Applications of Semantic Web Methodologies and Techniques to Social Networks and Social Websites

Sheila Kinsella[1], John G. Breslin[1], Alexandre Passant[2], and Stefan Decker[1]

[1] DERI, National University of Ireland, Galway, Ireland
`firstname.lastname@deri.org`
[2] LaLIC, Université Paris-Sorbonne, France
`alexandre.passant@paris4.sorbonne.fr`

Abstract. One of the most visible trends on the Web is the emergence of "Social Web" sites which facilitate the creation and gathering of knowledge through the simplification of user contributions via blogs, tagging and folksonomies, wikis, podcasts, and the deployment of online social networks. The Social Web has enabled community-based knowledge acquisition with efforts like the Wikipedia demonstrating the "wisdom of the crowds" in creating the world's largest online encyclopaedia. Although it is difficult to define the exact boundaries of what structures or abstractions belong to the Social Web, a common property of such sites is that they facilitate collaboration and sharing between users with low technical barriers, although usually on single sites. As more social websites form around the connections between people and their objects of interest, and as these "object-centred networks" grow bigger and more diverse, more intuitive methods are needed for representing and navigating the content items in these sites: both within and across social websites. Also, to better enable user access to multiple sites, interoperability among social websites is required in terms of both the content objects and the person-to-person networks expressed on each site. This requires representation mechanisms to interconnect people and objects on the Social Web in an interoperable and extensible way. The Semantic Web provides such representation mechanisms: it can be used to link people and objects by representing the heterogeneous ties that bind us all to each other (either directly or indirectly). In this paper, we will describe methods that build on agreed-upon Semantic Web formats to describe people, content objects, and the connections that bind them together explicitly or implicitly, enabling social websites to interoperate by appealing to some common semantics. We will also focus on how developers can use the Semantic Web to augment the ways in which they create,reuse, and link content on social networking sites and social websites.

Keywords: Social web, Semantic Web, social networks, social media, FOAF, SIOC, object-centred networks.

1 Introduction

Since the foundations of the Web, it has been used to facilitate communication not only between computers but also between people. Usenet mailing lists and web forums allowed people to connect with each other and communities to form, often

around topics of interest. The social networks formed via these technologies were not explicitly stated, but were implicitly defined by the interactions of the people involved. Later, technologies such as IRC, instant messaging and blogging continued the trend of using the Internet to build communities. Social networking sites - where explicitly-stated networks of friendship form a core part of the website - began to appear around 2002. Since then, the popularity of these sites has grown hugely and continues to do so.

Social networking sites such as Friendster, orkut, LinkedIn and MySpace have become part of the daily lives of millions of users, and generated huge amounts of investment. Boyd and Ellison [8] recently described the history of social networking sites (SNSs), and suggested that in the early days of SNSs, when only the SixDegrees service existed, there simply were not enough users: "While people were already flocking to the Internet, most did not have extended networks of friends who were online". A graph from Internet World Stats[1] shows the growth in the number of Internet users over time. Between 2000 (when SixDegrees shut down) and 2003 (when Friendster became the first successful SNS), the number of Internet users had doubled.

Content-sharing sites with social networking functionality such as YouTube, Flickr and last.fm have enjoyed similar popularity. The basic features of a social networking site are profiles, friend's listings and commenting, often along with other features such as private messaging, discussion forums, blogging, and media uploading and sharing. Many content-sharing sites, such as Flickr and YouTube also include some social networking functionality. In addition to SNSs, other forms of social websites include wikis, forums and blogs. Some of these publish content in structured formats enabling them to be aggregated together.

A limitation of current social websites is that they are isolated from one another like islands in a sea. For example, different online discussions may contain complementary knowledge and topics, segmented parts of an answer that a person may be looking for, but people participating in one discussion do not have ready access to information about related discussions elsewhere. As more and more Social Web sites, communities and services come online, the lack of interoperation among them becomes obvious: a set of single data silos or "stovepipes" has been created, i.e., there are many sites, communities and services that can not interoperate with each other, where synergies are expensive to exploit, and where reuse and interlinking of data is difficult and cumbersome. The main reason for this lack of interoperation is that for the most part in the Social Web, there are still no common standards for knowledge and information exchange and interoperation available. RSS could be a first solution for interoperability among social websites, but it has various limitations that make it difficult to be used efficiently in such a context, as we will see later.

However, the Semantic Web effort aims to provide the tools that are necessary to define extensible and flexible standards for information exchange and interoperability. The Scientific American article from Berners-Lee et al. [3] defined the Semantic Web as "an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation". The last couple of years have seen large efforts going into the definition of the foundational standards supporting data interchange and interoperation, and currently a well-defined
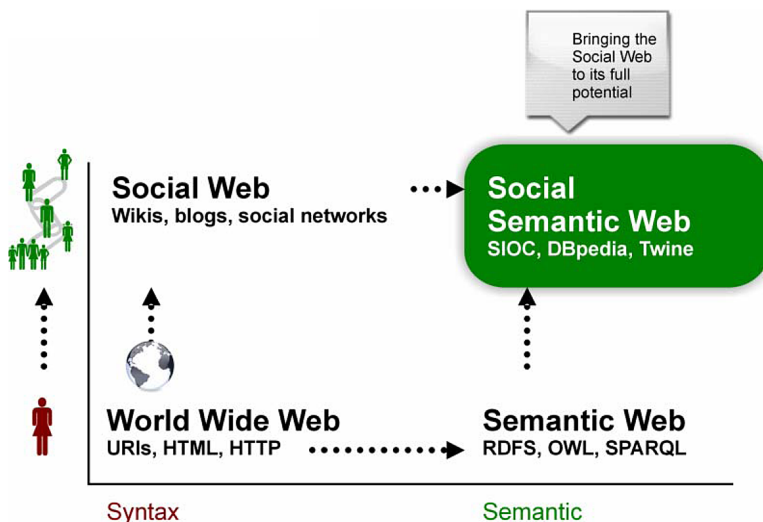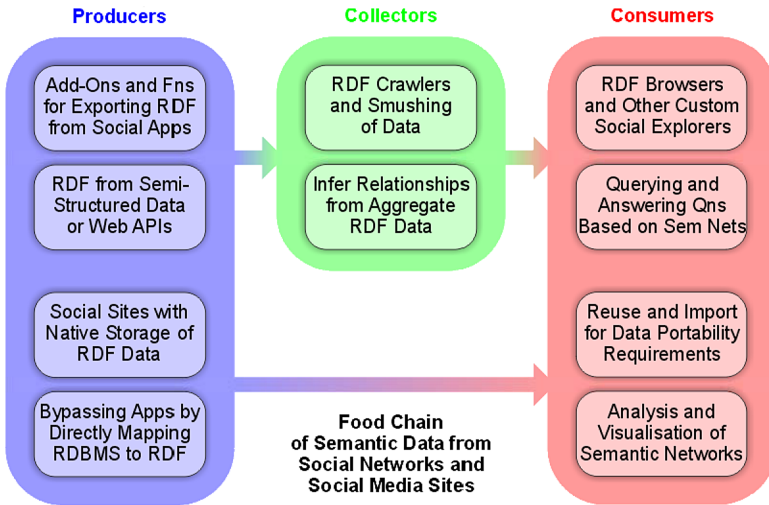
---

[1] http://www.internetworldstats.com/emarketing.htm

Fig. 1. The Social Semantic Web

Semantic Web technology stack exists, enabling the creation of defining metadata and associated vocabularies. The Semantic Web effort is in an ideal position to make Social Web sites interoperable. The application of the Semantic Web to the Social Web can lead to a "Social Semantic Web" (Figure 1), creating a network of interlinked and semantically-rich knowledge. This vision of the Web will consist of interlinked documents and data created by the end users themselves as the result of various social interactions, and it is modelled using machine-readable formats so that it can be used for purposes that the current state of the Social Web cannot achieve without difficulty.

A semantic data "food chain" (see Figure 2), i.e. producers, collectors and consumers of semantic data from social networks and social websites can lead to something greater than the sum of its parts: a Social Semantic Web where the islands of the Social Web can be interconnected with semantic technologies, and Semantic Web applications are enhanced with the wealth of knowledge inherent in user-generated content.

Applying semantic technologies to social websites can greatly enhance the value and functionality of these sites. The information within these sites is forming vast and diverse networks which can benefit from Semantic Web technologies for representation and navigation. Additionally, in order to easily enable navigation and data portability across sites, mechanisms are required to represent data in an interoperable and extensible way. These are termed semantic data producers.

An intermediary step which may or may not be required is for the collection of semantic data. In very large sites, this may not be an issue as the information in the site may be sufficiently linked internally to warrant direct consumption after production, but in general, may users make small contributions across a range of services which can benefit from an aggregate view through some collection service. Collection services can include aggregation and consolidation systems, semantic search engines or data lookup indexes.

