Towards the new generation of web knowledge.

Fefie Dotsika
Keith Patrick

Westminster Business School

This is an electronic, author-formatted version of an article published in VINE: the journal of information and knowledge management systems, 36 (4). pp. 406-422, 2006. The definitive version in VINE is available online at:

http://dx.doi.org/10.1108/03055720610716665

Towards the New Generation of Web Knowledge

Fefie Dotsika, Keith Patrick
Business, Information, Organisation & Process Management Research Centre, Westminster Business School, University of Westminster, London, UK
F.E.Dotsika@westminster.ac.uk & K.Patrick01@westminster.ac.uk

Abstract: The web has evolved and is evolving: both its purpose and the nature of its use are changing. Two significant aspects lie in its development from a static information area to a dynamic web of latent or potential knowledge, with a further shift into a corporate environment. This latency leaves us with the question of how can we access this web of knowledge and how can we manage this knowledge better. In brief, can the web provide for the competing stakeholders, who are similarly evolving and increasingly see it as a significant part of their business or market place?

This paper adopts an exploratory and reviewing approach to the emerging trends and patterns emanating from this changing use and explores the underpinning technologies and tools that facilitate this use and access. It examines the future and potential of web-based Knowledge Management and reviews the emerging web trends, tools, and enabling technologies that will provide the infrastructure of the next generation web. It investigates some of the requirements for effective web-based knowledge searching, retrieval and sharing and identifies the implications of integrating the two aspects of web-based knowledge management, namely the business-organisational-users' perspective and that of the enabling web technologies. Finally, it seeks to present a view that, whilst looking toward the leading edge of the wave, ensures that the previous waves or approaches are not being overlooked and that nothing significant is missed or has been insufficiently considered or examined.

Keywords: *Semantic Web, Web 2.0, Knowledge Search, Knowledge Sharing, Knowledge Management, Knowledge Modelling, Social Software.*

1. Introduction

Technology has been heralded as the answer to our information requirements, a charge that has been extended to meet our knowledge requirements as well. Intranets have been cited as examples of such a solution and success, so have web-enabled databases and portals. But to what extent does this address the needs of the user in the creation and particularly the ability to search for and share this information and knowledge? And how effectively does this facilitate the creation of new knowledge? It is our proposition that users and organisations need to beware of the balancing act of successful web-based search and sharing.

The Web was originally designed as a text and image repository for human use. Its unprecedented expansion however has triggered a significant increase in the expectations for web-based information retrieval, knowledge sharing and collaborative working. Search engine indices have become too large, with every search producing an enormous amount of results. Search engines are often limited by poor indexing, ranking of pages according to inappropriate metrics, the absence of keywords on relevant pages and inaccessibility to distributed information repositories of different formats, such as databases. At the end of every query the searchers are inundated with a great amount of links that they need to go through in order to gather the knowledge they require. Companies often try to second-guess the ‘magic’ words used by searchers, or employ search engine optimisers. Organisations often end up paying for content that could be found for free on the web.

Looking into web knowledge search and sharing, users can search for knowledge using a number of means: following a path of hypertext links, using search engines, web-directories or intelligent agent software. Based on the actor of the search two distinct approaches can be identified: the end-user practice and the automated approach. The first one (also termed 'cognitive' approach) is considered the traditional method and relies on the user going through websites in order to gather the required knowledge. The second method is the technical equivalent of the same process and relies on intelligent agents ('bots') for the gathering of knowledge. Each method has its advantages and disadvantages and, depending on the task at hand, each is associated with particular quality and suitability issues and/or specific limitations.

A further reflection is required when we observe the ubiquitous and pervasive nature of technology throughout our lives, which impacts on both our working practices and attitude toward that technology, in terms of how it is used and it continues to evolve. There is a shift from a specialised, centralised and controlled application and implementation of technological solutions to one that sees distributed, and multiple solutions that are local and enterprise-wide. An additional shift is from a smaller number of centralised specialists to increasingly involved and sophisticated end-users. This confronts the issues of technological design and development (what should be made to fit what or whom and should the user fit the system/software or the reverse) with intrinsic implications for developers, users and organisations alike. It is not untypical to organise around business processes with a tendency to embed them within rigid bureaucracies with the inherent procedures and rules, technology systems, and structures such as ERP and SAP systems. So if for some reason, the process needs to be changed, it becomes very difficult to make any adjustments because so much structure has been wrapped around it. Allee (1997) saw the need where *strategies are human-centred and not technology centred* and for a culture that addresses and supports knowledge creation, sharing and learning.

These technological shifts can be seen to be further reflected in the changing nature of the economy from manufacturing to knowledge and information based economics, which focus more toward productivity, new products and services, new modes of delivery/supply, time based competition, and shorter product life cycles. This economy is global in terms of both the market-place and the internet/web mediated market-space which is engendering a workforce characterised by three significant types of worker: Data Workers who process and disseminate organisation’s paperwork; Information Workers who primarily create and process information; and Knowledge Workers who design products or services, or create new knowledge for the organization (Laudon & Laudon 2005). This growth in knowledge work and knowledge workers requires not only the ability to find and access information and knowledge, but also ability to share this synchronously and asynchronously in terms of both time and location. Newell et al. (2002) saw the knowledge worker in a more evolved form than Laudon and Laudon (2005), characterised by higher levels of education, specialist skills and ability to apply these skills to identify and solve problems. According to them these workers effectively '*own the primary means of production'*, and have the knowledge, skills and ability to apply them.

