté pouvant fournir de l'aide. De plus, Protégé permet d'importer une ontologie exprimée via un fichier XML et est fournie avec une API Java.

En ce qui concerne le raisonneur, Protégé admet quelques raisonneurs comme plug-in. Parmi les open-sources, il y a HermiT [8], Pellet [9] et FaCT++ [10]. Ces trois raisonneurs fonctionnent sur des ontologies de type OWL-DL [11].

Nous n'utiliserons pas FaCT++, car il semble ne pas prendre en charge les règles ajoutées via SWRL [12]; Or, il se peut que nous devions en écrire quelques unes. De plus, la documentation relative à ce plugin est relativement rare. Par exemple, nous n'avons trouvé aucun document précisant clairement les limitations de l'outil. Entre HermiT et Pellet, notre choix s'est porté sur Pellet, car il semble nettement plus rapide. De plus, Pellet est parmi les plus utilisés, donc il est assez facile de trouver de la documentation.

7.2.7 Risque de connexion entre deux profils

Nous tenterons d'évaluer le risque de voir deux profils se trouvant dans deux réseaux distincts d'être connectés par les sites de réseaux sociaux, et ce, en fonction des données fournies par l'utilisateur. Cela peut se produire lorsque par exemple, un site est racheté par un autre.

Parfois, c'est l'utilisateur lui-même qui fait la connexion. En effet, actuellement, beaucoup de sites proposent de lier le compte d'un membre avec un ou plusieurs autres de ses comptes maintenus par un ou plusieurs autres sites de réseaux sociaux. Cela constitue une facilité pour l'utilisateur: il peut ainsi gérer et publier du contenu à partir d'un seul endroit.

Mais le lien pourrait se faire sans le consentement, ni à la connaissance de la personne. Ce lien pourrait ainsi être basé sur un identifiant de messagerie instantanée, sur une adresse email, ou une combinaison d'informations etc.

7. `http://protege.stanford.edu/overview/protege-owl.html`
8. `http://hermit-reasoner.com/`
9. `http://clarkparsia.com/pellet/`
10. Page officiel(?), `http://owl.man.ac.uk/factplusplus/`
11. `http://www.w3.org/2007/OWL/wiki/Implementations`
12. http://en.wikipedia.org/wiki/Semantic_reasoner

Les listes d'amis pourraient être utilisées afin de connecter des profils de différents sites de réseaux sociaux. A titre informatif, il est possible de trouver une algorithme de fusion dans le document [16].

Nous allons concentrer notre travail sur les caractéristiques d'un individu. Nous baserons le risque que nous cherchons à quantifier sur la rareté d'un set d'informations. Par exemple, considérant deux profils distincts:

- un individu précisant dans ses deux profils son nom, son prénom ainsi que sa date de naissance, a un risque total de voir ses données de profil connectées, car ces informations sont généralement considérées comme uniques.
- un individu dont l'unique information commune connue est la couleur de ses cheveux, en l'occurrence bruns, court un risque de connexion quasi nul; cette caractéristique étant très courante.

La rareté d'un set d'informations peut être estimée à partir d'un jeu de données en évaluant le pourcentage d'individus ayant ces caractéristiques. Par conséquence, des données statistiques relatives à la fréquence d'apparition des différents éléments d'un profil sont nécessaires. Les sites de réseaux sociaux possèdent ces informations. Dés lors, cette méthode est facilement réalisable par ces sites. Or ces données ne nous sont pas accessibles. Nous nous limiterons donc à évaluer de manière arbitraire - au seul fin de validation de l'outil - la rareté de quelques sets selon l'échelle: nul, faible, moyen, fort, total.

Notons que l'évaluation de la rareté d'un set d'informations peut être influencé par la variation de la transcription d'une information. Une même donnée peut ainsi être écrite différemment, soit parce qu'il existe plusieurs manières de l'orthographier, soit parce que le terme contient une ou des fautes d'orthographes. Par exemple, les noms suivants se rapportent tous à la même personne:

- J.F.K.
- John Kennedy
- J. Kennedy
- John Kenedy
- Kennedy John
- John Fitzgerald Kennedy
- Jack Kennedy
- ...

Il existe des outils, comme, par exemple Open Source ChoiceMaker Technology [13] qui permettent, en grande partie, de résoudre ces conflits. Nous supposerons dans notre travail qu'il n'y a pas de variation dans l'écriture des données.

Nous avons illustré le risque de connexion à deux niveaux différents. Au premier niveau, nous nous sommes attachés à rester au niveau de la TBox via l'écriture de questions posées au modèle. Par exemple, sachant qu'une combinaison particulière de données débouche sur un risque fort ou total de connexion de deux profils, peut-on utiliser cette combinaison afin de connecter un profil Facebook avec un profil LinkedIn? En d'autres mots, les éléments de la combinaison sont-ils collectés par Facebook et LinkedIn? Nous analyserons quelques sets d'informations.

Au deuxième niveau, nous avons envisagé quelques cas particuliers: étant donné deux profils concrets, quel est le risque de voir ces deux profils connectés? Nous l'étudierons via l'étude de quelques cas.

7.2.8 Conséquence de la connexion de deux profils

Nous avons essayé de découvrir les conséquences vis-à-vis de la connaissance sur un individu lorsque deux profils de réseaux sociaux différents sont associés.

Les informations laissées sur les deux profils peuvent correspondre à des facettes totalement différentes d'une personnalité. Un individu choisit un site en fonction d'objectifs spécifiques qui lui sont personnels. Il peut être sérieux sur un site professionnel, avec son identité dévoilée, son expérience professionnelle afin d'attirer des employeurs potentiels. Il peut être plus expansif, par exemple, sur un site de partage de photographies, dévoilant ses lieux de voyage etc. Il peut également avoir un comportement totalement décalé sur un autre en se cachant derrière un pseudo. Cette séparation permet de laisser libre cours à différentes facettes de l'utilisateur sans que l'une impacte sur l'autre. En réunissant les données liées à ces deux facettes, la connaissance sur un individu peut se voir fortement augmentée.

De plus, une plus grande connaissance sur un individu implique de plus grandes possibilités de déductions. Exemple:

`Si`

13. http://oscmt.sourceforge.net/content/intro/index.html

```
   sur Facebook, Anne est la femme de Jean
   sur Flickr, Jean habite à Namur
Alors
   Anne habite très probablement à Namur.
```

Nous avons étudié l'évolution du type de connaissances qu'un réseau social a de ses membres en cas de fusion.

Nous avons également essayé d'évaluer quelques-unes des conséquences suite à la connexion de deux profils, via les cas particuliers étudiés lors du calcul de risque de fusion. Il est, en effet, utopique de croire que l'on peut prévoir toutes les conséquences à partir de quelques inférences (voir section 7.2.5).

Chapitre 8

Résultats

En se contentant de relever les informations stockées directement par les différents sites, nous arrivons à plus de cinq cents éléments distincts, chacun d'eux pouvant être analysé et utilisé; ce qui est colossal, sachant qu'à partir de là, la connaissance peut encore être augmentée via différentes techniques (voir, par exemple, section 3.4).

Tous les sites ne collectent pas les cinq cents informations; mais, en cas de fusion, les cinq cents informations peuvent se retrouver réunies au même endroit. Nous allons essayer de décrypter les informations collectées, les risques de connexion, ainsi que ses conséquences.

8.1 Schémas

Via le plugin décrit dans la section 7.2.3, nous avons extrait de l'ontologie les schémas relatifs aux différents sites de réseaux sociaux: Facebook (8.1), Flickr (8.2), LinkedIn (8.3), LiveJournal (8.4), Youtube (8.5) et Match (8.6)