Producers

Add-Ons and Fns
for Exporting RDF
from Social Apps

RDF from Semi-
Structured Data
or Web APIs

Social Sites with
Native Storage of
RDF Data

Bypassing Apps by
Directly Mapping
RDBMS to RDF

Collectors

RDF Crawlers
and Smushing
of Data

Infer Relationships
from Aggregate
RDF Data

**Food Chain
of Semantic Data from
Social Networks and
Social Media Sites**

Consumers

RDF Browsers
and Other Custom
Social Explorers

Querying and
Answering Qns
Based on Sem Nets

Reuse and Import
for Data Portability
Requirements

Analysis and
Visualisation of
Semantic Networks

Fig. 2. A food chain for semantic data on the Social Web

The final step involves consumers of semantic data. Social networking technologies enable people to articulate their social network via friend connections. A social network can be viewed as a graph where the nodes represent individuals and the edges represent relations. Methods from graph theory can be use to study these networks, and we will describe how social network analysis can consume semantic data from the food chain.

Also, representing social data in RDF enables us to perform queries on a network to locate information relating to a person or people. Interlinking social data from multiple sources may give an enhanced view of information in distributed communities, and we will describe applications to consume this interlinked data.

In this paper, we will begin by describing various social networking sites and social websites, along with some of their limitations and initial approaches to leverage semantics in social networks, blogs and wikis. We will then describe each of the stages in the semantic data food chain in more detail, giving examples of queries that can be used to consolidate data or extract information from aggregates of data from social websites. Finally, we will give our conclusions and ideas for future work.

2   Social Websites and Approaches to Add Semantics

2.1   Social Networks

The "friend-of-a-friend effect" often occurs when someone tells someone something and they then tell you - linked to the theory that anybody is connected to everybody else (on average) by no more than six degrees of separation. This number of six degrees came from a sociologist called Stanley Milgram who conducted an experiment in the late 1960s. Random people from Nebraska and Kansas were told to send a letter (via intermediaries) to a stock broker in Boston. However, they could only give the

letter to someone that they knew on a first-name basis. Amongst the letters that found their target (around 20%), the average number of links was around 5.5 (rounded up to 6). Some other related ideas include the Erdös number (the number of links required to connect scholars to mathematician Paul Erdös, a prolific writer who co-authored over 1500 papers with more than 500 authors), and the Kevin Bacon game (the goal is to connect any actor to Kevin Bacon, by linking actors who have acted in the same movie).

It is often found that even though one route is followed to get in contact with a particular person, after talking to them there is another obvious connection that was not previously known about. This is part of the small-world network theory [28], which says that most nodes in a network exhibiting small-world characteristics (such as a social network) can be reached from every other node by a small number of hops or steps.

There has been a proliferation of social networking sites (SNSs) which Boyd and Ellison [8] define as a category of websites consisting of user profiles, which other users can comment on, and a traversable social network originating from publicly articulated lists of friends. The idea behind such services is to make people's real-world relationships explicitly defined online - whether they be close friends, business colleagues or just people with common interests. Most SNSs allow one to surf from a list of friends to find friends-of-friends, or friends-of-friends-of-friends for various purposes. While the majority of these sites are for purely social reasons, others have additional purposes such as LinkedIn which is targeted towards professionals.

Before 2002, most people networked using online services such as OneList, ICQ or eVite. The first big SNS in 2002 was Friendster; in 2003, LinkedIn (a SNS for professionals) and MySpace (a band-oriented service) appeared; then in 2004, orkut (Google's SNS) and Facebook (by a college student for college students) were founded; these were followed by Bebo (targeting both high school and college students) in 2005. Social networking services usually offer the same basic functionalities: network of friends listings (showing a person's "inner circle"), person surfing, private messaging, discussion forums or communities, events management, blogging, commenting (sometimes as endorsements on people's profiles), and media uploading. In general, these sites do not usually work together and therefore require you to re-enter your profile and redefine your connections when you register for each new site.

Some motivations for SNS usage include building friendships and relationships, arranging offline meetings, curiosity about others, arranging business opportunities, or job hunting. People may want to meet with local professionals, create a network for parents, network for social (dating) purposes, get in touch with a venture capitalist, or find out if they can link to any famous people via their friends.

A key feature of these sites is community-contributed content that may be tagged and can be commented on by others. That content can be virtually anything: blog entries, board posts, videos, audio, images, wiki pages, user profiles, bookmarks, events, etc. Already, sites are being proposed where live multiplayer video games will appear in browser-embedded windows just as YouTube does for videos, with running commentaries going on about the games in parallel. Tagging is common to many social networking websites - a tag is a keyword that acts like a subject or category for the associated content. Folksonomies - collaboratively generated, open-ended labelling systems - emerge from the use of tagging on a given platform and enable users of these sites to categorise content using the tags system, and to thereby visualise popular tag usages via

"tag clouds" (visual depictions of the tags used on a particular website, similar to a weighted list in visual design, that provides an overview of the different categories and topics used within a community).

Even in a small-sized SNS, there can be a lot of links available for analysis, and this data is usually meaningless when viewed as a whole, so one usually needs to apply some social network analysis (SNA) techniques[2]. Apart from comprehensive textbooks in this area [27], there are many academic tools for examining social networks and performing common SNA routines. For example, the tool Pajek[3] [2] can be used to drill down into various social networks. A common method is to reduce the amount of relevant social network data by clustering. One can choose to cluster people by common friends, by shared interests, by geography, by tags, etc.

In social network analysis, people are modelled as nodes or "actors". Relationships (such as acquaintanceship, co-authorship, friendship, etc.) between actors are represented by lines or edges. This model allows analysis using existing tools from mathematical graph theory and mapping, with target domains such as movie actors, scientists and mathematicians (as already mentioned), sexual interaction, phone call patterns or terrorist activity. There are some useful tools for visualising these models, such as Vizster[4] by Heer and Boyd [19], based on the Prefuse[5] open-source toolkit.

2.2   Leveraging Semantics in "Object-Centred" Social Networks

Jyri Engeström, co-founder of the micro-blogging site Jaiku, has theorised[6] that the longevity of social websites is proportional to the "object-centred sociality" occurring in these networks, i.e. the degree to which people are connecting via items of interest related to their jobs, workplaces, favourite hobbies, etc. On the Web, social connections are formed through the actions of people - via the content they create together, comment on, link to, or for which they use similar annotations. For many of the social websites, success has come from enabling communities formed around common interests, where the users are active participants who as well as consuming information also provide content and metadata. In this way, it is probable that people's SNS methods will continue to move closer towards simulating their real-life social interaction, so that people will meet others via something they have in common, not by randomly approaching each other- eventually leading towards more realistic interaction methods with friends à la virtual worlds like Second Life.

As more social networks form around connections between people and their objects of interest, and as these object-centred social networks grow bigger and more diverse, more intuitive methods are needed for representing and navigating the information in these networks - within and across social networking sites. Also, to better enable navigation across sites, interoperability among SNSs is required in terms of both the content objects and the person-to-person networks expressed on each site. That requires representation mechanisms to interconnect people and objects on the Web in an interoperable, extensible way [10].

---

[2] http://lrs.ed.uiuc.edu/tse-portal/analysis/social-network-analysis/

[3] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[4] http://jheer.org/vizster/

[5] http://prefuse.org/

[6] http://www.zengestrom.com/blog/2005/04/why_some_social.html

Semantic Web representation mechanisms are ideally suited to describing people and the objects that link them together in such object-centred networks, by recording and representing the heterogeneous ties that bind each to the other. By using agreed-upon Semantic Web formats to describe people, content objects, and the connections that bind them together, social networks can also interoperate by appealing to common semantics. Developers are already using Semantic Web technologies to augment the ways in which they create, reuse, and link content on social networking and social websites. These efforts include the Friend-of-a-Friend (FOAF) project[7], the Nepomuk social semantic desktop[8], and the Semantically-Interlinked Online Communities (SIOC) initiative[9]. Some SNSs, such as Facebook, are also starting to provide query interfaces to their data, which others can reuse and link to via the Semantic Web[10].

The Semantic Web is a useful platform for linking and for performing operations on diverse person- and object-related data gathered from heterogeneous social networking sites. In the other direction, object-centred networks can serve as rich data sources for Semantic Web applications. This linked data can provide an enhanced view of individual or community activity in localised or distributed object-centred social networks. In fact, since all this data is semantically interlinked using well-given semantics (e.g. using the FOAF and SIOC ontologies), in theory it makes no difference whether the content is distributed or localised. All of this data can be considered as a unique interlinked machine-understandable graph layer (with nodes as users and related data and arcs as relationships) over the existing Web of documents and hyperlinks, i.e. a Giant Global Graph as Tim Berners-Lee recently coined[11]. Moreover, such interlinked-data allows advanced querying capabilities, for example, "show me all the content that Alice has acted on in the past three months".

As Tim Berners-Lee said in a 2005 podcast[12], Semantic Web technologies can support online communities even as "online communities ... support Semantic Web data by being the sources of people voluntarily connecting things together". For example, SNS users are already creating extensive vocabularies and annotations through folksonomies [24]. Because a consensus of community users is defining the meaning, these terms are serving as the objects around which those users form more tightly-connected social networks.