During this period, with a significant causal effect from the introduction and spread of technology, organisations can also be seen to have evolved by changing their structure. There is evidence of flattening, reduction in the number of level of management and reporting within the structure and decentralisation. Satellite structures are often used that include geographic relocation of parts of either the organisation or particular activities or tasks, including their outsourcing or off-shoring. These changes sought to derive flexibility, location independence, low (lower) transaction and co-ordination costs, empowerment, and create the need for collaborative work.

Zack (1999) suggested: *'To remain competitive, organisations must efficiently and effectively create, locate, capture, & share their organisation's knowledge & expertise.'* This, results in a series of questions for any organisation: what do we know about our customers, services, products, markets and environment? How well do we know what we know? With the additional proviso 'is this known by the right people?', queries regarding information quality and information integrity, and with the final question 'how well do we act on what we know?'. This series of activities and actions is commonly identified as Knowledge Management, and aptly described by Elliott (2004) as: *'the coordination and management of human understanding and knowledge within an organization'.* Hence the need to be able to search and find the information and knowledge required at that time and to be able to share and reuse concurrently and at subsequent occasions.

The rest of the paper is organised as follows. Section 2 examines the different methods adopted when searching for knowledge on the web. In section 3 we explore the emerging trends and technologies that influence the future of the web, identify the ones most pertinent in knowledge search and share and assess their overall impact in web-based knowledge management. Section 4 identifies the problems of each approach, determines possible solutions and sets the foundation of the proposed framework. Finally in section 5 we sum up our conclusions and outline future work.

2. Searching for knowledge on the web

The web, as it stands, holds information using natural language, multimedia content and hypertext marking. Such information can be combined from various sites via a search engine and can then be processed either by humans (cognitive approach) or by intelligent agents (automated approach).

The cognitive method places the burden of knowledge discovery on the end-user who is required to go searching through a number of web pages. Equivalency of terminologies is not an issue - since humans can make associational mappings on the fly - neither is deduction. However, humans cannot process this information when it comes in overwhelming quantities. This is where the intelligent agents take over, though, for this to happen, a number of standards have to be met and appropriate technologies need to be adopted.

2.1. The Cognitive Approach

The web was designed primarily for human interpretation and use. The end-user searches across pages either through hyperlinks (free-style web surfing and/or use of hyperlink indices), subject directories, or search engine results. In order to be successful, this method requires strong and sound indexation that enhances navigation. It further relies on the searchers' mental model that is their personal experience and domain knowledge. Mental models influence the actual search and determine how the searchers will interpret the information gathered. Both pre-existing and found knowledge are mapped into a contextual cognitive structure, a 'schema'. Schemata facilitate the organisation of knowledge and incomplete information around a basic framework and affect future search behaviour and further evaluation of knowledge (Greve & Taylor 2000).

However, finding information on the web is not necessarily the direct result of searching. We identify three different factors that influence the users' search behaviour and the overall task success in the cognitive approach: (a) search strategy, (b) choice of keywords (associated with the user, not with the dynamics of the search engine, which are addressed in section 3) and (c) usability & navigation issues.

(a) Depending on the focus of the search, three strategies can be identified, that are directly linked to the task of the search (Navarro-Prieto et al. 1999):

- ✓ Top-down strategy, where the searchers start with a general area and proceed by narrowing down their search by following the links provided. Favoured when the topic of the search is contained within a general site consisting of a well-organised list of subtopics.
- ✓ Bottom-up strategy, where the choice of keyword(s) is specific. This method is chosen for precise, fact-finding searches.
- ✓ Mixed strategy, where both the above methods are used within the same search, either in parallel (multiple searches) or alternating strategies for better results. This strategy is typically used by the more experienced searchers.

(b) Depending on the choice of keywords, the searcher can opt for a plain keyword search, a boolean (a search using OR, NOT and AND operators), or an exact phrase search. Search engine findings provide additional help by including the results from fuzzy searches (matches which are returned even when words are partially spelled or misspelled) and precision indicators (how close the result link is to the original query in percentage of relevancy).

(c) Apart from the search strategy and keyword choice, some searches never return the information sought, simply because users often get lost in hyperspace. Quantitative assessment of the success rate of web navigation has identified a number of factors that influence the results. Successful searches are correlated with shallow hierarchical navigation (high compactness), while failure is related to a linear style of navigation (high stratum) (McEneaney 2001). A number of algorithms (longest repeated sequence, sequence alignment etc.) have been used to assess the similarity between optimal and user navigation paths (Pitkow & Pirolli 1999, Wang & Zaiane 2002) proving that the higher the similarity to the optimal path, the better the chances of finding information successfully.

