A Knowledge Services Roadmap For Online Learning

by

Anand Rajagopal

B.Tech, Civil Engineering
Indian Institute of Technology-Bombay, India, 2002

Submitted to the Department of Civil and Environmental Engineering
and the Engineering Systems Division
in Partial Fulfillment of the Requirements for the Degrees of

Master of Science in Civil and Environmental Engineering

and

Master of Science in Technology and Policy

at the
Massachusetts Institute of Technology
February 2005

Signature of Author.....
Civil and Environmental Engineering and Technology and Policy
January 14, 2005

Certified by.....
John R. Williams
Professor of Civil and Environmental Engineering and Engineering Systems
Thesis Supervisor

Accepted by.....
Dava J. Newman
Professor of Aeronautics and Astronautics and Engineering Systems
Director, Technology and Policy Program

Accepted by.....
Andrew Whittle
Chairman, Departmental Committee on Graduate Studies
Department of Civil and Environmental Engineering

A Knowledge Services Roadmap For Online Learning

by

Anand Rajagopal

Submitted to the Department of Civil and Environmental Engineering
and the Engineering Systems Division
on January 14, 2005 in Partial Fulfillment of the
Requirements for the Degrees of
Master of Science in Civil and Environmental Engineering
and Master of Science in Technology and Policy

ABSTRACT

In today's society, there is a need for organizations to have a robust knowledge infrastructure in place, so that they can create or acquire knowledge; store knowledge; disseminate knowledge, and protect and manage their knowledge assets. However, with advances in the publishing media, our ability to generate information has far exceeded our abilities to find, review and understand it, thus leading to "Information Overload". Information overload refers to the inability to extract needed knowledge from existing information due to the volume of information, or lack of understanding of information and its whereabouts, or efficient ways to locate relevant information. These issues could be addressed by having efficient **Knowledge Management Systems/Knowledge Services**, so that people can create and understand available information, and have services to help them learn effectively and make better decisions.

To tackle the new information needs, the use of technologies such as **Weblog Services** (weblog-enabled knowledge services) offer opportunities for decentralized knowledge creation and dissemination; as such tools put the authors in charge of knowledge creation process without any administration-enforced policies. Learning environments are also typically characterized by challenges such as barriers to use, quality control and relevance issues, or issues of credibility of information. These issues are effectively tackled by weblog services since weblogs are often open source and need no training for authoring. In addition, favorite blogs act as information filters or "bird dogs" and point at useful information. Feedback incorporated in weblog services makes people react and learn "interactively" and also enhances credibility and trust in information.

Weblog services can also share published content through the process of Content Syndication, and thus offer an insight into knowledge assets in the timeliest of ways. This thesis report describes certain weblog services implementations carried out at MIT. Results of such implementations have emphasized the applications of such weblog (knowledge) services in knowledge sharing and online learning. However, there are certain issues to be addressed in weblog services such as privacy and intellectual property issues, as well as resolution of organizational tussles in the domain of content syndication standards.

Thesis Supervisor: John R. Williams
Title: Professor of Civil and Environmental Engineering and Engineering Systems

Table of Contents

List of Figures

1. Introduction

In today's society, the success of organizations and individuals hinges upon their ability to "locate, analyze, and use information skillfully and appropriately." (Nelson) However, with advances in the publishing media (online and offline), our ability to generate information has far exceeded our abilities to find, review and understand it. In particular, the volume of information on the Internet has exceeded the ability of most people to find the information they need, thus giving rise to the concept of "Information Overload". Murray (1966) estimates the following: "In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would require 1.5 months, reading 8 hours every day, to get through 1 day's technical output, and at the end of that period, he would have fallen 5.5 years behind in his reading!" Wurman (1989) writes, "a weekday edition of The New York Times contains more information than the average person was likely to come across in a lifetime in seventeenth-century England."