2.3 Blogs

A blog, or weblog, is a user-created website consisting of journal style entries displayed in reverse-chronological order. Entries may contain text, links to other websites, and images or other media. Often there is a facility for readers to leave comments on individual entries. Blogs may be written by individuals, or by groups of contributors. A blog may function as a personal journal, or it may provide news or opinions on a particular subject.

---

[7] http://www.foaf-project.org/

[8] http://nepomuk.semanticdesktop.org/

[9] http://sioc-project.org/

[10] http://www.openlinksw.com/blog/~kidehen/?id=1237

[11] http://dig.csail.mit.edu/breadcrumbs/node/215

[12] http://esw.w3.org/topic/IswcPodcast

The growth and takeup of blogs over the past five years has been impressive, with a doubling in the size of the "blogosphere" every six or so months (according to statistics from Technorati[13]). Over 100,000 blogs are created every day, working out at about one a second. Nearly 1.5 million blog posts are being made each day, with over half of bloggers still contributing to their sites three months after the blog's creation.

RSS feeds are also a useful way of accessing information from your favourite blogs, but they are usually limited to the last 15 entries, and do not provide much information on exactly who wrote or commented on a particular post, or what the post is talking about. Some approaches like SIOC (more later) aim to enhance the semantic metadata provided about blogs, forums and posts, but there is also a need for more information about what exactly a person is writing about. Blog entries often refer to resources on the web and these resources will usually have a context in which they are being used could be described. For example a post which critiques a particular resource could incorporate a rating, or a post announcing an event could include start and end times.

When searching for particular information in or across blogs, it is often not that easy to get it because of "splogs" (spam blogs) and also because of the fact that the virtue of blogs so far has been their simplicity - apart from the subject field, every-thing and anything is stored in one big text field for content. Keyword searches may give some relevant results, but useful questions such as "find me all the Chinese res-taurants that bloggers reviewed in Dublin with a rating of at least 5 out of 10" cannot be posed, and you cannot easily drag-and-drop events or people or anything (apart from URLs) mentioned in blog posts into your own applications.

2.4 Adding Semantics to Blogs

There have been some approaches to tackle the issue of adding more information to blog posts, so that queries can be made and the things that people talk about can be reused in other posts or applications (because not everyone is being served well by the lowest common denominator that we currently have in blogs). One approach is called "structured blogging"[14] and the other is "semantic blogging".

Structured blogging is an open-source community effort that has created tools to provide microcontent (including microformats[15] like hReview) from popular blogging platforms such as WordPress and Moveable Type. Although the original effort has ta-pered off, structured blogging is continuing through services like LouderVoice[16]. In structured blogging, packages of structured data are becoming post components. Some-times (not all of the time) a person will have a need for more structure in their posts - if they know a subject deeply, or if their observations or analyses recur in a similar manner throughout their blog - then they may best be served by filling in a form (which has its own metadata and model) during the post creation process. For example, someone may be writing a review of a film they went to see, or reporting on a sports game they at-tended, or creating a guide to tourist attractions they saw on their travels. Not only do people get to express themselves more clearly, but blogs can start to interoperate with enterprise applications through the microcontent that is being created in the background.

---

[13] http://technorati.com/weblog/2007/04/328.html

[14] http://structuredblogging.org/

[15] http://microformats.org

[16] http://www.loudervoice.com/

Take the scenario where someone (or a group of people) is reviewing some soccer games that they watched. Their after-game soccer reports will typically include information on which teams played, where the game was held and when, who were the officials, what were the significant game events (who scored, when and how, or who received penalties and why, etc.) - it would be easier for these blog posters if they could use a tool that would understand this structure, presenting an editing form with the relevant fields, and automatically create both HTML and RSS with this structure embedded in it. Then, others reading these posts could choose to reuse this structure in their own posts, and their blog reading / writing application could make this structure available when the blogger is ready to write. As well as this, reader applications could begin to answer questions based on the form fields available – "show me all the matches from South Africa with more than two goals scored", etc.

At the moment, structured blogging tools provide a fixed set of forms that bloggers can fill in for things like reviews, events, audio, video and people - but there is no reason that people could not create custom structures, and news aggregators or readers could auto-discover an unknown structure, notify a user that a new structure is available, and learn the structure for reuse in the user's future posts.

Semantic Web technologies can also be used to ontologise any available post structures for more linkage and reuse. Blog posts are usually only tagged on the blog itself by the post creator, using free-text keywords such as "scotland", "movies", etc. (or can be tagged by others using social bookmarking services like del.icio.us or personal aggregators like Gregarius). Technorati, the blog search engine, aims to use these keywords to build a "tagged web". Both tags and hierarchical categorisations of blog posts can be further enriched using the SKOS framework. However, there is often much more to say about a blog post than simply what category it belongs in.

This is where semantic blogging comes in. Traditional blogging is aimed at what can be called the "eyeball Web" - i.e. text, images or video content that is targeted mainly at people. Semantic blogging aims to enrich traditional blogging with metadata about the structure (what relates to what and how) and the content (what is this post about - a person, event, book, etc.). Already RSS and Atom are used to describe blog entries in a machine-readable way and enable them to be aggregated together. However by augmenting this data with additional structural and content-related metadata, new ways of querying and navigating blog data become possible.

In structured blogging, microcontent such as microformats or RDFa is positioned inline in the HTML (and subsequent syndication feeds) and can be rendered via CSS. Structured blogging and semantic blogging do not compete, but rather offer metadata in slightly different ways (using microcontent and RDF respectively). There are already mechanisms such as GRDDL which can be used to move from one to the other and allows one to provide RDF data from embedded RDFa or microformats. Extracted RDF data can be then reused as would any native RDF data, and so it may be processed using common Semantic Web tools and services.

The question remains as to why one would choose to enhance their blogs and posts with semantics. Current blogging offers poor query possibilities (except for searching by keyword or seeing all posts labelled with a particular tag). There is little or no reuse of data offered (apart from copying URLs or text from posts). Some linking of posts is possible via direct HTML links or trackbacks, but again, nothing can be said about the nature of those links (are you agreeing with someone, linking to an interesting post, or

are you quoting someone whose blog post is directly in contradiction with your own opinions?). Semantic blogging aims to tackle some of these issues, by facilitating better (i.e. more precise) querying when compared with keyword matching, by providing more reuse possibilities, and by creating "richer" links between blog posts.

It is not simply a matter of adding semantics for the sake of creating extra metadata, but rather a case of being able to reuse what data a person already has in their desktop or web space and making the resulting metadata available to others. People are already (sometimes unknowingly) collecting and creating large amounts of structured data on their computers, but this data is often tied into specific applications and locked within a user's desktop (e.g. contacts in a person's address book, events in a calendaring application, author and title information in documents, audio metadata in MP3 files). Semantic blogging can be used to "lift" or release this data onto the Web, as in the semiBlog[17] application (now called Shift) which allows users to reuse metadata from Apple Mac desktops in blog posts. For example, Aidan can write a blog post which he annotates using metadata about events and people from his desktop calendaring and address book applications. He publishes this post onto the Web, and John, reading this post, can reuse the embedded metadata in his own desktop applications. As well as semiBlog, other semantic blogging systems have been developed by HP[18], the National Institute of Informatics, Japan[19] and MIT[20].

Also, conversations often span multiple blog sites in blog posts and their comments, and bloggers may respond to the entries of other users in their own blogs. The use of semantic technologies can also enable the tracking of these distributed conversations. Links between units of conversation could even be enhanced to include sentiment information, e.g. who agrees or disagrees with the initial opinion.

2.5 Wikis

A wiki is a website which allows users to edit content through the same interface they use to browse it, usually a web browser, while some desktop-based wikis also exist. This facilitates collaborative authoring in a community, especially since editing a wiki does not require advanced technical skills. A wiki consists of a set of web pages which can be connected together by links. Users can create new pages, and change existing ones, even those created by other members. As well as the Wikipedia online encyclopaedia, wikis are being used for free dictionaries, book repositories, event organisation, and software development. They have become increasingly used in enterprise environments for collaborative purposes: research projects, papers and proposals, coordinating meetings, etc. SocialText[21] produced the first commercial opensource wiki solution, and many companies now use wikis as one of their main intranet collaboration tools.

There are hundreds of wiki software systems now available, ranging from MediaWiki, the software used on the Wikimedia family of sites, and PurpleWiki, where fine grained elements on a wiki page are referenced by purple numbers, to OddMuse,

---

[17] http://semiblog.semanticweb.org/

[18] http://www.hpl.hp.com/personal/Steve_Cayzer/semblog.htm

[19] http://www.semblog.org/

[20] http://theory.csail.mit.edu/~dquan/iswc2004-blog.ppt

[21] http://www.socialtext.com/

a single Perl script wiki install, and WikidPad, a desktop-based wiki for managing personal information. Many are open source, free, and will often run on multiple operating systems. The differences between wikis are usually quite small but can include the development language used (Java, PHP, Python, Perl, Ruby, etc.), the database required (MySQL, flat files, etc.), whether attachment file uploading is allowed or not, spam prevention mechanisms, page access controls, RSS feeds, etc.