The main advantages of the cognitive approach include low costs (end-users instead of the increased expenditure of specialised, often tailor-made software and equally costly maintenance) and increased suitability when dealing with 'open' domains and community-based applications. The limitations are typically two-fold. On the one hand there is often a (potentially) overwhelming volume of results returned by the search engine. Going through them harvesting the knowledge sought is not always easy especially as time constraints are often involved. On the other hand, there are the search-engine bound problems: poor indexing, ranking of pages according to a range of not always appropriate metrics, the absence of keywords on relevant pages and inaccessibility to distributed information repositories of different formats.

2.2. The Automated Approach

The technical approach ranges from the use of general search engines to the specialised search of intelligent agents (bots). Search engines employ robot programs (known as spiders) that roam the Internet in search of information, rank the results according to relevance and list them for the user. In a similar way meta-search engines transmit user queries to multiple individual search engines and subject directories at once and then compile and consolidate the results into a uniform format and listing. The top search engines employ intelligent agent software, which navigates the Internet searching for information.

However, conventional web mark-up provides syntax but lacks semantics, a fact that severely limits the task of intelligent agents. The new generation of web standards add semantics and deduction capabilities to traditional mark-up. The Semantic Web is about sharing knowledge between communities, individuals and machines. It expands the web by supporting semantic mark-up, transforming it into a distributed Knowledge Base which provides the ideal environment for intelligent agents performing various automated tasks (McIlraith 2001). The linking of information is done by re-usable, task-specific, high-level generic procedures, featuring user-specific customising constraints over a framework of standards and an ontological approach that determines a shared common concept of a domain. This 'new' web, still relies on old technologies such as HTML and XML for looks and content structure, but it is further enabled with new languages and standards such as Resource Description Framework (RDF) (Brikley 1999) and a variety of ontology languages (Fensel 2001) such as the World Wide Web Consortium's standard, OWL (Web Ontology Language) (W3C OWL 2004). These languages are used to create vocabularies that add semantics, inference tools and formal specifications of contents and relationships. As a result, web content becomes process-able (and, thus, ultimately "understandable") by intelligent agents, that is autonomous, interactive and adaptable software programs that search, gather and filter information.

Placement for Figure i: Complexity of Agents

Figure i: Complexity of Agents.

The more the information, the greater the degree of complexity involved (and required). Benjamins et al. plot the dimensions of (web) information overload, (intelligent agent) task delegation and relevant complexity (Benjamins et al., 2004) as depicted in Figure i.

Overload of information corresponds to higher intelligence requirements (*y*). Equally, where intelligent agents are concerned, greater task delegation corresponds to greater autonomy (*x*), where autonomy represents the agents' primal characteristic of being able to operate on their own, without that is, human interference. The complexity of an intelligent agent can then be defined as a function f: x -> y. The rate of complexity multiplies with the increase of information and task delegation.

The automated approach fares best with 'closed' consensual domains of knowledge and when highly precise information needs to be retrieved automatically, especially when semantic mark-up, ontologies and intelligent agents are deployed. The limitations of this approach include storage and scalability problems but also requirements for specialist end-users (Dotsika & Patrick 2005a). However, the main drawbacks of the method arise from ontology quality issues, as we will see in the next section.

3. The Future of the Web

While the Web is an essential repository of information, the simple use of search engines often fails to capture and interpret the users' real information needs. It is said that '*a quality result is not a long list of links but the correct* *list'*. The semantic gap between the users' perception of the search domain and the results provided by the search may be the outcome of the sheer volume of answers returned, low quality, or plain irrelevance. Despite the fact that a part of corporate KM usually relies on web-based collaborative computing technologies by means of intranets, KM suites, corporate portals etc., the quality of information retrieval, reuse and sharing is rather disappointing. Organisations and individuals are looking into the emerging trends and technologies for a possible solution. As a consequence there has been much speculation about the future of the web and its use as an efficient knowledge management platform.

The idea of enhancing KM by enabling it to tap into the Semantic Web is to make a huge amount of electronically information more accessible by using ontologies to make searches more intelligent. The principle is simple: keyword searches are based on matching word patterns, whereas intelligent searches are based on answering questions. The Semantic Web supporters declare that the future lies in formal semantics, standardisation and intelligent agents. The Semantic Web key technology for managing knowledge is ontologies.

The Web 2.0 (O'Reilly 2005) enthusiasts on the other hand proclaim that the future should be all about collaboration, sharing and end-users. According to this scenario, the future lies in the tools supporting these activities, which are collectively known as Social Software.

Our framework proposal seeks to reconcile the two trends, since, although the sets of followers of the two camps seem disjoint in the first instance, they clearly have overlapping goals. It then furthers the notion of the web knowledge platform to include the 'invisible' web. This 'hidden' part of the web (referred to as the 'invisible', 'dark' or 'deep' web) contains a huge amount of information that is not accessible by search engines.