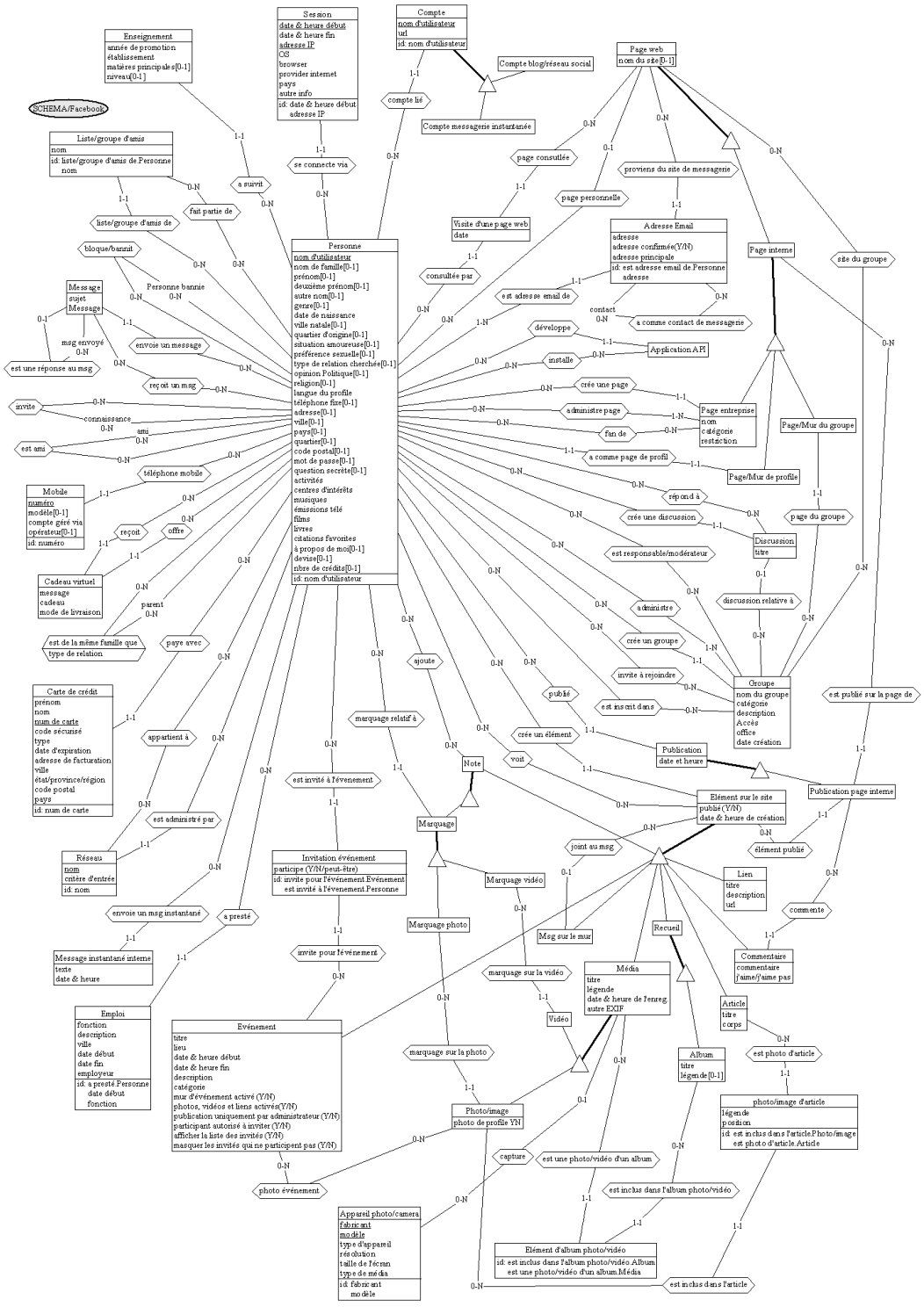

FIGURE 8.1 – Informations collectées par Facebook

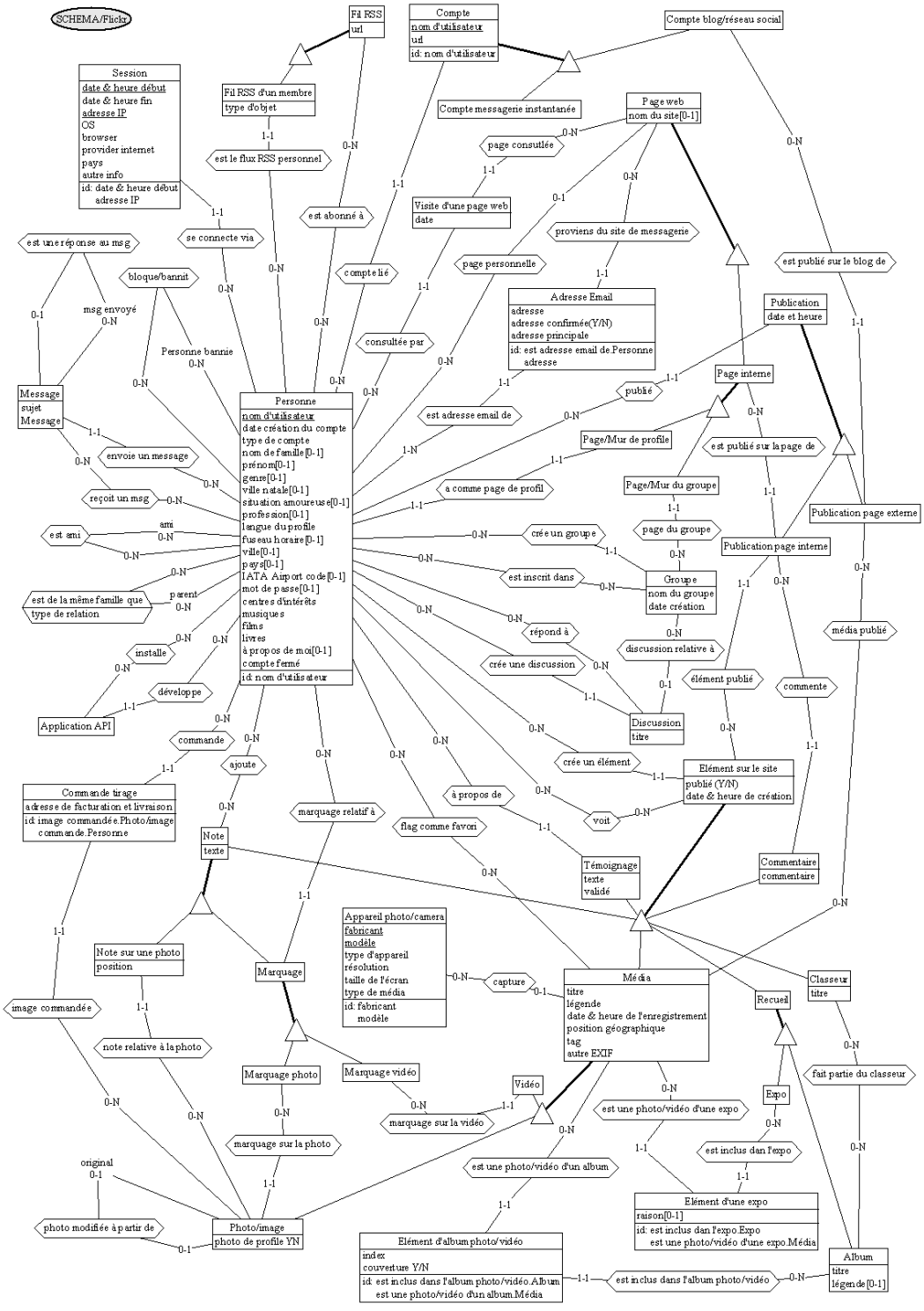

FIGURE 8.2 – Informations collectées par Flickr

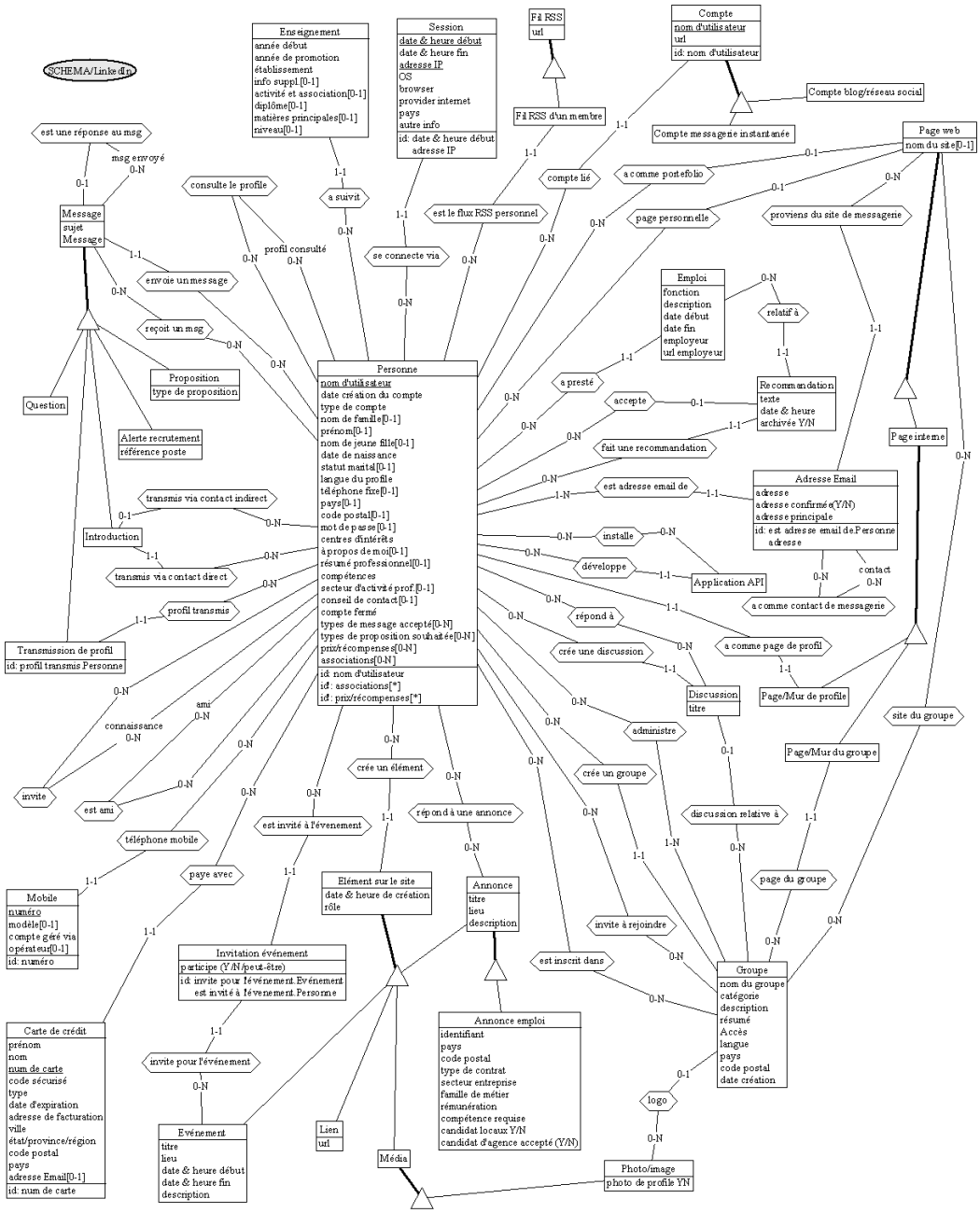

FIGURE 8.3 – Informations collectées par LinkedIn

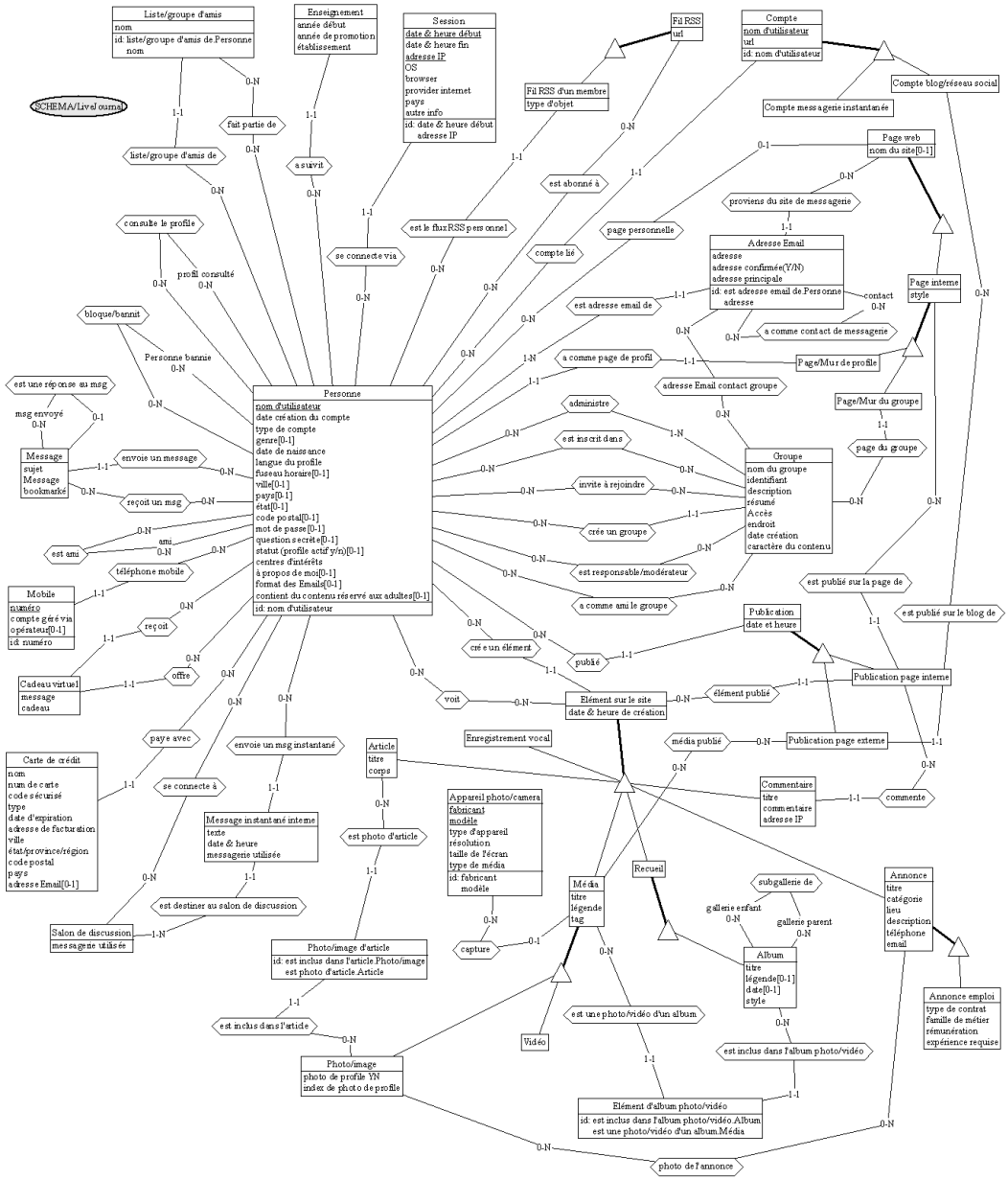

FIGURE 8.4 – Informations collectées par LiveJournal

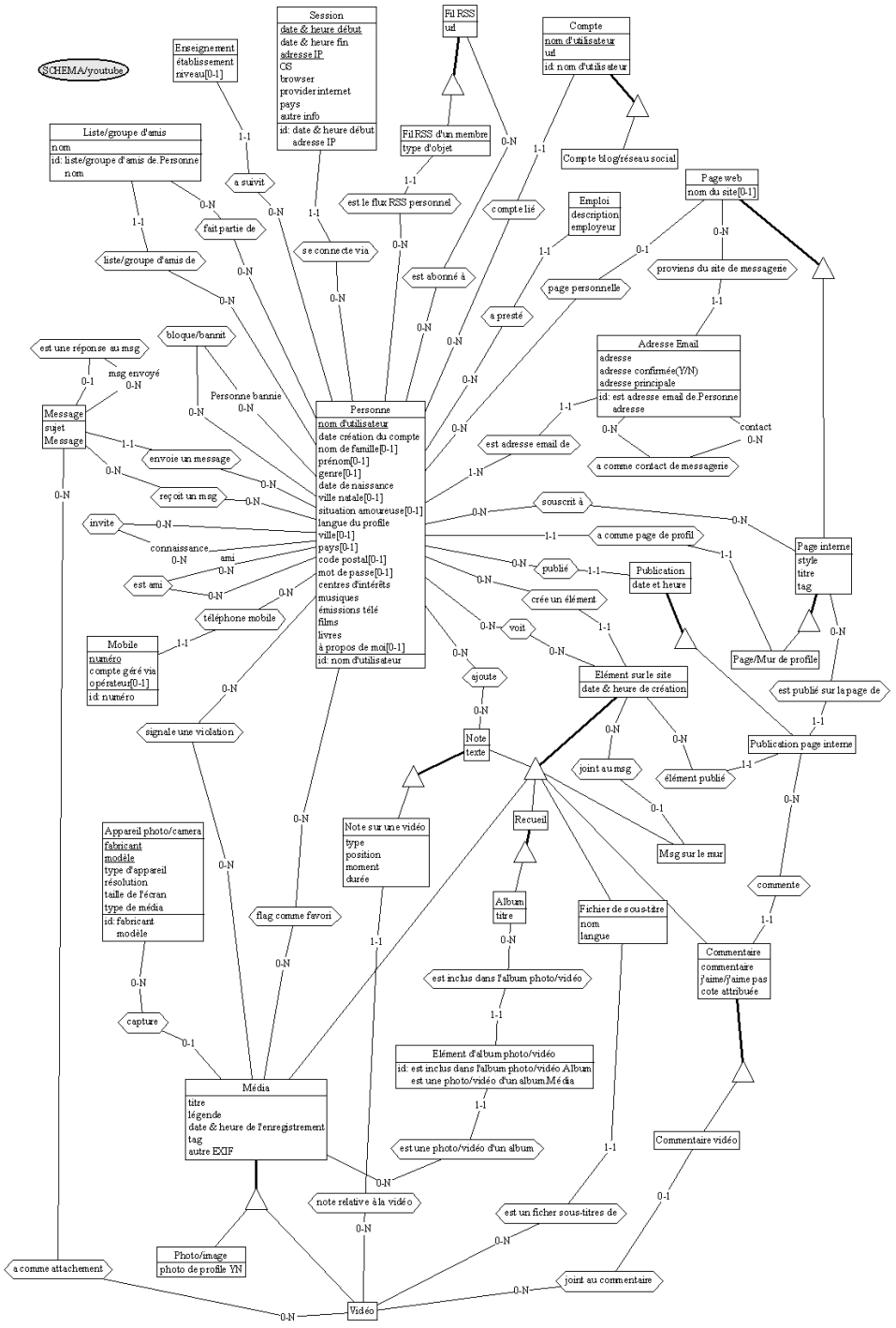

FIGURE 8.5 – Informations collectées par Youtube

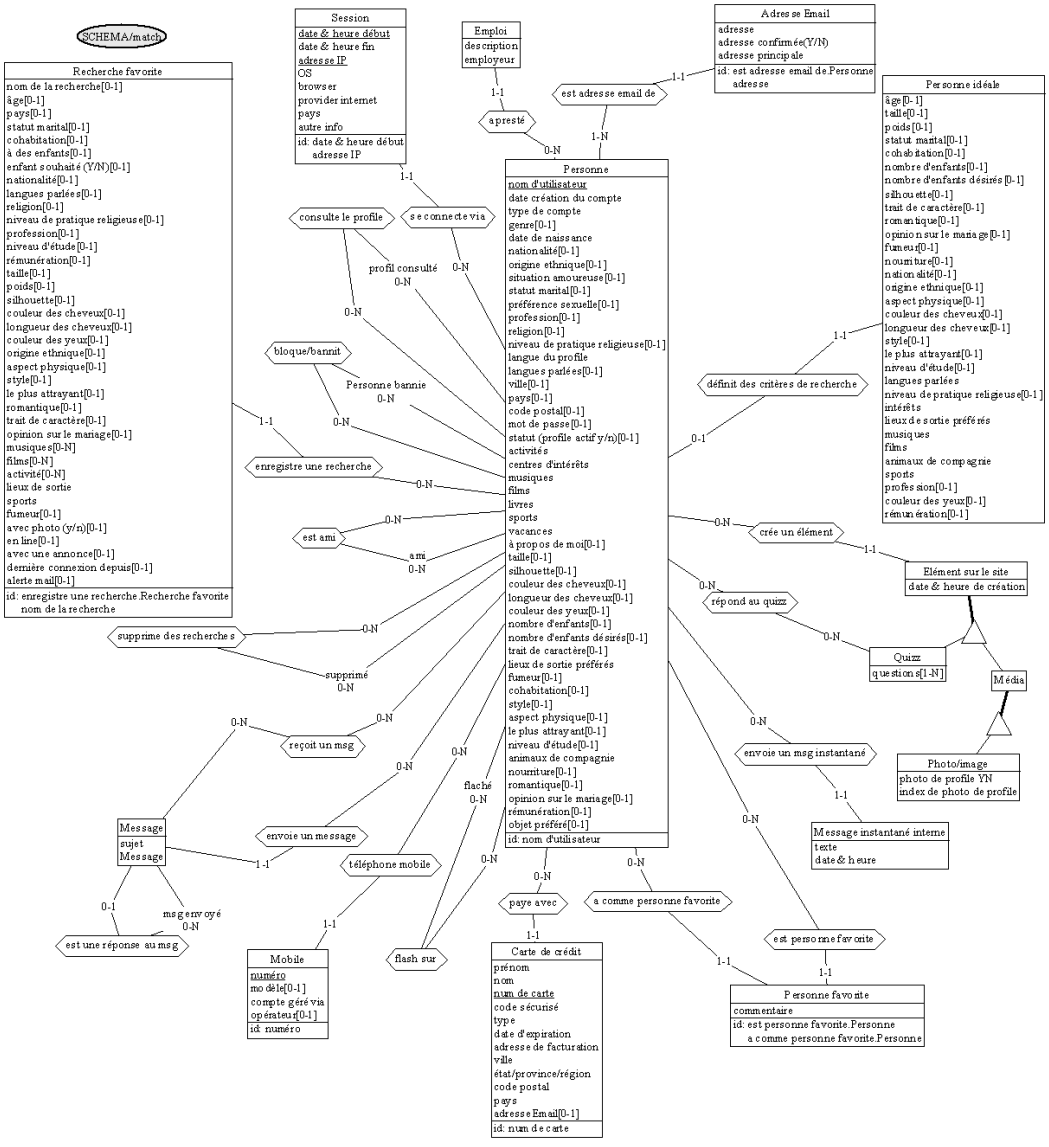

FIGURE 8.6 – Informations collectées par Match.com

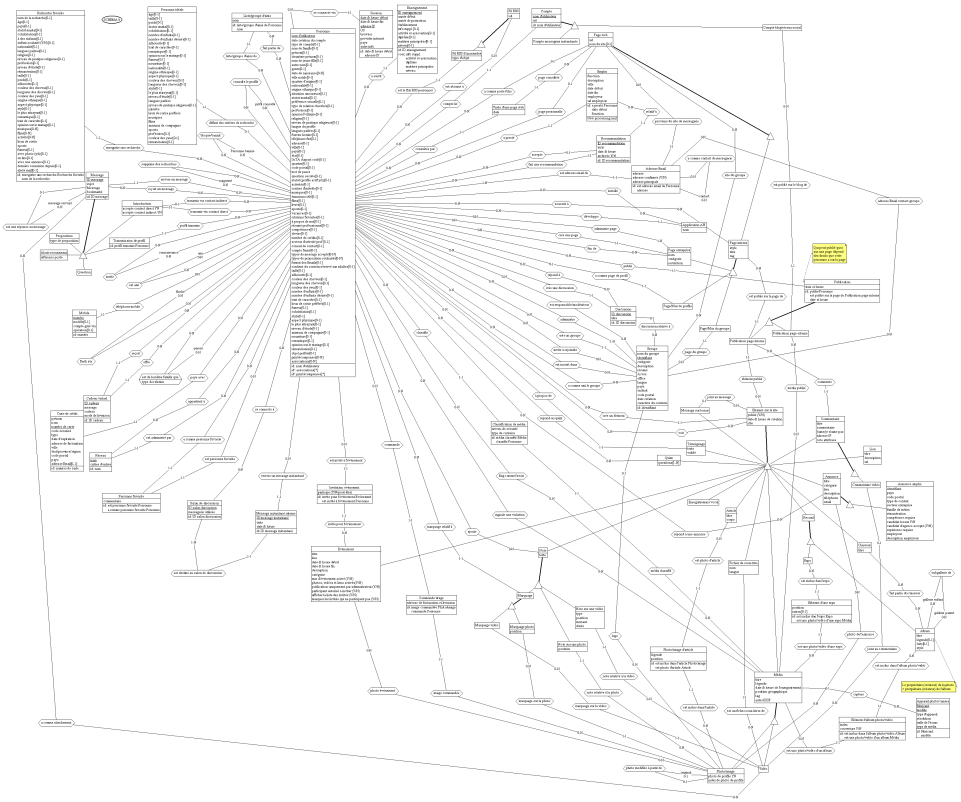

FIGURE 8.7 – Schéma global

8.2 Commentaire sur les informations collectées

Parmi les informations directes tirées d'un profil, seules quelques unes sont obligatoires: le nom ou le pseudo, l'adresse email, parfois la date de naissance, le genre (homme ou femme), le code postal, ou, dans le cas de Flickr, le compte Yahoo associé. Toutes les autres, donc la grosse majorité, sont facultatives.