The Wikipedia project consists of over 250 different wikis, corresponding to a variety of languages. The English-language one is currently the biggest, with over 2 million pages, but there are wikis in languages ranging from Gaelic to Chinese. A typical wiki page will have two buttons of interest: "Edit" and "History". Normally, anyone can edit an existing wiki article, and if the article does not exist on a particular topic, anyone can create it. If someone messes up an article (either deliberately or erroneously), there is a revision history so that the contents can be reverted or fixed by the community. Thus, while there is no pre-defined hierarchy in most wikis, content is auto-regulated thanks to an emergent consensus within the community, ideally in a democratic way (for instance, most wikis include discussions pages where people can discuss sensible topics). There is a certain amount of ego-related motivation in contributing to a wiki - people like to show that they know things, to fix mistakes and fill in gaps in underdeveloped articles (stubs), and to have a permanent record of what they have contributed via their registered account. By providing a template structure to input facts about certain things (towns, people, etc.), wikis also facilitate this user drive to populate wikis with information.

2.6 Adding Semantics to Wikis

Typical wikis usually enable the description of resources in natural language. By additionally allowing the expression of knowledge in a structured way, wikis can provide advantages in querying, managing and reusing information. Wikis such as the Wikipedia have contained structured metadata in the form of templates for some time now (to provide a consistent look to the content placed within article texts), but there is still a growing need for more structure in wikis. Templates can also be used to provide a structure for entering data, so that it is easy to extract metadata about the topic of an article (e.g. from a template field called "population" in an article about London). Semantic wikis bring this to the next level by allowing users to create semantic annotations anywhere within a wiki article text for the purposes of structured access and finer-grained searches, inline querying, and external information reuse. Generally, those annotations are designed to create instances and properties of domain ontologies (either explicit ontologies or ontologies that will emerge from the usage of the wiki itself), whereas other wikis use semantic annotations to provide advanced metadata regarding wiki pages. There are already about 20 semantic wikis in existence, and one of the largest ones is Semantic MediaWiki, based on the popular MediaWiki system. Semantic MediaWiki allows for the expression of semantic data describing the connection from one page to another, and attributes or data relating to a particular page.

Let us take an example of providing structured access to information in wikis. There is a Wikipedia page about JK Rowling that has a link to "Harry Potter and the Philosopher's Stone" (and to other books that she has written), to Edinburgh because she lives there, and to Scholastic Press, her publisher. In a traditional wiki, you cannot

perform fine-grained searches on the Wikipedia dataset such as "show me all the books written by JK Rowling", or "show me all authors that live in the UK", or "what authors are signed to Scholastic", because the type of links (i.e. the relationship type) between wiki pages are not defined. In Semantic MediaWiki, you can do this by linking with [[author of::Harry Potter and the Philosopher's Stone]] rather than just the name of the novel. There may also be some attribute such as [[birthdate:=1965-07-31]] which is defined in the JK Rowling article. Such attributes could be used for answering questions like "show me authors over the age of 40" or for sorting articles, since this wiki syntax is translated into RDF annotations when saving the wiki page. Moreover, page categories are used to model the related class for the created instance.

Since Semantic MediaWiki is completely open in terms of the wiki syntax for annotating content, extracted data may be subject to heterogeneity problems. For instance, some users will use [[author of:xxx]] while others will prefer [[has written:xxx]], leading to problems when querying data. Other wikis such as OntoWiki, IkeWiki or UfoWiki assist the user when modelling semantic annotations, in order to avoid those heterogeneity issues and provide data that is based on pre-defined ontologies.

Some semantic wikis also provide what is called inline querying. A question such as "?page dc:creator EyalOren" (or find me all pages where the creator is Eyal Oren) is processed as a query when the page is viewed and the results are shown in the wiki page itself. Also, when defining some relationships and attributes for a particular article (e.g. "foaf:gender Male"), other articles with matching properties can be displayed along with the article. Moreover, some wikis feature reasoning capabilities, for example, retrieving all instances of foaf:Person when querying for a list of all foaf:Agent(s) since the first class subsumes the second one in the FOAF ontology.

Finally, just as in the semantic blogging scenario, wikis can enable the Web to be used as a clipboard, by allowing readers to drag structured information from wiki pages into other applications (for example, geographic data about locations on a wiki page could be used to annotate information on an event or a person in your calendar application or address book software respectively).

2.7 Tags, Tagging and Folksonomies

Apart from providing a means to define and manage social networks, one of the most important features of social websites is the ability to upload and share content with others, either with anyone subscribed to (or just browsing) the website or else within a restricted community. Various media files can be shared, such as pictures, videos, bookmarks, slides, etc. In order to make this content more easily discoverable, users can add free-text keywords, or tags, to any content that they upload. For example, this chapter could be tagged with 'semanticweb', 'socialnetworks', 'sioc' on a scientific bibliography management system such as bibsonomy.org. While the same content can be tagged by various users on the same system, anyone can use their own tags. Yet, most services suggest existing tags for a given item when someone begins tagging it.

The main advantage of tagging for end-users is that one does not have to learn a predefined organisation scheme (such as a hierarchy or taxonomy) and one can use the keywords that exactly fit with his or her needs. Websites that support tagging benefit from the "wisdom of the crowds" effect. Tags evolve quickly according to the needs of the users, and these tags, combined with the tagging actions and the frequency with

which they are used, lead to the emergence of a folksonomy, i.e. a user-driven, open and evolving classification scheme. Moreover, tags can be used for various purposes and [17] has identified seven different functions that tags can play for end users, from topic definition to opinion forming and even self-reference.

In spite of its advantages when annotating content, tagging leads to various issues in information retrieval. Since a single tag can refer to various concepts, it can lead to ambiguity. For instance, 'paris' can refer to a city in France, a city in the USA or even a person. Moreover, various tags can be used to define the same idea, so that a user must run various queries to get the content related to a given concept. Such heterogeneity is mainly caused by the multilingual nature of tags (e.g. 'semanticweb' and 'websemantique') but also due to the fact people will use acronyms or shortened versions ('sw' and 'semweb'), as well as linguistic and morpho-syntactic variations (synonyms, plurals, case, etc.). Finally, since a folksonomy is essentially a flat organisation of tags, the lack of relationships between tags makes it difficult to suggest related content.

2.8   Adding Semantics to Tags and Related Objects

Numerous works related to the links between tags, the tagging process, folksonomies and the Semantic Web have been published during the last couple of years. We can divide these into two general approaches: the ones aiming to define, mine or automatically link to ontologies from existing folksonomies, and works based on defining Semantic Web models for tags and related objects (e.g., tagging, tag clouds, etc.).

The first set of approaches is based on the idea that emergent semantics naturally appears through the use of tags, relying on various methods to achieve this goal. For example, [26] combines automatic tag filtering, clustering and mapping with ontologies already available on the Web in order to extract ontologies from existing folksonomies in a completely-automated approach. Another approach involving a social aspect is the one defined by [24], which uses social network analysis to extract ontologies from the Flickr folksonomy, based on the way that the community shares and uses tags.

Regarding the second approach, various models have been proposed to define Semantic Web vocabularies for tagging. Representing tags using Semantic Web technologies offer various advantages: providing a uniform, machine-readable and extendable way to represent tags as well as other concepts such as tagging actions, tag clouds, the relationships between tags and the meanings that they carry. While tag-based search is the only way to retrieve tagged content at the moment (and leads to the aforementioned problems), these new models allow advanced querying capabilities such as "retrieve all the content tagged with something relevant to the Semantic Web field" or "give me all the tags used by Bob on Flickr and Alice on del.icio.us". Moreover, having tags and tagged content published in RDF allows one to easily link to it from other Semantic Web data, and to reuse it across applications.

The Tag Ontology[22] provides an initial model to represent tags and tagging actions in RDF, based on the ideas of Gruber [18] and on a common mathematical model of tagging that defines it as a tripartite relationship involving a "Tag", a "User", and a tagged "Resource". This ontology defines the Tag class by sub-classing skos:Concept,

---

[22] http://www.holygoat.co.uk/projects/tags/

which means that each tag has a given URI. This offers the ability to interlink tags together with semantic relationships, as this model permits. SCOT [20] aims to represent tag clouds, and so defines a model to represent the use and co-occurrence of tags on a given social platform, allowing one to move his or her tags from one service to another and to share tags with others. Finally, MOAT [30] aims to represent the meaning of tags using URIs of existing domain ontology instances from existing public knowledge bases (such as Geonames or DBPedia). It also provides a framework using this model, the goal of which is to let people easily bridge the gap between simple free-text tagging and semantic indexing.