3.1. Semantic Web and Ontologies

The application of ontologies as the conceptualisation of a given domain is well documented within the context of enterprise models (Fox & Gruninger 1998). With the arrival of the Semantic Web there is a growing demand for facilitating ontologies' re-use and deployment, coupled with an increasing concern about the quality and validity of the information provider. Re-use (and/or extension) of existing ontologies is possible, and knowledge engineers are called upon to determine their suitability and decide on the best possible choice. One way to develop new ontologies is to identify and adapt existing ones from a neighbouring field. This method can increase consistency while keeping costs low. But, regardless of the technique employed, the quality of the ontology is of the utmost importance. We identify the following quality issues: (a) ontology modelling features, (b) express-ability/re-usability and (c) application environment issues.

(a) In accordance with the principles of conceptual modelling, ontological quality comes in three flavours: syntactic, semantic and pragmatic (Lindland 1994). Syntactic quality reflects the syntactic correctness of the model. Semantic quality addresses the question ‘does the model cover the domain of interest?’ Finally, the pragmatic dimension indicates whether the model is comprehensible by the user.

(b) In modelling ontologies, express-ability is a synonym to complexity. Complexity hinders re-usability, one of the most important characteristics of ontologies. A high-quality ontology is specific in modelling the domain’s attributes, but should not be more specific than necessary.

(c) Ontologies should be able to integrate with a variety of applications and interfaces. They should therefore be language independent (not tied to a particular natural or programming language) an aspect that may affect the ontology’s express-ability.

The use of semantic mark-up and ontologies have led to the deployment of an increasing number of intelligent agent information retrieval systems. They often employ a combination of agent types (brokers, mediators and wrappers), search technologies (natural language understanding, filtering and domain modelling, conceptual search techniques) and architectures (simple or multi-agent, local or distributed). These systems tend to be task-specific and, consequently, the quality of the search results depends upon the particular assignment.

Nevertheless, information retrieval is not intelligent agents’ only suitable application. Agent software provides a specialised form of ‘push’ technology, a dynamic form of electronic publishing that automates the transfer of information to end-users. Push technologies are an increasingly popular type of sharing content as well as applications. The agents undertake the time-consuming task of monitoring web information resources and are controlled by end-users who can specify the type of information they want to receive.

There is a number of existing RDF tools, developers’ API’s and ontology editors that can be combined to provide Semantic Web-enabled KM platforms. The best known open-source ones are Protégé-2000 (Noy et al. 2001) and Sesame (2004), while OntoEdit (2002) and Jena 2 toolkit (HP Laboratories Research 2002) are commercial suites. Other products include OILEd (Bechhofer et al., 2001), Ont-O-Mat (Handschuh, 2001) and the more recent Swoop (Kalyanpur et al. 2005). They invariably offer ontology browsing and editing and may provide querying facilities (Sesame, Jena 2, etc.) and/or plug-ins (Sesame, Swoop, etc.)

A pick-and-mix combination of tools like the above has lead to complete ontology assisted KM platforms. KAON (Bozsak et al. 2002) and On-to-Knowledge (Davies et al. 2002) are the most comprehensive among them. KAON is an open-source ontology management platform targeted for business applications. KAON's front-end consists of the user-level applications and its core addresses the developer needs and comprises two APIs and a number of libraries. On-To-Knowledge comprises an ontology-based environment that provides tools for the support of knowledge management, a bottom layer of machine-processable metadata and a core repository that uses semantics to describe meanings for annotated data

3.2. Web 2.0 and Social Software

Whilst the technical/automated approaches have been viewed as the solution to meeting information requirements they do not represent a complete solution, as they do not follow the patterns of cognitive practice of individuals. Reflecting the question '*where is the fit?*' there are two possible and opposing views: 'technology to the user' and 'the user to the technology' (Dotsika & Patrick 2005b). Historical evidence however shows a pattern which is not always in line with theory or discussion. According to it, developer approaches typically take the former view, while practice echoes the latter. The result of the inherent compromise impacts upon the proposed efficiency gains of any solution and the current and future goodwill toward subsequent technology solutions. This is not necessarily an aversion to technology or technology solutions, but dissatisfaction with how the solution fails to meet or fit the requirements and behaviour of the proposed and potential user. This can be seen in the provision of information (and subsequent information overload) of a 24/7 technologically connected world, whose need to be able to 'pull' the information required is far greater than the overwhelming nature of the 'push' of the continuous stream of information broadcasting to customers and employees. The problem of the latter is the notion that it throws information by the bucket, when a glassful was all that was needed, with these buckets rarely being other than tangential to the actual need.

Two observations can be made in relation to this scenario regarding technology, users, and organisations. Organisations can be seen to have two fundamental structural components in their make-up. At the core there is the typically formal structure with its levels and responsibilities and reporting hierarchies. Then there is the informal organisation centred on personal connections, common interest or goals, '*…an invisible force influencing resource allocation…',* and *'…an antidote to inflexible bureaucracy…'* (Gabriel et all 2000). The second observation regards technology and how individuals can and will use it, seen in how they use technology to interact and cluster with other individuals, through mobile telephony, instant messaging, mailing list and groups, etc. It should perhaps be noted that typically this grows organically, and this aspect has significance in examining how to balance the technical and end-users aspects of search and sharing. This collection of technologies, being branded as Social Software, supports the desire of individuals to be pulled into groups to achieve goals (Boyd 2003). Figure ii below depicts the potential components of Social Software (adapted from Bryant 2003, 2005b).