En utilisant également le plugin de la section 7.2.3, les informations ont été triés en fonction de leurs types (voir figures 8.8, 8.9, 8.10 et 8.11). Le tableau 8.8 et le graphe 8.9 montrent que pour la majorité des sites, la plupart des informations demandent une analyse de contenu afin d'en extraire la connaissance, ce qui n'est pas toujours chose aisée; raison pour laquelle, probablement, Facebook essaye de préciser le contenu d'un document via l'insertion de tags dans différentes pages web (voir section 5.6).

	Facebook		Flickr		LinkedIn		LiveJournal		Match		Youtube		Total	
Depend du contenu	142	57,7%	100	60,6%	98	53,3%	106	56,4%	25	12,8%	73	51,1%	252	49,4%
Interet	111	45,1%	72	43,6%	55	29,9%	66	35,1%	99	50,8%	60	42,0%	234	45,9%
Relationel	77	31,3%	40	24,2%	63	34,2%	45	23,9%	111	56,9%	35	24,5%	195	38,2%
Identification	45	18,3%	43	26,1%	52	28,3%	48	25,5%	29	14,9%	45	31,5%	60	11,8%
Financier	45	18,3%	15	9,1%	37	20,1%	31	16,5%	28	14,4%	21	14,7%	58	11,4%
Geographique	32	13,0%	19	11,5%	25	13,6%	28	14,9%	23	11,8%	20	14,0%	42	8,2%
Contact	30	12,2%	20	12,1%	29	15,8%	27	14,4%	21	10,8%	18	12,6%	36	7,1%
Physique	21	8,5%	19	11,5%	8	4,4%	12	6,4%	22	11,3%	14	9,8%	34	6,7%
Profession/étude	16	6,5%	4	2,4%	26	14,1%	8	4,3%	10	5,1%	11	7,7%	33	6,5%
Religieux	11	4,5%	7	4,2%	6	3,3%	6	3,2%	10	5,1%	8	5,6%	16	3,1%
Sexualité	6	2,4%	4	2,4%	3	1,6%	4	2,1%	9	4,6%	4	2,8%	10	2,0%
Personnalité	2	0,8%	2	1,2%	2	1,1%	3	1,6%	6	3,1%	2	1,4%	7	1,4%
Politique	2	0,8%	1	0,6%	1	0,5%	1	0,5%	1	0,5%	1	0,7%	2	0,4%
Total	246	100%	165	100%	184	100%	188	100%	195	100%	143	100%	510	100%

FIGURE 8.8 – Répartition des éléments collectés par réseaux sociaux, en fonction du type. Le pourcentage représente le nombre d'éléments d'un type donné collectés par un réseau social versus le nombre total d'éléments collectés par le réseau social. Les trois types d'informations les plus collectées dans chaque réseau social ont été mis en évidence.

Tout les sites apportent également une grande importance à connaître les intérêts de leurs membres. Et ce pour deux raisons: premièrement pour leur donner envie de faire partie du site de réseau social en leur offrant des services personnalisés; deuxièmement afin d'analyser ces intérêts pour adapter la publicité de manière individuelle.

Vient ensuite tout ce qui est relationnel. Le relationnel, c'est la base d'un

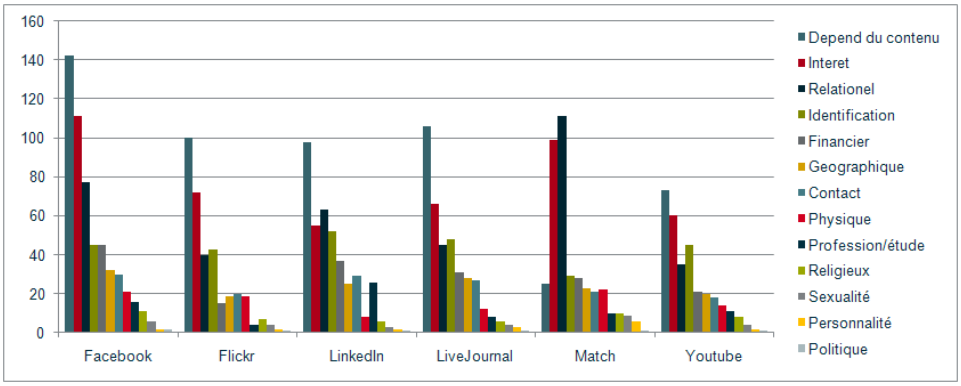

FIGURE 8.9 – Graphe du nombre d'éléments collectés par réseau, en fonction du type

réseau social; mais c'est également une grande source d'informations sur une personne, via l'adage « dis-moi qui sont tes amis et je te dirai qui tu es ». Notons que pour les sites basés principalement sur le contenu, tels Flickr, LiveJournal ou Youtube, le relationnel semble moins important que ce qui est publié.

Enfin, sur les différents sites, soixante informations pouvant intervenir, d'une manière ou d'une autre dans l'identification d'un membre ont été trouvées. C'est nettement moins que pour les types d'informations précédents, mais le sujet est également moins vaste. Parmi ces soixante, une bonne trentaine sont communes à Facebook, LiveJournal, Youtube, Flickr et LinkedIn, d'où un plus faible pourcentage global qu'individuellement dans chaque site dans le tableau 8.8. Dans les données de type identifiant, dont certaines sont obligatoires, nous avons, par exemple, une adresse email, une adresse IP, un compte de messagerie, la date de naissance etc. D'où, même si cela n'est pas impossible, il devient difficile, pour un individu, d'éviter la connexion de deux de ses profils, à condition bien sûr, que les informations fournies soient vraies. Nous approfondir quelques unes des informations identifiantes.

Parmi les informations identifiantes se trouve la photographie de profil. Tous les sites encouragent son utilisation. Un site tel Match.com vérifie même que celle-ci soit de suffisamment bonne qualité avant de l'accepter ou de la rejeter. Cela pourrait se justifier par la recherche d'une plus grande convivialité entre membres. Dans un site de rencontre, cela a même une grande importance. Néanmoins, une photographie identifiable représente une manière quasi certaine d'être identifié. Le membre perd donc son anonymat;

et ce n'est peut-être pas ce qu'il souhaite, justement sur un site comme Match.com. Par contre, lorsqu'on est membre d'un réseau du type de Facebook, avec l'utilisation de sa véritable identité, on pourrait arguer que cela ne fait que « compléter » l'identification. En fait, cela ajoute beaucoup plus que cela. Cela permet d'augmenter de façon significative la connaissance sur un individu (voir section 4.7)

Autre information identifiante, la date de naissance. Tous les sites, excepté Flickr et LinkedIn forcent l'introduction de la date de naissance exacte. Quant à Flickr, il force l'association à un compte Yahoo qui exige, entre autres, la date de naissance. D'un point de vue logiciel, cette donnée ne semble pas nécessaire. Mais d'un point de vue identification de la personne, elle est importante. C'est même une des données les plus discriminantes[22]. De plus, c'est également le début du numéro d'identification au registre national. Il suffit alors de connaître le sexe et l'heure approximative de naissance pour limiter la recherche à quelques nombres [1]

Sur les sites de réseaux sociaux que nous avons analysés, l'adresse email est utilisée comme identifiant à l'intérieur du site. Il ne peut donc y avoir qu'un seul profil à l'intérieur d'un même site ayant une même adresse email. Cette idée n'est pas unique. Par exemple, dans le monde du web sémantique, le projet FOAF [2] définit la propriété mbox [3] (c'est-à-dire l'adresse de messagerie internet) comme « Inverse Functional »(pour chaque mbox définie dans FOAF, il existe un et un seul agent correspondant à cette adresse email). Or, dans la pratique, une personne peut posséder plusieurs adresses email, et une adresse email peut être partagée par plusieurs personnes. Deux profils de réseaux sociaux différents partageant une même adresse pourront être connectés de manière immédiate, mais il n'y aura aucune certitude que ces deux profils correspondent effectivement à la même personne réelle. Néanmoins, vu la facilité avec laquelle il est possible de créer une adresse email, on peut supposer que le partage d'adresse, dans le cadre privé, est de plus en plus rare. De plus, même s'il ne s'agit pas de la même personne réelle, elles ne peuvent être que très proches.

1. Le numéro national est composé de 11 chiffres: les 6 premières positions forment la date de naissance en sens inverse, les 3 positions suivantes constituent le compteur journalier des naissances (ce chiffre est pair pour les femmes et impair pour les hommes), les 2 dernières positions constituent les chiffres de contrôle.
2. `www.foaf-project.org/`
3. `hxmlns.com/foaf/0.1/#term_mbox`

Il en va de même en ce qui concerne une adresse de messagerie instantanée. Il est tellement aisé de créer un compte qu'il semblerait logique que deux personnes ne partagent pas le même identifiant. Cette information peut être retrouvée sur les sites de Facebook, de Flickr, de LinkedIn ou de Livejournal, et peut donc, au même titre que l'adresse email, servir à connecter des profils de manière immédiate.

	Facebook		Flickr		LinkedIn		LiveJournal		Match		Youtube		Total	
Depend du contenu	142	56,3%	100	39,7%	98	38,9%	106	42,1%	25	9,9%	73	29,0%	252	100,0%
Interet	111	47,4%	72	30,8%	55	23,5%	66	28,2%	99	42,3%	60	25,6%	234	100,0%
Relationel	77	39,5%	40	20,5%	63	32,3%	45	23,1%	111	56,9%	35	17,9%	195	100,0%
Identification	45	75,0%	43	71,7%	52	86,7%	48	80,0%	29	48,3%	45	75,0%	60	100,0%
Financier	45	77,6%	15	25,9%	37	63,8%	31	53,4%	28	48,3%	21	36,2%	58	100,0%
Geographique	32	76,2%	19	45,2%	25	59,5%	28	66,7%	23	54,8%	20	47,6%	42	100,0%
Contact	30	83,3%	20	55,6%	29	80,6%	27	75,0%	21	58,3%	18	50,0%	36	100,0%
Physique	21	61,8%	19	55,9%	8	23,5%	12	35,3%	22	64,7%	14	41,2%	34	100,0%
Profession/étude	16	48,5%	4	12,1%	26	78,8%	8	24,2%	10	30,3%	11	33,3%	33	100,0%
Religieux	11	68,8%	7	43,8%	6	37,5%	6	37,5%	10	62,5%	8	50,0%	16	100,0%
Sexualité	6	60,0%	4	40,0%	3	30,0%	4	40,0%	9	90,0%	4	40,0%	10	100,0%
Personnalité	2	28,6%	2	28,6%	2	28,6%	3	42,9%	6	85,7%	2	28,6%	7	100,0%
Politique	2	100,0%	1	50,0%	1	50,0%	1	50,0%	1	50,0%	1	50,0%	2	100,0%
Total	246	48,2%	165	32,4%	184	36,1%	188	36,9%	195	38,2%	143	28,0%	510	100,0%

FIGURE 8.10 – Répartition des éléments collectés par types d'informations, en fonction du réseau social. Le pourcentage représente le nombre d'éléments d'un même type collectés par un réseau social versus le nombre total d'éléments du même type collectés. Le réseau social collectant le plus d'information d'un même type a été mis en évidence.

Comme nous le montre le tableau 8.10 et le graphe 8.11, le site émergeant comme collectant le plus d'informations de toutes sortes est Facebook. C'est un site généraliste, ce qui signifie qu'il est à la fois un lieu de rencontre et une plateforme de diffusion de photographies, de vidéos, etc. Près de 80% des informations collectées par une site tel que LiveJournal ou Youtube se retrouvent également dans Facebook. D'un point de vue commercial, Facebook est donc largement le plus intéressant. C'est celui qui possède le plus grand nombre de membres, ainsi que potentiellement le plus d'informations sur ces adhérents. Cela correspond bien à la stratégie de Facebook qui est de vendre des espaces publicitaires ciblés. En ce qui concerne l'identification de ses membres, elle est complète. En effet, le site exige que le profil corresponde à un individu de la vie réelle. Les informations telles que le nom, le prénom, la date de naissance, ainsi que le genre sont donc obligatoires.

Le site de rencontre Match.com vient après Facebook en ce qui concerne

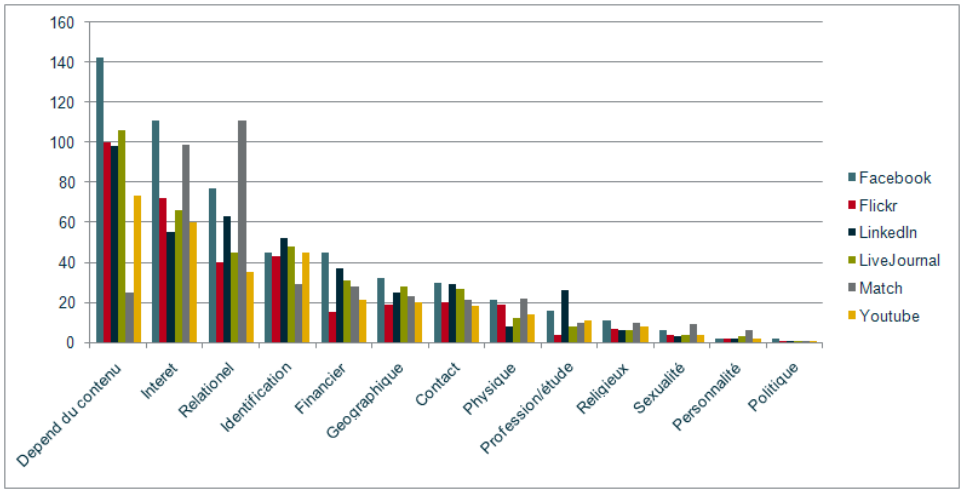

FIGURE 8.11 – Graphe du nombre d'éléments collectés par type, en fonction du réseau

le nombre d'éléments. Il fait donc partie des sites collectant le plus d'informations. Cette collecte peut se justifier par le thème du site qui nécessite de dessiner la personnalité des différents membres. Celle-ci se limite principalement à tout ce qui touche le relationnel, ainsi qu'aux intérêts, mais très peu ce qui dépend du contenu. En effet, ce genre de sites, contrairement aux autres, ne permet pas beaucoup d'activités comme des événements, des publications d'images, des commentaires, etc. En fait, les informations sont relativement différentes. Plus de la moitié des données collectées dans Match ne sont collectées par aucun autre site. De plus, vu le thème du site, Match laisse entendre qu'il veut préserver un certain anonymat à ses membres, d'où, on peut le supposer, moins d'informations d'identification collectées. Or, ce site réclame la date de naissance, le genre, ainsi que le code postal, informations critiques pour l'identification. De plus, les seules photographies de profil acceptées sont celles qui identifient clairement le membre.

La dynamique du site LiveJournal se situe principalement au niveau de la publication de textes et moins au niveau relationnel. Le centre de ce réseau social n'est donc pas le profil ou les amis, mais le contenu publié, ainsi que les activités inhérentes, raison probable pour laquelle il y a moins d'informations directes collectées. D'ailleurs, c'est le site qui a le moins de données d'identifications exigées: la date de naissance ainsi que le genre. Cela ne signifie pas qu'il connait moins ses membres, l'analyse du contenu

pouvant être très riche en informations. De plus, le fait de s'inscrire révèle, pour la majorité des membres, un goût de la lecture et/ou de l'écriture. LiveJournal peut donc offrir un espace publicitaire basé sur le contenu, mais également proposer des produits et services liés au thème du site.

