Community based annotation for the Semantic Web

Matthew Rowe
Supervisor: Prof. Fabio Ciravegna

Web Intelligence Technologies Lab
Natural Language Processing Group
Department of Computer Science
University of Sheffield

m.rowe@dcs.shef.ac.uk
Abstract

The current web trend of online communities consisting of users with diverse interests and social contexts known as social networks has become a phenomenon. The increase in the use of in-browser tools for annotating and bookmarking web resources has allowed more web resources to be shared amongst users of the web, providing useful links for research groups and those with similar interests. Attempting to combine the two areas would allow a greater generation of metadata for the semantic web.

This report presents research carried out in areas heavily influenced by providing annotations to a community of users by its members. The literature presented provides an insight into what current online communities exist and how they can be adapted, what current state of the art annotation tools are available and how they manage the knowledge they generate. The current Web 2.0 trend and what it entails, and what intelligent web systems currently exist, and how they can be used. How the Social Web has evolved and the known issues that are associated with this evolution,

The report describes the current work done to date and what technologies have been used, an overview of possible future research areas and possible domain applications, and finally, the conclusion of the research that has been conducted and a feasible analysis of work that I plan to do towards my thesis.
Table of Contents

Table of Figures ........................................................................................................... v

1. Introduction ............................................................................................................. 1

2. Literature Survey .................................................................................................... 3
   2.1 The Semantic Web ............................................................................................ 3
       2.1.1 Semantic Blogging .................................................................................... 4
       2.1.2 Semantic Portals ...................................................................................... 5
       2.1.3 Annotation Tools ...................................................................................... 6
   2.2 Web 2.0 ............................................................................................................. 12
       2.2.1 Folksonomies ............................................................................................ 13
       2.2.2 Identity 2.0 .............................................................................................. 15
   2.3 The Social Web .................................................................................................. 17
       2.3.1 Social Networks ....................................................................................... 17
       2.3.2 The Friend of a Friend Project (FOAF) ..................................................... 20
       2.3.3 Social Networking Sites .......................................................................... 22
       2.3.4 Online Privacy ......................................................................................... 25
       2.3.5 Lateral Surveillance .................................................................................. 26
       2.3.6 Social Recommendation Systems .............................................................. 27
       2.3.7 Social Bookmarking Tools ........................................................................ 29
       2.3.8 Social Annotations .................................................................................... 32

3. Progress to Date ..................................................................................................... 34
   3.1 Requirements Definition .................................................................................... 34
       3.1.1 The Semantic Web .................................................................................... 34
       3.1.2 Web 2.0 ..................................................................................................... 34
       3.1.3 Social Web ............................................................................................... 34
       3.1.4 Work to Date ............................................................................................. 35
   3.2 Meerkat .............................................................................................................. 36
       3.2.1 Toolbar ...................................................................................................... 37
       3.2.2 Annotating Web Content ......................................................................... 37
       3.2.3 Annotation Controller .............................................................................. 38
       3.2.4 Suggestive Tagging ................................................................................... 38
   3.3 Meervisage ......................................................................................................... 40
   3.4 Facebook Friend Plotter ..................................................................................... 41
   3.5 FOAF Generators .............................................................................................. 42
       3.5.1 Facebook FOAF Generator ...................................................................... 42
       3.5.2 MySpace FOAF Generator ..................................................................... 44

4. Proposed Research .................................................................................................. 46
   4.1 Social Networks as Annotation Communities .................................................... 46
       4.1.1 Inter-Annotator Agreement ........................................................................ 46
       4.1.2 Tag Based Analysis ................................................................................... 47
       4.1.3 User Analysis ............................................................................................. 47
   4.2 Identity Extraction ............................................................................................... 47
       4.2.1 Focussed Crawling ................................................................................... 49
       4.2.2 Indexing a Subset of the Web ................................................................... 50
4.3 Aggregation of Web 2.0 Information..................................................50
4.4 Semantic Trails.................................................................50

5. Conclusion .................................................................................52

References ....................................................................................54

Appendices .....................................................................................58
Appendix A: Meervisage Conference Paper ..................................58
Appendix B: Facebook FOAF Generator Output .........................73
Table of Figures

Figure 1 - Piggy Bank’s Architecture ................................................................. 8
Figure 2 - Tag Cloud taken from Meerkat............................................................ 13
Figure 3 - Relation Identification Method from [Mori et al, 2006]........................... 19
Figure 4 - Facebook's Photo Tagger .................................................................... 32
Figure 5 - Flickr's Photo Note Annotator .............................................................. 33
Figure 6 - Screenshot of the Meerkat Server....................................................... 36
Figure 7 – The Meerkat Toolbar ......................................................................... 37
Figure 8 - Highlighting content for annotation .................................................... 37
Figure 9 - Generating metadata for the annotation .............................................. 38
Figure 10 - Editing metadata .............................................................................. 39
Figure 11 - Context Menu within Meerkat .......................................................... 39
Figure 12 – Newspaper snippet of the Facebook Friend Plotter.......................... 41
Figure 13 - Sample output from Facebook FOAF Generator ............................... 44
Figure 14 – Identity Extraction Architecture ....................................................... 48
Figure 15 – Gantt chart of planned work ............................................................. 52
1. Introduction

The semantic web is an extension of the World Wide Web, a web that exists within a web, where both man and machine can understand the same information. Developers and web site creators encode information that is normally written in natural language as machine-readable data, allowing software agents and computer programs to process this information. By allowing information to be read by software agents, the original information is enriched with metadata; data about data. This metadata is described using the resource description framework, where concepts and the relationships between them can be specified.

The web 2.0 phenomena that began 2 years ago has steadily increased the creation of web applications offering a rich user experience to the user. Many ideas and theories have been implemented and have lead to great success in their respective domains. Youtube, for example, was the big success story of 2006 by allowing users to upload and watch one another’s videos using state of the art streaming software, youtube was sold for $1.6 billion to Google\(^1\). Youtube is also a good example of the community aspect of web 2.0 applications where users are able to comment on one another’s videos, review videos and post messages to users they have added as friends. By allowing users to add content to a site and offering a service in return, web 2.0 sites have grown exponentially both in size and popularity. Wikipedia, the online encyclopaedia has more than 60 million hits per day\(^2\), and contains information contributed completely free of charge by volunteers.

Following on from the web 2.0 bubble, the user has been the main focus of new sites and services, with community based platforms thriving from this user-centric viewpoint. The Social Web has become one of the most active areas of the World Wide Web in recent years, and has spawned research harnessing communities of users for analysis and reuse. Communities of users have chosen to share information using social bookmarking tools, and tag resources using folksonomies; labels that are domain independent free-text user interpretations.

Annotating the web is not a new feature. The term annotating simply refers to adding information to already existing information that is relevant to the current information being referenced; therefore enriching information with metadata. Many Social Web and Web 2.0 sites enrich information with metadata such as tags, descriptions and user assigned notes. Such mechanisms could be reused for Semantic metadata creation.

This report presents the work that I have carried out in the first year of studying for my PhD. I have structured this report to give a detailed account of the literature that I have studied, the work that I have done to date, what work I am planning on doing towards my thesis, and what I have concluded following my first year. I will now briefly explain the contents of this report in greater detail.

The goal of my work is to research and investigate topics associated with the generation of semantic metadata and the enrichment of web resources with semantic

\(^1\) http://en.wikipedia.org/wiki/Youtube
metadata through the use of a community of users. The following section documents the literature that I have read throughout the first year of research covering three areas of work: The Semantic Web, Web 2.0 and the Social Web. I have documented the current state of the art in each section, and the various issues associated with work performed within each area, scientifically and ethically.

Chapter three details the progress I have made so far. I begin by outlining the relevance of the reading I have done to my studies and explaining the relevance of the work I have produced both through publications I have written and the concepts I have attempted to prove. This section then describes in more detail the experimental work that I have done to date, explaining the reasons behind producing work.

Chapter four details what work I am planning on doing in the future, and documents the ideas that I have come up with that would constitute work towards by thesis. The proposals set out in this section are fairly diverse as a result of the influence from the wide spectrum of literature that has been covered. The final chapter describes my conclusions following the completion of one years research within this field. I analyse the work that I have done, and what further work I have carried out alongside my PhD research.
2. Literature Survey

This chapter describes what literature I have read to date throughout the first year of my PhD. This section is broken down into three sections: The Semantic Web, Web 2.0 and the Social Web.

2.1 The Semantic Web

A semantic web would be a data centric web where information could be automatically extracted by software agents by reading metadata attached to this information. The metadata would be rich with semantics capable of describing exactly what information was being extracted. The Semantic web can be regarded as a web of data allowing software agents to exchange data, this is in contrast to the original World Wide Web which was simply a web of documents connected by hyperlinks.

In 2001, the founder of the World Wide Web and Semantic Web visionary Tim Berners-Lee proposed the web of data capable of offering a powerful medium for the exchange of knowledge. He described how software agents would be able to automatically book doctors appointments by accessing data from several online resources and intelligently aggregating the data to perform the necessary task. There have been many critics keen to point out the shortfalls of such a scenario both at low-level functionality and high-level ethics with regards to privacy of information. However, there is evidence to suggest that the growth in popularity in Social Web and Web 2.0 technologies has given the semantic web access to the information it needs to perform such tasks as mentioned in the previous scenario. This transfer report presents the state of the art in this area.

When considering the extraction of information from the web, it is important to note the processes that were carried out primarily to prepare the extracted information prior to extraction. To place this in the context of the above scenario, the software agent must know when it has found the piece of information that it requires, or more simply, the information to be extracted must be annotated in order for it to be recognised as being the intended data for extraction. An ontology containing concepts and the relationships between those concepts is used to annotate information located on the web. The generated metadata will then contain the semantic information that can be recognised by software agents, inferences can then be made about the information based on the ontology used for the annotations.

This section contains examples of several areas of work within the semantic web that are linked to the work I have carried out. Semantic blogging explains the process of enriching blogs with metadata to aid the discovery of similar blogs and facilitate the extraction of information. Semantic Portals present an approach to knowledge management within a web portal, where semantically rich information can be aggregated. The next section describes both automatic and manual annotation tools used to generate the semantic metadata which is paramount for the generation of a semantic web.

---

3 http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21
4 http://www.shirky.com/writings semantic_syllogism.html
2.1.1 Semantic Blogging

Blogging has become one of the most popular Internet pastimes of recent years. With many blogging sites to choose from (blogger, blogspot, live journal) Internet users are publishing more blog posts than ever with great social implications [Cayzer, 2004]. This has supported the web 2.0 philosophy of placing power back into the hands of the user by allowing them to publish regularly and have their blogs indexed and viewed using tool such as Technorati\(^5\). Some blogging services allow other data to be published such as videos and pictures, not restricting users simply to text. Blogs are easy to manage and give users the chance to arrange blog posts as they wish and edit them at a later time. Due to the creation of RSS (really simple syndication), users can subscribe to blogs written by bloggers and keep up to date with recent blog posts. By reading blog posts by other bloggers, users are able to comment about a specific blog post by another blogger in their own blog where the blog is referred to as a link, therefore creating a blogosphere [Moller & Decker, 2005].

The term Semantic blogging relates to adding metadata to blog posts, as with the semantic web where information is enriched with metadata. Blog posts can therefore also include other forms of information that can be downloaded and shared, such as contact information in the form of a FOAF description about the author [Moller et al, 2005], and images with annotations attached to them. In order to provide a semantic blogging service, several characteristics must be adhered to; a reference mechanism must be chosen that allows simple access to objects (ie. DOM structure for elements), a metadata representation format must be selected, and usability must also be considered as the service should function like any other blogging service. It is also suggested that the generation of metadata should be an automatic process that is done as the user posts the blog.

A blog post can be thought of not simply as a journal post or an item of interest but also as a viable means of providing metadata [Karger & Quan, 2004]. Blogs can be used as annotations; commenting on existing web sites and therefore annotating a specific URL with metadata. Blogs can be message chains; commenting on another users blog with a reference to a specific blog post could cause the owner of the blog to reply with a link to the other users blog, therefore creating a message chain. Blogs can also be used as a table of contents; providing links to several other blogs that are related to a particular domain.

**semiBlog**

Written for Mac OSX, semiBlog is a standalone desktop application designed to make semantic blogging simple and effective [Moller et al, 2005]. The system uses application wrappers to access information in several other Mac OS X applications so that it can be added to the blog post, for example; accessing the address book for people’s details, or a web browser to wrap the content of web pages. The core of semiBlog then creates entities from the wrapped content, and generates and associates the metadata that has been generated when writing the blog post. SemiBlog also allows the metadata to be exported in several formats to allow versatile uploading to external blogs as required.

\(^5\) http://www.technorati.com
2.1.2 Semantic Portals

The management of metadata generated by users interactions with annotation tools is evident in a number of web based systems (flickr, del.icio.us). By allowing users to edit what they have produced at a later date, users are able to gain a better experience from the use of the system. Semantic portals offer a similar experience by incorporating ontologies to retrieve information and guide the presentation and interaction with metadata [Halaschek-Wiener et al, 2006]. As metadata is collected in a semantic portal, it can be audited, edited and deleted as required by the user.

Provenance information can also be edited through the use of a semantic portal. The examples of web sites offering management of metadata largely concern the management of provenance information; such as the username, comments, and timestamp of the metadata. A semantic portal would also offer this functionality and allow users to edit their provenance information at a later stage. One of the key features of a semantic portal is the ability to allow third party software applications to interface with the portal through the use of an API or web services. Metadata can be retrieved by accessing the semantic portal and using the portal as a repository to retrieve the required information. The majority of successful and innovative web 2.0 sites (last.fm, facebook) offer third party software the available functionality to access information within their system, therefore also making them semantic portals.

a-PIE [Maneewatthana et al, 2005] is an example of a community portal, created to support a community of interest. Although, the same as a semantic portal in definition, a community portal makes use of more refined ontologies that are specialised for that particular community. Users are able to browse the information within the portal, which can be tailored to their needs, as a semantic portal would. When using a-PIE, each user of the system is able to store information they have found in their own personal repository, where it can be then used, as they require.

In both of the portals that have been demonstrated key functionality is evident: The manipulation of metadata by the user is important to define each system as a portal, allowing a user to interact with the portal to add, edit and delete metadata offers a more interactive experience to the user with control over the metadata they have generated. The ability to access data within a portal using an API allows third party software applications to interface with the portals and retrieve useful data.
2.1.3 Annotation Tools

The open source and third party development nature of modern browsers such as Mozilla Firefox\(^6\) has led to an increase in the number of extensions and plug-ins available. Several institutions and companies have created in browser systems capable of annotating existing web documents with varying degrees of success. This section presents an overview of the most popular and comparable semantic annotation tools available for download and use, both from research communities and the commercial development sector.

**Automatic Annotation Tools**

**S-CREAM**

[Handscho, 2002] presents a S-CREAM, a trainable framework for automatic generation of metadata. S-CREAM includes an implementation of CREAM; the annotating and authoring framework for metadata creation known as Ont-O Mat. S-CREAM also includes Amilcare for adaptive information extraction from text. S-CREAM runs in two modes; training where the system adapts to a new subset of data, and extraction; for annotation of text. For the training phase, a training corpus is supplied that has been manually annotated for learning either by hand or using Ont-O Mat. Using the training corpus, S-CREAM induces rules capable of extracting the annotated information.