Although this tag can be applied to many types of software, there are several key elements, such as a means for conversational activity that is both synchronous and asynchronous, and feedback in the form of contributions and comments from others, with evidence of the personal relationships of the participants, who together form the *social network.*

Placement for Figure ii: Social Software

Figure ii: Social Software

Social Software effectively is a convergence of the thinking of the domains of Social Networks, Human-Computer Interaction (HCI) and web services. In relation to the question of the technology-to-user fit, Social Software adapts to its environment, as opposed to the environment being required to adapt to the software. Successful software can be seen to be intuitive so that it enables the user to adapt and continue to use it. An additional feature in relation to the organisation is the duality of its informality and typically bottom-up development. The more interesting aspect and relevant to the examination of the balance between end-user and bots is how the adoption of Social Software in organisations is also seeing a different approach, drawing on the ethos and nature of Social Software itself, with vendors and proponents (like Headshift) seeking to shift from IT-centric solutions and implementations to building on the information and knowledge store within the organisation (Table i, source Bryant 2005a).

Traditional Solutions	*Social Software*
Top-down command and control One-to-many, impersonal Formal, bloated, inflexible Coporate voice Large, slow, expensive Owned by the vendors or IT	Bottom-up, devolved Many-to-many, personal Informal, lightweight, flexible Human voice Small, iterative, cheap Owned by you and your people

Table i: Traditional Solutions vs. Social Software

A characteristic of this approach is the centring on the users without over-burdening them from above. The key population of taxonomy or ontology is from the bottom, although within a top-down framing or seeding. There is additional support for the lateral bridging of elements across groups, rather than the traditional/typical top-down constraining, enabling collaboration with the users instead of shaping them to the technology. In general, this technique seeks to join across the differing and diverse individuals and workgroups within an organisation, but also to allow for the re-factoring of stored information and knowledge around the current and changing needs, creating flexibility and scope for innovation. Core to this approach is the encouragement and stimulation of the social networks and interaction, especially the conversational aspects. These elements seek to expand user attitudes, from single-loop learning and rigid focusing upon direct problem solving, to the adoption of double-loop learning.

Web 2.0 is a reference to perceptions of what the next generation web will look like and can be seen in aspects of the Social Software, services like *flickr* (the online photo sharing community site) or *technorati* (the blog Internet search engine), places/spaces for sharing, an environment providing users with web based applications and collaborative environments and resources that are accessible from any computer and location, regardless of operating systems or software installed on that machine. It reflects a coming of age of aspirations underpinning the thin-client and network-computer approaches proposed in the 1990's. In essence, Web 2.0 is a development from the wellsprings that fed the Social Software movement but increasingly involving larger technology and web-focused organisations like Yahoo and Google. Yahoo purchased *flickr,* while Google followed with the acquisition of *writlely*, the web word-processor environment that enables the sharing of documents and collaboration in real-time, with the ability to limit access and edit documents from anywhere (Ukn Google blog, 2006). Google has recently launched a web based collaborative spreadsheet application (Ukn BBC, 2006) and an online sharable calendar, with further linked support through RSS (Really Simple Syndication), enabling links to content deemed relevant to and for the collaborating users.

3.3. The Invisible Web (IW)

In 2001, BrightPlanet, a search technology company, speculated that IW possibly contained 550 billion documents, perhaps 500 times the content of the conventional Web, when Google – which claims to index the most comprehensive collection of documents on the Internet – had identified 1.2 billion documents and was actually capable of searching a mere 600 million of those (Bergman 2001).

The IW comprises content that search engines either cannot or will not index. Most of the IW is made up of the contents of specialized databases that can be queried via the Web. The results are then delivered in dynamically generated web pages, whose storage is expensive and are therefore discarded as soon as the user reads them. Technical barriers related to the design and functionality of spiders mean that search engines cannot find or create these pages. Spiders navigate the Web by following hyperlinks (a page with no links becomes 'invisible') but can neither type nor 'think'. Hence, specialised databases that are searchable over the Web are inaccessible if they have no static pages with links containing information, so are Web sites that require login. The rest of the IW consists of the so-called excluded pages. They are certain types of pages that the search engines exclude by policy.

They either contain special formats that hinder indexing (e.g. contents in Flash, Shockwave, images only etc.), or script-based pages (e.g. sites with URLs that contain the '?' sign).

Although there are not general tools for searching the IW, there are an increasing number of links and subject directories to invisible Web databases, such as The Invisible Web Directory (IWD 2005). Integration of the traditional and the invisible Web is, of course, problematic. Directed query technology and pre-assembled storehouses provide some (far from seamless) support. The former is cumbersome and places the burden on the user, who has to download the appropriate software and issue effective queries. The latter supports selected content and query customisation which disadvantages general requests and needs.