Some tools already used some of these models to provide advanced and more precise querying tag-based capabilities to their users, including Gnizr, SweetWiki and int.ere.st.

3   Producers of Social Semantic Data

Applying Semantic Web technologies to online social spaces allows for the expression of different types of relationships between people, objects and concepts. By using common, machine-readable ways of expressing individuals, profiles, social connections, and content, they provide a way to interconnect people and objects on the Web in an interoperable, extensible way.

On the conventional Web, navigation of social data across sites can be a major challenge. Communities are often dispersed across numerous different sites and platforms. For example, a group of people interested in a particular topic may share photos on Flickr, bookmarks on del.icio.us and hold conversations on a discussion forum. Additionally, a single person may hold several separate online accounts, and may have a different network of friends on each. The information existing in these spaces is generally disconnected, lacking in semantics, and centrally controlled by single organisations. Individuals generally lack control or ownership of their own data.

Social spaces on the Web are becoming bigger and more distributed. This presents new challenges for navigating such data. Machine-readable descriptions of people and objects, and the use of common identifiers, would allow for linking diverse information from heterogeneous social networking sites. This would create a starting point for easy navigation across the information in these networks.

The use of common formats allows interoperability across sites, enabling users to reuse and link to content across different platforms. This also provides a basis for data portability, where users could have ownership and control over their own data and could move profile and content information between services as they wish. Recently there has been a push within the web community to make data portability a reality.

Additionally, the Social Web and social networking sites can contribute to the Semantic Web effort. Users of these sites often provide metadata in the form of annotations and tags on photos, links, blogs posts etc. social networks and semantics can complement each other. Already within online communities, common vocabularies or folksonomies for tagging are emerging through of a consensus of community members.

There are also a number of semantically-enabled social applications appearing that have been enhanced with extra features due to the rich content being created in social

software tools by users. The Twine application from Radar Networks is a recent example of a system that leverages both the explicit (tags and metadata) and implicit semantics (auto tagging of text) associated with content items. Twine is a "knowledge networking" application that allows users to share, organise, and find information with people they trust. People create and join "twines" (community containers) around certain topics of interest, and items (documents, bookmarks, media files, etc., that can be commented on) are posted to these containers through a variety of methods. The underlying semantic data can be exposed as RDF by appending "?rdf" to any Twine URL. The DBpedia represents structured content from the collaboratively-edited Wikipedia in semantic form, leveraging the semantics from many social content contributions by multiple users. DBpedia allows you to perform semantic queries on this data, and enables the linking of this socially-created data to other datasets on the Web by exposing it via RDF. Revyu.com combines Web 2.0 interfaces and principles such as tagging with Semantic Web modelling principles to provide a reviews website that is integrated with Linked Data principles. Anyone can review objects defined on other services (such as a movie from DBpedia), and the whole content of the website is available in RDF, therefore it is available for reuse by other applications.

3.1 FOAF

Semantic Web technologies allow for a more expressive description of a social network, enabling the use of heterogeneous nodes and link denoting different types of objects and different types of relationships. This enables us to express a model of an object-centred network where content and other items of interest can be described along with people.

The Friend-of-a-Friend (FOAF) project was started in 2000 and defines a widely-used vocabulary for describing people and the relationships between them, as well
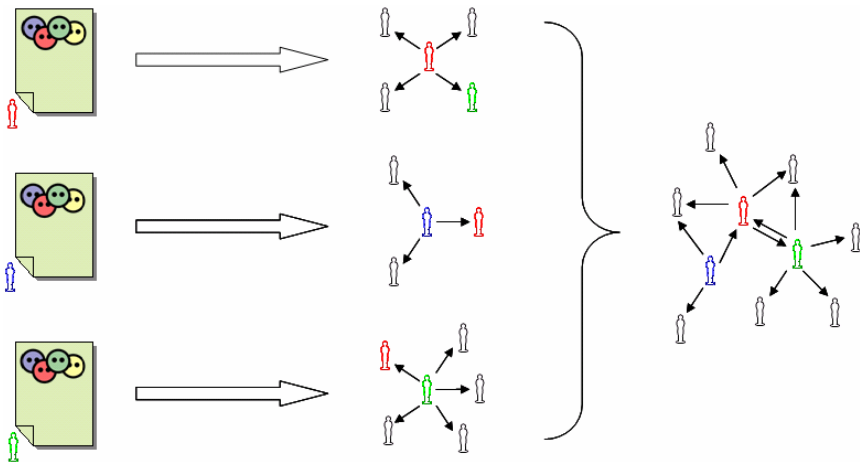
Fig. 3. Integrating social networks by using FOAF as a common representation format and having unique URIs for people

as the things they create and do. Anyone can create their own FOAF file describing themselves and their social network, and the information from multiple FOAF files can easily be combined to obtain a higher-level view of the network across various sources, as shown in Figure 3. This means that a group of people can articulate their social network without the need for a single centralised database.

FOAF can be integrated with any other Semantic Web vocabularies, such as SIOC, SKOS, etc. Some prominent social networking services that expose data using FOAF include hi5, LiveJournal, Vox, Pownce and MyBlogLog. People can also create their own FOAF document and link to it from their homepage, and exporters are available for some major social websites as Flickr, Twitter an Facebook. Such FOAF documents usually contain personal information, links to friends, and other related resources.

The knowledge representation of a person and their friends would be achieved through a FOAF fragment similar to that below.

```
<foaf:Person rdf:about="#JB">
  <foaf:name>John Breslin</foaf:name>
  <foaf:mbox rdf:resource="mailto:john.breslin@deri.org" />
  <foaf:homepage rdf:resource="http://www.johnbreslin.com/" />
  <foaf:nick>Cloud</foaf:nick>
  <foaf:depiction
rdf:resource="http://www.johnbreslin.com/images/foaf_photo.jpg" />
  <foaf:interest>
    <rdf:Description rdf:about=" http://dbpedia.org/resource/SIOC"
rdfs:label="SIOC" />
  </foaf:interest>
  <foaf:knows>
    <foaf:Person>
      <foaf:name>Sheila Kinsella</foaf:name>
      <foaf:mbox rdf:resource="mailto:sheila.kinsella@deri.org" />
    </foaf:Person>
  </foaf:knows>
  <foaf:knows>
    <foaf:Person>
      <foaf:name>Stefan Decker</foaf:name>
      <foaf:mbox rdf:resource="mailto:stefan.decker@deri.org" />
    </foaf:Person>
  </foaf:knows>
</foaf:Person>
```

The evolving requirement for distributed social networks and reusable profiles, as highlighted by efforts such as DataPortability.org, DiSo and Google's Social Graph API, can be realised through open standards like FOAF. There have been a lot of complaints in recent years about the walled gardens that are social network sites. Some of the most popular SNSs would not exist without the walled garden approach, but some flexibility would be useful. Users may have many identities on different social networks, where each identity was created from scratch. A reusable profile would allow a user to import their existing identity and connections (from their own homepage or from another site they are registered on), thereby forming a single global identity with different views.

The structure of the social network formed by relations expressed in FOAF documents on the Web has been studied in [11], particularly the small-world characteristics of the graph.

3.2 SIOC

The SIOC initiative is aimed at interlinking related online community content from platforms such as blogs, message boards, and other social websites. In combination with the FOAF vocabulary for describing people and their friends, and the Simple Knowledge Organisation Systems (SKOS) model for organizing knowledge, SIOC lets developers link discussion posts and content items to other related discussions and items, people (via their associated user accounts), and topics (using specific "tags" or hierarchical categories). As discussions begin to move beyond simple text-based conversations to include audio and video content, SIOC is evolving to describe not only conventional discussion platforms but also new Web-based communication and content-sharing mechanisms.

Since disconnected social websites require ontologies for interoperation, and due to the fact that there is a lot of social data with inherent semantics contained in these sites, there is potential for high impact through the successful deployment of SIOC. Many online communities still use mailing lists and message boards as their main communication mechanisms, and the SIOC initiative has created a number of data producers for such systems in order to lift these communities to the Semantic Web. As well as having applications to social websites, there is a parallel lack of integration between social software and other systems in enterprise intranets. So far, SIOC has been adopted in a framework of 50 applications or modules[23] deployed on over 400 sites.