The usage scenario presented in [Handscho, 2002] demonstrates how information within a web page can be automatically annotated with semantic information from an ontology. As mentioned previously, an annotated corpus must be supplied as training data to aid the semi-automatic annotation process. Users of S-CREAM are also able to verify annotations that are suggested, therefore allowing rules to be re-induced by the system. S-CREAM presents an automatic approach for annotating information by combining two different approaches to include information extraction with the current rich annotation features of CREAM.

**KIM**

Although many annotation tools rely on the user to annotate web pages manually, KIM uses named entity recognition to automatically parse a web page and identify all the entities in the page [Kiryakov et al, 2004]. The named entities that have been recognised are then annotated and assigned hyperlinks that reference named entities that KIM already has corresponding references to, therefore creating semantic metadata automatically. As KIM has previously found documents that are relevant to a given entity, documents can be indexed and retrieved using that entity. By semantically linking between a new identified entity and corresponding documents associated with that entity, KIM is able to provide indexing and retrieval. However, should KIM find an entity that it does not have a reference for then it generates a new URI for the entity along with a description and stores the information.

\(^6\) http://www.mozilla.com
In order to carry out the named entity recognition, KIM employs an ontology where named entities are matched against concepts in the ontology. The ontology used by KIM is a lightweight ontology that offers simple definitions of entity classes and allows efficient management of knowledge by reducing the complexity of relations and the number of axioms. In turn the ontology is linked to KIM’s knowledge base containing references to other documents that contain that entity and descriptions of the entity. The ontology is used as a schema for the knowledge base and it’s reasoning.

Knowledge is represented as RDF, allowing simple access using APIs and web services. The knowledge base can store two different types of knowledge; pre-populated and automatically extracted. The former is knowledge that has been acquired from a trusted source, and the latter is knowledge that has been discovered when carrying out semantic annotation. KIM comes bundled as a plug-in for Microsoft Internet Explorer; it is light weight and stable although it does slow down the loading of web pages once they have been browsed to as it is parsing the content.

**Magpie**

Similar to KIM, Magpie automatically annotates entities found in a web page. It defines this process as associating a Semantic layer with a web resource to allow simple access to any background knowledge that might be relevant [Domingue et al, 2004]. The function of Magpie is similar to KIM, upon browsing to a new web page, Magpie uses a populated ontology to identify named entities within the web page, once identified the entities are annotated and added to the semantic log knowledge base comprising of context entities that have been found on the given page. Unlike KIM, Magpie does not annotate each entity to be hyperlinked to an item in the knowledge base. Instead, each item is assigned a context menu, which is triggered by right clicking the identified entity. The context menu then offers services depending on the class of the entity within the current ontology being used.

Several automatic annotation tools offer a mono-sequential approach to annotating web resources; as an entity is found in a page and annotated, the scope of the system limits functions that are available to the user so that when they click on the annotated component very little happens (hyperlinked to a web page, popup menu). In contrast Magpie offers a multi-dimensional approach, when an annotation is generated the service uses an ontology to trigger other services depending on the concept that has been used for annotations. These services differ depending on the concept from the ontology that has been used for the annotation.

Magpie uses a browser-based approach to annotating web resources. As each page is loaded the Magpie Microsoft Internet Explorer plug-in parses the HTML in the web page using predefined parsing rules, the parsed content is then processed by the lexicon on the Magpie server to find any possible annotations. The lexicon simply looks for any entities that correspond to those in the ontology. Found entities are annotated accordingly and the parsed HTML is returned to the plug-in where it is displayed to the user within the browser window, and the relevant services are loaded into the context menu [Dzbor et al, 2004].
Piggy Bank

Piggy Bank is another example of an automatic annotation tool able to retrieve information represented as RDF from several web sites [Huynh et al., 2005]. Should the user of the tool come across a web page containing an information item that is useful, the item can be highlighted and saved to the piggybank server. Piggy bank is bundled with several screen scrapers, JavaScript scripts that can be run on certain web sites to generate the RDF automatically. If a screen scraper can be used to generate ‘purer’ RDF then the user is notified of this and the option to use the corresponding screen scraper can be selected.

As extracted RDF is saved to a central semantic bank, this extracted content can be shared with other users, either those users within a group or globally among all users. Unlike other annotation tools, Piggy Bank does not annotate content within the browser window; instead it allows access to the semantic bank where the RDF is stored through a web interface. Within the semantic bank, tags can be assigned to the saved information items, and the information items can be saved to the user’s local semantic bank for later use. An important feature of Piggy Bank is the use of tags as

Figure 1 - Piggy Bank’s Architecture

As extracted RDF is saved to a central semantic bank, this extracted content can be shared with other users, either those users within a group or globally among all users. Unlike other annotation tools, Piggy Bank does not annotate content within the browser window; instead it allows access to the semantic bank where the RDF is stored through a web interface. Within the semantic bank, tags can be assigned to the saved information items, and the information items can be saved to the user’s local semantic bank for later use. An important feature of Piggy Bank is the use of tags as

URIs rather than simply as folksonomies. By treating each tag as a URI, tags with the same name can be differentiated from one another by the relations and properties they have.

**Potluck**

Following on from work presented in [Huynh et al, 2005], work by the same group presented in [Huynh et al, 2007] takes inspiration from the Web 2.0 phenomena of ‘mash-ups’ to allow users to aggregate semantic information using an advanced web interface front end to handle the data. [Huynh et al, 2007] presents a system that uses two RDF data sources to retrieve semantic information from. Users are then able to merge information from these sources to generate integrated data that can be output to a web browser for users to view. Facets can also be generated, which present subsets of information derived from the semantic relationships that the user has defined when using potluck.

Pot Luck is a system capable of automatically integrating semantic information to produce merged information that can be explored according to predefined semantic facets. In order to generate this data, the supplied resources must be able to have semantic data extracted from them automatically. The same annotation process is used by piggy bank, RDF is extracted from web resources that have been supplied to Potluck and the metadata is then extracted automatically.
Manual Annotation Tools

Annotea
Annotea is a manual annotation tool and framework that allows users the choice to annotate and store their annotations locally or on an already existing Annotea server, where local annotations are private to the user and remote annotations stored on the server are shared with other users. Multiple annotation servers can be deployed to allow maximum access to already existing annotations, and to store annotations in more than one place.

Annotea [Kahan et al, 2002] proposed that there were two approaches to web annotation systems:

- Proxy based - a proxy server is used to store and merge annotations with a web document.
- Browser based - a browser plug-in is used to merge the annotations from a store or server with a web document.

Annotea adopted the latter of the two approaches and relies on a simple architecture that consisted of a browser plug-in communicating with a server. When saving annotations the browser plug-in would post RDF data to the annotation server hosting an RDF store, and then retrieve RDF data from the server when previous annotations were to be retrieved.

Developments of the basic Annotea architecture have lead to an evolution in the definition of exactly what defines an annotation. The notion of Annotea objects [Koivunen, 2005] suggests that Annotea objects are not restricted to annotations (equivalent of notes in the margin of a textbook), they can also be replies, bookmarks, and topics; all of which have a URI, and contain some metadata. Annotea allows collaborative groups to be created to allow users to work together and annotate specific resources with annotations that only members of that group are able to view. Annotea topics can be created that, to begin with, are vague and not well refined, covering possible other topics. However, as more annotations are made, the user of the system is able to fine-tune the topics until they can be linked to a specific ontology concept.

Diigo
With the development of the web, the increase of online communities and the exchange of information (see web 2.0 section) both for collaboration and sharing, the web user has been encouraged to become more involved. Diigo8 is an example of this evolution of the average web user by offering a lightweight Mozilla Firefox extension that allows users to annotate socially. Users are able to browse to any web page, highlight a portion of text they wish to annotate, and then select from Diigo’s context menu to save the annotation. Annotations can have descriptions and tags assigned to them to make referencing easier.

8 http://www.diigo.com
One of the key features that differentiate Diigo from similar annotation tools is the user interface when annotating. As users are able to form and join groups using Diigo, members of that group can view what annotations other group users have made. When browsing a web page and finding an annotation that another group user has made, Diigo’s advanced user interface offers the functionality to add further information to the annotation, comment on the annotation, or edit the annotation, if the user has sufficient privileges.

Blue Organizer

By automatically recognising items within given web resources, Blue Organizer can save useful pieces of information and annotate web content. When a web site is browsed to, predefined JavaScript screen scrapers are able to extract the content from the page and save it in a local repository. Blue organizer comes bundled with several screen scrapers that can wrap content from many commercial sites selling products such as DVDs and video games. If the user browses to a site that Blue Organizer does not have a screen scraper for, the content can be annotated manually and saved in a similar fashion to Diigo’s annotation mechanism by highlighting the content and choosing to save the annotation from the context menu. Blue organizer defines the saved annotation items as ‘BlueMarks’, it also offers users additional functionality such as Blue Badges; widgets that can be placed in any web page which when accessed display the given user’s previous BlueMarks, therefore sharing the annotated items with other users.

---

9 http://www.adaptiveblue.com
2.2 Web 2.0

As the web has developed since the dot com bubble burst, web users have been more interested in using web sites that offer good services and a rich and interesting user experience. Over the past 3 years the term ‘web 2.0’ has become synonymous with many web sites, but defining exactly what it is has lead to disagreements between various parties. There are several characteristics that web 2.0 sites exhibit; they treat the web as a platform [O’Reilly, 2005] offering web services that can be used to access data they have stored through an API\(^{10}\). Users are encouraged to become trusted co-developers by adding content to sites\(^{11}\). Users can become part of a community, which in turn is a social network allowing users to communicate with one another. Users are able to ‘tag’ information for other users to easily find using folksonomies. Web logging or ‘blogging’ as it has come to be known is offered by several sites to allow users to post blogs regularly and easily.

With regards to developments in technology, web 2.0 is not a technology but a set of principles. It has brought about the increased use of certain technologies such as asynchronous JavaScript and xml (Ajax) and similar remote procedure call technologies. By using remote procedure calls developers have been able to build web applications that offer a rich user experience that function more like a desktop application by loading components within a web page rather than the whole web page. The increase of web services offered by sites has also led to a lot of web sites publishing RSS (Really Simple Syndication) feeds [Millard & Ross, 2006], offering users the chance to subscribe to feeds from various web sites in order to keep up to date with their news or products by re-downloading the RSS feed when an alteration has been made.

---

\(^{10}\) http://www.flickr.com

\(^{11}\) http://www.wikipedia.org
2.2.1 Folksonomies

Otherwise known as ‘tags’, folksonomies are essentially keywords that are used by users to annotate a variety of media formats for referencing either by themselves or other users. Rather than using a taxonomy of predefined keywords that a user can select, folksonomies are keywords that a user can define as they wish [Guy & Tonkin, 2006]. This led to an increase in the use of folksonomies throughout a wide range of web sites. An interesting feature of folksonomies is the ability to generate a tag cloud; an area, normally a box, containing the most popular tags used on the site hosting the cloud. The tag cloud displays the tags in varying sizes, where the larger the tag the more popular it’s use has been (figure 2).

![Tag Cloud](http://www.dcs.shef.ac.uk/~mrowe/meerkat.html)

Figure 2 - Tag Cloud taken from Meerkat

Although easier to use for most people and offering an open service to users to define tags how they wish, folksonomies suffer from various problems. ‘Sloppy tags’ are bad tags that are spelt wrongly, contain compound word groupings, do not follow conventions in use (use of plural when singular is correct form), and single use tags that are specific to the information in question [Guy & Tonkin, 2006].

Folksonomies also have semantic difficulties regarding synonymy and polysemy: The former term referring to occurrences where different words have the same meaning; this can cause problems when searching for information and users have tagged the same piece of information with different tags based on their own interpretation and opinion regarding what it should be tagged with. The latter term relates to the ambiguity associated with a word having two separate meanings [Marlow et al, 2006]. Problems like the ones mentioned would arise with the use of an open vocabulary dependent on the user, the only real method to control differences is to rely on a controlled vocabulary that only allows certain keywords to be used and the user is aware of the definitions of each keyword.

Several practices to improve tagging have been suggested, but until all users adopt these at the level of assigning tags then they will have little affect on improving the quality of tags used. It has been suggested [Guy & Tonkin, 2006] that using plurals rather than singular terms would improve consistency, so too would using all lower case lettering, and grouping words using underscores to allow easier parsing to break up terms if needed. Common sense approaches to looking at how other users have tagged and following their tag conventions would help. Users of del.icio.us have begun employing their own tag conventions that are similar to directory structures, for example; tagging football would be tagged as ‘sports/football’, and similarly cricket.

---

12 http://www.dcs.shef.ac.uk/~mrowe/meerkat.html

13 http://del.icio.us
would be tagged as ‘sports/cricket’. Systems could be implemented to verify the tags being input into the system at the lowest level to check for any errors that they may contain, therefore avoiding any erroneous tags that could appear.

Although folksonomies are, by their nature, a user centric free text approach to the common problem of labelling something of interest, there has been several pieces of work to formalise tags and define them explicitly using predefined ontologies. [Al-Khalifa & Davis, 2006] uses folksonomies from del.icio.us as a source for tags, the tags are then matched against concepts in an appropriate ontology. This approach works by first normalising the tags using natural language processing techniques to remove generality, stemmed words and clustering of similar tags. The tags are then processed to map them with concepts from several ontologies included in the system. The mapping process is simply carried out using string matching between the ontology concepts label, and the tags entered into the system.

Another example of work carried out by members of the Semantic Web community to align the popular action of tagging with the formal approaches of the Semantic Web is demonstrated in [Specia & Motta, 2007]. In a similar fashion to [Al-Khalifa & Davis, 2006] the process of mapping tags to ontology concepts begins with the pre-processing, or ‘normalisation’ of the tags obtained from a useful resource, in this case; flickr and del.icio.us. Tags are clustered and each cluster is analysed using concept and relation identification. This technique takes a pair of related tags and tries to identify the relation that forms the relationship between the pair. Semantic Web search engines, Wikipedia and Google are all used to find the mappings between the tags as knowledge bases and retrieval services.
2.2.2 Identity 2.0

Work presented by Dick [Hardt, 2005] first coined the term identity 2.0 to tackle issues associated with the distributed identity and presence problems web users face today. Each web site that requires handling of sensitive information, typically depicted in a profile, requires a user to input this information for processing. As a given user accesses more sites, each requiring the user to input sensitive information, the workload increases and so too does the growth of accounts a user has to their name. [Hardt, 2005] argues that credibility and online presence that has been built on one site should be transferable to another. For example; a user’s feedback on an online auction site should be useful to another auction site. At present there is no scalable method to transfer this information. The identity of a user is lost outside of a given domain.

In the context of the Social Web, identity 2.0 is less corporate than previously expected. It is who we are, what we do, what we like and who we are related to, family, friends or otherwise. Social networking sites offer users the chance to customise how an individual is perceived, tailor their identity to depict who they are in the real world, or maybe an alter ego.