Quality issues are similar to those encountered in the conventional Web: matters of availability, quality of information and duplication. Duplication is particularly difficult to assess, though the guidelines are similar to those of the 'traditional' web-sites. Sites whose content is unique include topical and scientific databases, library holdings, satellite imaging data and internal site indices. Duplicated sites (and information) include product listings, software, press releases, mirrored sites and search engine results. Nonetheless, assessing IW's overall information quality can be tricky, as there is no standardisation of retrieval methods and no availability of proper statistics of depth and volume. Similarly, the sharing of the information retrieved from the IW is not straight-forward: resulting pages are dynamic and lack of relevant organisational strategy in their storing for sharing purposes can well lead to storage inefficiency. Moreover, neither the Semantic Web nor the Web 2.0 tools and methodologies can be applied here.

4. Towards a New Framework

The frameworks visited lack a number of tools/facilities that we deem essential for supporting KM. Our proposal criticises both approaches by pinpointing their respective advantages (features we need to retain) and disadvantages (issues we need to resolve). Our framework therefore differs in the following points:

(a) Knowledge modelling tools of existing or proposed systems are usually editor and/or form based. As such they are largely counter-intuitive and require expertise not always present where end-users are concerned. The alternative to editor-based schema design is *conceptual modelling*: the process of constructing a model of the information at hand that is independent of the implementation details, application programs and software/hardware considerations. As a concept it applies to the modelling of information and knowledge and plays a central role in the creation of any information repository, from web content to knowledge management systems. Conceptual modelling tools fitted with a graphical user interface have proved to be more appropriate than editor-based environments (Dotsika & Watkins 2004). They facilitate knowledge capture by hiding complexity, are user friendly and can be cost-effective since they automatically generate code.

(b) Whatever the future of the web, there always will be information repositories residing outside the boundaries of the new technologies. Therefore, an integrated approach should try to maintain interoperability with such sources for as long as needed (Dotsika 2003). Current systems provide some access to existing sources, such as KAON's access to relational data sources via OntoMat-REVERSE (Boszak et al 2002), however a full integration with legacy systems would require a more flexible approach that transcends schema architectures.

(c) The idea of enhancing KM by enabling it to tap into the Semantic Web is to make a huge amount of electronically information more accessible by using ontologies to make searches more intelligent. The adoption of a common ontology language has been considered a must for the support of semantic interoperability, resulting in the Web Consortium's OWL recommendation [9]. Ontology language standardisation however is inversely proportional to ontology content design. The quality criteria particularly relevant to semantic web ontologies are *accuracy* (inaccurate ontologies would produce wrong results), *transparency* (opaqueness would affect reusability) and *reason-ability* (otherwise inference would be disabled) (Svatek 2004). There are a number of methods offering ontology content quality support, such as meta-properties, pre-fabricated patterns support, collected hints etc. (Svatek 2004). While most methods fare well with accuracy control, their performance in controlling transparency and reason-ability varies significantly depending on the application area.

(d) However, this typically top-down approach runs the risk of failing in capturing the detail required. This detail usually resides at the bottom, where the key people often find themselves constrained by technology, rigid software support and bad system design. Inability to engage and involve the end-user results in systems that don't get employed efficiently and can potentially lead to system failure. The solution is to combine the flexible top-down framing / bottom-up populating of Social Software with the formal semantics of the Semantic Web.

(e) When it comes to semantic mark-up, storage, scalability and retrieval are problematic areas. Storing semantic web data has led to the debate over the implementation architecture (relational vs. graph-based), while scalability and constant increase of storage requirements have given birth to further storage concerns. The storage debate is well timed as it coincides with the launching of the new file system implementations brought out by the major operating system vendors (Sun Microsystems with ZFS as part of their OS Solaris 10 and Microsoft with WinFS as part of Longhorn). On the retrieval front, query languages at present do not always have the flexibility required (eg. query across multiple graphs and sub-graphs).

(f) The SW framework has been described as overestimating the value of deductive logic, while underestimating the difficulty of a shared worldview (Shirky 2003). Even if the automation of web information retrieval by means of intelligent agents is successful, web contents will always be used and processed by humans as well as agents, with or without the involvement of some partial automated tasks. In this 'traditional' use of the web, indexation takes precedence over formal semantic mark-up, as navigation is more pertinent than inference. Although this approach lacks the advantages of computational deduction it may nevertheless prove enduring due to its low-cost, easily maintenance and no requirements for specialist end-users. Therefore new systems should take this into consideration and look into integrating the cognitive approach with the automated one.

Figure iii sums up the Proposed Framework.

Placement for Figure iii

Figure iii: Web-based KM.

5. Conclusions and Future Work

Our exploration identified several non- exclusive trends that represent views on how the next generation of web could evolve and how the latency of web knowledge can be unlocked. However, there is the inherent problem that each trend may overwhelm the previous one and not allow its full exploration. Indeed computer history is littered with ideas left behind which remain unfulfilled and never fully explored: a problem associated with technology is the penchant for riding the front of the wave, the cutting edge.