Le centre du site LinkedIn n'est pas le contenu, mais les liens tissés entre des professionnels. Les informations ne relevant pas de la vie professionnelle sont peu ou pas collectées. Il n'y a pas d'activité de publication d'images ou de vidéos : la publication se fait au niveau des offres d'emplois, de partenariats, etc. Le site collecte dés lors à peine moins d'informations dépendant du contenu que LiveJournal. LinkedIn, se voulant une référence dans le domaine, cherche à offrir des services liés au monde des affaires, comme, par exemple, le recrutement. Cela nécessite de stocker toutes informations relatives au parcours professionnel d'un individu, ainsi que son identité; d'où la nécessité de collecter le nom et le prénom. Notons que la date de naissance, qui pourrait identifier complètement un individu, n'est pas requise.

Comme LiveJournal, Flickr est basé essentiellement sur les publications, le profil ne venant qu'au second plan. Une site comme Flickr collecte ainsi moins d'informations directes. Mais, la connaissance pouvant être extrapolées est relativement grande. Par exemple, par le simple fait de s'inscrire, les membres dévoilent déjà leur goût pour la photographie. Par conséquence, Flickr sait qu'il peut proposer des biens ou services sur ce thème. De plus, à travers les images publiées, leur sujet, leur localisation, une description assez riche de la personne peut être dessinée. Flickr a donc la possibilité d'offrir un espace publicitaire assez riche, basé tant sur les personnes que sur le contenu publié. Enfin, un compte Flickr ne peut être créé que s'il est associé à un compte de messagerie Yahoo, qui lui exige nettement plus d'informations directes telles que nom, prénom, date de naissance, genre et code postal. Par conséquence, la personne derrière le profil est parfaitement identifiée.

Le site collectant le moins d'informations est Youtube. En effet, sur ce site, le profil n'a que peu d'importance. C'est une fonctionnalité qui est assez accessoire. Il n'y a d'ailleurs pas beaucoup de possibilité de gérer la privacité de ses données. Les relations sont également accessoires. Tout est basé sur les publications. Le but est d'amener les gens à publier et à regarder un maximum de vidéos, afin de vendre un maximum d'espaces publicitaires, ces derniers étant ajoutés sur les vidéos. Dans le tableau 8.8, le pourcen-

tage d'informations d'identifications semble grand sur Youtube. En fait, le nombre est plus ou moins le même que sur les autres sites, mais vu que le nombre d'éléments collectés est plus faible, les données d'identifications sont proportionnellement plus grandes. Le site impose néanmoins de fournir la date de naissance, le genre, ainsi que le pays. Notons qu'actuellement, il n'est plus possible de créer un compte sur Youtube sans associer un compte Google, qui ne peut être activé qu'en donnant son numéro de téléphone.

8.3 Risque de connexion : illustration de l'outil

Afin d'illustrer l'utilisation de l'ontologie dans l'estimation des possibilités de connexion de profils, nous avons interrogé le modèle sur quelques combinaisons de données.

8.3.1 Combinaison nom, prénom, date de naissance

Le premier set d'informations que nous avons envisagé est le plus fréquemment utilisé dans les bases de données afin d'identifier un individu et se compose du nom, du prénom et de la date de naissance. Notons que ce triplet n'est pas entièrement fiable. En effet, outre les variations d'écriture possibles déjà discutées, le nom peut changer. Par exemple, une femme se marie et décide de porter le nom de son mari. Deux profils d'une même personne pourraient donc afficher des informations différentes, et ainsi ne pas pouvoir être connectés. Mais si deux profils affichent le même nom, prénom, date de naissance, alors on peut dire que ces deux profils appartiennent certainement à la même personne.

Afin de déterminer si il est possible d'utiliser ce triplet dans le but de connecter, par exemple, un profil de Facebook avec un profil de Flickr, nous avons interrogé le modèle via un DL-query recherchant les instances de la classe Personne ayant un nom, un prénom et une date de naissance associés:

```
Facebook and Flickr
and hasPersonne_date_de_naissance some date_de_naissance
and hasPersonne_nom_de_famille some nom_de_famille
and hasPersonne_prenom some prenom
```

Le raisonneur a calculé que la classe owl:Nothing (i.e. le concept vide) était équivalente à la classe exprimée par notre query, ce qui signifie qu'il

est impossible de joindre un profil Facebook avec un profil Flickr basé sur la date de naissance, le nom et le prénom. En effet, la date de naissance n'est pas collectée par Flickr.

Par contre, en réécrivant le query pour tous les autres sites de réseaux sociaux, nous n'avons plus d'incompatibilité; en d'autre terme, le triplet pourrait être utilisé afin de connecter les profils venant de n'importe quel site parmi Facebook, Match, Youtube, Livejournal et LinkedIn.

8.3.2 Combinaison code postal, date de naissance, genre

Aux États-Unis, la majorité de la population peut-être identifiée en utilisant conjointement le code postal, la date de naissance et le genre [9]. Nous avons donc écrit le DL-Query suivant dans le cas Facebook et LinkedIn:

```
Facebook and Linkedin
and hasPersonne_date_de_naissance some date_de_naissance
and hasPersonne_genre some genre
and hasPersonne_code_postal some code_postal
```

Le résultat fourni par le raisonneur indique que cette combinaison ne peut pas servir à connecter deux profils dans les réseaux sociaux LinkedIn et Flickr. En effet, d'une part, LinkedIn ne collecte pas le genre, et d'autre part, Flickr ne collecte pas la date de naissance ni le code postal. Néanmoins, le genre d'une personne peut être déduit à partir du prénom dans la majorité des cas. Quant au code postal, les renseignements relatifs à la ville permettent d'en affiner la connaissance. Notons également que Match, le site récoltant le moins d'information d'identification (voir figure 8.10),collecte ce triplet; il pourrait, par conséquence, avoir suffisamment de données permettant l'identification de certains de ses membres.

8.3.3 Photographie

Comme nous l'avons vu (voir section 4.7), deux photographies d'un même individu peuvent, si la qualité est suffisante, être assez facilement associées. C'est donc un très bon moyen de connecter deux profils. Un individu peut se retrouvé sur une photographie de deux manières différentes: soit via l'image de profil, soit parce qu'il est marqué sur une photographie ou une vidéo. Le query recherchant les individus ayant une photographie de

profil ou marqués sur une photographie dans e les différents sites s'exprime comme suit:

```
Facebook and Flickr and Youtube and Match and Linkedin
and Livejournal and (
cree_un_element some (hasPhoto_image_photo_de_profil_YN
some ( isPhoto_image_photo_de_profil_YNfrom value
photo_de_profil_y))
or
est_marque_sur some Photo_image)
```

Ce query ne renvoie pas d'incompatibilité. En effet, dans tous les sites de réseaux sociaux que nous avons examinés, il y a au moins la possibilité d'associer une photographie à son profil.

8.4 Conséquences de la connexion

Nous avons tenter d'évaluer l'augmentation possible de la connaissance dans le cadre de la connexion de deux profils venant de deux sites de réseaux sociaux différents. À cette effet, nous avons compté le nombre d'éléments collectés par couple de réseaux et par type d'informations (voir 8.12 et 8.13).

En regardant les tableaux, nous pouvons voir que le couple Facebook-Match est le leader dans le nombre de données collectées. Ètant donné qu'individuellement, Facebook est le champion, il est logique qu'il apparaisse dans la liste des couples collectant le plus de renseignements. De plus, le site Match, dont les informations sont relativement différentes, est complémentaire à Facebook. Il est donc plus que normale qu'à eux deux, Facebook et Match arrive en tête. De même, si l'on regarde le graphe 8.14, on se rend compte que les huit couples en tête correspondent à une fusion soit avec Facebook, soit avec Match.

On peut même aller plus loin : au vu du tableau 8.15, que ce soit LinkedIn, Flickr, LiveJournal ou Youtube, c'est l'association avec Match qui augmente le plus la connaissance sur un individu, suivie de près par l'association avec Facebook.

	Depend du contenu	Interet	Relationel	Identification	Financier	Geographique	Contact	Physique	Profession/étude	Religieux	Sexualité	Personnalité	Politique	Total
Facebook - Flickr	176	126	83	50	50	36	34	21	17	11	6	2	2	290
Facebook - Match	145	197	163	49	49	36	31	34	20	15	10	6	2	362
Facebook - LiveJournal	176	124	84	52	47	36	32	22	18	12	7	3	2	293
Facebook - LinkedIn	181	123	103	56	51	34	31	21	28	11	7	2	2	307
Facebook - Youtube	164	122	79	50	45	32	30	21	17	11	6	2	2	273
Flickr - Match	113	161	137	49	40	31	33	33	11	13	9	6	1	315
Flickr - LiveJournal	156	93	65	50	36	33	33	21	9	9	5	3	1	253
Flickr - LinkedIn	165	103	84	54	51	34	34	20	28	8	5	2	1	276
Flickr - Youtube	128	87	54	46	26	24	23	20	12	8	4	2	1	209
Match - LiveJournal	112	158	134	51	40	32	28	24	16	11	9	7	1	313
Match - LinkedIn	108	144	156	56	41	29	30	22	30	12	9	6	1	314
Match - Youtube	83	147	128	49	39	30	28	27	15	13	9	6	1	284
LiveJournal - LinkedIn	162	96	84	56	46	30	31	12	27	8	5	3	1	270
LiveJournal - Youtube	135	85	54	50	36	31	29	15	14	9	5	3	1	229
LinkedIn - Youtube	147	98	73	54	46	30	30	14	28	8	5	2	1	255
Total	252	234	195	60	58	42	36	34	33	16	10	7	2	510

FIGURE 8.12 – Nombre d'éléments collectés en cas de fusion de deux réseaux sociaux, en fonction du type. Les couples de réseaux sociaux collectant le plus d'informations, par type d'informations, ont été mis en évidence.

	Depend du contenu	Interet	Rela tionel	Identificati on	Finan cier	Geogra phique	Contact	Physi que	Profes sion/étude	Religieux	Sexua lité	Person nalité	Politique	Average
Facebook - Flickr	69,84%	53,85%	42,56%	83,33%	86,21%	85,71%	94,44%	61,76%	51,52%	68,75%	60,00%	28,57%	100,00%	56,86%
Facebook - Match	57,54%	84,19%	83,59%	81,67%	84,48%	85,71%	86,11%	100,00%	60,61%	93,75%	100,00%	85,71%	100,00%	70,98%
Facebook - LiveJournal	69,84%	52,99%	43,08%	86,67%	81,03%	85,71%	88,89%	64,71%	54,55%	75,00%	70,00%	42,86%	100,00%	57,45%
Facebook - LinkedIn	71,83%	52,56%	52,82%	93,33%	87,93%	80,95%	86,11%	61,76%	84,85%	68,75%	70,00%	28,57%	100,00%	60,20%
Facebook - Youtube	65,08%	52,14%	40,51%	83,33%	77,59%	76,19%	83,33%	61,76%	51,52%	68,75%	60,00%	28,57%	100,00%	53,53%
Flickr - Match	44,84%	68,80%	70,26%	81,67%	68,97%	73,81%	91,67%	97,06%	33,33%	81,25%	90,00%	85,71%	50,00%	61,76%
Flickr - LiveJournal	61,90%	39,74%	33,33%	83,33%	62,07%	78,57%	91,67%	61,76%	27,27%	56,25%	50,00%	42,86%	50,00%	49,61%
Flickr - LinkedIn	65,48%	44,02%	43,08%	90,00%	87,93%	80,95%	94,44%	58,82%	84,85%	50,00%	50,00%	28,57%	50,00%	54,12%
Flickr - Youtube	50,79%	37,18%	27,69%	76,67%	44,83%	57,14%	63,89%	58,82%	36,36%	50,00%	40,00%	28,57%	50,00%	40,98%
Match - LiveJournal	44,44%	67,52%	68,72%	85,00%	68,97%	76,19%	77,78%	70,59%	48,48%	68,75%	90,00%	100,00%	50,00%	61,37%
Match - LinkedIn	42,86%	61,54%	80,00%	93,33%	70,69%	69,05%	83,33%	64,71%	90,91%	75,00%	90,00%	85,71%	50,00%	61,57%
Match - Youtube	32,94%	62,82%	65,64%	81,67%	67,24%	71,43%	77,78%	79,41%	45,45%	81,25%	90,00%	85,71%	50,00%	55,69%
LiveJournal - LinkedIn	64,29%	41,03%	43,08%	93,33%	79,31%	71,43%	86,11%	35,29%	81,82%	50,00%	50,00%	42,86%	50,00%	52,94%
LiveJournal - Youtube	53,57%	36,32%	27,69%	83,33%	62,07%	73,81%	80,56%	44,12%	42,42%	56,25%	50,00%	42,86%	50,00%	44,90%
LinkedIn - Youtube	58,33%	41,88%	37,44%	90,00%	79,31%	71,43%	83,33%	41,18%	84,85%	50,00%	50,00%	28,57%	50,00%	50,00%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100,00%

FIGURE 8.13 – Répartition des éléments collectés en cas de fusion de deux réseaux sociaux, en fonction du type. Le pourcentage représente le nombre d'éléments collectés versus le nombre total d'éléments collectés par les deux réseaux sociaux fusionnés. Les couples de réseaux sociaux collectant le plus d'informations, par type d'informations, ont été mis en évidence