Work presented in [Cameron 2005] defines the ‘Laws of Identity’ explaining issues present in digital identity. The laws state that user control and consent have been compromised in several web 2.0 web sites, where information about a user is divulged without their consent and prior knowledge. As described in more detail in the next section about online privacy, there have been several cases where information that has been written on social networking sites has had a detrimental effect. Another requirement is for information to be released in its most minimal format, therefore preserving the remaining information that could be sensitive, and only releasing information to parties that have a justifiable means for using this information, avoiding identity theft and fraud.

[Cameron 2005] also argues that the current use of identity management in the digital domain consists of a patchwork of identity one offs without the use of standard methodology to validate if a person is who they say they are. In order to combat this, the ‘Laws of Identity’ describe the need for a universal identity layer that incorporates various existing external technologies for authenticating identities. This would allow sensitive information to be shared between various web 2.0 sites, containing semantic metadata about a given identity, aligning ontologies that have been created to describe the different information on those sites.

Work presented in [Jordan et al 2003] outlines a detailed proposal for developing the next generation of the Internet, paying particular attention to user centricity and the role of identity within online social circles and cliques. The theory of persistent identity is presented which has become more prevalent in recent years with the increased usage of Web 2.0 and Social Web sites. [Jordan et al 2003] describes treating users not as anonymous avatars but as real people, who play a role in society, something they describe as a ‘civil society digital profile’. An individual’s interests and concerns stem from their role in society, effecting how their identity is perceived amongst peers both online and offline. Individuals are to be encouraged to show their affinities and capabilities, therefore aiding the discovery of individuals with similar
interests. Introduction protocols would be used to allow individuals with similar concerns and interests to communicate and interact for the first time. This forecast of augmented social networks demonstrates a shift from treating individuals as consumers, to treating individuals as citizens.

Although the work presented by [Jordan et al 2003] is now four years old, the identity criteria that was presented has been fulfilled and has lead in a large rise in the user base of sites such as Facebook and MySpace and Flickr that treat users as citizens of a particular community. This has lead to a shift in the market following the web 2.0 movement away from consumerism and towards a greater user centric trend, benefiting the individual.

Openid 2.0 [Cameron & Reed 2006] is a possible solution to the issues associated with the current patchwork of identity one offs. By developing the original work behind openid, openid 2.0 provides a framework of authentication offering useful semantic services once a user has been authenticated. Address based identifiers are used to identify a user through the use of a unique web address, and card based identifiers are used to identify a user using digital tokens. This framework could be adapted to include existing authentication technologies to use one single identifier for an individual, therefore reducing the need for a patchwork of identity one offs. The popularity in the use of openid amongst developers has lead to an increase in the number of web 2.0 sites adapting this technology.
2.3 The Social Web

The Social Web is the fastest growing incarnation of the Internet in today’s society. The Wikipedia philosophy that should one be given the power and the trust then one will contribute and flourish has stemmed from successful web startups such as Youtube and Flickr by allowing any user of those services to contribute. As a consequence the Social Web has seen a dramatic increase in both size and value over the past 2 years.

This section of my transfer report documents the various areas within the social web and the work that has been carried out within those areas to date. It begins by explaining what a social network is in the sense of the social web and provides an analysis of a Semantic Web formalisation for describing social networks effectively. Social networking sites are then explained by providing examples of popular sites, their characteristics and the issues that are associated with such services. Several social tools are described to demonstrate how semantics play a role within the social web, and finally providing examples of socially generating metadata.

2.3.1 Social Networks

“A social network consists of people or groups connected by a set of social relationships, such as friendship, co-working or information exchange.”

[Finin et al, 2005]

The above quote encapsulates the true definition of a social network. From forming relationships with peers, friends and colleagues, individuals are able to build a bond of trust and understanding. In terms of the Semantic Web, trust is a key factor when considering the validity and credibility of generated metadata. According to [Ding et al, 2005], the social-network model and the Semantic Web support one another: The Semantic Web allows social information to be defined in a formal and explicit manner through the use of ontologies. The social-network model allows knowledge to be accessed and acquired through the use of social networks.

The capturing of social networks, until recent years, has always become a challenge for the scientific and anthropological communities. The study of human interaction and understanding is of great social importance, and helps researchers understand the way groups of individuals behave and interact.

Capturing the demographic and structure of social networks has been achieved with success in work presented by [Matsuo et al, 2006] and [Hamasaki et al, 2006]. In both pieces, the same methodology is applied to harness the information within social networks. Their technique is comprised of a three-step approach: Firstly, mining the web for social network information identifying links between two individuals, and once found these links are then stored. Secondly, monitoring real world interactions between individuals to obtain confirmation of the relationship between two individuals. And thirdly, interactions between users on the web, capturing online communications between individuals.
Mining the web for social network information is an approach that has also been applied in [Mika, 2004]. This was one part of a two part methodology to gather social network information from the web by both mining information from the web and crawling the web for semantic documents containing information described using the FOAF\textsuperscript{14} ontology, which will be discussed in more detail in the next section.

[Mika, 2004] mines social network information from the web by querying the search engine ‘Google’, with pairs of names of individuals that are considered to be friends. The number of pages returned containing both names co-occurring is the count for that pair; this gives the strength of the relationship between the two individuals. [Mika, 2004] uses the obtained values to plot a graph using a Java visualisation package to provide a graphical interpretation of the social network.

**Relation Identification**

The use of co-occurrence to measure links between individuals is criticised in [Jin et al, 2007]. This is due to certain domains having universally co-occurring entity pairs. [Jin et al, 2007] argues that this characteristic is apparent in the majority of homogeneous communities where the members are all involved in a specific area reducing the scope for external involvement. This stems from two assumptions that can be made about the social network extraction methodologies used in [Mika, 2004] and [Finin et al, 2005]:

i. Web pages are created to define the collaboration of two or more actors in an event.
ii. Community of individuals in the form of a social network is homogeneous.

The first assumption could be said to be true when considering the quality of web pages representing research work carried out by an individual. Any work that has been collaborated on is represented along with the co-workers, therefore depicting a relationship between two of more individuals, depending on the number of other co-workers. The second assumption assumes that any people in the same social network exhibit the same interests and could therefore be considered to be the same type. This cannot be true in the case of social networking sites where an individual’s interests make it impossible for them to be labelled as belonging one particular group of individuals, instead users are more likely to span multiple groups.

[Jin et al, 2007] attempts to build on these underlying assumptions of previous work to improve the mining of social network information by using only a search engine, to reduce the complexity of information aggregation evident in [Mika, 2004] and [Finin et al, 2005], by using two addition functions: Relation identification and threshold tuning.

Relation identification attempts to identify labels for relations between two entities, the label word along with the two entities then form the query to be entered into the search engine. This methodology allows ontologies to be generated using the entities and relations that link them together. However, [Jin et al, 2007] does not fully explain the methodology for generating the possible keywords for relations. Threshold tuning

\textsuperscript{14} http://xmlns.com/foaf/spec/
defines the importance between two entities using objective and subjective criteria. This allows relations to be analysed upon discovery to decide whether they are valid or not, depending on their strength in comparison with the objective criterion. The objective criterion is constant, whereas the subjective criterion is domain dependent.

Another example of relation identification is presented in work by [Mori et al, 2006]. Again the goal is to produce the link between two entities in the form of a relation: (entity, relation, entity). Unlike [Jin et al, 2007] this methodology attempts to use collective contexts as a source for similar entities. A local context is found containing two entities from a predefined entity list, using the collective context of each pair of entities, pairs are then clustered based on similar contexts. This then produces labels capable of defining the relationships between the entities. While this process is being performed, social network information is extracted concurrently using the hit counts of the query containing the entities and their labels. Figure 3 demonstrates the methodology provided by [Mori et al, 2006] to extract relation identification in greater detail.

The work presented in [Liu & Maes, 2005] extracts information directly from social network profiles of individuals. The information that is extracted is primarily concerned with the individual’s interests, using the nature of social networks to allow an individual to create an online identity depicting them. Interests are clustered together and visualised along with the identities of the individuals to show the network of interests labelled the ‘InterestMap’. As correlations are found, these are represented by links between interests and individuals.

This use of social network mining presents an interesting methodology for extracting important information that could be used to recommend individuals from outside of an individual’s social network that the individual may have similar interests to. Although this approach offers good results due to the large data set mined to perform the primary analysis, it does pose several questions and issues: How could information from separate social network sources be aggregated and analysed together? And what sources would be the best to carry out this analysis of personal information?
2.3.2 The Friend of a Friend Project (FOAF)

Applying semantics to information relating to social networks is perfectly exemplified in FOAF\(^{15}\). The friend of a friend project uses a lightweight specification to define a person and the various useful information items associated with that given individual. What is important about FOAF is its ability to define the friends and acquaintances of a given individual using formal semantics. By using formal semantics to define an individual and their attributes, FOAF offers a useful format for aggregating information together, essentially compiling more explicit semantic knowledge about a given individual from more than one source.

[Finin et al, 2005] presents an approach to social network extraction using FOAF files. By employing a crawler, FOAF files are extracted from the web using ‘Swoogle’, the semantic search engine. Information is extracted from each FOAF file and aggregated with information from other FOAF files. This fusion of information allows the identity of individuals to be found from one or more corresponding FOAF files by making assertions about the individual using the supplied semantic information.

[Mika, 2004] also includes work that has been carried out to extract social network information from the web. The previous section contained information about the mechanism for crawling the web manually for information. [Mika, 2004] also uses a technique similar to [Finin et al, 2005] to crawl the semantic web and find FOAF files containing personal information. Unlike [Finin et al, 2005] the crawler uses a scatter approach to find other resources containing personal information. The scatter works by accessing the `rdf:seeAlso` property contained in each FOAF file, if present. This provides a URL to another FOAF file containing more information. This process can be repeated until all discovered FOAF files URLs have been completely exhausted.

The vocabulary of FOAF uses classes and their properties to explicitly define information about a given individual and their associated friends. The following is an example of FOAF generated using the Facebook Foaf Generator\(^{16}\).

```xml
<foaf:Person rdf:ID="me">
  <foaf:name>Matthew Rowe</foaf:name>
  <foaf:givenname>Matthew</foaf:givenname>
  <foaf:family_name>Rowe</foaf:family_name>
  <foaf:gender>male</foaf:gender>
  <foaf:img rdf:resource="http://profile.ak.facebook.com/s61109237.jpg"/>
  <foaf:based_near>
    <geo:Point geo:lat="53.990615" geo:long="-1.539111"/>
  </foaf:based_near>
  <foaf:holdsAccount>
    <foaf:OnlineAccount>
      <foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
      <foaf:accountName>61109237</foaf:accountName>
    </foaf:OnlineAccount>
  </foaf:holdsAccount>
</foaf:Person>
```

\(^{15}\) http://www.foaf-project.org/

\(^{16}\) http://ext.dcs.shef.ac.uk/~u0057/FoafGenerator
<foaf:interest>Football</foaf:interest>
<foaf:interest>music</foaf:interest>
</foaf:Person>

This example demonstrates how FOAF is able to semantically define several pieces of important information in a machine and human readable format. The individual has their name defined as foaf:name and the an image depicting the individual defined as foaf:img with the property described as the rdf:resource containing the URL of the image. FOAF also contains a property to define the location of the individual, this can be a text string of the given place name, or can also contain a geocoded property describing the location. The geo:Point property contains sub properties geo:lat and geo:long to define the latitude and longitude of the individuals location, similar to work carried out in [Mika, 2004] where geocoded locations are created using a different geocoded service.

Recent modifications to the FOAF specification have allowed web 2.0 and social web properties to be included in the vocabulary. As the above example shows, the property foaf:holdsAccount and foaf:OnlineAccount contains information about an account an individual has online. FOAF describes the details of the account, using foaf:accountServiceHomepage and the subproperty rdf:resource to define the URL of the web site where the individual has an account, and foaf:accountName to describe the username or user id of the individual.

Predictions stated in 2004 from [Downes, 2004] predicted that social networking sites would offer the functionality to create FOAF files easily and automatically from the sites metadata. In truth, this has not happened. FOAF has become a useful, but not widely used description language compared to the amount of users who are members of social networks today. Many social networking sites offer little functionality to automatically export information from their sites due to their perspective that the information is sensitive.
2.3.3 Social Networking Sites

As the growth of web 2.0 sites has increased so has the influx of web sites offering social networking capabilities. This sharp rise in the popularity of social networking sites has led to large investments and in some cases social networking sites being sold for vast amounts of money (MySpace sold for $580 million\(^{17}\)). Social networks can be split into two categories; primary social networks and secondary social networks\(^ {18}\). The former network consists of those friends around a person that a person sees regularly and socialises with the most, for example, classmates at school. The latter network consists of friends that are acquaintances that a person does not socialise with a great deal or at all, yet maintains a relationship of some degree with the person. This section presents a brief overview of four of the most popular social networking sites at present, and focuses on the various features and development solutions each site has to offer, each of the sites mentioned are associated with primary social networks.

**Facebook**

Facebook was originally developed by three university students based at Harvard in the United States\(^ {19}\). Although only one member of the original team was technically able to develop the site, the remaining two members were more concerned with the social networking aspect of offering interactive services to users. Each user who signs up for an account with Facebook is given an area where they are able to add information about themselves, they have a wall where other users who are their friends are able to leave them messages, they are given areas where they can post notes, and most importantly they are given a large amount of space to upload photos to. Facebook uses an ontology driven paradigm when specifying relations between users. If a user would like to add another user as their friend, they are able to select a variety of options that defines the relationship they have with that person, similar to the relation that would be specified between two concepts in an ontology.

A Facebook user is also able to tag people in pictures they have uploaded using a Java applet that runs within any web browser. By selecting an area within a picture, the user can then tag their friend in the picture. Upon doing so the system informs the user that they have been tagged in a picture by another user, and should the tagged user wish, they could navigate to the picture and remove the tag. This semantic annotating of images is a popular component of other web 2.0 sites and allows semantic metadata to be applied to multimedia content.

What sets Facebook apart from a lot of other social networking sites, is its definition and restrictiveness of exactly what a network entails. Upon signing up for an account with Facebook the user is prompted to join a network, once joined the user can browse all the users within that network but not all the users on Facebook’s system, therefore restricting an average user’s accessibility.

Facebook, like many web 2.0 applications, offers an API to developers for integration with third party applications. This opens the door for exciting applications, and as has

\(^{17}\) http://freemyspace.com
\(^{18}\) http://chimprawk.blogspot.com/2006/01/situational-relevance-in-social.html
\(^{19}\) http://en.wikipedia.org/wiki/Facebook
been accustomed to the web 2.0 phenomenon, mashups; where several technologies are merged together to produce a new tool or product. The Facebook API comes in a PHP format and a java format, both of which use authentication tokens to authenticate the user’s session before any further functions or procedures can be accessed.

Work presented by [Lampe et al 2006] examines the use of Facebook in a social context to discover the disclosure of information online compared to offline. The use of Facebook sees two primary roles; social searching and social browsing. Both methods help to build previous established relationships with peers offline by allowing a less inhibited method of interaction through a mediating service such as Facebook. Although, as I will describe later in the section titled ‘Lateral Surveillance’, this can have a detrimental effect on social cliques as users become more concerned with how other users are acting and begin to develop stalking characteristics in severe cases20.