It is possible to observe several patterns in how these trends and ideas are driven; from within the existing web-developer environment and from the collaboration and swarming of IT-literate web users seeking to build or help build a shared vision of a web that is customisable and delivers what users want and not what developers think they want. At the same time technology companies seek to create and/or exploit the commercial benefits of the next wave. This can be discerned in the interests of the significant players within the web environment, such as Yahoo, E-bay, Google, Microsoft, etc. that seek to harvest the commercial benefit of the web. This behaviour is shown through their own development, acquisitions and manoeuvring in the marketplace. The significant patterns lie in the collaborative views of the social software movement, which are now solidifying in the Web 2.0 framework and being consolidated into web applications and services. Another significant trend is that of assisting the management of the exponential growth of the web, in relation to the data, information and latent knowledge, which is the base of the

Semantic Web infrastructure with its established potential in information retrieval and knowledge discovery. To this extent we present a framework that could reenergise the development of the potential that lies within the Semantic Web and support the creation of a web of knowledge that is no longer a latent hope.

Based on the above we investigated the main requirements for the support of Knowledge Management in the next generation of web, looked into existing developments and solutions and provided an independent framework for the capturing, accessing and distributing of web knowledge. This framework retains the semantic mark-up, a feature that we deem indispensable for the future of KM, employing web ontologies to structure organisational knowledge and semantic text processing for the extraction of knowledge from websites. Furthermore, our proposal accommodates the collaborative tools and services offered by Web 2.0, acknowledging the fact that knowledge-based systems are shared, dynamic, evolving resources, whose underlying knowledge model requires careful management due to its constant changing.

However, web search and sharing is only part of the problem. An increasing problem lies in user expectation, as more systems are clothed in web-based front-ends that mask the underlying disparate nature of the information repositories, legacy systems and databases that are at the back-end. This suggests to users of all levels functionality that is neither realistic nor practicable, with consequences for systems developers, administrators and managers. It further indicates the need for proactive management of users and has an impact on how their expectations are encouraged and supported.

While our research was based upon web-based knowledge, the next step should include non-web-based sources of information, such as office documents, e-mail messages and news feeds. A recent Butler Group Review (Thornton 2005) reports that anywhere up to 80% of a knowledge worker's time is spent hunting for information and 80% of corporate information is held on users' desktop PCs. Search strategy and practice should include desktop search, thus integrating web servers, file servers, DBMSs and e-mail storage. There are currently a number of desktop search environments that do just that, with Google, Copernic, Yahoo! and MSN Toolbar Suite leading the market.

6. References

Allee V. (1997), *The Knowledge Evolution: expanding organisational intelligence*, Butterworth-Heinemann, Boston.

Bechhofer S., Horrocks I., Goble C., Stevens R., (2001) *OilEd: a reason-able ontology editor for the Semantic Web*, Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence. 19–21 September 2001, Vienna, Austria, Aicommunications, Volume 14, Number 2, 2001, pp. 125-126(2).

Benjamins V.R., Contreras J., Corcho O. and Gomez-Perez A., (2004), *Six Challenges for the Semantic Web*, SIGSEMIS Bulletin, Vol 1, Issue 1, April 2004.

Bergman, M.K. (2001), The Deep Web, Surfacing Hidden Value, The Journal of Electronic Publishing, August, 2001, Volume 7, Issue 1 University of Michigan, Michigan.

Berners-Lee, T., Hendler J., Lassila O. (2001) *The Semantic Web*, Scientific American, May 2001, pp.35-43.

Bozsak E. et al, (2002), KAON – *Towards a Large Scale Semantic Web, E-Commerce and Web Technologies*, Proceedings of the Third International Conference, EC-Web 2002, Aix-en-Provence, France, September 2-6, 2002, volume 2455 of Lecture Notes in Computer Science, pp. 304-313. Springer.

Boyd, S., (2003), *Are you ready for Social Software?* [online], http://darwinmag.com/read/050103/social.html

Bryant, L., (2003), *Smarter, Simpler, Social*, [online] http://headshift.com/moments/archive/sss2.html

Bryant, L., (2005a), M*aking Knowledge Work*, [online], http://headshift.com, June 2006.

Bryant, L., (2005b), *Introduction to Social Software for the Networked Social Enterprise*, [online], http://headshift.com, June 2006.

Brickley D. (1999), *Semantic Web History: Nodes and Arcs 1989-1999 The WWW Proposal and RDF*, [online], http://www.w3c.org/1999/11/11-WWWProposal/, November 2005.

Davies J., Fensel D., and F. van Harmelen, ed*itors,* (2002), *On-To-Knowledge: Semantic Web enabled Knowledge Management*. J. Wiley and Sons.

Dotsika F., (2003) *From data to knowledge in e-health applications: An integrated system for medical information modelling and retrieval*, International Journal of Medical Informatics and the Internet in Medicine, Vol 28, issue 4,. pp231-251, December 2003.

Dotsika F., Patrick K. (2005a), *Knowledge Capture, Sharing and Maintenance in the Semantic Web age: a Framework Proposa*l, Proceedings of the 4th International ISOneWorld Conference, Las Vegas, April 2005.