A sample fragment of SIOC RDF is shown below, representing a blog post, its metadata and associated follow-up comments.

```
    <sioc:Post
rdf:about="http://johnbreslin.com/blog/2006/09/07/creating-connections-
between-discussion-clouds-with-sioc/">
    <dc:title>Creating connections between discussion clouds with
SIOC</dc:title>
    <dcterms:created>2006-09-07T09:33:30Z</dcterms:created>
    <sioc:has_container
rdf:resource="http://johnbreslin.com/blog/index.php?sioc_type=site#weblo
g"/>
    <sioc:has_creator>
      <sioc:User rdf:about="http://johnbreslin.com/blog/author/cloud/"
rdfs:label="Cloud">
        <rdfs:seeAlso
rdf:resource="http://johnbreslin.com/blog/index.php?sioc_type=user&sioc_
id=1"/>
      </sioc:User>
    </sioc:has_creator>
    <sioc:content>SIOC provides a unified vocabulary for content and
interaction description: a semantic layer that can co-exist with exist-
ing discussion platforms.</sioc:content>
    <sioc:topic rdfs:label="Semantic Web"
rdf:resource="http://johnbreslin.com/blog/category/semantic-web/"/>
    <sioc:topic rdfs:label="Blogs"
rdf:resource="http://johnbreslin.com/blog/category/blogs/"/>
    <sioc:has_reply>
```

```
      <sioc:Post
rdf:about="http://johnbreslin.com/blog/2006/09/07/creating-connections-
between-discussion-clouds-with-sioc/#comment-123928">
         <rdfs:seeAlso
rdf:resource="http://johnbreslin.com/blog/index.php?sioc_type=comment&si
oc_id=123928"/>
      </sioc:Post>
    </sioc:has_reply>
  </sioc:Post>
```

So far, work on SIOC has focussed on producing social semantic data, but the augmentation of this data with rules to aid with reasoning is the next step (for example, as discussed by the ExpertFinder initiative[24]). By combining information from one's explicitly defined social network and from implicit connections that may be derived through common activities (e.g. commenting on each other's content, participating in the same community areas), the suggestion of experts can be enhanced.

4   Collectors of Social Semantic Data

The semantic social data available on the web is distributed across numerous sources and is stored in many different formats. In some cases, this data may be published in such a way that it can be consumed directly by applications, for example in an RDF store with a SPARQL endpoint. Alternatively it may be necessary to first gather and process the data, for example when it is stored in documents which need to be crawled and indexed. In the following we describe issues with interpreting social data from mined the web, inferring relations from semantic data, and technical aspects of collecting data.

4.1   The Web as a Source of Social Network Data

Common traditional methods of collecting social network information include administering questionnaires, conducting interviews or performing observational studies, and studying archival records. There are some fundamental differences between the networks acquirable by these methods and the networks retrievable from the Internet. Extracting data from the Web presents a different set of challenges but also offers some advantages over traditional methods.

A major advantage of mining online social networks for analysis is the much lower cost of acquiring data due to the reduced time and effort involved. Also, the scale of the social information available online is unprecedented. In the past, acquisition of social network data of the order of millions of nodes would have been impossible; with the social data now freely available on the Internet it is easy. In addition, networks collected from the Web are evidence-based and objective. Unlike interviews or questionnaires, results are not dependant on the accurate recall of the subjects, who may interpret questions differently, or may be unwilling to cooperate. Furthermore, while it is unlikely you will get a 100% participation rate in a survey, especially on a large network, if you have access to a full web dataset you can analyse a whole

---

[24] http://expertfinder.info/

network. Finally, electronic data collection easily enables longitudinal studies, allowing the dynamics of networks to be investigated, as opposed to surveying, where repeated data collection would be time-consuming and maybe impossible if the subjects are unwilling or unable to repeat the survey.

However, the accuracy of social network data mined from the Internet can be highly questionable. People can easily misrepresent themselves or others. Depending on Internet usage habits, some people will have far more information available about them online than others. This means that the social networks extracted from the Web may not give a balanced representation of real-life social networks. There is also the question of how exactly to interpret information from the Internet, e.g. the strength of the relationship implied. The people on an individual's contact list on a social networking site may encompass a spectrum from close friends to distant acquaintances or even strangers. Another problem is that there are likely to be errors in Web data, for example resulting from typos, inconsistent spelling of names, and variations on names.

Semantic Web technologies can greatly assist the process of harvesting social networks. The use of common, structured formats means that social network data can easily be aggregated from multiple, heterogeneous sources. References to the same person or resource can be identified across multiple sources and consolidated. Much of the effort needed to construct a model of a social network is removed and the need for human effort is lessened. It is possible to do reasoning on the data and infer relations from certain properties. Additionally, it is possible to extract a network of typed nodes and links.

Harvesting and analysing social data from the Web raises important ethical issues. It involves using data for purposes which were not intended by the users who uploaded for their use and that of their friends. Trust and provenance of information are important aspects that should be taken into consideration. At a technical level, the ability to confirm the origin of data is important, and at a more social level, a means to express trust in sources is also required [16].

4.2 Collecting and Aggregating Data

Data on the Semantic Web is published in different ways, so different methods may be required to collect it. Additional processing may also be required to merge data from multiple sources.

Crawling. Due to the linked nature of social networks, given URIs to seed members of the network, we can follow links from these nodes to their friends, and then their friends-of-friends and so on. This can be done by simply following rdf:seeAlso links. Additional knowledge about the structure of the data can be used to improve the task. For example, the SIOC Crawler [4] uses knowledge of the ontology's structure to incrementally retrieve new SIOC data in threads.

Exporters. For some platforms, exporters are available which generate a structured RDF representation of the data. These allow information in a relational database or other structured stores to be automatically transformed into RDF. Exporters make it easy for users to maintain semantic representations of their data. For example, there are SIOC exporters available for platforms including mailing lists [12], web forums and blogs [9], and existing Web 2.0 services such as Flickr.
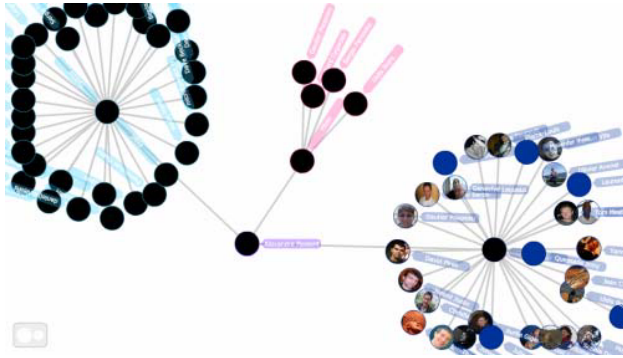
Fig. 4. Identity consolidation and social network browsing using data exported from various social websites[25]

Object Consolidation. An important task in extracting social data from the web is merging identifiers of equivalent instances occurring across different sources. This involves identifying instances representing the same object, and unifying them into one entity. Object consolidation (or "smushing") can be performed for instances which share the same value for inverse functional properties, for example foaf:mbox [23][26]. Another option is to provide explicit identification using owl:sameAs links between various resources that identify the same person or data, in spite of various URIs. This best practice allows one to unify all of their identities from various exporters (e.g. Flickr, Twitter, Facebook, etc.) and to then query their complete social network with a single entry point, as the schema below shows. Finally, it can also be achieved by considering various alternative criteria and if a certain threshold is reached in similarity between two instances, they can be considered equal [1]. Yet, while one can define such rules within his or her own restricted social graph, it may lead to unexpected results on the complete Web (for instance, since different people will sometimes have the same name) and identity management on the Semantic Web is a vast research topic.

4.3   Inferring Relationships from Aggregate Data

The simplest way of extracting a social network from the Web is to look at explicitly stated connections. Social networking sites and other types of social software allow users to express lists of friends. Blogging platforms may allow users to add a blogroll which is a list of favourite blogs. Depending on the platform, these connections may indicate a directed or undirected link between users. For example, blogroll links are frequently unreciprocated, and are therefore directed, but many social networking sites require both users to consent to the link, creating undirected ties. A sample query

---

for extracting the social network formed by explicit foaf:knows relationships follows using the SPARQL query language.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?s ?o
WHERE {
    ?s rdf:type foaf:Person .
    ?o rdf:type foaf:Person .
    ?s foaf:knows ?o .
}
```

In addition to explicitly stated person-to-person links, there are many implicit social connections present on the Web. Links between people may be inferred due to links to some common objects, for example appearing in the same pictures, tagging the same documents, replying to each others blog posts. These connections indicate relationships of varying strengths - for example, e-mail communication may be interpreted as stronger evidence of a real tie than the case of one person replying to another's blog post. Co-occurrence of names in documents would be an even weaker sign of a relation. A sample query for extracting the implicit social network formed by replies to posts follows.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?author1 ?author2
WHERE {
    ?post1 rdf:type sioc:Post .
    ?post1 foaf:maker ?author1 .
    ?post1 sioc:has_reply ?post2 .
    ?post2 rdf:type sioc:Post .
    ?post2 foaf:maker ?author2 .
}
```

Instead of running queries to retrieve those implicit relationships, we can define rules to make them explicit and to state the acquaintance of users on a weblog. For instance, we can consider that there is a formal agreement relationship between two users (modelled with an arg:agreedWith relationship) as soon as one replies to a post from the other one using "I agree" in his or her answer[27]. To model this rule, we rely here on the SPARQL CONSTRUCT pattern, which can be used to produce new statements from existing ones. Thus, we can apply the following query on our triple store, and then put the created RDF graph in the store itself, so that the relationship will become explicit. The produced statements may then be used to extract a more precise social network within a blogging community when querying data.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
```

---

[27] Ideally, more advanced pattern matching and NLP methods should be used to define agreement between two users on a weblog.