MySpace
Along with Facebook, MySpace is one of the largest social networking sites used in the world today. It offers several features similar to Facebook such as a user page where a user can add personal information, a comments area where friends can leave messages, a blogging area, large space for images to be uploaded to, and large space for videos to be uploaded to [Snyder et al, 2006]. MySpace is more orientated to cater for bands and artists as it offers a free music player which bands and artists can uploaded their songs onto, and users browsing to their page can listen to them, this has led to several bands being signed to record labels in the past21.

Although MySpace is more popular than Facebook, it has suffered from hack attacks and security scares on several occasions, and spammers have been able to send messages to all users of the site due to insecure source code22. MySpace does not offer any networks as such, every user is visible to the public, unless they make their profile private and they only their friends can view their page. Instead, MySpace offers groups, which users can join, and at present these only cover schools. From a developers perspective, MySpace is disappointing as it does not have an API which allows access to its features, therefore making it hard to develop third party tools and applications for.

The differences between Facebook and MySpace are explained in detail in [Croft, 2007], differentiating between the social demographic of users. [Croft, 2007] believes that socially Facebook is more reflective of a person’s real world personality and demeanour, how they wish to be perceived from day to day. In contrast, [Croft, 2007] argues that MySpace contains more profiles that could be classed as containing alter egos of people. An interesting comparison is drawn by comparing 100 Facebook accounts from the authors friend list, with the same individual’s MySpace account. This produces the realisation that MySpace contains different profile information from the Facebook equivalent; therefore Facebook could be seen as being the formal social networking site which is indicative of public perception.

---

20 http://observer.case.edu/Archives/Volume_38/Issue_26/Story_898/
21 http://en.wikipedia.org/wiki/MySpace
22 http://mashable.com/2006/07/17/myspace-hacked/
Bebo

The formula for social networking sites seems to be a predictable one by covering all of solutions that a user might require. Bebo, like Facebook and MySpace, offers a customisable user space where a user can add personal information, photos, blogs, and the ability to add friends too. Bebo, similar to MySpace, also encourages users to join groups, which at present are restricted to schools and universities only.

A new feature of Bebo is Bebo widgets, these are small tools that can be integrated into a Bebo user’s page to display third party content such as a slide show of images hosted on a separate site\(^\text{23}\), or similar third party tools. This feature shows that bebo is also allowing users to create their mashups between technologies and provide hosting for the products. Unfortunately, Bebo does not offer an API allowing developer to access information from Bebo.

Friendster

Possibly the smallest of the four social networking sites in this section, Friendster has been running as a commercial product the longest, since 2002. Using the popular circle of friends model, Friendster was able to file for a patent of its adaptation of the model with a web of contacts where two individuals that were related in the network in some way could be graphically displayed. Friendster, due to its intrinsic nature of being a self-contained web site, does not offer an API for development and integration with third party software.

[Downes, 2004] describes Friendster as having serious issues with fake users, where users have signed up to the site and impersonated some other person, in several cases a famous person. [Downes, 2004] argues that is due to the lack of functionality to allow consistent digital identities to be maintained online. This refers to the multitude of social networking sites and the redundant accounts that users create on each site. However, in the case of Friendster, [Downes, 2004] believes that the site simply offers a hollow shell, allowing any user to assume any identity, and therefore making it “a prime candidate for spoofing”.

\(^\text{23}\) http://www.bebo.com/OurBlog.jsp
2.3.4 Online Privacy

As the number of social network users grows, so too does the amount of personal and possibly sensitive information collected about individuals. The theft of this information has become a global scale issue and due to the ease of access of information, personal user details can be retrieved. Sites such as Facebook regularly include addresses and telephone numbers of users, available to any user within the same social network, and in some cases all users. These details can be readily exploited to commit identity theft, leading to serious implications for the unfortunate victim\(^\text{24}\). [Gross et al, 2005] presents a worrying simple approach to generate US social security numbers using personal information entered into Facebook. The use of social security numbers when committing identity theft in the US is a large scale problem\(^\text{25}\), any information that can be obtained to generate these numbers will aid the process, and Facebook provides this.

Work presented in [Atkinson et al, 2007] describes the issue of vulnerability of two selected groups of users. The former group consisted of survivors of abuse, mainly women who were seeking refuge away from harm, and the latter group consisted of teenagers, who were in general web-savvy. The group of survivors demonstrates a real need for isolation, and protection from being found. Although after seeking refuge it is extremely unlikely that one would divulge information about their current location. However, should the information appear on a social networking site, this information could be accessed easily and used by the abuser to track down the abused.

The teenagers were less hesitant when inputting personal information into a social networking site. They were more concerned when they were contacted by anonymous and unknown users who had clearly accessed their profile information without them knowing. By divulging personal information such as an msn messenger login name, teenagers could be contacted externally to the social network at a much more obtrusive level via instant messaging in real time. [Gross et al, 2005] states that there are a significant number of ‘weak’ ties between people on social networking sites, where users who are friends may not have spoken previously but are in fact friend of a friend of a friend, linked through several nodes within the social network.

The semantic web does not deal with the issue of online privacy. The freedom of information is an important component of realising a truly machine readable web, allowing automatic extraction of information with no intent to harm. Sadly this is not the case with many web users. According to [Atkinson et al, 2006] the use of personal information and the privacy of its owner are not important parts of the semantic web, but are instead attributes of trust. By accessing personal information and using reasoning tools, semantic web applications have the ability, wrongly or rightly, to aggregate personal information about a large number of oblivious users.

To tackle the issue of online privacy, privacy enhancing technologies (PETs) must be developed to protect users from inputting sensitive information that will not be hidden from unknown users. At present the current technologies do not offer the functionality

\(^{25}\) http://en.wikipedia.org/wiki/Social_Security_(United_States)#
to restrict, or warn the user of entering sensitive information into a publicly accessible web resource.

2.3.5 Lateral Surveillance

Facebook stalking\textsuperscript{26} is a term given to the action of regularly checking a users profile on Facebook without actually communicating with the person. By offering the functionality to allow anonymous users to view personal information about any user within the same social network is a surprising characteristic of several social networking sites. Several social networking sites as default set profiles to be viewable by all users, rather than the converse to restrict access. From the sites perspective, this is commercially just, allowing people to view the service as more interactive rather than restrictive. However, ethically this restricts the privacy of users, and promotes lateral surveillance of other users.

The majority of Facebook users commonly include several pictures depicting themselves, therefore enabling re-identification to occur; identifying a user without using any explicit identifiers such as name, or date of birth. Recent advances in technology have allowed this process to become automated through the use of Semantic Web technologies such as knowledge acquisition and management, and content based image retrieval. [Gross et al, 2005] explains how a cross analysis could be done to match students from university directory web pages with Facebook networks containing the same users. Images could be compared and personal information could then be retrieved and aggregated from both sources.

There have been several cases where companies have performed lateral surveillance\textsuperscript{27} to effectively vet prospective employees, and check up on current employees. Social networking sites offer the perfect uncomplicated platform to carry out this process. It has become such a regular process that [Andrejevic, 2006] describes several cases of lateral surveillance being performed by socialites, and party goers to access information about prospective dates. This ‘market research’ allows the general public to generate a possible second impression about any given person.

Work presented in [Gosling et al, 2007] explains how personality impressions based on Facebook profiles are an important indication of what the user is like in real life. Tests were conducted to compare how users are perceived offline in the real world, and how they are perceived online based on their Facebook profile. The findings suggest the motion that online social networks are an important aspect of modern day communication, allowing communication to occur through another layer.

\textsuperscript{26} http://observer.case.edu/Archives/Volume_38/Issue_26/Story_898/
\textsuperscript{27} http://technology.timesonline.co.uk/tol/news/tech_and_web/the_web/article2087306.ece
2.3.6 Social Recommendation Systems
The increase of ubiquitous systems has given birth to better and more intelligent web based systems, capable of adapting to suit the user’s needs and requirements. Systems are focused on using intelligence to personalise aspects of the system by customising ontologies for specific areas of a system [De Bra et al, 2004]. The key to this process throughout the following systems that are to be presented is the ability to adapt to suit the user with the minimum amount of user interaction. By presenting several social recommendation systems, this section demonstrates different solutions to the task of recommending media to users. The nature of social recommendation systems is to use prior knowledge from other users to provide accurate recommendations based on similar interests.

Last.fm
Like many web 2.0 web phenomena; last.fm\(^\text{28}\) has become a well-known internet service. By accessing the site users are able to listen to a vast amount of music. Last.fm keeps track of all the songs that a user has listened to and allows other users to access and view their personalised playlist. New bands and artists can be discovered when using last.fm as the system suggests other songs to listen to that other users have classed as being similar to the song is being currently listening to, therefore through the use of recommendations.

Last.fm has invented its own term for the process of listening to a song on their system and saving the song to your profile; ‘scrobbling’. This process is reinforced with last.fm’s easy to use interface that offers various functions when a song is being listened to. A user can select if they love or hate this song, which in turn adds more information to be saved during the scrobbling process. Over time a user’s profile can be built up to consist of their favourite music, and therefore users with a similar taste in music are able to network with the user and find new bands or artist that they have yet to discover through the user. Like most web 2.0 applications, last.fm also offers an API which can be used to access data about users who have signed up to use the site. Using the API a users details can be found along with their top tracks, their most common tags, and their chart.

\(^{28}\) http://www.last.fm
Pandora
One of the most exciting examples of recommendation systems is Pandora\textsuperscript{29}, an in browser flash application that contains a comprehensive database of songs by a wide variety of artists. What makes Pandora so special is it’s technique for recommending songs and identifying songs that are similar. Rather than relying solely on user feedback to specify the similarity between two songs, Pandora uses a special technique developed by the Music Genome Project\textsuperscript{30}. Every song added to Pandora’s database is analysed for harmony, rhythm, instrumentation, orchestration, arrangement, lyrics, and vocal harmony. By assigning metrics and values to every attribute that can be taken from a song, similarities can be drawn between various songs on a linear level.

When a user loads Pandora, the system prompts the user for an artist or band to start the system with. It loads a song by that artist, and once the song has finished playing, it loads another song that it believes to be similar using the relations it has saved. The user can also specify whether they believe the song to be similar or not, and the system takes this into account when updating the relation between the two songs; to either strengthen the relation should the user deem the song to be similar, or to weaken the relation should the user deem the song to be different in a similar manner to how last.fm functions.

Amazon
Amazon’s popular and commonly replicated idea to suggest products that other users had bought based on the current product being viewed by a user was one of the forerunners for recommendation systems\textsuperscript{31}. The recommender system learns what products users of Amazon have bought together and can therefore draw relations between products for suggestions. This model has been copied by rival sites in a bid to allow a more tailored interactive experience for the user. Rather than learning to suggest products by associating products that users have also bought, several web sites allow users to rate products which affects the recommendations used by the system.

\begin{footnotesize}
\begin{enumerate}
\item http://www.pandora.com
\item http://www.pandora.com/mgp.shtml
\item http://www.amazon.com
\end{enumerate}
\end{footnotesize}
2.3.7 Social Bookmarking Tools

The growing use of folksonomies throughout various web applications has brought about the creation of various new tools and techniques throughout the web. Social bookmarking tools are an advancement of the regular bookmarking functionality that many web browsers come equipped with. As pages are browsed to, they are bookmarked and their location is saved so they can be viewed later. The same principle is used with social bookmarking tools, pages are saved but along with the address of the page, tags and a description of the page are also saved to a central store so other users are able to view what pages have been bookmarked and how they have been tagged [Hammond et al 2005].

This sections presents four of the most used social bookmarking tools on the Internet today. Like most web 2.0 applications they are described by their functionality, the user experience they offer and what they have to offer as a web service.

**Del.icio.us**

Del.icio.us is the most popular of the social bookmarking tools. It runs as an in browser Mozilla Firefox plug-in and is less subtle than other social bookmarking tools. Once a user has signed up for an account they are able to browse to any web page and bookmark the page, the page is then saved to a central store along with the tags and description used. The user can then browse to the del.icio.us web site and view what annotations they have made along with the tags they have assigned. Biddulph [Biddulph, 2004] explains how del.icio.us is broken up into three main axis; users, tags and URLs. By combining the three axis the system is then able to return the bookmarks that are for a given URL, have been tagged with the given tags and were bookmarked by the given user, therefore providing a simple query mechanism that users can use in the browser’s query string.

There are various interesting features that del.icio.us has to offer, and there are APIs available to develop mashups and software to integrate with del.icio.us. Using JavaScript users are able to access del.icio.us and download bookmarks that they have made, or bookmarks made by specific friends in XML format.

Research work presented in [Mason & Thomas, 2007] provides an interesting analysis of the tagging trend used throughout del.icio.us. The nature of tagging is analysed by using groups of users to tag literature they have found on the Internet about a specific assigned topic, this allows key tags to be found for certain information items. [Mason & Thomas, 2007] describes two kind of tagging mechanisms being used when tagging; Ontotagging and Idiotagging. Ontotagging commonly uses a taxonomy of tags, where a user may have a restricted set of tags that they believe defines the majority of the resource they need to label. Idiotagging uses more descriptive tags that may contain sub classifications within the tag itself, or unique words that differentiate the resource. [Mason & Thomas, 2007] argues that within del.icio.us the use of Idiotagging is restricted to experienced regularly users who are keen to allow the community of users to benefit from their tags.
CiteULike
This web site is concerned with the sharing and social aspects of suggesting papers to other users\(^3\). When browsing the web for research papers users are able to bookmark papers they have read and find interesting to CiteULike where they are stored in the user’s library. CiteULike also allows users to create research groups, which users can join and papers can be added to as they are found to be relevant to the group. As papers are read they can also be commented on and rated either as been relevant to the group or rated as a good paper. Papers can also be tagged and have descriptions saved about them as other bookmarking tools also offer. CiteULike does not offer an API to use for development of third party applications and tools.

Connotea
Connotea\(^3\) is a hybrid of the technologies used in both CiteULike and del.icio.us, by moulding together the saving bookmark interface popup that comes with del.icio.us, the user area CiteULike contains with a library section and the ability to create research groups to share bookmarks with. Connotea uses online storage of references to allow easier citations of bookmarked papers, it also has automatic discovery of bibliographic information that retrieves information about a particular paper that has been bookmarked to the system [Lund et al, 2005]. This information that is saved can be called upon if a user attempts to bookmark a page that has already been saved into the system, Connotea recognises the URL being saved and retrieves the metadata by querying the Connotea web service.

Connotea also offers a web API that can be used to query Connotea and retrieve information as needed. Once a user is authenticated by the system, the API can then retrieve a list of tags used by the user or the most recent bookmarks saved to Connotea. The result of this query is returned either as raw XML or as RDF if required.

Technorati
The increase in the number of ‘bloggers’ who write daily online journals called web logs or blogs and publish them to the web brought about a problem when sharing this mass influx of information [Hammond et al 2005]. Due to the current nature of blogs search engines were too slow to crawl through and could only index one blog in a million so Technorati was developed to fill the void. This social bookmarking tool and web site offers users the chance to save their favourite blog posts to the Technorati server and assign tags and a description for other users to view. In comparison with the other social bookmarking tools, Technorati relies heavily on user involvement and community interactions. As new blog posts are saved on to the system, users are encouraged to comment and rate the blog posts, which are then tracked for changes. As a blog has more posts added to it, which are then in turn added to Technorati, it gains a good rating from the users of the tool.