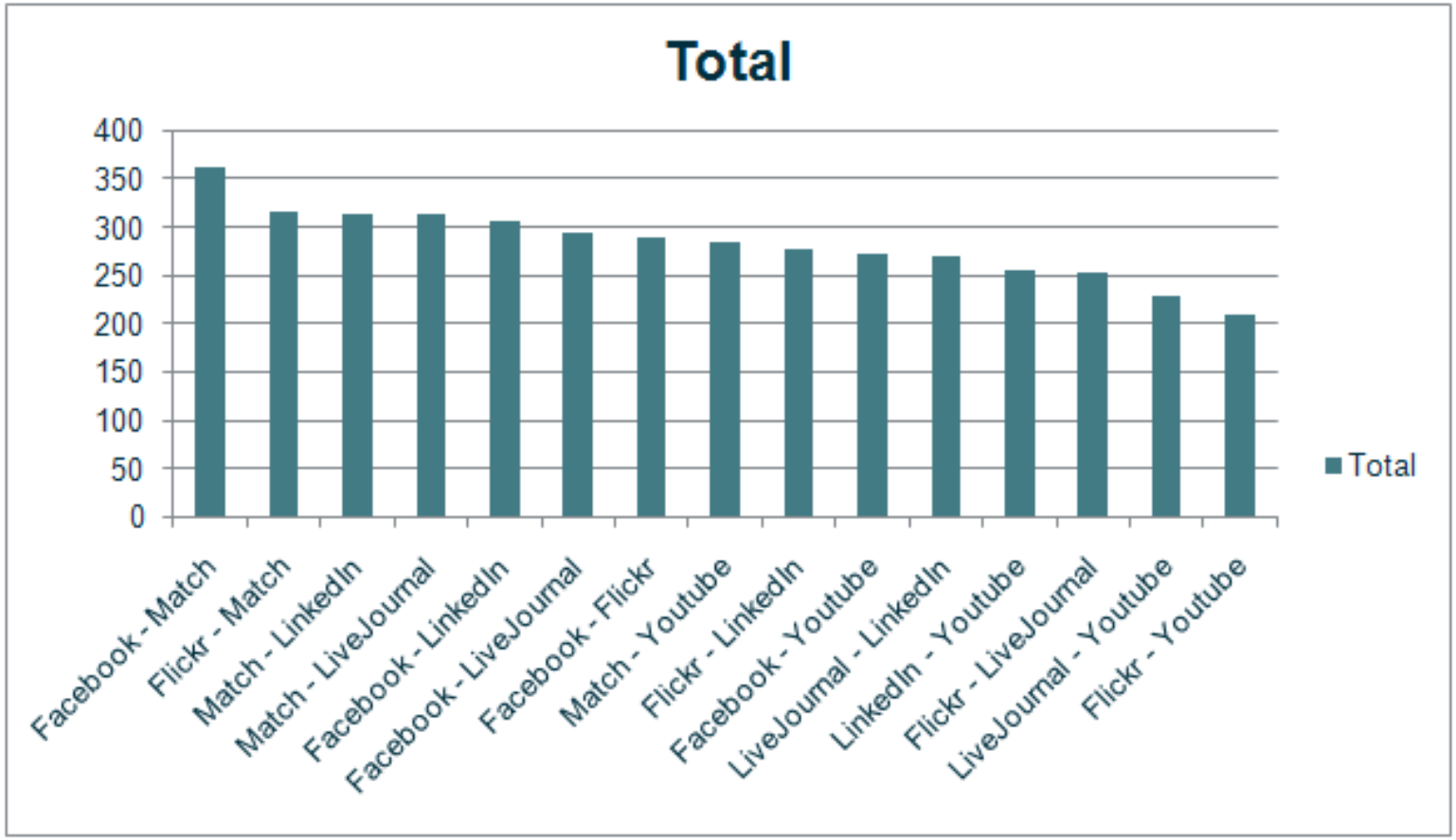

FIGURE 8.14 – Nombre d'éléments collectés par couple de réseaux sociaux

	Depend du contenu	Interet	Rela tionel	Identificati on	Finan cier	Geogra phique	Contact	Physi que	Profes sion/étude	Religieux	Sexua lité	Person nalité	Politique	Total
Facebook + Flickr	13,5%	6,4%	3,1%	8,3%	8,6%	9,5%	11,1%	0,0%	3,0%	0,0%	0,0%	0,0%	0,0%	8,6%
Facebook + Match	1,2%	36,8%	44,1%	6,7%	6,9%	9,5%	2,8%	38,2%	12,1%	25,0%	40,0%	57,1%	0,0%	22,7%
Facebook + LiveJournal	13,5%	5,6%	3,6%	11,7%	3,4%	9,5%	5,6%	2,9%	6,1%	6,3%	10,0%	14,3%	0,0%	9,2%
Facebook + LinkedIn	15,5%	5,1%	13,3%	18,3%	10,3%	4,8%	2,8%	0,0%	36,4%	0,0%	10,0%	0,0%	0,0%	12,0%
Facebook + Youtube	8,7%	4,7%	1,0%	8,3%	0,0%	0,0%	0,0%	0,0%	3,0%	0,0%	0,0%	0,0%	0,0%	5,3%
Flickr + Facebook	30,2%	23,1%	22,1%	11,7%	60,3%	40,5%	38,9%	5,9%	39,4%	25,0%	20,0%	0,0%	50,0%	24,5%
Flickr + Match	5,2%	38,0%	49,7%	10,0%	43,1%	28,6%	36,1%	41,2%	21,2%	37,5%	50,0%	57,1%	0,0%	29,4%
Flickr + LiveJournal	22,2%	9,0%	12,8%	11,7%	36,2%	33,3%	36,1%	5,9%	15,2%	12,5%	10,0%	14,3%	0,0%	17,3%
Flickr + LinkedIn	25,8%	13,2%	22,6%	18,3%	62,1%	35,7%	38,9%	2,9%	72,7%	6,3%	10,0%	0,0%	0,0%	21,8%
Flickr + Youtube	11,1%	6,4%	7,2%	5,0%	19,0%	11,9%	8,3%	2,9%	24,2%	6,3%	0,0%	0,0%	0,0%	8,6%
Match + Facebook	47,6%	41,9%	26,7%	33,3%	36,2%	31,0%	27,8%	35,3%	30,3%	31,3%	10,0%	0,0%	50,0%	32,7%
Match + Flickr	34,9%	26,5%	13,3%	33,3%	20,7%	19,0%	33,3%	32,4%	3,0%	18,8%	0,0%	0,0%	0,0%	23,5%
Match + LiveJournal	34,5%	25,2%	11,8%	36,7%	20,7%	21,4%	19,4%	5,9%	18,2%	6,3%	0,0%	14,3%	0,0%	23,1%
Match + LinkedIn	32,9%	19,2%	23,1%	45,0%	22,4%	14,3%	25,0%	0,0%	60,6%	12,5%	0,0%	0,0%	0,0%	23,3%
Match + Youtube	23,0%	20,5%	8,7%	33,3%	19,0%	16,7%	19,4%	14,7%	15,2%	18,8%	0,0%	0,0%	0,0%	17,5%
LiveJournal + Facebook	27,8%	24,8%	20,0%	6,7%	27,6%	19,0%	13,9%	29,4%	30,3%	37,5%	30,0%	0,0%	50,0%	20,6%
LiveJournal + Flickr	19,8%	11,5%	10,3%	3,3%	8,6%	11,9%	16,7%	26,5%	3,0%	18,8%	10,0%	0,0%	0,0%	12,7%
LiveJournal + Match	2,4%	39,3%	45,6%	5,0%	15,5%	9,5%	2,8%	35,3%	24,2%	31,3%	50,0%	57,1%	0,0%	24,5%
LiveJournal + LinkedIn	22,2%	12,8%	20,0%	13,3%	25,9%	4,8%	11,1%	0,0%	57,6%	12,5%	10,0%	0,0%	0,0%	16,1%
LiveJournal + Youtube	11,5%	8,1%	4,6%	3,3%	8,6%	7,1%	5,6%	8,8%	18,2%	18,8%	10,0%	0,0%	0,0%	8,0%
LinkedIn + Facebook	32,9%	29,1%	20,5%	6,7%	24,1%	21,4%	5,6%	38,2%	6,1%	31,3%	40,0%	0,0%	50,0%	24,1%
LinkedIn + Flickr	26,6%	20,5%	10,8%	3,3%	24,1%	21,4%	13,9%	35,3%	6,1%	12,5%	20,0%	0,0%	0,0%	18,0%
LinkedIn + Match	4,0%	38,0%	47,7%	6,7%	6,9%	9,5%	2,8%	41,2%	12,1%	37,5%	60,0%	57,1%	0,0%	25,5%
LinkedIn + LiveJournal	25,4%	17,5%	10,8%	6,7%	15,5%	11,9%	5,6%	11,8%	3,0%	12,5%	20,0%	14,3%	0,0%	16,9%
LinkedIn + Youtube	19,4%	18,4%	5,1%	3,3%	15,5%	11,9%	2,8%	17,6%	6,1%	12,5%	20,0%	0,0%	0,0%	13,9%
Youtube + Facebook	36,1%	26,5%	22,6%	8,3%	41,4%	28,6%	33,3%	20,6%	18,2%	18,8%	20,0%	0,0%	50,0%	25,5%
Youtube + Flickr	21,8%	11,5%	9,7%	1,7%	8,6%	9,5%	13,9%	17,6%	3,0%	0,0%	0,0%	0,0%	0,0%	12,9%
Youtube + Match	4,0%	37,2%	47,7%	6,7%	31,0%	23,8%	27,8%	38,2%	12,1%	31,3%	50,0%	57,1%	0,0%	27,6%
Youtube + LiveJournal	24,6%	10,7%	9,7%	8,3%	25,9%	26,2%	30,6%	2,9%	9,1%	6,3%	10,0%	14,3%	0,0%	16,9%
Youtube + LinkedIn	29,4%	16,2%	19,5%	15,0%	43,1%	23,8%	33,3%	0,0%	51,5%	0,0%	10,0%	0,0%	0,0%	22,0%

FIGURE 8.15 – Ecart de répartition entre les éléments collectés par un réseau social X avant la fusion (tableau 8.10) et par les informations collectées par le couple de réseaux sociaux $X + Y$ après la fusion (tableau 8.13). Les réseaux fusionnés dont l'augmentation est la plus importante pour un type donné ont été mis en évidence

Chapitre 9

Etude de cas

Ce chapitre sert à illustrer le principe selon lequel les risques de connexion sont en relation avec la rareté démographique d'un set d'informations. Quelques couples de profils venant de réseaux sociaux différents ont donc été examinés. Les profils avec une ou plusieurs photographies identifiables, ou avec une adresse email commune ne sont pas envisagés. En effet, dans ces deux cas, le risque est déjà connu: il est quasi total (voir section 8.2). Afin d'estimer le nombre d'individus pouvant correspondre à un set d'informations spécifiques, des données démographiques réelles ont été utilisées. Quand cela n'a pas été possible, elles ont été devinées en fonction du « bon sens ».

9.1 Cas Facebook/Livejournal

Voici un profil venant de Facebook:

```
Nom et prénom : Julie Dumoulin,
Date naissance : 18 janvier 1995,
Sexe : fille,
Vit en Belgique, dans la province de Namur,
Hétérosexuelle,
Aime : écouter de la musique Rock,
Aime : se maquiller, faire du shopping,
```

Est étudiante au Collège St-Louis de Namur,
Est célibataire,
À un frère: Louis,
Est fan de Grey's anatomy,
Fait partie du groupe : grand feu de Bovesse,
Fait partie du groupe : protection des baleines,
Langue : français

Ainsi qu'un profil venant de LiveJournal:

Pseudo : Juju25,
Date naissance : 18 janvier 1995,
Sexe : fille,
Vit en Belgique, dans la province de Namur,
Est étudiante au Collège St-Louis de Namur,
Centre d'intérêt : Grey's anatomy, Lady gaga,
Langue : français,
Article publié sur le site : contre la graisse des baleines utilisée dans les cosmétiques,
Photographie de l'article : baleine avec un jet d'eau (dessin trouvé sur internet)

A partir des informations explicites des profils, on peut déjà déduire quelques informations implicites. Par exemple:

- le nom de famille spécifié sur Facebook - Dumoulin - laisse entendre que la famille à un ou plusieurs ancêtres originaires d'un pays francophone.
- via Facebook ou Livejournal, on sait que la demoiselle étudie dans une école catholique. Donc, soit ses parents sont catholiques, soit ils pensent que l'éducation de leur fille, dans cette école, est plus importante que la religion enseignée.

Afin d'évaluer le risque de connexion de ces deux profils, nous allons déterminer la rareté des informations communes aux deux profils. Ces dernières sont: fille, née le 18 janvier 1995, étudiante au Collège Saint-Louis de Namur, avec comme centre d'intérêt : la série télévisée « Grey's anatomy », Lady gaga et la protection des baleines.

Les informations - fille et étudiante au Collège Saint-Louis de Namur - peuvent correspondre à près de sept cent individus [1]. Le fait qu'elle ait

1. Informations trouvées sur le site de l'école, http://www.saintlouisnamur.be/fichiershtml/index2.htm

seize ans réduit le nombre à une centaine. Si on ajoute qu'elle est née le 18 janvier, il ne reste plus que quelques individus. L'intérêt pour la série télévisée « Grey's anatomy »et pour Lady Gaga pouvant être courant, ça ne modifie en rien le nombre de possibilités. Par contre, un certain engagement pour les baleines risque de limiter les possibilités à un ou deux individus. Les informations communes aux deux profils constituent un set d'informations très rare car applicable uniquement à un ou deux individus; donc, nous estimons que le risque de connexion est quasi total.

A partir de la fusion des connaissances explicites des deux profils, on peut déduire de nouvelles informations. Par exemple, elle a publié, sur LiveJournal, un article contre l'utilisation de la graisse de baleine dans les cosmétiques; par Facebook, on apprend qu'elle adore se maquiller. Elle prend donc sûrement grand soin de choisir ses produits de maquillage. C'est le genre d'individu qui sera sûrement intéressé par des produits de maquillage respectant les animaux.

9.2 Cas LinkedIn/Match

Voici un profil venant de Match:

```
Pseudo : Charlolte77,
Date naissance : 05 mars 1983,
Sexe : fille,
Code postal : 5080,
Localité : Villers-lez-Heest - Namur,
Statut : jamais marié, pas d'enfant,
Recherche homme entre 26 et 33 ans,
Caractère : insouciant,
Fume régulièrement,
Taille : 1m79,
Sportive ,
Cheveux blonds ,
Yeux noisettes,
Lunettes,
Aime les sorties au cinéma,
Hobbies :internet ,
Activité: jogging, ski, fitness, gym, VTT
```

Ainsi qu'un profil venant de LindkedIn:

```
Nom : Caroline Dupont,
Date naissance : 05 mars 1983,
Sexe : fille,
Code postal : 5080,
Etat civil : célibataire,
Centre d'intérêt : internet, sport ,
Profession : chef de projets junior,
Etude : licence en informatique au FUNDP, diplômée
en 2006
```

A partir des informations explicites des profils, on peut déduire quelques informations implicites. Par exemple:

- la couleur des cheveux et des yeux décrits sur Match - blonde au yeux bleus - laisse supposer que l'individu est de type européen
- le lieu des études ainsi que le code postal spécifié sur Linkedin implique que la personne fréquente la région namuroise depuis plusieurs années.

Les informations communes aux deux profils sont: fille, née le 05 mars 1983, code postal (5080), célibataire et centre d'intérêt: internet, sport. Le code postal, la date de naissance et le genre sont donc connus, d'où, on peut supposer que le risque de connexion est quasi total (voir 8.3.2). Ce risque est encore renforcé par les informations supplémentaires telles que célibataire (seul 44% des femmes sont célibataires à 28 ans [2]) ou les centres d'intérêts.

A partir de la fusion des connaissances explicites des deux profils, on peut tenter diverses déductions. Par exemple, les activités décrites sur Match - fitness, ski, sorties au cinéma - ainsi que le fait d'être une jeune chef de projets célibataire à 28 ans laissent supposer que son niveau de vie doit être relativement élevé. Si on associe son intérêt pour le cinéma, on peut présumer qu'elle sera probablement intéressée, par exemples, par des DVD collector ou éditions spéciales.

9.3 Cas Youtube/Flickr

Voici un profil venant de Youtube:

2. Institut national de statistique, `http://appsso.eurostat.ec.europa.eu/nui/show.do`

```
pseudo : Bibi209
Femme
Date naissance : 5 octobre 1987
Pays : Belgique
```

Ainsi qu'un profil venant de Flickr:

```
pseudo : Bea209
Compte Yahoo : beatricepalmene@yahoo.fr
```

Le profil de Flickr ne contient pas beaucoup d'informations directement exploitables; par contre, le comte Yahoo associé en contient beaucoup et est évidement accessible à Yahoo:

```
Nom : Beatrice Palmene
Femme
Date naissance : 5 octobre 1987
Pays : Belgique
Question secrète n°1 : Quel est le prénom de votre
oncle préféré ? Toni
Question secrète n°2 :Quel est votre livre
préféré ? Dracula
```

Sans le compte Yahoo associé, le risque de connexion est nul. Par contre, en tenant compte du profil Yahoo, les informations communes sont: femme, date de naissance, Belgique. Si le code postal avait été spécifié, le risque aurait été quasi total. Or, seul le pays est fourni. Le risque est donc moindre, mais sans être nul. Selon l'institut national de statistique, en Belgique, il y a approximativement 65.000 femmes de 24 ans. Il y a donc, en moyenne 365 fois moins de femmes nées le 5 octobre 1987, ce qui correspond à approximativement 180 individus. Via l'adresse IP, il est possible de cerner géographiquement le lieu de ou des ordinateurs utilisés lors des différentes connexions. Si l'on suppose que la province a pu être cernée, le nombre de personnes pouvant correspondre à ce profil est de l'ordre de 15-20. La moindre information commune supplémentaire pourrait alors permettre la connexion de manière certaine. Le risque est donc élevé.

Chapitre 10

Conclusions

Le but de ce mémoire est, dans un premier temps, de cerner la connaissance que les réseaux sociaux ont à propos de leurs membres via la création d'une ontologie; pour ensuite, dans un deuxième temps, examiner les risques, ainsi que les conséquences sur cette connaissance en cas de fusion de deux sites de réseaux sociaux.

Le manque d'accès aux serveurs des différents sites, ainsi que le manque de documentation relative à l'ensemble des données collectées imposent une recherche par rétro-ingénierie, c'est-à-dire à partir de la face visible des logiciels supportant les sites de réseaux sociaux. La détermination de la connaissance stockée sur les sites a donc dû se limiter à la partie visible pour les membres, c'est-à-dire, aux informations de profils, ainsi qu'à l'activité au sein même du site. En effet, il est difficilement envisageable de `deviner` quelles données sont collectées en arrière-plan, de manière invisible pour les membres du sites.

Notre ontologie a donc été construite à partir de six sites sélectionnés en fonction de leurs caractéristiques. Rien ne dit qu'un choix de sites différents n'aurait pas donner un autre résultat. Néanmoins, il est probable que les résultats auraient été similaires, car les sites ont été choisis de manière à représenter au maximum les différents types de réseaux.