```
CONSTRUCT {
   ?author2 arg:agreedWith ?author1 .
} WHERE {
   ?post1 rdf:type sioc:Post .
   ?post1 foaf:maker ?author1 .
   ?post1 sioc:has_reply ?post2 .
   ?post2 rdf:type sioc:Post .
   ?post2 foaf:maker ?author2 .
   FILTER REGEX(?post2, "I agree", "i") .

}
```

While the above examples result in simple networks of people and untyped ties, more complex social networks consisting of multiple node and link types can also be studied. These examples are only possible through linking people and content in and across sites. Traditional, non-semantic queries like in SQL would be limited to one site and would require some kind of join on a user / content table. However the use of shared semantically-rich vocabularies makes it possible to perform operations like these on data originating from many different sources.

5   Consumers of Social Semantic Data

Once data has been collected and aggregated, or made directly accessible through a SPARQL endpoint, it can be studied or used in applications. As the information is in a structured format, it can easily be converted into the formats required by popular social network analysis and visualisation tools. RDF data can also be queried directly to return some set of items that fit certain criteria that a user is interested in. In the following we describe these two ways of using semantic social data.

5.1   Social Network Analysis

Social network analysis uses methods from graph theory to study networks of individuals and the relationships between them. The individuals are often referred to as nodes or actors, and they may represent people, groups, countries, organisations or any other type of social unit. The relations between them can be called edges or ties, and can indicate any type of link, for example acquaintance, friendship, co-authorship and information exchange. Ties may be undirected, in which case the relationship is symmetric, or directed, in which case the relationship has a specific direction and may not be reciprocated.

The nodes in a social network can be seen as analogous to entities in an RDF graph, where a <subject, predicate, object> triple indicates a directed tie from the subject node to an object node, and the predicate indicates the type of the relationship. While social network analysis methods are generally applied to social networks, they can be used to analyse any kind of networked data.

We can apply mathematical measures from social network analysis to get interesting information about a social network. The more complex methods of network analysis cannot be performed directly on a graph in RDF format, but must be converted to a representation more suited to network analysis. An RDF graph can be loaded into a

network analysis program such as Pajek or UCINET [7] which can perform various measures and visualisations. Alternatively, a library like JUNG [25], which provides analysis and visualisation methods, can be used to develop custom analytic or visual tools.

Locating important individuals. Centrality measures can be used to locate key players in a network [27]. Degree centrality is based on the number of connections a person has. This measure locates individuals who are connected to a large number of others. In a directed graph, indegree is the number of incoming connections and outdegree is the number of outgoing connections. Closeness centrality is calculated based on the total shortest distance to all other nodes in the network. This measure can be an indicator of people who can most quickly communicate information to the whole network. Betweenness centrality is based on the number of shortest paths on which a node lies. A node which scores highly according to this metric may occupy a strategic position and function as a bridge between different parts of the network. Flink [23] applies these measures to a social network of Semantic Web researchers in order to investigate whether the network position of a scientist is related to their performance.

Extracting communities. We may be interested in finding subgraphs or small communities within a larger graph. This enables the restriction of network to a manageable size for performing further analysis. Algorithms exist for partitioning a network into different groups, for example that of Girvan and Newman [15]. Alternatively, if there is a particular individual of interest we can extract their ego network, the area of the graph focussed around them. For example, spreading activation algorithms can activate an input node or nodes, and propagate the activation from these in order to locate those individuals which are most strongly connected and therefore receive the most activation [21].

Characterising a social network. There are some interesting whole network properties that can be investigated in order to gain an understanding of the overall structure of the network [27]. Centralisation measures the degree to which the network has a leader. Cohesiveness measures the well-connectedness of the network. These measures can also be used to make comparisons between different networks.

Visualising a social network. By creating a pictorial image of a social network, it may be possible to get an improved insight into the structure of the graph. A visual representation can help analysts to understand the network better themselves, and also aid in explaining features of the network to others [13]. Flink provides visualisations of the ego-networks of individual researchers and allows users to browse members of the Semantic Web research community.

5.2 Querying an RDF Graph

By representing social data in RDF and putting it in a store with a SPARQL endpoint, we can perform queries to extract interesting information about users, communities and content. In the following we discuss some example scenarios and illustrate them with sample queries.

Finding a person's ego-network. Identifying an ego-centric network centred around a focus person involves finding all people to whom they are connected to online. This means searching over all their accounts, and across all social networking sites of

which they are a member. Below is a simple example query over FOAF data to get all friends of Persons with a particular e-mail address sha1sum. We use the hash of an e-mail address as an identifier (since the foaf:mbox_sha1sum is defined as an owl:InverseFunctionalProperty in FOAF), as the focus person is likely to have different URIs on different sites.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?o
WHERE {
    ?s foaf:mbox_sha1sum "9a348bd34fe67b15f388c95c2cb9b4bfc9073797" .
    ?s foaf:knows ?o .
}
```

Finding a person's implicit social links. While locating a person's explicitly stated connections goes some way to locating their social network, they may have more acquaintances with whom they are implicitly linked. It is possible to identify additional potential acquaintances of a person via objects to which they are both connected. The example below shows a query to find all people with the same workplace, school or project as the focus person. We could also consider people who are co-authors of some documents, or who have replied to each others SIOC-enabled posts.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?s
WHERE { {
    <http://sw.deri.org/~sheila/foaf.rdf#me> foaf:workplaceHomepage ?o
.
    ?s foaf:workplaceHomepage ?o .
} UNION {
    <http://sw.deri.org/~sheila/foaf.rdf#me> foaf:schoolHomepage ?o .
    ?s foaf:schoolHomepage ?o .
} UNION {
    <http://sw.deri.org/~sheila/foaf.rdf#me> foaf:project ?o .
    ?s foaf:project ?o .
} }
```

We can carry out simple reasoning by expressing a set of rules to describe when such implicit links create a social connection between people and when they may not. For example, we may decide that two people are socially connected if one posts a comment on someone else's blog post; alternatively, we may conclude that a weak link exists if two people posted on the same lengthy discussion thread and that no social connection exists.

Aggregating a person's web contribution. This means retrieving content that a person has contributed to various sources on the web; for example, all blog posts and comments on other blogs, chat logs, mailing list and forum posts. This is a difficult problem to perform with a normal search engine as people may share their name with other people, or may use different account names on different sites. A sample query over SIOC data is shown below, to get all posts created by a particular user.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX sioc: <http://rdfs.org/sioc/ns#>

SELECT DISTINCT ?post
WHERE {
      ?post rdf:type sioc:Post .
      ?post sioc:has_creator
<http://www.mindswap.org/blog/author/hendler/#foaf> .
  }
```

Yet, since this query is based on a precise URI, it will not retrieve content created by the same user while using another URI (for instance, http://example.org/hendler). One option to retrieve this content is to define owl:sameAs statements between this URIs and other URIs of the same user, such as:

```
    <http://example.org/hendler> owl:sameAs
<http://www.mindswap.org/blog/author/hendler/#foaf> .
```

Then, by adding these statements in the triple store that holds the data, and assuming it supports reasoning based on owl:sameAs, the query will also retrieve posts that have http://example.org/hendler as a sioc:has_creator.

A second way to do retrieve the person's contributions is to run the query not based on the URI, but based on an inverseFunctionalProperty, such as the foaf:mbox or foaf:openid. Since OpenID aims to become a standard for authentication on the web, this can be a useful way to retrieve all the contributions of a given user no matter which social website it comes from - providing the person signs in using the same OpenID URL - and this method is shown in the following query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX sioc: <http://rdfs.org/sioc/ns#>

SELECT DISTINCT ?post
WHERE {
      ?post rdf:type sioc:Post .
      ?post sioc:has_creator ?user .
      ?user foaf:openid <http://example.org/hendleropenid> .
}
```

Locating a community around a topic. We may be interested in extracting a community centred around a certain topic, using tags, keywords and other metadata to find people who are talking about a certain thing. The query below locates posts with the topic "semantic web" and returns the URIs of the authors of these posts.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?author
WHERE {
   ?post rdf:type sioc:Post .
   ?post foaf:maker ?author .
   ?post sioc:topic ?post_topic .
   ?post_topic rdfs:label "semantic web" .
}
```