The inclusion of APIs written in several different languages helps users to integrate their software with Technorati. The online system can be queried to retrieve

\(^3\) http://www.citeulike.org
\(^3\) http://www.connotea.org
information regarding the most highly rated blog posts, the most previous blog posts, and various other useful pieces of information in either RSS or XML formats.
2.3.8 Social Annotations

Facebook Photo Tagger

Annotating images within a web browser has become a useful component of Facebook’s photo sharing system. Images can be tagged socially by annotating images using a given person's name. This name is selected from the taxonomy of friends that a user is acquainted with on the site. This innovative feature allows images to be assigned semantic metadata from community containing information about the inherent content of the image. A key component of this approach is the agreement that must be achieved over the annotation using a community. Figure 4 demonstrates how the image has a region selected and then the user is prompted to select the person from their list of friends. In this case I would select Emily Rockett, she would then be notified that she has been tagged in a photo, and so too would all her friends as Facebook publishes a notification of this happening to her community of users. If she did not wish to be tagged in a photo, then the tag can be removed and cannot be re-applied.

![Figure 4 - Facebook's Photo Tagger](image-url)
One of the most interesting features of Flickr’s photo viewing system is the ability to annotate portions of the image within a web browser. Flickr allows users to add ‘notes’ to portions of an image by selecting areas of the image and then entering notes within the textbox provided. Unlike the Facebook photo tagger, Flickr does not offer annotation using existing semantic information, instead free text is used to describe the image portion. Using Flickr it is possible to create groups that depict a common interest or field of work within photography. Once a note has been made the annotator is able to share the image with these groups, and members are able to view the annotations for each image.

It is important to consider the role that both Flickr and Facebook play. The former is designed primarily as a way to store and share images that have been taken by a wide variety of users. The key component that brings users together is the want and need to share photos and interact with other users by commenting on one another’s photos. The latter; Facebook, is primarily concerned with the interaction and connections that can be made by networking socially. This is in comparison to Flickr which implements it’s social network as a second feature to the site. Flickr allows users to annotate images using folksonomies, therefore enabling referencing of images in a straight-forward manner. Facebook allows users to annotate images using personal identifiers for users within their social network, not using folksonomies.
3. Progress to Date

3.1 Requirements Definition
The literature that I have read covers three distinct areas that relate to the title of my PhD. I believe that the research I have made in these areas has given me a good understanding about which direction I wish to take my research and the work I am going to do in the future. This section explains the literature I have read so far and its relevance to the work I have done to date. The work I have done to date is then described in brief, explaining the reasoning behind it along with papers that have been produced as a result.

3.1.1 The Semantic Web
The literature presented covering the Semantic Web addresses the issues associated with the generation of semantic metadata, a fundamental aspect of my work. The described annotation tools present various approaches to the automatic, semi-automatic and manual annotation of web resources with semantic metadata. Semantic blogging and portals present an alternative social perspective to generating semantic metadata with the purpose of sharing this information socially.

3.1.2 Web 2.0
The literature that covers Web 2.0 based research includes work addressing the current trend of the use of folksonomies as annotation labels. This research has influenced the work I have carried out when producing Meerkat and Meervisage, both of which will be explained in more detail to follow in the subsequent sections. Literature describing the issues associated with identity information is included to document the issues of identity management, and distributed identity data. The research I have carried out in this area has given me ideas for future work investigating identity extraction and the sensitivity of personal stored online by information individuals.

3.1.3 Social Web
The work that I have documented about the Social Web is mainly concerned with the extraction of and utilization of social networks from online resources. The primary research that was carried out in relation to social networks was to analyse the involvement of online communities when generating semantic metadata. Work presented in this section tackles various issues:

a) The extraction of social networks from online resources using various techniques to model the relationships between individuals, and presents a formal approach for semantically defining these relations using FOAF. This literature has also contributed to ideas for future work, mainly concerning the extraction of identity information, and particularly the technical challenges associated with such a research challenge.

b) The current rise of social networking sites by presenting various examples of such sites and literature that has been presented linked to studies conducted about such
services. Such analysis allowed me to develop several services that access user bases from social networking sites, either to generate formal definitions of the relationships of individuals, or to use the individuals as generators of semantic metadata using an annotation tool.

d) Online privacy, the accessibility of ones personal information, and the rise in lateral surveillance; the behaviour associated with the ability to view online profiles relating to an individual and the marked increase in this occurring. This literature has demonstrated to me that the abundance of accessible personal information could be easily applied for research purposes, particularly when researching the extraction of identity information from web resources, but could also be exploited easily for malicious purposes.

3.1.4 Work to Date

Experimental Applications

The literature that I have read throughout the first year of my studies has allowed me to produce several pieces of work that each tackle separate issues. Research performed in several areas inspired me to create a lightweight annotation tool called Meerkat that allows users to annotate content within a web page using folksonomies. Further research concerning the Social Web then allowed me to focus on adapting this tool for use within social networks, using the already established online communities as trusted user bases for the generation of metadata.

Other work that I have created includes the Facebook Friend Plotter, influenced by the reading performed within the area of the Social Web. This application addresses the issues of online privacy and the accessibility of personal information by globally positioning individuals on a map of the UK using extracted information. This work will be reused for research in the area of identity extraction.

Similar work has also been produced to investigate the privacy of data, and the openness of certain online services to divulge personal information. The Facebook FOAF generator and the MySpace FOAF generator are both examples of applications that can extract sensitive identity information. I plan to reuse this work for further research in the area of identity extraction.

Papers

This year I have written one paper entitled ‘Meervisage – A Community Based Annotation Tool’. As I will explain in further on within this chapter, Meervisage uses established communities of users to generate semantic metadata. I presented this paper at the ‘Towards a Social Science of Web 2.0’ conference that took place at York University on the 5th and 6th September 2007. A video of my presentation is available online on my university web site34.

34 http://www.dcs.shef.ac.uk/~mrowe
3.2 Meerkat

A social annotation tool designed and built as a Mozilla Firefox extension. The tool allows easy annotation of text on any given web page with the user being able to assign a colour to the highlighting, and tags and a description to the annotation. The semantic annotation is then saved into an RDF store where other users of the tool can easily access it. As users browse the Internet with the tool installed in their version of Mozilla Firefox they are also able to view annotations that have been made by other users, and share semantic information across platforms.

Meerkat employs a centralised RDF store of annotations. At present this only provides a useful and lightweight solution to reviewing all the annotations that have been saved on to the system. Users of the system are able to access the store and view all the annotations made by other users and find annotations that have been tagged with certain keywords. The components and features of the system will now be explained.

Figure 6 - Screenshot of the Meerkat Server
3.2.1 Toolbar

The toolbar contains all the useful functions and tools necessary to use Meerkat to annotate web content semantically. One of the primary issues concerning annotations that are displayed to the user on a web page is the differentiation between annotations. Meerkat provides a palette of colours for the user to select when highlighting the text they wish to annotate as shown in figure 6. The toolbar also indicates to the user to the number of annotations that have been saved for the current page by all users of the tool.

3.2.2 Annotating Web Content

Highlighting the text region, and selecting ‘Annotate Text’ from the context menu annotates content of any given web page as shown in figure 7.

Figure 7 – The Meerkat Toolbar

Figure 8 - Highlighting content for annotation
3.2.3 Annotation Controller

Once content from a web page has been selected for annotation, the annotation controller loads the initial annotation ‘Annotate’ form. The user is prompted to enter metadata for the annotation, this can either be tags or a description about the text. Once the user has entered the tags and description, the annotation is saved in to the RDF store. The annotation controller uses Asynchronous Javascript and XML (AJAX) to communicate with the Meerkat server. This provides a lightweight mechanism that reduces the client-side computation required to process the annotations, the Meerkat server is used to carry out these computations and remote calls to the RDF store for annotation information, therefore using a 3 tier architecture where the Meerkat Firefox plugin simply presents the information to the user.

3.2.4 Suggestive Tagging

In order to aid the process of generating metadata for the annotation, folksonomies are suggested by Meerkat that are associated to the URL of the web page that is currently being annotated. As the annotation controller loads up the ‘Annotate’ form, Meerkat uses a third party tag source to access tags associated to the URL. This is done by accessing the del.icio.us Javascript Object Notation (JSON)\(^{35}\) and returning the most common tags that have been assigned to a given URL. By using a third party source for tags, the user is provided with established pieces of metadata to be reused.

\(^{35}\) http://del.icio.us/help/json/
Figure 10 - Editing metadata

Figure 11 - Context Menu within Meerkat
3.3 Meervisage

Meervisage is a semantic annotation application capable of using existing social networks as user bases for the manual annotation process. Semantic metadata is created using Meerkat, which has been adapted to allow annotations to be annotated using any social networking site. It builds upon the Meerkat framework excluding the functionality to share annotations globally, and instead confining the tool to allow users to share annotations with their peers within the same social network.

This work embodies the theme that has been presented throughout the literature review and the title of my PhD: Community Based Annotation for the Semantic Web. This statement can be broken down into two distinct problems:

i. The use of a community as a user base for metadata generation.

ii. The generation of semantically rich metadata.

A possible solution to part ii has been presented in the previous section with the overview of the Meerkat annotation tool. Generating semantic metadata in the form of RDF that can be reused by other users of the tool, but without the use of a formal ontology to define the semantics. Part i can be tackled by adapting the solution presented for part ii but with the added functionality to include a group of users rather than a single user as the user base for the generation of semantic metadata.

In order to extend this user base Meervisage uses existing social networks as users of the system. The framework provides an abstract component to allow Meerkat to be integrated with any social network resource. In this case of Meervisage; Meerkat has been integrated with the social networking site; Facebook. Facebook contains useful components that are important when considering the needs for semantic metadata generation from a community.

As a user of Facebook one can create groups within the social networking site, these groups commonly depict a common interest or topic. The group can have image, urls and videos posted to it, it also has a discussion board and an area where users are able to discuss anything. Increasingly Facebook groups are created to encapsulate useful resources that have been found by users centred around some topic, i.e. Research based topics. Meervisage uses these established social networking groups to annotate the web and provide a repository of annotations that all members of the group can access.

As a group member begins to annotate web resources, they are able to share their annotations with groups that they are part of. Annotations can be assigned to multiple groups. Once an annotation has been completed, and has been assigned a group within Facebook, all members of that group are notified that annotation has been made using Facebook’s notification service. Upon logging into Facebook, group members are then shown the annotation that has been made along with the semantic metadata that has been assigned to it; folksonomies used. Each group also has its own tag cloud that depicts the tags that have so far been used to annotate resources by group members. Possible development of this would cluster together groups that have similar tags to suggest collaboration.
3.4 Facebook Friend Plotter

The Facebook Friend Plotter is a light weight web 2.0 application combining three of the most up to date and used web 2.0 APIs. The principle behind the application is to attempt to assign rich geographical information to a user’s social network profile. This would allow an individual to discover the true distribution of their peers in an easy to conceptualise format. The system works by plotting all friends that are associated to a specific user of the system onto a map, together with their name, a picture and a link to their Facebook page. The friend plotter includes a useful framework for combining two web services that could be reused to develop other more advanced web applications.

The application works by logging the user into their Facebook account, and extracting all the locations of all the friends that are listed within that account. Using Facebook’s API the tool accesses the friends details that are associated with that user, it, and then retrieves the hometown location of each user. Using the Google Maps API\(^{36}\) and Google Ajax API\(^{37}\) the tool queries Google Ajax for a geocoded point object corresponding to the hometown location. Once the object is returned it is plotted on the Google map of the United Kingdom. This tool could become useful if Meerkat is integrated with Facebook to provide annotation functionality to social networks.

There has been keen interest in the friend plotter, and at the time of writing it has over 2000 users. The Independent\(^{38}\) Newspaper included the application in their ‘Extra’ supplement in a feature describing useful Google Maps based applications.

| Figure 12 – Newspaper snippet of the Facebook Friend Plotter |

\(^{36}\) http://www.google.com/apis/maps/

\(^{37}\) http://code.google.com/apis/ajaxsearch/

\(^{38}\) http://www.independent.co.uk
Future work for this application will include expanding the current scope of the plotting mechanism to include placing friends on a map of the world in the short term. Work will also include placing geographical distances from the principle user of the system to help the user to visualise how far they are from their friends. Long term, the friend plotter will also include metrics that cluster users into specific interests to find interest demographics. An example of this would be to choose a specific band or music artist, and attempt to highlight the areas of the country with the strongest affinity to this concept.

3.5 FOAF Generators

Based on previous work by [Mika, 2004], [Finin et al, 2005] and [Jin et al, 2007] I decided to work towards aggregating information together from various social networks to address issues such as identity theft, redundant information and also suggesting individuals with similar interests and affinities whilst also suggesting alternative social networks to individuals. In order to tackle these issues, it is important to consider the format of the current information dispersed across many social networking sites.

The majority of social networking sites do not allow information to be downloaded or accessed in a machine processable format. Facebook offers an API to allow third party applications to interface with the service and retrieve information following the correct authentication of a registered Facebook user. Even in that instance, the individual can only access information from Facebook about themselves and their friends. Information is only accessible in raw data form using a Java or a Php API to access the service, therefore requiring the information to be imported into a machine readable format. This is also the case for MySpace which does not offer any functionality to interface with the site and access a user’s information. However, MySpace does allow any user who is registered on the site to be able to browse to any other users profile outside of their social networks.

In order to aggregate information from two or more social networking sites, it is important that the information is converted into a common format containing semantic information about the given user. Based on research presented in chapter two, the FOAF ontology was chosen to define the extracted social network information. Two generators were created to generated the FOAF, both of which will now be described in more detail.

3.5.1 Facebook FOAF Generator

By using the current Facebook API\(^{39}\) the FOAF Generator provides a Facebook account holder with all the important information that is contained within their profile. Most importantly though, it also exports semantic information about their friends also. According to the FOAF specification, the ideal unique identifier for any given person is the email address that they use for regular correspondence. As we do not share email address, this becomes the one piece of information that makes an individual unique, and would be specified by the `foaf:mbox` property. Upon

\(^{39}\)http://developers.facebook.com/
registering with Facebook users must submit an email address, however Facebook does not allow this to be accessed by the API.