Broadly defined, "**Information overload**" is the inability to extract needed knowledge from an immense quantity of information for one of many reasons. Wurman (1989) explains that information overload can occur when there is:

- Inability (on part of the person) to understand available information.
- Overwhelming volume of available information.
- People do not know if certain information exists.
- People do not know where to find information.
- People know where to find information, but do not have the key to access it.

One result of having all of the Internet information available is an increased difficulty in finding the particular information for which we are searching. We have to determine which information is useful, which is not, and where to look next when necessary. Fine and Newman distinguish between *Information* and "*Real Need Knowledge.*" There is a difference between what information is available to us and what information we need or can use. In addition to our own perception of useful - "real need knowledge", it is crucial to acknowledge that information considered as real-need knowledge and useful by one person might be deemed as unusable by another person. If all information was of equal value to everyone, controlling the volume in a more meaningful way might be easier. Unfortunately, this is not the case and so a person must search for *personally meaningful information.* To reduce the

amount of time used to search for "real need knowledge" we must find a way to successfully overcome this issue in the information overload problem.

The volume of information on the Internet creates more problems than just trying to search an immense collection of data for a small and specific set of knowledge. Large volumes of data are fraught with inconsistencies, errors and useless data. When we try to retrieve or search for information, we often get *conflicting information* or information which we do not want. Therefore, validating information is another important aspect of information overload. Individuals searching for information want to maximize the quantity and quality of relevant information retrieved. Curtis and Rosenberg (1965) contend that a search request consists of two components which determine the accuracy of search results. The first is the point of view of the person seeking information and the second deals with system functionality. "Users must take on the responsibility of designing and structuring the parameters of their searches to match their own points of view. Depending on a user's ability to select search parameters, one hopes to eliminate a large portion of false drops, while not excluding relevant documents." (Nelson) False drops are those documents which match given search criteria, but are irrelevant to the user's needs. According to Curtis and Rosenberg (1965), an individual must be willing to sacrifice the possibility of "instantaneous response" for greater accuracy of retrieval.

There appear to be two major factors affecting the ability of people to access information effectively: **information literacy** and **application usability**. Information literacy is "the ability to effectively access and evaluate information for a given need." (Breivik) Application usability refers to the interactive environment which a software application or system provides to a user searching for information. These two factors complete the information ecology in the Internet life-cycle today and lay down the foundations for establishing Knowledge Services/Knowledge Management Systems. Information literacy is the "people" aspect of information access. (Nelson) Horton (1983) described the purpose of information literacy as:

...raising the levels of awareness of the knowledge explosion and involving understanding as to how computers can help identify, access, and obtain data and documents needed for problem solving and decision making. Acquiring the ability to understand how to find/search information is an important element in the process of overcoming information overload. "Some skills which contribute to information literacy are problem solving, decision making,

critical thinking, information gathering and interpretation" (Breivik). These skills are needed in addition to a basic competence and familiarity with computers. The application usability side of information access requires that computer-based information systems be designed for ease of use. Hence, keeping these factors in mind, the need is for organizations to have a *robust knowledge infrastructure* in place so that they can *create or acquire knowledge; retain and store knowledge; disseminate and use knowledge, and protect and manage their knowledge assets*. This can be brought about by having efficient **Knowledge Management Systems/Knowledge Services** that address the issues of information literacy and application usability, so that people can create and understand available information, and have efficient services to help them learn effectively and make better decisions.

In the last few years, an increasing number of studies have been published examining issues related to the development and use of Knowledge Management Systems. "In general, most of these studies have either been conceptual studies that define terms, studies that identify important issues or anecdotal case studies that have described a particular development of a knowledge management system." (Gallupe, 2000) However, the terms defined in such studies still mean different things to different people and there's no systematic framework to guide KMS/Knowledge service research. This thesis report aims at explaining some of these concepts and builds the foundations for knowledge services/systems to enhance learning in organizations.