L'ontologie contient la connaissance explicite, c'est-à-dire ce qui est collecté directement par le site. Elle pourrait être complétée par la connaissance implicite, c'est-à-dire la connaissance pouvant être déduite, via différentes techniques, à partir de la connaissance explicite. Le datamining, c'est à dire

l'extraction de connaissances à partir de données [1], domaine en pleine effervescence, permet régulièrement la création de nouvelles connaissances implicites. Il est donc illusoire de vouloir lister de manière exhaustive toutes les connaissances implicites pouvant être établies.

Une fois l'ontologie élaborée, quelques résultats ont pu être extraits. D'abord, le nombre d'informations collectées est beaucoup plus grand que ce qui avait été envisagé de prime abord. Ensuite, les sites analysés ont la possibilité d'en connaître beaucoup sur l'identité de leurs membres. En effet, même parmi les sites prônant l'utilisation d'un pseudo, la majorité (trois sur quatre) collecte, d'une manière ou d'une autre, des données d'identifications fortement discriminantes, telles que date de naissance ou code postal. Enfin, la plupart des sites ont besoin de connaître leurs membres afin de vendre des espaces publicitaires ciblés. Notons que toutes les publications, telles que photographies, textes, messages, etc. demandent une analyse de contenu afin d'extraire la connaissance, ce qui n'est pas toujours évident, d'où, par exemple, l'Open Graph Protocol de Facebook.

De cette ontologie ont également été extraits quelques résultats relatifs aux conséquences sur la connaissance en cas de fusion de deux réseaux sociaux différents. D'une part, les deux sites collectant le plus d'informations étant Match et Facebook, il est normal qu'ils apparaissent dans la liste des couples de sites de réseaux sociaux collectant le plus d'informations. D'autre part, le site Match récoltant des informations relativement différentes des autres sites, c'est la fusion avec ce dernier qui augmente le plus la connaissance.

Enfin, l'ontologie a été modifiée de manière à devenir un outils permettant d'interroger le modèle sur la possibilité d'user d'un set d'informations spécifique pour connecter deux profils, avec un risque associé à la fréquence d'apparition des éléments communs aux deux profils. Les sites de réseaux sociaux possèdent ces données. Malheureusement, ce n'est pas notre cas. Le rareté d'un set de données à donc du être évalué de manière arbitraire. Il en ressort néanmoins qu'il est difficile de ne pas risquer la fusion au vu des informations exigées sur ces sites.

1. Définition sur Wikipedia, `fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es`

Bibliographie

[1] Foaf vocabulary specification 0.98. http://xmlns.com/foaf/spec/.

[2] Le trésor de la langue française. http://atilf.atilf.fr/tlf.htm.

[3] Protégé wiki: Swrl language faq. http://protege.cim3.net/cgi-bin/wiki.pl?SWRLLanguageFAQ.

[4] Réseautage social , wikipedia. http://fr.wikipedia.org/wiki/Réseautage_social.

[5] W3c : Owl web ontology language, overview. http://www.w3.org/TR/owl-features/.

[6] Boyd, danah m.and Ellison, N. B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication 13*, 1 (2008), 210–230. http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html.

[7] Dagmara, K. Web 2.0: pas à l'abri de la justice. *Le Journal en Ligne de l'Ecole Universitaire de Journalisme de Bruxelles* (May 2011). http://webjournal.ulb.ac.be/index.php?option=com_content&view=article&id=2223:web-20-pas-a-labris-de-la-justice&catid=42:dossier&Itemid=68.

[8] Gómez-Pérez, A., Fernández-López, M., and Corcho-Garcia, O. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[9] Gross, R., Acquisti, A., and Heinz, H. John, I. *Information revelation and privacy in online social networks*. ACM, New York, NY, USA, 2005. http://portal.acm.org/citation.cfm?doid=1102199.1102214.

[10] Hainaut, J.-L., and Brogneaux, A.-F. Cours d'ingénierie des bases de données (cinquième édition). Master à horaire décalé en informatique.

[11] Hendler, J., and Allemang, D. *Semantic Web for the Working Ontologist*. Morgan Kaufmann, 2008.

[12] Lazega, E. Analyse de reseaux et sociologie des organisations. *Revue française de sociologie 35*, 2 (Apr. - Jun 1994), 293.

[13] Mercklé, P. Les réseaux sociaux. les origines de l'analyse des réseaux sociaux. *CNED-ENS/Ish* (2003-2004). http://eco.ens-lsh.fr/sociales/reseaux_merckle_03_origines.pdf.

[14] Motik, B., Cuenca Grau, B., Horrocks, I., and Sattler, U. Representing ontologies using description logics, description graphs, and rules. *Artificial Intelligence 173*, 14 (2009), 1275 – 1309.

[15] Mushtaq, A. Privacy in online social networks. http://www.cse.hut.fi/en/publications/B/1/papers/Mushtaq_final.pdf.

[16] Narayanan, A., and Shmatikov, V. De-anonymizing social networks. *Security and Privacy, IEEE Symposium on 0* (2009), 173–187. http://doi.ieeecomputersociety.org/10.1109/SP.2009.22.

[17] OutspokenMedia.com. Online reputation management guide. http://outspokenmedia.com/downloads/ORM-Guide.pdf.

[18] Parrochia, D. Quelques aspects historiques de la notion de réseau. *Flux*, 62 (Octobre - Décembre 2005), 10–20. http://olegk.free.fr/flux/Flux62/pdffl62/02Parrochia10-20.pdf.

[19] Parsia, B. Understanding swrl (part 1). http://weblog.clarkparsia.com/2007/08/12/understanding-swrl-part-1/.

[20] Parsia, B. Understanding swrl (part 2): Dl safety. http://weblog.clarkparsia.com/2007/08/27/understanding-swrl-part-2-dl-safety/.

[21] Parsia, B. Understanding swrl (part 3): Some tricky bits. http://weblog.clarkparsia.com/2007/09/13/understanding-swrl-part-3-some-tricky-bits/.

[22] Quantin, C., Binquet, C., Bourquard, K., Allaert, F.-A., Gouyon, B., Ferdynus, C., Pattisina, R., Harmenil, G., Pequignot, S., and Gouyon, J.-B. Estimation de la valeur discriminante des traits d'identification utilisés pour le rapprochement des

données d'un patient. *Revue d'Épidémiologie et de Santé Publique 52*, 5 (October 2004), 431–440.

[23] Randall, D., and Richards, V. Facebook can ruin your life. and so can myspace, bebo... *The independent* (Feb 2008). http://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-can-ruin-your-life-and-so-can-myspace-bebo-780521.html.

[24] Schmidt-Schauss, M., and Smolka, G. Attributive concept descriptions with complements. *Artificial Intelligence 48* (1991), 1–26.

[25] Schneier, B. A taxonomy of social networking data. *IEEE Security and Privacy 8*, 4 (July-August 2010), 88.

[26] Sege, I. Where everybody knows your name. *The Boston GLobe* (2005). http://www.boston.com/news/globe/living/articles/2005/04/27/where_everybody_knows_your_name/.

[27] Suply, L. Définition: Réseau social. *Le Figaro* (janvier 2008). http://blog.lefigaro.fr/hightech/2008/01/definition-reseau-social.html.

[28] Zheleva, E., and Getoor, L. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *18th International World Wide Web conference (WWW)* (April 2009). cs.northwestern.edu/~akuzma/classes/CS495-s09/doc/To%20Join%20or%20Not%20to%20Join.pdf.

Table des figures

Annexe A

Plugin : Mise en évidence de l'information, mode d'emploi

Ce plugin offre à l'utilisateur la possibilité d'utiliser soit le marquage, soit les couleurs (voir figures A.1 et A.4), afin de mettre en évidence les informations relatives à un ou plusieurs réseaux, à une ou plusieurs catégories etc.(voir figures A.2 et A.3). Il permet également de lister les informations colorées ou marquées. Ces dernières peuvent être conservées via « un copy & past », par exemple, dans Excel.

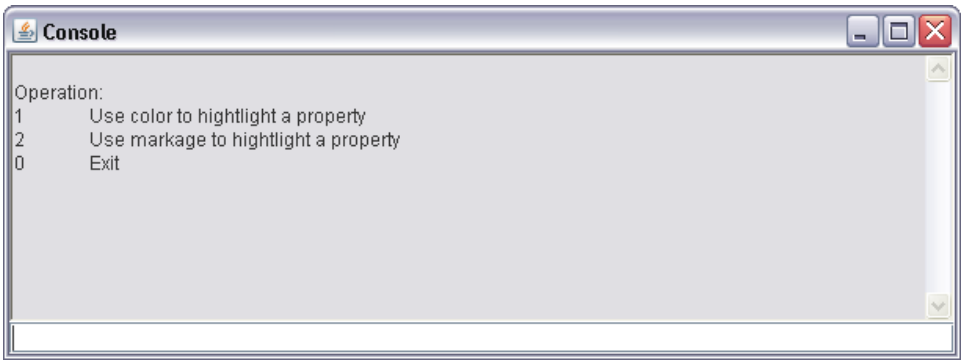

FIGURE A.1 – Le plugin propose soit d'utiliser les couleurs, soit d'utiliser le marquage

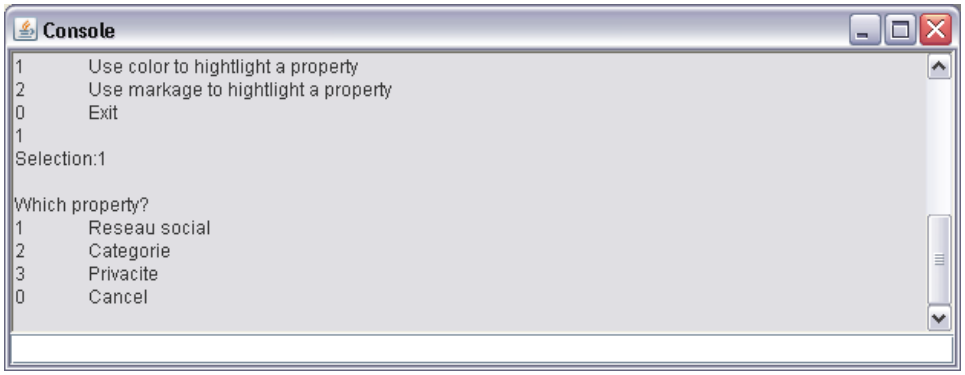

FIGURE A.2 – Le plugin permet de choisir la propriété utilisée pour la mise en évidence

FIGURE A.3 – Le plugin permet de mettre en évidence une ou plusieurs valeurs relatives à une méta-propriété.
La « , » correspond à un « OR »; Le « & » correspond à un « AND »; Le « ! » correspond à un « NOT »

FIGURE A.4 – Le plugin permet de choisir la couleur et/ou le marquage qui sera utilisé

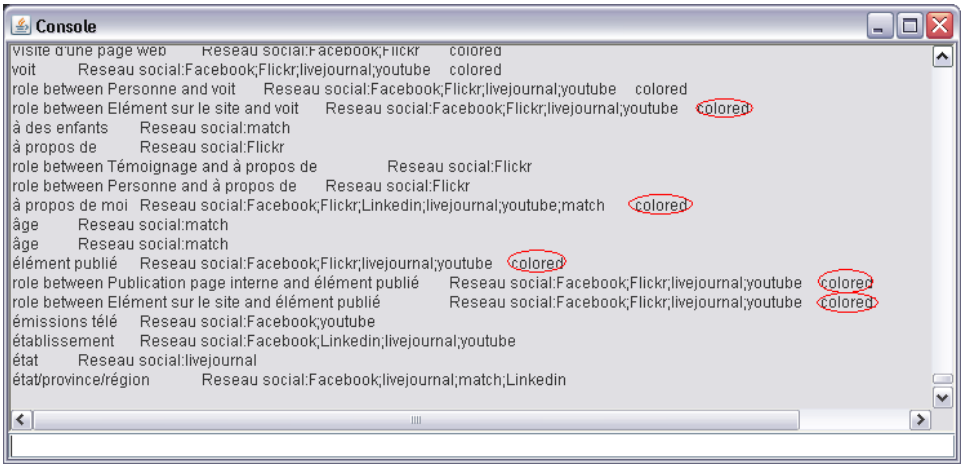

FIGURE A.5 – Le plugin affiche les transformations faites au schéma.

Annexe B

Concepts dégagés par l'ontologie

Voici les principaux concepts dégagés via les recherches dans les différents sites:

Adresse email

Description : Système de messagerie sur internet permettant de recevoir ou d'envoyer du courrier électronique. Une adresse email est confirmée lorsque l'utilisateur a cliqué (ou fait un copier-coller dans un browser) sur un lien reçu via un email à destination de l'adresse. L'adresse principale est l'adresse email qui sera utilisée en premier pour recevoir des emails du site de réseau social. Elle peut également être utilisée comme nom d'utilisateur de la fenêtre de connexion pour certains sites. L'adresse email est donc considérée comme identifiant d'un individu pour ces sites. Une adresse email permet de contacter la personne. Elle permet, dans certains cas de lister les personnes en contact avec l'utilisateur via cette adresse email. De plus, si cette adresse est relative à une société - par exemple Jean.dupont@forstis.be -, elle peut fournir des renseignements sur la société employant l'individu.

Attribut : adresse, adresse confirmée (Y/N), adresse principale.

Domaine des relations : provient du site de messagerie (Page web), a comme contact de messagerie (Adresse email), est adresse email de (Personne).

Range des relations : (Adresse email) a comme contact de messagerie, (Groupe) adresse email contact groupe.

Album

Supertype : Recueil

Description : Un album est un recueil de vidéos et/ou photographies appartenant au « créateur »de l'album. Dans LiveJournal, il est appelé galerie. Un album peut avoir un titre, une légende et un style. Le thème d'un album peut indiquer un intérêt pour un sujet donné. Par exemple, si un album contient des photographies de motos, on peut supposer que l'individu n'est pas indifférent à la moto. Un album peut également correspondre à une date, cette date étant la date à laquelle les photographies/vidéos ont été enregistrées. Si le thème de l'album est: « mes vacances en Grèce », alors on peut déduire qu'à cette date, l'individu était en Grèce. D'autres types d'informations peuvent également être déduites, mais cela dépend du contenu. Par exemple, si le thème de l'album est « dernier meeting du parti socialiste », l'information déduite sera de type politique.

Attribut : titre, légende, date, style.

Domaine des relations : est inclus dans l'album photo/vidéo (Elément d'album photo/vidéo), subgallerie de (Album), fait partie du classeur (Classeur).

Range des relations : (Album) subgallerie de.

Alerte recrutement

Supertype : Message

Description : C'est un message d'un utilisateur de LinkedIn dont le but est d'avertir un autre utilisateur d'une offre d'emploi. Ce message contient donc la référence de l'offre d'emploi dont il est question. Une alerte de recrutement est envoyée en général à une personne connue. Elle donne donc des indications sur les relations entre deux membres du réseau. On peut également supposer que le poste concerné correspondra plus ou moins aux études/compétences du membre. En effet, on enverra probablement pas une alerte exigent un universitaire à un diplômé en hôtellerie, ni une offre pour un cuisinier à quelqu'un n'ayant jamais approché un fourneau.

Attribut : référence poste.

Annonce

Supertype : Elément sur le site

Description : Annonce passée sur le site, dans le but de vendre, d'acheter, de proposer des services etc. aux autres membres du réseau. Le fait de répondre à une annonce permet de déduire que l'objet de l'annonce fait partie des intérêts du membre du réseau. D'autres types d'informations peuvent être déduites, mais cela dépend du contenu. Par exemple, une offre pour des cours d'anglais suppose que l'émetteur de l'annonce à quelques connaissances en anglais, une proposition de covoiturage informe du lieu de résidence et du lieu de travail, etc.

Attribut : titre, catégorie, lieu, description, téléphone, email.

Domaine des relations : photo de l'annonce (Photo/image).

Range des relations : (Personne) répond à une annonce.

Annonce emploi

Supertype : Annonce

Description : C'est une annonce dont le but est de recruter une ou plusieurs personnes pour un emploi spécifié. Les personnes répondant à ces annonces donnent ainsi des informations sur elles-même. On peut, par exemple, supposer que le poste concerné correspondra plus ou moins aux études/compétences du membre. De plus, si des candidats locaux sont exigés, celui qui répondra à cette annonce indiquera sa localité

Attribut : identifiant, pays, code postal, type de contrat, secteur entreprise, famille de métier, rémunération, compétence requise, candidat locaux Y/N, candidat d'agence accepté (Y/N), expérience requise, employeur, description employeur.