In terms of Facebook the unique identifier for an individual is the user identification number assigned to an individual upon registering. This piece of information can be handled using the `foaf:holdsAccount` and `foaf:OnlineAccount` properties to define an account with an internet service. `foaf:accountServiceHomepage` defines the URL of the site with which the user has the account, and `foaf:accountName` defines the user identification for that account, which in Facebook, is the id number.

```xml
<foaf:Person rdf:ID="me">
  <foaf:name>Matthew Rowe</foaf:name>
  <foaf:givenname>Matthew</foaf:givenname>
  <foaf:family_name>Rowe</foaf:family_name>
  <foaf:gender>male</foaf:gender>
  <foaf:img rdf:resource="http://profile.ak.facebook.com/s61109237_8593.jpg"/>
  <foaf:based_near>
    <geo:Point geo:lat="53.990615" geo:long="-1.539111"/>
  </foaf:based_near>
  <foaf:holdsAccount>
    <foaf:OnlineAccount>
      <foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
      <foaf:accountName>61109237</foaf:accountName>
    </foaf:OnlineAccount>
  </foaf:holdsAccount>
  <foaf:interest>Football</foaf:interest>
  <foaf:interest>music</foaf:interest>
  <foaf:knows>
    <foaf:Person>
      <foaf:name>Matthew Solomon</foaf:name>
      <foaf:givenname>Matthew</foaf:givenname>
      <foaf:family_name>Solomon</foaf:family_name>
      <foaf:img rdf:resource="http://profile.ak.facebook.com/s6200736_8002.jpg"/>
      <foaf:holdsAccount>
        <foaf:OnlineAccount>
          <foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
          <foaf:accountName>6200736</foaf:accountName>
        </foaf:OnlineAccount>
      </foaf:holdsAccount>
      <foaf:knows>
        <foaf:Person>
          <foaf:name>Marcus Walsh</foaf:name>
          <foaf:givenname>Marcus</foaf:givenname>
          <foaf:family_name>Walsh</foaf:family_name>
          <foaf:holdsAccount>
            <foaf:OnlineAccount>
              <foaf:accountServiceHomepage...>
            </foaf:OnlineAccount>
          </foaf:holdsAccount>
        </foaf:Person>
      </foaf:knows>
    </foaf:Person>
  </foaf:knows>
</foaf:Person>
```
Figure 13 - Sample output from Facebook FOAF Generator

Figure 6 shows the output that the Facebook FOAF Generator returns. Information is encapsulated along with the semantic descriptions of each piece of information. The RDF is further enriched by providing geographically coded information. The location string is extracted using the API and then submitted to the Google AJAX Local Search API, which in turn returns a geocoded Javascript object corresponding to the location.

3.5.2 MySpace FOAF Generator

Unlike Facebook, MySpace does not offer an API for information extraction from the site. Instead information must be mined from MySpace manually using screen scrapers to harvest the site’s content. There are several tools already available that perform this functionality, however all are written in Perl and have bugs and errors causing problems when running tests.

Instead, I developed a simple Java package capable of automatically extracting information about an individual and information about their friends. The application extracts information from MySpace using wrapper technology to define the constraints and delimiters within a given web page surrounding the content to be extracted based on the work of [Kushmerick, 1997]. This efficient methodology produces a concise and easy to use format for extracting information by specifying several delimiters surrounding the content to be extracted within a web page.

Three wrapper were written for the FOAF generator. The first wrapper was responsible for extracting information from a regular user profile page, containing information based round the foaf:person property and the underlying sub-properties. The second wrapper is written for extracting information from an artist profile page. MySpace has two kinds of users; regular users and bands. The band page contains extra features and is therefore a different format to the regular user page and requires a different wrapper. The extracted information still populates the foaf:person property and sub-properties.

The final wrapper is used to extract the friends from each type of profile. It recursively accessed the friends page of a given user and extracts links to each persons page. The FOAF generator then accesses each page and extracts the information populating the foaf:Knows property with the foaf:person property containing information about the user.

40 http://code.google.com/apis/ajaxsearch/
In a similar manner to the Facebook FOAF Generator, the MySpace FOAF Generator extracts the username of each user’s account to represent their unique identity within MySpace. Also, similarly location details are extracted and geocoded using the Google AJAX Local Search API.
4. Proposed Research

For the final two years of my PhD I wish to carry out research centred around the title of my PhD; *community based annotations for the Semantic Web*. This will include carrying out research largely in two fields; the social web and web 2.0, to compliment work that has already been conducted in the field of the Semantic Web.

As described in the description of the Meervisage system in chapter 3, my current work can be broken down into two distinct research areas. The first consists of annotating the web with Semantic metadata to allow automatic extraction of information using software agents. The second area investigates the user base used when generating the metadata, ie. Who is best suited for generating this information. The literature review presents state of the art work for generating semantic metadata using a single user, however no work has been carried out using a community of users to generate semantic metadata.

This chapter presents several possible areas of work that focus on the use of a community of users as a user base for semantically annotating documents.

4.1 Social Networks as Annotation Communities

Very little work has been done to attempt to use social networks as established community sources for annotating web content. Work has been carried out to mine social network resources to use as knowledge base information when performing automatic annotations [Reeve & Han, 2005]. Several social networks allow the creation of groups and networks that can be used as communities when annotating, annotations by that group could then be saved in a central repository for members of the group to access and use. In particular using Facebook, groups contain useful functionality such as private areas where users are able to post discussion threads, submit photos which can then be tagged, and provide RSS feeds for users to keep up to date with the group. As annotations are made, the group could be automatically informed, and hyperlinks could be generated for the repository entry.

The work that I have carried out to date includes a system called Meervisage. Although only at the initial alpha prototype stage, this system uses existing social networks extracted from social networking sites as user bases for annotating web content. Two different research areas are now described that analyse the difference in annotations when comparing a user base of a single user to an existing social network:

4.1.1 Inter-Annotator Agreement

Annotation quality must be assessed in order to derive any possible benefits that are apparent in the annotation of web resources using an established community of users compared to a single user. One common NLP procedure is to use inter-annotator agreement to compare one set of annotations against another, metrics such as the Kappa statistic [Carletta, 1996] could be used to compare sets of annotations for likenesses. In order to make use of this statistic and analysis effectively, several case studies would be created in order to generate annotations that can be comparable.
Each case study would include annotation tasks that involve finding and annotating key web resources about a particular domain in a given amount of time.

This work would attempt to prove or disprove the hypothesis that community based annotations are of a higher standard than single user annotations. Forming a crucial initial experiment for my thesis. Upon proving this comparison future work could build on experiments using communities of users and the semantic annotations that they are able to generate. A key reinforcement of this hypothesis would be presented by tag based analysis:

4.1.2 Tag Based Analysis
The current usage of folksonomies as metadata has been a key component of several Web 2.0 web sites and their features. Following on from analysing the hypothesis that community based annotations are of a higher standard than single user annotations the further analysis of generated tags during this process could play an important role in emphasising or disproving the initial hypothesis results. Work has been carried out by [Mason & Thomas, 2007] which presents an important study of tags, their nature and tagging trends. Such analysis could be performed using the data collected from experiments carried out using Meervisage to analyse the differences in tags generated by an established community of users compared to single user generated tags.

4.1.3 User Analysis
Using an established community of users would allow the further analysis of the users themselves within the community and their annotating and tagging trends. Annotations could also have a mechanism to generate feedback to the author, allowing cumulative feedbacks to be generated for the author affecting their trust level within the community. Analysis would entail identifying key community contributors and then studying their annotations compared with other annotators within the group. This study could lead to the generation of a trust metric able to prioritise semantic annotations produced by one individual over another.

4.2 Identity Extraction
As the number of users of various social networking sites grows, so too does the issue of distributed identity information throughout those sites. As users log in to each site they duplicate a lot of data that they have entered manually on previous occasions. The problem also arises where the user enters personal information into the site without the prior knowledge of how the information is going to be stored and who is able to access it. Several possible research challenges arise. Firstly; looking at how identity information from various social networking sites can be integrated in an effective manner. Secondly, investigating the distribution of this information throughout the web, how this information is being duplicated with or without the knowledge of the author.

Information can be extracted from each social networking site using a predefined meta-ontology, which would contain the concepts and relationships between them and
is populated with the relevant information. As pure RDF cannot be automatically retrieved from the majority of social networking sites, a framework could be used that includes an abstract semantic services layer. This layer consists of services that can be constructed to connect to a specific web service or resource. Each service would ideally contain methods for gathering information from the external resource, and an ontology specifying the information to be retrieved. The ontology specifying the retrieved information from the external resource could then be matched with the meta-ontology. The services layer can be bypassed if pure RDF can be retrieved from a source (i.e. LiveJournal), although an ontology must still be used to describe the RDF and provide a mapping with the meta-ontology.

As more information is retrieved inferences could be carried out using the information to suggest social networks that other users are members of, and tackle semantic problems such as ontology matching and alignment to match up the RDF being returned from different services. As each network can be visualised to contain islands of sparse data, matching the ontologies that represent these islands would allow the data to become related.

![Identity Extraction Architecture](image)

Figure 14 – Identity Extraction Architecture

Recently Google has launched Google Opensocial\(^{41}\) a set of standards that several online social networking sites have agreed to adhere to. This means that creating bespoke web application for one social networking site does not specifically reduce the portability of the application, therefore allowing it to be placed on other social networking sites, each offering access to profile information about a given individual which includes properties of their profile and their friends.

Investigating the distribution of this information throughout the Internet poses a much larger problem than its primary acquisition. Once information has been extracted about a given individual and integrated to form a common description that includes key properties that distinguish an individual, this forms a seed for finding duplications of this data. The issues and problems of the duplication of this information have been linked to identity theft\(^{42}\) due to the inherent visibility of personal information. Finding

\(^{41}\) [http://code.google.com/apis.opensocial/](http://code.google.com/apis/opensocial/)

duplications of this information has been attempted in the past with varying degrees of success. Garlik\(^{43}\) is an online service that crawls the Internet searching for information items about a given person using a seed set created when the user signs up for the service. This has varying degrees of success and regularly returns information items not linked to the seed individual.

The seed set that would be used for identity extraction in this case would contain more semantic information about a given individual such as geographical locations, relationships with other individuals and personal details that make that person unique. Extracting identity information from the web would be carried out using either of two approaches which will now be explained:

### 4.2.1 Focussed Crawling

As the World Wide Web contains such a large expanse of information it would be impossible to crawl all this information searching for occurrences of key information items. Instead focussed crawling would crawl specific areas of the web that have been derived from analysis of key information sources and query mechanisms. At present, to find information about a specific individual it is important to disambiguate the individual from others of the same name. Using the semantic metadata that has been extracted from the social networking sites that a user has become a member of, it would possible to derive key queries using the individuals name coupled with certain items of metadata. The queries would be assessed realistically to find which return the most accurate results relating to the individual.

In order to focus the crawling further it might be the case that certain information sources have a higher level of entropy than others. For instance, blogs and other social networking sites might contain the majority of information about an individual outside of the seed set. It would therefore be beneficial to concentrate the crawling on this cluster of information sources prior to broadening the search for further results.

Geographical locations would also play an important role in differentiating one individual from another. Social networking sites such as Facebook and MySpace divulge this information and it can be used to narrow the area of search dramatically. For example, if an individual called John Smith was from Harrogate in North Yorkshire, it is unlikely that the individual in question would have information stored about themselves where their hometown was outside of the UK, or indeed outside of North Yorkshire.

Using a combination of both searching and crawling using the semantic metadata mined from social networking sites as a seed set could produce interesting results. One of the key research challenges associated with this work would be identifying the correct balance of information integration and deciding what pieces of information are truly relevant to yielding good, consistent results.

\(^{43}\) http://www.garlik.co.uk
4.2.2 Indexing a Subset of the Web
Following on from focussed crawling, another possible challenge to the scalability of retrieving relevant identity information would be to index a portion of the web. Based on the results of focussed crawling it makes sense to index the cluster of relevant information sources that have been found in order to aid the easier extraction of information.

4.3 Aggregation of Web 2.0 Information
Following on from the semantic portals described in the literature review, web 2.0 portals would be developed to view and manage metadata from web 2.0 sites. Due to the large number of web 2.0 sites that offer APIs for third party web developers the exchange and linking of data between these applications can be challenging process. A web 2.0 portal would offer the functionality to allows users to manage their web 2.0 metadata in one location, managing bookmarked pages retrieved from del.icio.us, photos from flickr, and videos from youtube.

The aggregated folksonomies could be analysed and shared between users of the portal, which are associated through a social network’s API. It could be possible to analyse what a user and their friends have been tagging with most and create similarity metrics to analyse their use of folksonomies across the various domains for traits and trends when tagging.

4.4 Semantic Trails
With the increase of social networks and sharing of metadata; social bookmarking, annotation tools, there have been very few recommendation systems that make use of user’s browser trails as an indication of similarity between several concepts. To put this in context, a user could be considered a successful reviewer who recommends products based on their similarity to a baseline product. By starting off on a specific artist’s MySpace page, the user has decided to browse to all other artists that are listed in that particular artist’s friends. As the user accesses the artist’s page, a user specified threshold is assigned to how similar the artist is to the original artist. After completing this task several times, the user has built up a trail of similarity scores that define how similar they believe those artists are to the original artist.

Recording trails in this manner is not restricted to artists, but could also be applied to several domains. In order to record the user’s trail, a browser plugin could be adapted to perform the task. Once a trail is completed it could then be saved for later referencing or sharing with friends, as the trail is comprised of metadata, this metadata could therefore be shared with other users of the tool or by third party applications.

Current state of the art technology does not include the functionality to store trails of sites visited. Instead, similar work includes Hooey44; a tool able to record the browsing history of a web session and store this information remotely. It can then be

44 http://www.hooeey.com/index.php
aggregated with other saved browsing sessions and key pieces of information can be extracted from the history.
5. Conclusion

This transfer report presents work carried out in various areas of research that are relevant to my PhD title; *community based annotation for the Semantic Web*. Upon beginning research for my PhD I analysed the topic I was given and broke the work down that I had to do into three distinct areas: The Semantic Web, Web 2.0 and the Social Web. The Semantic Web is the main focus of the research that I have presented so far. It was important to familiarise myself with the state of the art annotation tools and become accustomed to how semantic metadata was produced.

The second half of my first year of reading was mainly focussed on the Social Web. I wanted to understand what work had been done to date analysing communities of users, what social web sites are the most popular and how the online communities within these sites function. As I have mentioned previously in the explanation of how Meervisage operates, I found it appropriate to separate the research I am currently doing into two distinct parts, the first looking at the mechanism for generating semantic metadata and the second focusing on who would generate this metadata, i.e. The community or a single user?

The annotation tools that were described in the literature review helped the development of Meerkat, the annotating and saving of the annotations functionality was heavily influenced by Diigo, and the backend store for the annotations using RDF was influenced by Piggy Bank, along with the use of Java within a Mozilla Firefox Extension. When annotating the user is able to specify tags, once the Meerkat web server is accessed the user is able to view their annotations and the folksonomies they have assigned to each annotation.

Following on from Meerkat, Meervisage uses the existing technology to use groups from social networks as networks of users, offering users the ability to annotate resources collectively and post details of their annotations as they are made. I believe that Meervisage and the paper that I wrote about the system (see Appendix A) is the first real steps that I have made towards creating a system that would be able to produce significant results for work towards my thesis. At present I am currently working on a paper for the European Semantic Web Conference that carries out evaluation of annotations produced by a group of users within the same social network against annotations produced by a single user.

From the Social Web research that I have carried out I found several issues associated with the level of accessibility that an individual has to sensitive information about another individual. It became apparent that users of several social networking sites were not concerned about the information they were allowing others to view, and more importantly what information they divulged. Following on from this I have described a possible avenue of research associated with identity extraction. At present the only research that has been carried out to address issues such as lateral surveillance and online privacy have been narratives of the issues and commentaries on identity theft and potential pitfalls associated with social networking sites.