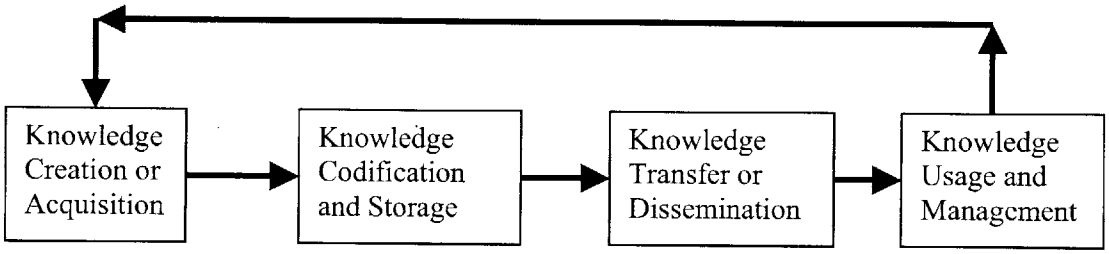


Figure 1: Knowledge Life-Cycle (Gallupe, 2000)

In terms of defining such **knowledge management systems or knowledge services**, a starting point would be the definition of knowledge itself. Knowledge is information combined with experience, context, interpretation and reflection (Davenport, 1998). For the most basic classification, knowledge can be thought of being either **Explicit** or **Tacit** (Polanyi, 1966 and Takeuchi). *Explicit knowledge* (or *Know-What*) is the kind we are familiar

with — information that is easily stored and shared. When we listen to a lecture, read a newspaper, watch the evening news, we are encountering explicit knowledge. Such knowledge can be easily codified, stored in machines, or documented and archived, for instance patents, trademarks, business plans, marketing research and customer lists. *Tacit knowledge* (or *Know-How*) is fuzzier, more subjective, and historically it has been less manipulatable. It is knowledge that arises less from the mind than from human experience. It includes emotional experience, dreams, belief systems, sudden insights, and flashing intuition. In Takeuchi's classification, tacit knowledge is gained in just about every experience, from tasting ice cream to witnessing a sunset. Traditional **Knowledge management systems** (KMS), which are essentially a federation of **knowledge services**, typically focused on explicit knowledge, but increasing attention is now being given to KMS supporting tacit knowledge capture and transfer.

1.1 Knowledge Services/KMS and Learning Networks

The fundamental concept providing the basis for KMS is the *Systems* concept. As explained earlier, the two main issues of Information Literacy and Application Usability lay down the foundations of knowledge services/systems. A *Knowledge management system* or a federation of interacting *Knowledge Services* is *information ecology*, comprising of the interacting components of people (knowledge workers, managers etc.), technologies, and knowledge itself. Extending the same definition, knowledge management systems or knowledge services are services designed and developed to give decision makers/users in organizations the knowledge they need to make their decisions and perform their tasks. These services go beyond traditional information systems in that they provide the "context" for the information presented, and are thus useful to a variety of different settings focused on learning and collaboration in the academia and the industry. However, a big problem with Knowledge Management (KM) itself is that the term has come to mean many different things to different people, and hence nothing at all. (Pollard, 2003)

In most organizations KMS/knowledge services is epitomized by the corporate intranet, the extranet, community-of-practice tools, sales force automation tools, customer relationship management tools, data mining tools, decision support tools, databases purchased from outside vendors, and sometimes business research and analysis. In other words, it's certain specialized technologies and information processing roles, with a thin wrapper of 'knowledge creating' and 'knowledge-sharing' processes. Most of the organizations that have implemented

KM/knowledge services bemoan their people's inability to find stuff, the lack of demonstrable productivity improvement, the complexity of the technology, and the absence of significant reusable 'best practice' content. The reasons for failure of such KMS or Knowledge Services were unrealistic expectations for bringing about changed human organizational behavior, without any compelling argument for people to use the complex KM tools/Knowledge Services. (Pollard, 2003) At the same time, the field of Knowledge management over the years has led to notions of **Social networks** or **Learning Networks.**