Yet, this query will not retrieve posts written in French, for example, using a "web seman-tique" string instead of the "semantic web" phrase. However, if people were encouraged to use a precise URI instead of the simple tag, such as http://dbpedia.org/resource/Category: Semantic_Web, we would then be able to retrieve all related posts. Moreover, using those URIs, we can run even more advanced queries, as in the example of retrieving all posts re-lated to the Semantic Web, we could also show those for which the topic is directly related to this URI (e.g. RDFa, SKOS, etc.), as the following query does, emphasising the benefits of combining data from various datasets, interlinked together in the whole Semantic Web graph.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?author
WHERE {
    ?post rdf:type sioc:Post .
    ?post foaf:maker ?author .
    ?post sioc:topic ?topic .
    ?topic ?rel <http://dbpedia.org/resource/Category:Semantic_Web> .
}
```

As with the example queries in Section 4, the queries above can be performed on data originating from various diverse sources.

6   Future Work

A key feature of the new Social Web is the change in the role of user from just a con-sumer of content, to an active participant in the creation of content. For example, Wikipedia articles are written and edited by volunteers; Amazon.com uses information about what users view and purchase to recommend products to other users; Slashdot moderation is performed by the readers. One area of future work in relation to social networks on the Semantic Web is the application of semantic techniques to take even more advantage of community input to provide useful functionality. As an example, we will look at the area of multimedia management.

There is an ever increasing amount of multimedia of various formats becoming available on the Internet. Current techniques to retrieve, integrate and present these me-dia to users are deficient and would benefit from improvement. Semantic technologies make it possible to give rich descriptions to media, facilitating the process of locating and combining diverse media from various sources. Making use of online communities can give additional benefits. Two main areas in which social networks and semantic technologies can assist multimedia management are annotation and recommendation.

Social bookmarking systems like del.icio.us allow users to assign shared free-form tags to resources, thus generating annotations for objects with a minimum amount of ef-fort. The informal nature of tagging means that semantic information cannot be directly inferred from an annotation, as any user can tag any resource with whatever strings they wish. However, studying the collective tagging behaviour of a large number of users

allows emergent semantics to be derived [29]. Through a combination of such mass collaborative "structural" semantics (via tags, geo-temporal information, ratings, etc.) and extracted multimedia "content" semantics (which can be used for clustering purposes, e.g. image similarities or musical patterns), relevant annotations can be suggested to users when they contribute multimedia content to a community site by comparing new items with related semantic items in one's implicit / explicit network.

Another way in which the wisdom of crowds can be harnessed in semantic multimedia management is in providing personalised social network-based recommender systems. Liu et al. [22] presents an approach for semantic mining of personal tastes and a model for taste-based recommendation. [14] explores how a group of people with similar interests can share documents / metadata and can provide each other with semantically-rich recommendations. The same principles can be applied to multimedia recommendation, and these recommendations can be augmented with the semantics derived from the multimedia content itself (e.g. the information on those people depicted or carrying out actions in multimedia objects[28]).

Some challenges must also be overcome regarding the online identity aspect and authentication / privacy for users of social websites. An interesting aspect of social networking and media sharing websites is that most people use various websites because they want to fragment their online identity: uploading pictures of friends on MySpace, forming business contacts on LinkedIn, etc. While the Semantic Web and in particular reasoning principles (such as leveraging IFPs) allow us to merge this data and provide vocabularies, methods and tools for data portability among social websites [5], [6], this identity fragmentation must be taken into account. It implies a need for new ways to authenticate queries or carry out inferencing, by delivering data in different manners depending on which social subgraph the person requesting the data belongs to.

7   Conclusions

In this paper, we have described the significance of community-oriented and content-sharing sites on the Web, the shortcomings of many of these sites as they are now, and the benefits that semantic technologies can bring to social networks and social websites. Online social spaces encouraging content creation and sharing have resulted in the formation of massive and intricate networks of people and associated content. However the lack of integration between sites means that these networks are disjoint and users are unable to reuse data across sites. Semantic Web technologies can solve some of these issues and improve the value and functionality of online social spaces. The process of creating and using semantic data in the Social Web can be viewed as a sort of food chain of producers, collectors and consumers. Semantic data producers publish information in structured, common formats, such that it can be easily integrated with data from other diverse sources. Collectors, if necessary, aggregate and consolidate heterogeneous data from other diverse sources. Consumers may use this data for analysis or in end-user applications.

In this way, it becomes possible to integrate diverse information from heterogeneous sites, enabling improved navigation and the ability to query over data. There are

---

[28] http://acronym.deri.org/

also advantages for those interested in studying social networks, as the Semantic Web makes freely available large-scale, multi-relational datasets for analysis. In this paper, we described some methods by which consolidated facts and content can be extracted from people and content networks aggregated from multiple social networks and social websites, and we presented our ideas for future work as the focus of these sites moves more towards the provision of multimedia content.

References

1. Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A.P., Arpinar, I.B., Joshi, A., Finin, T.: Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection. In: Proceedings of the 15th International Conference on the World Wide Web, Edinburgh, Scotland (2006)
2. Batagelj, V., Mrvar, A.: Pajek - Program for Large Network Analysis. Connections 21(2), 47–57 (1998)
3. Berners-Lee, T., Hendler, J.A., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
4. Bojārs, U., Heitmann, B., Oren, E.: A Prototype to Explore Content and Context on Social Community Sites. In: The SABRE Conference on Social Semantic Web (CSSW 2007), Leipzig, Germany (September 2007)
5. Bojārs, U., Breslin, J.G., Finn, A., Decker, S.: Using the Semantic Web for Linking and Reusing Data Across Web 2.0 Communities; Special Issue on the Semantic Web and Web 2.0, The Journal of Web Semantics (2008)
6. Bojārs, U., Passant, A., Breslin, J.G., Decker, S.: Social Network and Data Portability using Semantic Web Technologies. In: Proceedings of the BIS 2008 Workshop on Social Aspects of the Web, Innsbruck, Austria (May 2008)
7. Borgatti, S.P., Everett, M.G., Freeman, L.C.: UCINET for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard (2002)
8. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. The Journal of Computer-Mediated Communication 13(1) (2007)
9. Breslin, J.G., Harth, A., Bojārs, U., Decker, S.: Towards Semantically-Interlinked Online Communities. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 500–514. Springer, Heidelberg (2005)
10. Breslin, J.G., Decker, S.: The Future of Social Networks on the Internet: The Need for Semantics. IEEE Internet Computing 11, 86–90 (2007)
11. Ding, L., Zhou, L., Finin, T., Joshi, A.: How the Semantic Web is Being Used: An Analysis of FOAF Documents. In: Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 2005) (2005)
12. Fernandez, S., Berrueta, D., Labra, J.E.: Mailing Lists Meet the Semantic Web. In: Proceedings of the BIS 2007 Workshop on Social Aspects of the Web, Poznan, Poland (April 2007)
13. Freeman, L.C.: Visualizing Social Networks. Journal of Social Structure 1(1) (2000)
14. Ghita, S., Nejdl, W., Paiu, W.R.: Semantically Rich Recommendations in Social Networks for Sharing, Exchanging and Ranking Semantic Context. In: Proceedings of the 4th International Semantic Web Conference, Galway, Ireland (November 2005)

15. Girvan, M., Newman, M.E.J.: Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002)
16. Golbeck, J., Parsia, B., Hendler, J.: Trust Networks on the Semantic Web. In: Proceedings of Cooperative Intelligent Agents, Helsinki, Finland (August 2003)
17. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging Systems. Journal of Information Sciences 32(2), 198–208 (2006)
18. Gruber, T.: Ontology of Folksonomy: A Mash-up of Apples and Oranges. International Journal on Semantic Web and Information Systems 3(2) (2007)
19. Heer, J., Boyd, D.: Vizster: Visualizing Online Social Networks. In: IEEE Symposium on Information Visualization (InfoVis 2005), Minneapolis, Minnesota (October 2005)
20. Kim, H.L., Yang, S.K., Breslin, J.G., Kim, H.G.: Simple Algorithms for Representing Tag Frequencies in the SCOT Exporter. In: The IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pp. 536–539. IEEE Computer Society, Los Alamitos (2007)
21. Kinsella, S., Harth, A., Troussov, A., Sogrin, M., Judge, J., Hayes, C., Breslin, J.G.: Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects. In: The 4th Conference on Applications of Social Network Analysis (ASNA 2007), University of Zurich, Switzerland (accepted, September 2007)
22. Liu, H., Maes, P., Davenport, G.: Unraveling the Taste Fabric of Social Networks. International Journal on Semantic Web and Information Systems 2, 42–71 (2006)
23. Mika, P.: Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks. Web Semantics: Science, Services and Agents on the World Wide Web 3(2-3), 211–223 (2005)
24. Mika, P.: Ontologies are Us: A Unified Model of Social Networks and Semantics. In: International Semantic Web Conference. LNCS, pp. 522–536. Springer, Heidelberg (2005)
25. O'Madadhain, J., Fisher, D., White, S., Boey, Y.: The JUNG (Java Universal Network/Graph) Framework. University of California, Irvine (2003)
26. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)
27. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
28. Watts, D.J., Strogatz, S.H.: Collective Dynamics of 'Small-World' Networks. Nature 393(6684), 409–410 (1998)
29. Wu, X., Zhang, L., Yu, Y.: Exploring Social Annotations for the Semantic Web. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland (May 2006)
30. Passant, A., Laublet, P.: Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data. In: Proceedings of the WWW 2008 Linked Data on the Web Workshop (LDOW 2008), Beijing, China (April 2008)