The past two years have seen the steady rise and growth of the social networking phenomena, by harnessing this powerful paradigm; communities of already existing users could be used to generate annotations for the semantic web. Metadata could be
generated by enticing users to annotate resources by providing interesting tools with which the users are able to annotate, and interesting and stylish widgets displaying what a user’s past annotations have been to be placed on their page on a social networking site. The community aspect of my research has lead me to believe that the adaptation and use of third party social networking sites would allow the general public outside of the semantic web community to generate and use metadata, but without the knowledge of the underlining backend and the knowledge representation being used.

As RDF is now the ‘language’ of knowledge generated for the semantic web, it should be used throughout when annotations are saved to allow other third party applications access to the same store if they require. The generation of annotations should also be a seamless process with minimum effort; the current trend of annotation tools demonstrates this through their popularity by offering a simple to use interface that offers users the functionality to view, edit and delete their annotations once they have been made.

Other work that I have done alongside my research has included creating an application for geographically plotting individuals within the same social network; Facebook Friend Plotter, and exporting social network information into a simple to aggregate semantic format; Facebook Foaf Generator. The former achieving critical acclaim in a national newspaper with currently over 2300 users45, and the latter being widely used within the FOAF development community. I have also setup a Social Web and Web 2.0 reading group within the department of Computer Science to discuss papers, blogs and the current state of the art46.

Over the next two years I wish to continue my research in the three distinct areas that I have previously outlined: The Semantic Web, Web 2.0 and the Social Web. I believe that continuing work that I have carried out on Meervisage would constitute a large portion of my PhD thesis. I wish to carry out further analysis of social networks, and research the discovery of trusted users within these networks and their role in the generation of semantic metadata. This analysis would follow on from the work carried out on identity extraction to investigate the extraction of trust measures and factors from online profiles that make up a trusted community member.

45 http://www.facebook.com/apps/application.php?id=2255996210
46 http://www.dcs.shef.ac.uk/~mrowe/social-web-reading-group.html
References


Appendices

Appendix A: Meervisage Conference Paper

Meervisage – A Community Based Annotation Tool
Matthew Rowe
Natural Language Group
Web Intelligence Lab
Department of Computer Science
University of Sheffield, UK
m.rowe@dcs.shef.ac.uk

ABSTRACT
The social web has become one of the fastest growing areas of the internet. The increase in use of sites such as Facebook, MySpace, Flickr and other social network based Web sites has been phenomenal. The semantic web is the Web that contains information with underlying semantics, capable of being read automatically by software agents enabling information integration. This paper attempts to tackle the issues associated with the Semantic Web, and how it can be benefited from the incorporation and integration with the Social Web, adding a social aspect to semantics. By investigating the collaborative generation of metadata for the semantic web using social networks as a user base a system named ‘Meervisage’ is presented, capable of socially annotating web resources with semantic metadata.

KEYWORDS
Semantic Web, Social Networks, Social Web, Semantic Annotations, Online Communities, Knowledge Sharing

1. INTRODUCTION
The recent growth of interest in online social networks has been staggering, leading to a massive number of users signing up to use many of the major social networking sites. Facebook reported hitting the 30 million user mark this year 47, making it the most active social networking site in the world today, and one of the most sought after with various offers to buy the site.

Users of Facebook are encouraged to become involved with other users through the various interactive services such as messaging, sharing photos and videos, and communicating on various levels through applications and plug-ins that have been developed by third parties using Facebook’s open API ‘F8’, which was launched this

year\textsuperscript{48}. Facebook offers users the ability to create their own groups, and invite their friends and other users of Facebook to join. The group contains various interactive features including message sharing through the use of a wall which users are able to write on, along with a discussion board and a photo sharing facility.

Research groups from various institutions and organisations have created groups on Facebook to enable collaboration and organisation of research. Members of the group are able to discuss research topics and post relevant items of information that they have found whilst conducting research. By sharing information in this way, the members of the group are able to interact with their colleagues and discuss work in a distributed manner, similar to the practice of research group meetings. This benefits the members of the research group as they are able to socially interact in a private space, but they are also provided with a useful resource when researching. Group members are able to use items posted to the group by other members as reference points for further research, and as a checklist of important research information that has been found.

By posting links to external resources that contain information the research group might find useful, this process could be classed as annotating a specific resource. Any comments or discussions made about the link, along with the user’s description of the link is generating metadata. Although the content of the resource is not internally annotated, the specific URL is annotated in the sense that more information is generated about what it contains by external users.

A semantic web consists of machine understandable metadata located within web resources, however generating this information is one of the key issues for the semantic web. By semantically annotating existing information within web resources, machine understandable metadata is generated allowing information to be used automatically.

The user centric web 2.0 movement has created a shift away from automation towards manual data generation through users. There has been a dramatic increase in the use of folksonomies to annotate web resources and information of interest, leaving behind the old classification schemes of taxonomies.

Socially annotating a resource requires annotating a web resource and sharing the annotation with other users. Various social bookmarking tools are now available, allowing users to bookmark interesting URLs and share them with other users. When bookmarking, Folksonomies are used to annotate the web resource, and the bookmarks are saved to a central repository for global access.

This paper provides a solution to the issue of metadata generation for the semantic web through the use of collaborative annotations using social networks, essentially socially annotating web resources. In chapter two a set of requirements are presented that outline what is required of the annotations produced from socially annotating web resources. Chapter three presents related annotation tools and analyses them according to the list of requirements. Social annotations are then described, and the role of folksonomies when annotating the web. A system is then presented capable of

\textsuperscript{48} http://mashable.com/2007/05/24/facebook-f8-live/
annotating web resources using social networks, allowing the annotations to be shared throughout the network with other users.

2. MOTIVATION
Generating metadata for already existing content on the web has been one of the main focuses of the semantic web. This process of metadata generation is commonly carried out by annotating the content of a web resource with semantics. These semantics must be machine understandable; one of the key components for the realisation of a semantic web, therefore the semantic metadata must be generated according to a predefined formal ontology. The ontology contains explicitly defined concepts and relations, as content from a web resource is annotated, the ontology the concepts within the ontology become populated, and the relations between the concepts become applicable.

This rigid formal structure is in comparison to folksonomies. An ontology contains taxonomy of possible concepts and relations that a web resource can be annotated with. In contrast, folksonomies, or tags, are free words, created from user inputs that do not conform to any predefined specification. Folksonomies are dependent on a user’s viewpoint or perspective of what they are labelling. Folksonomies have become an important part of various web 2.0 sites and applications. Social bookmarking tools use folksonomies as a means of identifying the resource that is being bookmarked. Photo sharing and video sharing websites such as Flickr\(^\text{49}\) and Youtube\(^\text{50}\) allow users to tag photos and videos using folksonomies, therefore assigning metadata to content.

Folksonomies have become a common means of annotating content as they allow the annotator the scope to choose any labels they deem to be relevant. This freedom to choose is argued in [17] as being a very strong component of usability. Offering a taxonomy of possibly concepts is a restrictive method of classification, and one that inhibits the ability of the annotator to think freely.

In order to generate semantic metadata for web content, a system is required that conforms to certain requirements. This section will now describe the requirements for the system.

- **Annotations are shared amongst a community**
  Annotations can be shared and distributed amongst a previously established online community of users. This community could be a social network or an internet message board. The community itself must contain members who are familiar with one another and would benefit from sharing the annotations with one another.

- **Annotations can be reviewed and edited**
  By reviewing annotations, the process of auditing the metadata can take place. The metadata can be verified, and validated, and if required it can then be edited.

\(^{49}\) http://www.flickr.com
\(^{50}\) http://www.youtube.com
Members of the same social network are able to edit annotations posted or created for that group.

- **Collaborative Annotations**

Following on from reviewing and editing annotations, it is also important to allow the content of web resources to be collaboratively annotated. The community of users should be able to work together to generate metadata and then audit the metadata and edit if required.

- **Annotations are stored in a central easily accessible repository**

Annotations will be stored in an easily accessible central repository, not locally. By storing annotations remotely they become globally accessible providing there is access provided to the repository. This allows distributed annotations to be created and shared through the repository.

- **Annotations contain semantic metadata**

Annotations should generate metadata about web content according to predefined semantics. The predefined semantics will not be specified using an ontology, and will instead use folksonomies.

- **Content within web resources is annotated**

The URL of the web resource will not be annotated; instead it will be the content within the web resource allowing specific pieces of information in the form of textual snippets to be annotated. This allows more metadata to be generated about a specific URI, and a greater range of annotated content.

- **Provide a communication layer**

A communication layer would allow users to interact with one another without inhibiting the annotation process. Interactions could take place using external services therefore freeing up resources for their primary use of semantic metadata generation.

3. RELATED WORK

This chapter presents research work closely related to the work presented in this paper. The first section presents the state of the art annotation tools and analyses them according to the requirements mentioned previously, each tools own definition of an annotation is also explained.

Social annotations use folksonomies to generate semantic metadata about the web resource being annotated. The use of folksonomies aids the referencing of retrieved information for other users in the social network. This section presents research work analysing the underlying semantics of these folksonomies, and their social characteristics.
3.1 Annotation Tools

3.1.1 Annotea
The Annotea system [9] uses an open source framework to share annotations amongst other Annotea users. An annotation is part of an infrastructure designed to facilitate metadata generation. In [9] an annotation can take various forms, but in it’s most simplest form it is a “remark about a document identified by a URI, made by the author of the document or by a third party, with or without the author knowledge”.

Further work on the Annotea system [13] states that annotations would not strictly be limited to annotations of content within a page but would also include topics of discussion about the resource, replies to those topics and bookmarks, each annotation type containing metadata referenced by a specified URI. Annotea allows collaborative groups to be created to allow users to work together and annotate specific resources with annotations that only members of that group are able to view.

Although Annotea offers users the ability to annotate as part of a group, the level on interaction is restricted to within an annotation object in the form of discussions and threads. This allows annotations to be shared amongst the community of users of the tool. There is no communication layer to incorporate other levels of interaction, therefore not portraying a community as such and more of a collection of users.

Annotea allows users to review annotations that have been made by other users. Annotations can only be edited by the author, therefore not meeting the requirement for allowing editing of existing annotations by members of the same group. However, the system allows annotations to be reviewed using discussion threads where opinions could be stated about the annotation.

Annotations can be stored on both a central repository, and locally if required. Annotea servers can be created as private repositories allowing only certain users access to the server. Each annotation that is created uses the resource description framework (RDF) to depict the annotation, and the metadata that has been generated. The RDF is then saved to the repository containing semantic information about the author, the date of creation, the URI, associated comments and discussions, and the type of annotation.

3.1.2 Piggy Bank
Piggy Bank generates annotations automatically and manually in the form of RDF from several web sites [10]. Piggy bank is bundled with several screen scrapers; JavaScript scripts that can be run on certain web sites to generate the RDF automatically. If a screen scraper can be used to generate ‘pure’ RDF then the user is notified of this and the option to use the corresponding screen scraper can be selected. If pure RDF cannot be generated then content within the page can be highlighted and annotated. Either method allows users to annotate content within a web page.

Piggy Bank describes an annotation as an individual information item within a page. Unlike Annotea, Piggy Bank annotations are simply metadata relating to some content within a given web page, and are only of one type. Generated RDF is stored in a ‘Semantic Bank’, allowing extracted content to be shared amongst other users, either those users within a group or globally amongst all users. Within the semantic
bank, tags can be assigned to the saved information items, and the information items can be saved to the user’s local semantic bank for later use, not to a central repository.

Semantic metadata is generated following the creation of the annotation containing the author, the date, and folksonomies assigned by the author. In the system folksonomies are regarded as URIs allowing folksonomies with the same name to be differentiated from one another by the difference in context of application.

Annotations can be reviewed by the author and other users of Piggy Bank, however annotations may not be edited by any user other than the author. This inhibits Piggy Bank from meeting the requirement to annotate collaboratively due to the lack of interaction, and functionality to edit existing annotations from different authors. Although there is a global community of users of Piggy Bank, it does not allow sharing of annotations amongst closed groups of users such as a social network.

### 3.1.3 KIM

KIM uses automatic named entity recognition to identify all named entities within a web page, and linking the named entities to a populated knowledge base using an ontology [14]. KIM has a central repository containing all the stored annotations. Recognised named entities are then annotated and assigned hyperlinks that reference named entities that KIM already has corresponding references to in the repository. As KIM has previously found documents that are relevant to a given entity, documents can be indexed and retrieved using that entity. By semantically linking between a new identified entity and corresponding documents associated with that entity, KIM is able to provide indexing and retrieval.

When KIM finds an entity that it does not have a reference for, a new URI is generated for the entity along with a description. Annotations within KIM contain references to semantic metadata within a knowledge base, however annotations are only limited to a singular entity therefore not allowing annotating of large portions of text within a web document.

Annotations are shared globally amongst users of the tool, but not with closed communities of users. Annotations cannot be altered, although they can be removed and KIM can be retrained if an entity is wrongly annotated and must be corrected to stop future incorrect annotations. Although KIM demonstrates how a tool generates metadata automatically using an ontology it is also limited by the lack of social involvement that users of the tool experience.

### 3.1.4 Magpie

Similar to KIM, Magpie automatically annotates named entities found in a web page. It defines this process as associating a Semantic layer with a web resource to allow simple access to any background knowledge that might be relevant [11]. Magpie uses a populated ontology to identify named entities within the web page, identified entities are then annotated and added to the semantic log knowledge base comprising of context entities that have been found on the given page. Unlike KIM, Magpie does not annotate each entity to be hyperlinked to an item in the knowledge base. Instead, each item is assigned a context menu which is triggered by right clicking the
identified entity. The context menu then offers services depending on the class of the entity within the current ontology being used.

One of the key advantages of Magpie is the rich semantic annotations that are generated using an ontology and knowledge base. Annotations, like KIM, are simply named entities within a web resource. The saved metadata contains semantic information about the entity according to the concept it has been resolved to within the ontology. Annotations are shared globally with all users of Magpie, and cannot be edited by users of the tool.

Magpie is another example of automatically annotating web content using an ontology with a populated knowledge base. However, Magpie is limited when analysed from a social web perspective. There is no way of assigning folksonomies to an annotation, or sharing the annotations with other members of the same social network, and therefore no functionality for collaborative annotating.

3.2 Social Annotations

The work by [1] focuses on using already existing information to derive semantics from folksonomies generated using a popular social bookmarking tool. Bookmarks are a form of annotation as they create metadata about a specific resource using folksonomies that the user finds relevant. The work presented in [1] investigates the semantics of folksonomies and their concurrence with other folksonomies. Using this methodology, semantically related bookmarks can be discovered. Although this approach does not inherently use social networks to annotate web resources, it does use an exiting social bookmarking tool; del.icio.us[^1], which contains a community of users and the dataset extracted from it.

Work carried out in [15] investigates the knowledge that is generated from social bookmarking tools and how it could be used to aid resource discovery. Their methodology uses the folksonomies that web resources have been annotated with in order to build a probabilistic model of how the resources are annotated. Similar to the work carried out in [1], the del.icio.us site is used as the source for the folksonomies. The social aspect of this resource is implemented as an ideal knowledge base based on the assumption that users are inclined to write descriptive, yet structured folksonomies to depict the resource they are annotating. The more accurate the tagging, the more helpful the resource will become to the community using the site.