Appareil photo/camera

Description : Appareil photo ou caméra pouvant servir à la capture de photographies et/ou de vidéos. L'utilisation d'un appareil professionnel indiquera probablement une certaine compétence dans le domaine. Un appareil d'un bon niveau peut mettre en évidence un goût pour la photographie et/ou la vidéo. Un appareil onéreux peut permettre de dire que le membre a les moyens financiers d'acheter cet appareil, etc.

Attribut : fabricant, modèle, type d'appareil, résolution, taille de l'écran, type de média.

Domaine des relations : capture (Média).

Application API

Description : Programme utilisant l'API fournit par le site de réseau social. Le thème de l'API peut révéler l'intérêt de l'internaute qui l'utilise, mais les différents types d'informations pouvant être déduites dépendent de l'application, ainsi que de son utilisation.

Attribut : nom.

Range des relations : (Personne) développe, (Personne) installe.

Article

Supertype : Elément sur le site

Description : Texte publié par un membre du réseau social, pouvant traiter d'un sujet particulier. Le thème abordé donne des indications sur les intérêts d'une part, de la personne ayant écrit l'article et d'autre part, des personnes le lisant. Mais beaucoup d'autres indications pourraient être déduites. Tout dépend du contenu. Par exemple, un article expliquant les raisons de croire en Allah permet de supposer que l'écrivain est musulman.

Attribut : titre, corps.

Domaine des relations : est photo d'article (Photo/image d'article).

Cadeau virtuel

Description : Envoi payant d'une image accompagné d'un texte d'un membre à un autre membre.

Attribut : ID cadeau, message, cadeau, mode de livraison.

Range des relations : (Personne) offre, (Personne) reçoit.

Carte de crédit

Description : Une carte de crédit virtuelle reflète les informations se trouvant sur une carte de crédit réelle, ainsi que les informations nécessaires à son utilisation.

Livejournal spécifie que les données de la carte ne sont pas conservées. Est-ce que cela inclus toutes les données, ou juste le numéro de carte et le code de vérification? Une carte de crédit révèle des d'informations du type nom, prénom et adresse. Son utilisation peut refléter certaines habitudes de l'individu: achats réguliers de petits montants, achats occasionnels de gros montants, etc.

Attribut : prénom, nom, numéro de carte, code sécurisé, type, date d'expiration, adresse de facturation, ville, état/province/région, code postal, pays, adresse email.

Range des relations : (Personne) paye avec.

Classeur

Supertype : Elément sur le site

Description : Un classeur est un ensemble d'albums. Comme un album, le thème peut indiquer un intérêt pour un sujet donné. D'autres informations peuvent être déduites en fonction du contenu.

Attribut : titre.

Range des relations : (Album) fait partie du classeur.

Classification de média

Description : Utiliser lors de recherche de photographie avec filtre

Attribut : niveau de sécurité, type de contenu.

Range des relations : (Média) média classifié, (Personne) classifie.

Commande tirage

Description : Impression de photographies, calendriers,.... Lors d'une commande, le nom, l'adresse, etc. doit être fourni. Les éléments commandés donnent des indications sur les intérêts de l'utilisateur. Le montant des commandes peut informer sur la santé financière. Par exemple, l'achat de produits onéreux permet de supposer que l'individu en a les moyens.

Attribut : adresse de facturation et livraison.

Range des relations : (Photo/image) image commandée, (Personne) commande.

Commentaire

Supertype : Elément sur le site

Description : Un commentaire est une critique, une explication, une remarque, une réflexion,... relative à un élément publié sur le site. Flickr propose au « propriétaire »de la publication de retenir l'adresse IP des personnes déposant les commentaires. Beaucoup d'informations différentes peuvent être collectées dans les commentaires. Cela peut aller d'une adresse

email, en passant pas les opinions politiques, religieuses, etc. Mais il dénote au moins un intérêt pour l'élément publié

Attribut : titre, commentaire, j'aime/je n'aime pas, adresse IP, cote attribuée.

Domaine des relations : commente (Publication page interne).

Commentaire vidéo

Supertype : Commentaire

Description : Commentaire contenant une vidéo. Les informations pouvant être déduites sont les informations collectées via le commentaire, plus les informations attachées à la vidéo.

Domaine des relations : joint au commentaire (Vidéo).

Compte

Description : Il s'agit d'un compte d'accès à un site internet. C'est le genre d'informations qui permet à coup sûr de faire le lien entre deux profils. De plus, si le compte est relatif à un site de réseau social ayant un thème spécifique, cela suppose un certain intérêt pour ce thème. Connaître le compte sur un autre site permet également de contacter cette personne à l'intérieur du site de réseau social correspondant.

Attribut : nom d'utilisateur, url.

Range des relations : (Personne) compte lié.

Compte blog/réseau social

Supertype : Compte

Description : Compte dont le service correspond à un site de réseau social et/ou de blog.

Domaine des relations : est publié sur le blog de (Publication page externe).

Compte messagerie instantanée

Supertype : Compte

Description : Compte dont le service correspond à un site de messagerie instantanée, tel que Google, MySpace, YAHOO, MyOpenId, verisignPIP, OpenID

Discussion

Description : Question ou débat proposé dont le but est l'échange de point de vue, la discussion, etc. Dans les sites de réseaux sociaux, une discussion est liée à un groupe. Elle est donc souvent en rapport avec le thème du groupe. Comme pour les commentaires, beaucoup d'informations différentes peuvent être collectées. Cela dépend de ce que l'utilisateur inclut dans le texte relatif à la discussion.

Attribut : ID discussion, titre.

Domaine des relations : discussion relative à (Groupe).

Range des relations : (Personne) répond à, (Personne) crée une discussion.

Elément d'album photo/vidéo

Description : Média en tant qu'élément d'un album photo/vidéo, accompagné des informations relatives à sa position dans l'album. Il hérite des informations attachées à l'album et au média correspondant, tel que thème, position géographique, etc.

Attribut : index, couverture Y/N.

Range des relations : (Album) est inclus dans l'album photo/vidéo, (Média) est une photo/vidéo d'un album.

Elément d'une expo

Description : Média en tant qu'élément d'une exposition, accompagné des informations relatives à sa position dans l'exposition, ainsi que des éventuelles raisons de sa sélection pour l'exposition. Cela peut révéler l'intérêt du membre pour le thème du média. D'autres informations peuvent également être collectées via le média. Par exemple, une expo relative à la fête d'anniversaire d'un membre suppose que le propriétaire de l'expo connaît le membre.

Attribut : position, raison.

Range des relations : (Expo) est inclus dan l'expo, (Média) est une photo/vidéo d'une expo.

Elément sur le site

Description : Objet virtuel ajouté sur le site par un utilisateur. Il peut s'agir d'un commentaire, d'un lien, d'une photo,... Les types d'informations pouvant être déduites correspondent aux différentes types d'informations pouvant être collectées via une photographie, une vidéo, un commentaire, etc.

Attribut : publié (Y/N), date & heure de création, rôle.

Range des relations : (Publication page interne) élément publié, (Personne) crée un élément, (Personne) voit, (Message sur le mur) joint au message.

Emploi

Description : Emploi exercé par l'utilisateur. Il peut s'agir d'un emploi actuel ou passé. Cette donnée peut permettre de déduire des informations géographiques (le membre travaillait à cette adresse), financière (cette profession est bien payée), relationnel (il connaît probablement les personnes qui travaillaient pour cette société à la même époque), étude (diplôme/compétence nécessaire), etc.

Attribut : fonction, description, ville, date début, date fin, employeur, url employeur.

Range des relations : (Recommandation) relatif à, (Personne) a presté.

Enregistrement vocal

Supertype : Elément sur le site

Description : Un enregistrement vocal est un élément pouvant être ajouté sur le site par un utilisateur. Il s'agit d'un enregistrement fait via un appel téléphonique à un numéro spécifique. Tout comme les commentaires et les discussions, les informations pouvant être collectées dépendent de ce qui se trouve dans l'enregistrement

Enseignement

Description : Enseignement donné à l'utilisateur du réseau dans une école, pendant un certain laps de temps et ayant un thème. Les types d'informations recueillies peuvent être d'ordre géographique (la personne vivait près de l'établissement), relationnel (il connaît probablement les personnes

ayant suivi les mêmes études, au même endroit et au même moment), financier (certains diplômes, par exemple, permettent l'accès à des professions bien payées.)

Attribut : ID enseignement, année début, année de promotion, établissement, info suppl., activité et association, diplôme, matières principales, niveau.

Range des relations : (Personne) a suivit.

Evénement

Supertype : Elément sur le site

Description : Evénement, manifestation ayant un début et une fin dans le temps. Il peut se dérouler à un endroit précis. Plusieurs participants peuvent prendre part à l'événement. L'événement peut être décrit par des commentaires, des photographies, etc. Une personne participant à un événement se déroulant dans un lieu précis indique sa situation géographique au moment de l'événement. Les événements peuvent parfois donner des indications sur les liens entre les participants (par exemple, un événement de type anniversaire). Mais il peut encore donner plus d'informations en fonction du contenu. Par exemple, une réunion d'anciens de telle école.

Attribut : titre, lieu, date & heure début, date & heure fin, description, catégorie, mur d'événement activé (Y/N), photos, vidéos et liens activés(Y/N), publication uniquement par administrateur (Y/N), participant autorisé à inviter (Y/N), afficher la liste des invités (Y/N), masquer les invités qui ne participent pas (Y/N).

Domaine des relations : invite pour l'événement (Invitation événement).

Range des relations : (Photo/image) photo événement.

Expo

Supertype : Recueil

Description : Une expo est recueil de vidéos et/ou photographies publics appartenant à différents utilisateurs. Des informations peuvent être déduites à partir des éléments de l'expo, ainsi que du titre de l'expo.

Domaine des relations : est inclus dan l'expo (Elément d'une expo).

Fichier de sous-titre

Supertype : Elément sur le site

Description : Elément sur le site correspondant à un fichier de sous-titre lié à une vidéo. Il peut indiquer une connaissance de la langue utilisée dans le fichier.

Attribut : nom, langue.

Domaine des relations : est un ficher sous-titres de (Vidéo).

Fil RSS

Description : Les fils RSS (Really Simple Syndication) sont des liens web qui, lorsqu'ils sont saisis dans un lecteur RSS, informent des actualités relatives à un site. Cela donne bien évidement des informations sur l'intérêt de l'utilisateur abonné à ce flux. D'autres informations supplémentaires peuvent être également déduites, en fonction du contenu du site relatif au flux. Lorsque le flux se rapporte à l'activité d'un individu, des informations du type relationnel ou d'identification peuvent également se rajouter.

Attribut : url.

Range des relations : (Personne) est abonné à.

Fil RSS d'un membre

Supertype : Fil RSS

Description : Fils RSS qui informe de l'actualité relative à un membre du réseau et à un type d'objet (vidéo, photographie, etc.). Un individu abonné à un flux relatif à un membre peut indiquer un intérêt pour le contenu du flux, ou la personne relative au flux. De plus, une personne spécifiant cette adresse de flux comme la sienne permet de lié cette personne au membre relatif à ce flux.

Attribut : type d'objet.

Range des relations : (Personne) est le flux RSS personnel.

Groupe

Description : Ensemble des membres d'un réseau social se regroupant, en général, autour d'un thème. L'accès à un groupe peut être réglementé. La participation au groupe indique généralement un intérêt pour le thème du groupe. Il est parfois possible d'en déduire bien plus. Par exemple, le groupe des anciens de l'université donne des informations relatives aux études des membres du groupe.

Attribut : nom du groupe, identifiant, catégorie, description, résumé, Accès, office, langue, pays, endroit, code postal, date création, caractère du contenu.

Domaine des relations : site du groupe (Page web), adresse email contact groupe (Adresse email), page du groupe (Page/Mur du groupe), logo (Photo/image).

Range des relations : (Personne) invite à rejoindre, (Personne) est responsable/modérateur, (Personne) crée un groupe, (Personne) administre, (Personne) a comme ami le groupe, (Discussion) discussion relative à, (Personne) est inscrit dans.

Introduction

Supertype : Message

Description : Les introductions permettent de contacter un membre en demandant à une relation, qui sert de relais, de présenter l'utilisateur. Les relations peuvent être de niveau deux (une connaissance commune entre l'émetteur et le destinataire de l'introduction) ou de niveau trois (un amis de l'utilisateur à une connaissance commune avec le destinataire). Un relais peut accepter ou refuser de transmettre l'introduction. Cela indique un souhait d'une relation avec un autre membre. L'acceptation de transmettre ou non l'introduction peut également donner des indications sur la force du lien avec celui qui demande une introduction et/ou de celui qui la reçoit.

Attribut : accepte contact direct YN, accepte contact indirect YN.

Domaine des relations : transmis via contact indirect (Personne), transmis via contact direct (Personne).

Invitation événement

Description : Invitation proposant à une personne de participer à un événement. La personne peut indiquer sa participation ou non ou son indécision. Le fait de recevoir une invitation indique déjà un lien relationnel avec l'émetteur de l'invitation. De plus, si un membre reçoit une invitation pour un événement, c'est que quelqu'un suppose que la personne sera intéressée. D'autres informations relatives à l'événement peuvent alors parfois être appliquées au membre ayant reçu l'invitation.

Attribut : participe (Y/N/peut-être).

Range des relations : (Evénement) invite pour l'événement, (Personne) est invité à l'événement.

Lien

Supertype : Elément sur le site

Description : Un lien est une référence vers une page web. Elle s'exprime sous la forme de l'adresse URL de la page considérée. L'utilisateur ayant publié un lien suppose un intérêt pour le thème du site.

Attribut : titre, description, url.

Liste/groupe d'amis

Description : Liste contenant certains amis d'un individu. Permet de classer, grouper ses amis, afin, par exemple, de limiter l'accès à certaines publications. Livejournal, les listes d'amis suivantes sont prédéfinies au départ: Mobile View, Family, School, Online Friends, Local Friends, Work. Ces catégories donnent bien évidement des informations sur la famille, les études, l'emploi,...

Attribut : nom.

Domaine des relations : fait partie de (Personne).

Range des relations : (Personne) liste/groupe d'amis de.

Marquage

Supertype : Note

Description : Un marquage est une annotation d'un média signalisant de la présence d'une personne sur un média. Il y a une certaine probabilité que deux personnes se connaissent si soit elles sont toutes deux marquées, par exemple, sur une même photographie ou soit si une des deux est marquée sur un média appartenant à l'autre. De plus, le marquage sur une photographie peut permettre l'application à la personne de la connaissance déduite à partir du média.

Domaine des relations : marquage relatif à (Personne).

Marquage photo

Supertype : Marquage

Description : Un marquage photo signale la présence d'une personne sur une photographie, ainsi que sa position sur celle-ci.

Attribut : position.

Domaine des relations : marquage sur la photo (Photo/image).

Marquage vidéo

Supertype : Marquage

Description : Un marquage vidéo signale de la présence d'une personne sur une vidéo

Domaine des relations : marquage sur la vidéo (Vidéo).

Message

Description : Un message est un texte envoyé par un membre du réseau à un ou plusieurs membres. Le message sera déposé dans la boite aux lettres virtuelles des destinataires, qui pourront ensuite lire le message lors de leur prochaine connexion au site, la boite aux lettres étant gérée par le site. Remarque: cette option est payante pour Match.com. Un message indique souvent une connaissance entre l'émetteur et le receveur du message. Beaucoup d'autres types d'informations peuvent être déduites en fonction du contenu du message.

Attribut : ID message, sujet, Message, bookmarké.

Domaine des relations : a comme attachement (Vidéo), est une réponse au messageg (Message).

Range des relations : (Personne) reçoit un message, (Message) est une réponse au messageg, (Personne) envoie un message.

Message instantané interne

Description : Message envoyé à un salon de discussion dans le cadre d'une discussion instantanée (chatte), le site de messagerie instantanée étant géré par le site de réseau social. Souvent, les membres connectés au même salon de discussion se connaissent. D'ou possibilité de déduire des informations de type relationnel. Beaucoup d'autres types d'informations peuvent être déduites en fonction des discussions échangées via le salon de discussion

Attribut : ID message instantané, texte, date & heure.