Dotsika F., Patrick K. (2005b), *From end-users to bots: the balancing act of web-based knowledge search and sharing* Proceedings 2nd International Conference on Intellectual Capital, Knowledge Management and Organisational Learning, Dubai, November 2005, pp143-150.

Dotsika F., Watkins A., (2004), Can conceptual modelling save the day: a unified approach for modelling information systems, ontologies and knowledge bases, Proceedings of the 15th IRMA International Conference, May 2004.

Elliott G, (2004), *Global Business Information Technology: An Integrated Systems Approach*, FT Prentice Hall.

Fensel D. (2001), *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlag, Berlin.

Fox M.S. and Gruninger M., (1998), *Enterprise Modelling*, AI Magazine, AAAI Press, Fall 1998, pp. 109-121.

Gabriel Y., Fineman, S., & Sims, D., (2000), *Organizing & Organisations* (2nd Ed), Sage Publications, London.

Greve, H. Taylor, A. (2000), *Innovations as catalysts for organizational change: shifts in organizational cognition and search*, Administrative Science Quarterly, vol. 45, pp 54-80.

Handschuh, S, Staab, S, Mädche, (2001) A. *CREAM - Creating relational metadata with a component based, ontology driven annotation framework,* Proceedings of First International Conference on Knowledge Capture ACM K-CAP 2001, October 2001, Vancouver.

HP Laboratories Research (2002), *Jena 2 – A Semantic Web Framework*, [online], http://www.hpl.hp.com/semweb/jena.htm, July 2002.

Kalyanpur A., Parsia B., Sirin E., Cuenca-Grau B., Hendler J., (2005) *Swoop - a web ontology editing browser*, Journal of Web Semantics, 4(1), 2005.

Laudon K,C., & Laudon J,P., (2005), *Management Information Systems Managing the Digital Firm* 8th Ed, (Intl Ed), Pearson/Prentice Hall New Jersey.

Lindland O.I. et al. (1994) *Understanding quality in conceptual modelling,* IEEE Software 11 vol.2, pp. 42-49.

IWD (2005), *The Invisible Web Directory*, [online], http://www.invisible-web.net/, October 2005.

McEneaney, J.E. (2001) *Graphic and numerical methods to assess navigation in hypertext*, Intl. Journal of Human-Computer Studies, 55, pp 761-786.

McIlraith S.A., Son T.C., Zeng H. (2001), *Mobilizing the Semantic Web with DAML-Enabled Web Services*, Proceedings of the 2nd International Workshop on the Semantic Web, Hong Kong.

OntoEdit (2002), *Knowledge Modelling with OntoEdit, OntoPrise, Semantics for the Web*, http://www.ontoprise.de/products/ontoedit_en, July 2002.

Navarro-Prieto, R., Scaife, M., Rogers, Y., (1999), *Cognitive strategies in web searching*, Proceedings of 5th Conference on Human Factors and the Web, http://zing.ncsl.nist.gov/hfweb/proceedings/proceedings.en.html.

Newell, S,. Robertson, M., Scarbrough, H,. & Swan, J,. (2002), *Managing Knowledge Work*, Palgrave.

Noy N.F., Sintek M., Decker S., M. Crubezy, R. W. Fergerson, & M. A. Musen. (2001), Creating Semantic Web Contents with Protege-2000. IEEE Intelligent Systems 16(2):60-71.

W3C OWL (2004), *OWL Web Ontology Language Overview*, [online], http://www.w3.org/TR/owl-features/, April 2005.

Phillips N, & Patrick K, (2003), *Personality Type and the Natural Development of Knowledge Evolution* in Knowledge Management in the *Sociotechnical World: The Graffiti Continues*, eds E Coakes, D Willis, & S Clarke, Springer.

Pitkow, J. E., Pirolli, P. (1999) *Mining longest repeated subsequence's to predict World Wide Web surfing*. Proceedings Second USENIX Symposium on Internet Technologies and Systems, 11-14 October 1999.

Sesame (2004) *openRDF.org*, [online], Home of Sesame, http://www.openrdf.org, July 2004.

Shirky Clay, *The Semantic Web, Syllogism, and Worldview, Economics & Culture, Media & Community*, [online], www.shirky.com, November 2003.

Svatek V., (2004), *Design Patterns for Semantic Web Ontologies: Motivation and Discussion*. In: 7th Conference on Business Information Systems, Poznaň, 21-23 April 2004.

Thornton, R, (2005), *On the Road to Business Integration*, Butler Group Review Journal, February 2005, [online], http://www.butlergroup.com/review/default.asp, September 2005.

Ukn BBC (2006) *Google launches web spreadsheet*, [online], http://news.bbc.co.uk/1/hi/business/5051610.stm, June 2006.

Ukn Google blog (2006) *Googler insights into product and technology news and our culture*, http://googleblog.blogspot.com/2006/03/writely-so.html, June 2006.

Wang, W. and Zaïane, O. R. (2002) *Clustering Web Sessions by Sequence Alignmen,* Proceedings of DEXA Workshops, pp 394-398 IEEE Computer Society 2002, ISBN 0-7695-1668-8.