Using the generated folksonomies from social bookmarking tools other work has investigated the distributed nature of folksonomies [16], and how search mechanisms can be improved to find useful web resources. The work presented in [16] states that categorisation using tags is difficult to navigate and is not as rigid as the categorisation system employed using specific taxonomies. This rigidity can lead to problems with expression and interpretation of how to label or class something due to different perspectives from one person to the next [17].

One problem that is apparent with the metadata generated by social annotation tools is the sheer mass. Browsing this amount of data is time-consuming, and laborious. A browsing methodology described in [18] allows easier browsing of large scale social

[^1]: http://del.icio.us
annotation sets using the explicit semantics of the annotations, i.e. The proximity of resources based on assigned folksonomies.

In terms of closed group social annotations [18], and [16] evaluate large scale data sets from a global community. Del.icio.us does not offer the functionality to create groups and share specific annotations with specific groups. All the work that has been carried out using data generated by social annotation tools uses each annotation as an annotation of one web resource, and not sections of the content within the resource. This limits the functionality of the systems, and the ability of the user to generate multiple pieces of metadata about portions of the same page.

4. MEERVISAGE
This chapter presents the Meervisage system, describing an overview of the system and it’s components, and what Meervisage classes as an annotation. The annotation tool used to generate semantic metadata is then described, along with the annotation management component of the system.

4.1 Overview
Meervisage is an annotation tool capable of generating annotations containing semantic metadata using social networks.

To aid the process of collaborative annotating of web documents, Meervisage allows annotations to be shared with users of the same online community within a social networking web site. Unlike [9] communities of users are created using the social networking site, reducing the complexity and management of users. Annotations can be reviewed and edited by members of the same social network, providing a truly collaborative annotation process.

Annotations are stored within central annotation store, allowing remote access to annotations as Figure 1 demonstrates.

Meervisage generates annotations containing semantic metadata describing the author, the social network the annotation has been posted to, the folksonomies assigned to the annotation, and the creation date. Annotations are made from content within any web page, rather than the entirety of the page. End users of Meervisage are able to interact through the use of the social networking functionality provided by the tool.

Meervisage is comprised of two components; Meerkat, and Meervisage as figure 1 shows. Meerkat is responsible for generating semantic metadata by annotating external web resources. Meervisage is responsible for the management of these annotations within the social network.
4.2 Meervisage Annotation

In the Meervisage system an annotation is a semantically rich piece of metadata. The metadata describes the date the annotation was made, the author of the annotation denoted by their username, and the description assigned to the annotation by the author. The metadata also contains folksonomies that the user has entered to label the annotation. The annotation is stored in RDF in the annotation store where it can be accessed by both components of the Meervisage system.

4.3 Meerkat

Meerkat is a browser based annotation tool written for Mozilla Firefox. It works by loading a web page in the browser, and allowing the user to annotate text content within the page using a simple interface integrated into the browser window. To generate metadata about the page, the user is then prompted to enter folksonomies that relate to the content, along with a description.

In order to generate social annotations, the tool accesses Facebook and prompts the user to associate the annotation with a group which they are a member of. If no group is selected then the annotation becomes private to the user and only that user is able to view the annotation.

The tool allows users to browse to web pages that have been previously annotated by other users and view their annotations. Figure 2 demonstrates how this is displayed to the user with highlighted text and a menu allowing the user to edit, view or remove the annotation. As a basis for development, Meerkat allows any user to edit other users’ annotations if they are members of the same group on Facebook.

4.4 Meervisage

Meervisage is a Facebook application that runs within the Facebook platform developed using the Facebook API. By connecting to the annotation repository, it is able to retrieve all annotations for a given user depending on the social networks they belong to. Meervisage presents the semantic metadata assigned to each annotation in a legible format displaying the date the annotation was made, who made the annotation, and the folksonomies and description assigned to the annotation.

---

52 http://www.mozilla.org/firefox
Each Facebook group that a user is a member of has a unique page within the Meervisage application containing all the annotations that have been assigned to this group. The page mimics the standard Facebook group page and contains a wall for posting messages to, a discussion board for thread posting, and a picture area for posting photos to. This gives users an area in which to conduct their research and keep up to date with the annotations that have been posted to the group.

When viewing an annotation, Meervisage allows users to discuss the annotation by starting a discussion thread about the annotation, similar to [9]. The URL of the annotation denotes the subject of the thread. As users discuss more and more about a given URL, the popularity rating of the resource increases effecting how Meerkat functions: If a heavily discussed URL is accessed within a web browser using Meerkat, Meerkat will inform the user that this page has been heavily discussed in a given Facebook group aiding the user in finding relevant information.

Meervisage also provides an RSS feed that users are able to subscribe to. The feed contains information about recently submitted annotations and their related discussion threads. This functionality helps group members to stay up to date with any information that has been retrieved for the group, or any discussions that might have been started.
4.5 Evaluating Meervisage

In order to evaluate the generated semantic metadata from using Meervisage test cases could be created where a single user and a group of five users are informed to annotate as many resources as possible in an allotted time scale. Information Retrieval metrics can then be used to compute the precision and recall of the retrieved information, together with the resulting F measure. Using these formulae (Figure 3-5) is the best way to analyse the quality of the annotations returned and therefore assess the difference between social annotations and single user confined annotations.

\[
\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}
\]

Figure 3 - Precision Formula

\[
\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}
\]

Figure 4 - Recall Formula

\[
F = 2 \cdot (\text{precision} \cdot \text{recall}) / (\text{precision} + \text{recall})
\]

Figure 5 - F-Measure Formula

5. CONCLUSION

This paper presents a solution to the issues associated with generation of metadata for the semantic web and the underlining issue of generating this information socially. This section analyses the requirements stated in chapter two to assess if they have been met, and analyses the system against the state of the art.

- Annotations are shared amongst a community

Meervisage allows users to annotate the content of web resources and share the annotations with other members of the same social network. Annotations can be shared with multiple social networks at once, if the author should wish. Meervisage is limited to only sharing annotations privately and with members of a social network, there is no functionality to allow global annotations to be generated.

Meervisage is unable to publish annotations publicly. It was decided that access to all the annotations from a public store would be laborious and resource intensive on client side computations; highlighting many annotations within the browser at one time. It would also challenge the privacy of users who could be using the tool for their own private research, which they wish no one else to view. Instead a single user is able to save annotations to their own personal store for their own use if they do not

select a group within the social network to assign the annotation to. Although Meervisage demonstrates social annotating of web documents, it is only annotations from a closed group within a social network, and not publicly.

- **Annotations can be reviewed and edited**

The system presented allows annotations to be reviewed and edited as needed. The semantic metadata can be altered to change the folksonomies, or the description, or the social networks the annotation has been assigned to.

- **Collaborative Annotations**

In order to allow collaborative annotating of web resources, Meervisage allows annotations to be edited by all the members of the social network that the annotation has been posted to. This allows social collaboration to take place, and refine annotations if needed. As other members of the social network can review annotations, annotations can therefore be rated according to usefulness and carry more weight when compared to other less rated annotations.

Being able to annotate web resources with an already established community will produce enough metadata to help the semantic web grow. By using established communities within social networks the members are more familiar with one another, lacking the anonymity of Internet message boards and newsgroups.

Meervisage recognises a given resource as being an important information point for the group, affecting how the system operates to inform the user of the usefulness of the current resource. This helps to overcome the problem of redundant data, where already found information is submitted to the group and therefore duplicated.

- **Annotations are stored in a central easily accessible repository**

Meervisage uses a central annotation store storing annotations in RDF. Meervisage allows users to access the annotation store using Ajax\(^{54}\) requests, and returns an XML response containing the details of the annotations. Rather than using a local store for the annotations, Meervisage is therefore able to download annotations remotely.

The most up to date annotation tools use a centralised annotation repository as Meervisage offers. The main difference between Meervisage and state of the art annotation tools is at a low level by using light weight Ajax programming to make the tool as resource liberating as possible.

- **Annotations contain semantic metadata**

As Meervisage annotates web content within the browser, it generates semantic metadata. This metadata contains important information about the annotation describing the author of the annotation, a description of the annotation and the data the annotation was created. The annotation can also include folksonomies, as discussed in chapter three, although folksonomies are not as rigid as ontology concepts they do provide underlining semantics.

\(^{54}\) [http://en.wikipedia.org/wiki/Ajax_(programming)]
Although folksonomies have become a wide spread element of web 2.0 applications, they rely on user perspective, something which may cause problems with other users. Several pieces of work have attempted to define individual folksonomies as concepts of a formal ontology [20]. This is one limitation of Meervisage. It does not provide a formal explicit annotation using an ontology concept, and is limited to a users viewpoint.

- **Content within web resources is annotated**

Unlike social bookmarking tools, and several annotations tools described in chapter three, the system presented allows a more integrated approach to annotating web resources. Content within any given web page can be annotated allowing textual snippets to be assigned a unique URI, and generated metadata.

Several annotation tools are available, capable of annotating the content within a web page, but none offer the functionality to annotate with users from the same social network.

- **Provide a communication layer**

By using an existing social networking site, Meervisage harnesses the functionality available to provide a layer of communication external to the annotation layer. This allows more focus on annotating documents and sharing these annotations. The communication layer is comprised of the interactive services offered by Facebook. Users are able to communicate with other members of the same social network to discuss and review annotations.

This approach allows users to interact within their social clique through a social network group in order to discuss annotations that have been made. Further discussions indicate that the content within the web resource contains useful information, or information that is of great interest to the group.

The current state of the art provides very little work that is equivalent to Meervisage. The closest work is presented in [9] and [13] where annotations can be assigned to groups, however these groups are not already established networks, and do not offer a communication layer for interaction as Meervisage does.

### 7. FUTURE WORK

Annotating textual content within the browser is fairly straight-forward and one dimensional. Annotating multimedia content such as images and videos, to some extent, would provide interesting metadata. Semantically annotating images would be a feasible next stage for the development of the system. The majority of social networking sites offer photo hosting facilities, using the exiting functionality to access information from the social networks would provide a large source of images to be annotated.

As mentioned previously, one of the limitations of the current system is the lack of formal semantics in the annotations that are made. This could be alleviated using an ontology to annotate text. Work has already been carried out investigating the annotation of text using an ontology using a desktop application [19], and through automatic annotation using named entity recognition [12].
Following on from the possibility of annotating web resources using a formal ontology, matching assigned tags with concepts from ontologies mined from the web could produce interesting annotation results. As folksonomies are assigned to particular content, or a particular web resource, the system could learn exactly which ontology concept the defined folksonomies are resolved to by using past resolutions.

Evaluating the annotations produced by Meervisage is work that will be carried out in the future. Using the evaluation metrics described in chapter four will provide a good indication of the efficiency of Meervisage as a semantic metadata generation tool through social collaborations compared to a single user.

8. ACKNOWLEDGEMENTS
I want to thanks the EPSRC for funding my studentship for the entirety of my PhD with the University of Sheffield, and the Web Intelligence Lab within the department of Computer Science at the University of Sheffield.

REFERENCES


Appendix B: Facebook FOAF Generator Output

Output generated by the Facebook FOAF Generator.

<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE rdf:RDF [  
<!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#"  
<!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#"  
<!ENTITY foaf "http://xmlns.com/foaf/0.1/"  
<!ENTITY base "http://www.dcs.shef.ac.uk/~mrowe/foaf.rdf#"  
<!ENTITY contact "mailto:m.rowe@dcs.shef.ac.uk"  
<!ENTITY w11 "http://whatilike.org/ontology" ]>
<rdf:RDF  
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"  
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"  
xmlns:foaf="http://xmlns.com/foaf/0.1/"  
xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"  
xmlns:sn="http://www.dcs.shef.ac.uk/~mrowe/ontologies/social-networks.owl#"  
xmlns:admin="http://webns.net/mvcb/">  
<foaf:PersonalProfileDocument rdf:about="">  
<foaf:maker rdf:resource="#me"/>
<foaf:primaryTopic rdf:resource="#me"/>
<admin:generatorAgent rdf:resource="http://www.dcs.shef.ac.uk/~mrowe/foafgenerator.html"/>
<admin:errorReportsTo rdf:resource="mailto:m.rowe@dcs.shef.ac.uk"/>
</foaf:PersonalProfileDocument>
<foaf:Person rdf:ID="#me">  
<foaf:name>Matthew Rowe</foaf:name>
<foaf:givenname>Matthew</foaf:givenname>
<foaf:family_name>Rowe</foaf:family_name>
<foaf:gender>male</foaf:gender>
<foaf:based_near>  
<geo:Point geo:lat="53.990615" geo:long="-1.539111"/>
</foaf:based_near>
<foaf:holdsAccount>  
<foaf:OnlineAccount>  
<foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
<foaf:accountName>61109237</foaf:accountName>
</foaf:OnlineAccount>
</foaf:holdsAccount>
<foaf:interest>Football</foaf:interest>
<foaf:interest>music</foaf:interest>
<foaf:interest>thai boxing</foaf:interest>
<foaf:interest>gyming it</foaf:interest>
<foaf:interest>tom foolery</foaf:interest>
<foaf:interest>banter</foaf:interest>
<foaf:knows>  
<foaf:Person>
<foaf:name>Matthew Solomon</foaf:name>
<foaf:givenname>Matthew</foaf:givenname>
<foaf:family_name>Solomon</foaf:family_name>
<foaf:holdsAccount>  
<foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
<foaf:accountName>6200736</foaf:accountName>
</foaf:holdsAccount>
</foaf:Person>
</foaf:knows>
</foaf:Person>
</foaf:knows>
<foaf:knows>
  <foaf:Person>
    <foaf:name>Marcus Walsh</foaf:name>
    <foaf:givenname>Marcus</foaf:givenname>
    <foaf:family_name>Walsh</foaf:family_name>
    <foaf:holdsAccount>
      <foaf:OnlineAccount>
        <foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
        <foaf:accountName>10622037</foaf:accountName>
      </foaf:OnlineAccount>
    </foaf:holdsAccount>
  </foaf:Person>
</foaf:knows>

<foaf:knows>
  <foaf:Person>
    <foaf:name>James Jordan</foaf:name>
    <foaf:givenname>James</foaf:givenname>
    <foaf:family_name>Jordan</foaf:family_name>
    <foaf:holdsAccount>
      <foaf:OnlineAccount>
        <foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
        <foaf:accountName>44303603</foaf:accountName>
      </foaf:OnlineAccount>
    </foaf:holdsAccount>
  </foaf:Person>
</foaf:knows>

<foaf:knows>
  <foaf:Person>
    <foaf:name>Caralyn Purvis</foaf:name>
    <foaf:givenname>Caralyn</foaf:givenname>
    <foaf:family_name>Purvis</foaf:family_name>
    <foaf:img rdf:resource="http://profile.ak.facebook.com/v52/1603/10/s61101066_5436.jpg"/>
    <foaf:holdsAccount>
      <foaf:OnlineAccount>
        <foaf:accountServiceHomepage rdf:resource="http://www.facebook.com/"/>
        <foaf:accountName>61101066</foaf:accountName>
      </foaf:OnlineAccount>
    </foaf:holdsAccount>
  </foaf:Person>
</foaf:knows>
</rdf:RDF>