Domaine des relations : est destiné au salon de discussion (Salon de discussion).

Range des relations : (Personne) envoie un message instantané.

Mobile

Description : Le mobile représente le numéro de téléphonie mobile. Un numéro de téléphone peut être lié à un modèle de GSM, un opérateur de téléphonie, etc. Un numéro de téléphone mobile peut également être utilisé afin de gérer son compte sur le réseau social. Le modèle du GSM peut donner des indications à propos du membre: par exemple, un GSM très cher permet parfois de déduire que l'utilisateur en a les moyens.

Attribut : numéro, modèle, compte géré via, opérateur.

Range des relations : (Personne) téléphone mobile.

Message sur le mur

Supertype : Elément sur le site

Description : C'est un élément de type texte qui est publié directement sur la page d'un membre du réseau ou d'un groupe. Il ne peut être publié qu'une seul fois. Il peut être accompagné d'un autre élément type photographie, lien, etc. Ce qui peut en être déduit dépend du contenu du message. Si un membre publie un message sur le mur d'un autre membre, on peut supposer que les deux membres se connaissent.

Domaine des relations : joint au message (Elément sur le site).

Média

Supertype : Elément sur le site

Description : Un média est soit une image, soit une vidéo. Le thème d'un média indique souvent un intérêt pour ce thème, que ce soit lors de la publication ou du visionnage du média. Il donne des indications sur l'aspect physique des personnes apparaissant dans le média.

Attribut : titre, légende, date & heure de l'enregistrement, position géographique, tag, autre EXIF.

Domaine des relations : est une photo/vidéo d'une expo (Elément d'une expo), média classifié (Classification de média), média publié (Publication page externe), est une photo/vidéo d'un album (Elément d'album photo/vidéo).

Range des relations : (Personne) signale une violation, (Personne) flag comme favori, (Appareil photo/camera) capture.

Note

Supertype : Elément sur le site

Description : Information insérée à une photographie ou à une vidéo. Ce peut être le marquage d'une personne ou toutes autres informations relatives au média. Le fait d'insérer une note suppose un intérêt pour le média. Des informations supplémentaires peuvent éventuellement être déduites, en fonction du média et de la note

Attribut : texte.

Range des relations : (Personne) ajoute.

Note sur une photo

Supertype : Note

Description : Note insérée à une photographie.

Attribut : position.

Domaine des relations : note relative à la photo (Photo/image).

Note sur une vidéo

Supertype : Note

Description : Note insérée à une vidéo.

Attribut : type, position, moment, durée.

Domaine des relations : note relative à la vidéo (Vidéo).

Page entreprise

Supertype : Page interne

Description : A l'origine, les pages entreprises de Facebook sont des pages destinées à promouvoir une organisation, une société, une célébrité, etc. Selon Facebook, seuls les représentants officiel de l'organisation, de la société, etc. ont le droit de créer une page entreprise. Dans la pratique, n'importe qui peut créer une page avec n'importe quel thème. Par exemple, il y a plus de 500 pages sur Harry Potter... Être fan d'une de ces pages indique l'intérêt pour le thème de la page. D'autres types d'informations peuvent également être déduites en fonction du contenu de la page.

Attribut : nom, catégorie, restriction.

Range des relations : (Personne) administre page, (Personne) crée une page, (Personne) fan de .

Page interne

Supertype : Page web

Description : Page interne au site de réseau social, maintenue par un ou plusieurs utilisateurs et relatif à un utilisateur, un groupe, ou un thème spécifié par l'utilisateur. Les informations dévoilées dépendent du contenu, mais la page peut servir d'identification lorsqu'elle est considérée comme page personnelle d'un membre du réseau.

Attribut : style, titre, tag.

Range des relations : (Publication page interne) est publié sur la page de, (Personne) souscrit à.

Page web

Description : Une page Web est une ressource du World Wilde Web conçue pour être consultée par des visiteurs à l'aide d'un navigateur Web. Les informations pouvant être collectées dépendent bien évidement du contenu de la page. Si elles correspondent à une page personnelle, elles peuvent être identifiantes.

Attribut : url, nom du site.

Domaine des relations : page consultée (Visite d'une page web).

Range des relations : (Groupe) site du groupe, (Adresse email) proviens du site de messagerie, (Personne) a comme porte folio, (Personne) page personnelle.

Page/Mur de profil

Supertype : Page interne

Description : Page correspondant au mur d'un membre du réseau. Cette page montre les informations du profil selon le paramétrage de la confidentialité. Elle sert également de support aux différentes publications.

Range des relations : (Personne) a comme page de profil.

Page/Mur du groupe

Supertype : Page interne

Description : Page correspondant au mur d'un groupe. Cette page montre les informations relatives au groupe selon le paramétrage de la confidentialité. Elle sert de support aux publications relatives au groupe. Elle permet également de faire se rencontrer les différents membres du groupe

Range des relations : (Groupe) page du groupe.

Personne

Description : Être humain, individu. Sont rassemblées ici toutes les informations relatives à la personne directement. Il y a des informations du type politique, religieux, physique, intérêt, etc.

Attribut : nom d'utilisateur, date création du compte, type de compte, nom de famille, prénom, deuxième prénom, nom de jeune fille, autre nom, genre, date de naissance, ville natale, quartier d'origine, nationalité, origine ethnique, situation amoureuse, statut marital, préférence sexuelle, type de relation cherchée, profession, opinion Politique, religion, niveau de pratique religieuse, langue du profil, langues parlées, fuseau horaire, téléphone fixe, adresse, ville, pays, état, IATA Airport code, quartier, code postal, mot de passe, question secrète, statut (profil actif y/n), activités, centres d'intérêts, musiques, émissions télé, films, livres, sports, vacances, citations favorites, à propos de moi, résumé professionnel, compétences, devise, nombre de crédits, secteur d'activité prof., conseil de contact, compte fermé, types de message accepté, types de proposition souhaitée, format des emails, contient du contenu réservé aux adultes, taille, silhouette, couleur des cheveux, longueur des cheveux, couleur des yeux, nombre d'enfants, nombre d'enfants désirés, trait de caractère, lieux de sortie préférés, fumeur, cohabitation, style, aspect physique, le plus attrayant, niveau d'étude, animaux de compagnie, nourriture, romantique, opinion sur le mariage, rémunération, objet préféré, prix/récompenses, associations.

Domaine des relations : répond à (Discussion), répond au quizz (Quizz), reçoit un message (Message), a comme porte folio (Page web), offre (Cadeau virtuel), invite à rejoindre (Groupe), est le flux RSS personnel (Fil RSS d'un membre), est responsable/modérateur (Groupe), est abonné à (Fil RSS), envoie un message instantané (Message instantané interne), enregistre une recherche (Recherche favorite), définit des critères de recherche (Personne idéale), crée un élément (Elément sur le site), crée un groupe (Groupe), administre (Groupe), supprime des recherches (Personne), est personne favorite (Personne favorite), a comme personne favorite (Personne favorite), flash sur (Personne), classifie (Classification de média), signale une violation (Média), se connecte via (Session), souscrit à (Page interne), a comme ami le groupe (Groupe), bloque/bannit (Personne), se connecte à (Salon de

discussion), répond à une annonce (Annonce), administre page (Page entreprise), fan de (Page entreprise), crée une page (Page entreprise), a comme page de profil (Page/Mur de profil), publié (Publication), crée une discussion (Discussion), profil transmis (Transmission de profil), envoie un message (Message), développe (Application API), installe (Application API), accepte (Recommandation), fait une recommandation (Recommandation), consulte le profil (Personne), ajoute (Note), flag comme favori (Média), voit (Elément sur le site), page personnelle (Page web), est inscrit dans (Groupe), commande (Commande tirage), est de la même famille que (Personne), consultée par (Visite d'une page web), est ami (Personne), invite (Personne), liste/groupe d'amis de (Liste/groupe d'amis), reçoit (Cadeau virtuel), est invité à l'événement (Invitation événement), téléphone mobile (Mobile), paye avec (Carte de crédit), appartient à (Réseau), compte lié (Compte), a presté (Emploi), a suivit (Enseignement).

Range des relations : (Marquage) marquage relatif à, (Introduction) transmis via contact indirect, (Personne) supprime des recherches, (Personne) flash sur, (Personne) bloque/bannit, (Témoignage) à propos de , (Adresse email) est adresse email de, (Introduction) transmis via contact direct, (Personne) consulte le profil, (Personne) est de la même famille que, (Personne) est ami, (Personne) invite, (Liste/groupe d'amis) fait partie de, (Réseau) est administré par.

Personne favorite

Description : Personne que l'on souhaite indiquer comme favorite

Attribut : commentaire.

Range des relations : (Personne) est personne favorite, (Personne) a comme personne favorite.

Personne idéale

Description : Critère de recherche afin de trouver la personne idéale.

Attribut : âge, taille, poids, statut marital, cohabitation, nombre d'enfants, nombre d'enfants désirés, silhouette, trait de caractère, romantique, opinion sur le mariage, fumeur, nourriture, nationalité, origine ethnique, aspect physique, couleur des cheveux, longueur des cheveux, style, le plus attrayant, niveau d'étude, langues parlées, niveau de pratique religieuse,

intérêts, lieux de sortie préférés, musiques, films, animaux de compagnie, sports, profession, couleur des yeux, rémunération.

Range des relations : (Personne) définit des critères de recherche.

Photo/image

Supertype : Média

Description : Document de type image. Le thème de cette image peut dévoiler un intérêt pour ce thème. Les différents types d'informations pouvant être déduites dépendent du contenu. Lorsque que la photographie montre des individus, elle apporte des informations précises quand à l'aspect physique des personnes se trouvant sur cette photographie

Attribut : photo de profil YN, index de photo de profil.

Domaine des relations : photo modifiée à partir de (Photo/image), photo événement (Evénement), image commandée (Commande tirage), est inclus dans l'article (Photo/image d'article).

Range des relations : (Marquage photo) marquage sur la photo, (Note sur une photo) note relative à la photo, (Annonce) photo de l'annonce, (Groupe) logo, (Photo/image) photo modifiée à partir de.

Photo/image d'article

Description : Photographie ou image en tant qu'élément d'un article, accompagné des informations relatives à sa position, sa légende, etc. Souvent, elle apporte un complément d'information, mais cela dépend du contenu de la photographie.

Attribut : légende, position.

Range des relations : (Photo/image) est inclus dans l'article, (Article) est photo d'article.

Proposition

Supertype : Message

Description : Message dont le but est de proposer des opportunités de carrière, des demandes d'expertise, des offres de missions de conseil, des opportunités d'affaires, de nouveaux projets, etc. Envoyé une proposition indique une relation entre l'envoyeur et le receveur du message. En fonction du contenu, d'autres informations peuvent éventuellement être déduites.

Attribut : type de proposition.

Publication

Description : C'est l'action de publier un document sur une page web. Cela dénote un intérêt pour l'élément publié. D'autres types d'informations peuvent éventuellement être collectées, mais cela dépend du contenu de l'élément publié et/ou du site sur lequel il est publié.

Attribut : date et heure .

Range des relations : (Personne) publié.

Publication page externe

Supertype : Publication

Description : Publication d'un document sur une page externe au site de réseau social

Range des relations : (Compte blog/réseau social) est publié sur le blog de, (Média) média publié.

Publication page interne

Supertype : Publication

Description : Publication d'un document sur une page interne au site de réseau social, c'est-à-dire une page de groupe, d'entreprise ou de profil d'un membre

Domaine des relations : élément publié (Elément sur le site), est publié sur la page de (Page interne).

Range des relations : (Commentaire) commente.

Question

Supertype : Message

Description : Message dont le but est de poser une question.

Quizz

Supertype : Elément sur le site

Description : Elément contenant une liste de questions auxquelles toute personne visitant la page ou cette liste peut répondre. Des informations sur l'individu pourraient être extraites en fonction des questions posées et des réponses données.

Attribut : questions.

Range des relations : (Personne) répond au quizz.

Recherche favorite

Description : Enregistrement de critère de recherche de personnes. Cela permet la réutilisation lorsqu'un individu veut trouver une liste de membres correspondant à ses critères favoris

Attribut : nom de la recherche, âge, pays, statut marital, cohabitation, à des enfants, enfant souhaité (Y/N), nationalité, langues parlées, religion, niveau de pratique religieuse, profession, niveau d'étude, rémunération, taille, poids, silhouette, couleur des cheveux, longueur des cheveux, couleur des yeux, origine ethnique, aspect physique, style, le plus attrayant, romantique, trait de caractère, opinion sur le mariage, musiques, films, activité, lieux de sortie, sports, fumeur, avec photo (y/n), en ligne, avec une annonce, dernière connexion depuis, alerte mail.

Range des relations : (Personne) enregistre une recherche.

Recommandation

Description : Recommandation relative à un emploi qu'un membre du réseau fait à propos d'un autre membre du réseau. Faire une recommandation à propos d'un membre indique que celui qui écrit la recommandation connaît la personne visée par la recommandation. On peut supposer que les informations contenues dans le texte sont d'ordre professionnels, mais peut-être est-il possible dans déduire d'autres en fonction du contenu du message.

Attribut : ID recommandation, texte, date & heure, archivée Y/N.

Domaine des relations : relatif à (Emploi).

Range des relations : (Personne) accepte, (Personne) fait une recommandation.

Recueil

Supertype : Elément sur le site

Description : Un recueil est un ensemble de vidéos et/ou photographies. Cette classe a été créée afin de faire le lien entre une expo et un album

Réseau

Description : Groupe géré par Facebook qui relie les membres d'une même école ou d'une même entreprise. L'appartenance au réseau se fait

sur demande de l'utilisateur et en fonction du respect d'un ou plusieurs critères (par exemple, pour le réseau des FUNDP, avoir une adresse email du format xxx@fundp). Le type d'informations pouvant être collectées dépend du réseau.

Attribut : nom, critère d'entrée.

Domaine des relations : est administré par (Personne).

Range des relations : (Personne) appartient à.

Salon de discussion

Description : Lieu de rencontre entre les différents intervenants d'une discussion via une messagerie instantanée, celle-ci étant gérée via le site de réseau social. Les personnes d'un même salon de discussion se connaissent souvent. Quand aux autres types d'informations, elles dépendent du contenu de la discussion.

Attribut : ID salon discussion, messagerie utilisée.

Range des relations : (Message instantané interne) est destiné au salon de discussion, (Personne) se connecte à.

Session

Description : Interaction entre l'utilisateur et le serveur du réseau social. La session commence lorsque l'utilisateur se connecte au réseau social en entrant son nom d'utilisateur et son mot de passe. Différentes informations telles que adresse IP, fournisseur d'accès internet, etc. sont associées à la session et peuvent être collectées par le réseau social.

Attribut : date & heure début, date & heure fin, adresse IP, OS, browser, provider internet, pays, autre info.

Range des relations : (Personne) se connecte via.

Transmission de profil

Supertype : Message

Description : Message dont le but est la transmission du profil d'un membre du réseau à un autre membre. Cela indique l'existence d'une relation entre les trois intervenants.

Range des relations : (Personne) profil transmis.

Témoignage

Supertype : Elément sur le site

Description : Texte écrit par un membre du réseau et relatif à un autre membre. Il peut être validé ou nom par le destinataire du témoignage. Le témoignage est écrit par une relation du membre et les informations pouvant être collectées dépendent du contenu du texte.

Attribut : texte, validé.

Domaine des relations : à propos de (Personne).

Vidéo

Supertype : Média

Description : Document de type vidéo. Les différents types d'informations pouvant être déduites dépendent du contenu de la vidéo. Lorsque que la vidéo montre des individus, elle apporte des informations précises quand à l'aspect physique des personnes se trouvant dans cette vidéo.

Range des relations : (Marquage vidéo) marquage sur la vidéo, (Note sur une vidéo) note relative à la vidéo, (Fichier de sous-titre) est un ficher sous-titres de, (Commentaire vidéo) joint au commentaire, (Message) a comme attachement.

Visite d'une page web

Description : Correspond à la visite d'une page web par un membre du réseau. Différentes informations peuvent être stockées telles que la date et l'heure de visite de cette page. Les pages visitées permettent, entre autre, de donner des informations sur les intérêts du visiteur.

Attribut : date.

Range des relations : (Page web) page consultée, (Personne) consultée par.