The concept of *Social or Learning networks* can provide the essential context needed to make knowledge sharing possible, valuable, efficient and effective. (Pollard, 2003) These networks are the circles in which we make a living and connect with other people – essentially the ways in which we connect with others and learn (interactive learning). Every individual is the source of knowledge creation and it's this knowledge that gets shared between the others connecting with this individual. These networks transcend strict delineation between personal and business (there's often overlap between the two). They transcend organizational boundaries and hierarchies (we often trust and share more with people outside our companies, and outside our business units, than those inside, and often get better value from the exchange to boot). So, in many ways, such learning networks are characterized by "user-centered learning" (Anderson, 2004) - the instructor takes a back seat; users are empowered to learn on their own and teach one another. (Rajagopal et al, 2004) However, it has been realized that knowledge is sometimes very effectively conveyed to people by experts in a given subject area, and in the "context" of a given business/academic problem, for instance the weblog of Don Box (Don Box's Spoutlet) is considered as a vital source of knowledge in the domain of XML messaging and distributed computing. Thereafter, emphasis in learning networks should now be on capturing the specialized *Know-Who*, i.e. the granular identification of experts inside and outside the organization whose expertise can be quickly brought to bear to solve specific problems, and the best means of contacting those experts just-in-time. In addition to this implication for the KMS/Knowledge Services, the focus should be on capturing, codifying and sharing the *Know-What* and *Know-How* as well. Hence, learning networks/environments need KMS/Knowledge Services that capture 'Know-What', 'Know-How' and 'Know-Who' - all at the same time. (Pollard, 2003)

It has also been realized that traditional KMS/Knowledge Services offer people too many complex tools and technologies to use, but people rarely know how to use those tools

effectively and are misinformed about the overly ambitious objectives of the KMS/Knowledge Services. Hence, there is a need to have a *diverse set of knowledge services that allow personalized, customizable and independent usage*. In other words, the capture, organization, recall and dissemination of documents, messages and other personal knowledge has to be in an intuitive, transparent, automatic, personally customizable and simple manner.

New tools such as **Weblogs**, wikis and discussion group forums fulfill this requirement to a great extent. So, organizations now use a combination of technologies to address the issues in Knowledge sharing and management. Such systems allow people to participate locally rather than following enforced technology-based KM Policies. This thesis report focuses on the use of **Weblog Services** for enhancing such learning networks. The use of technologies such as **Weblog-Enabled Knowledge Services** or just **Weblog Services** offer opportunities for decentralized knowledge creation especially in learning and business applications (industry) and dissemination as such tools put the authors/users in charge of knowledge creation process without any administration-enforced policies.

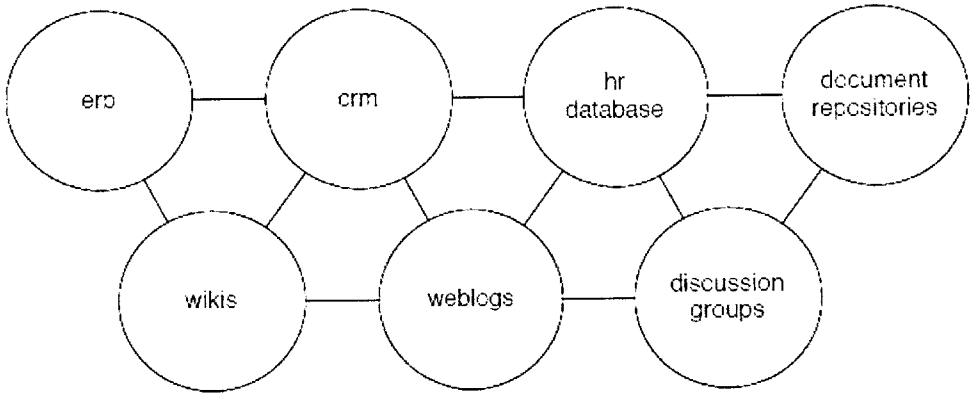


Figure 2: Contemporary KMS/Knowledge Services (Angeles, 2004)

1.2 Learning Challenges and Weblog Services – A Preview

- "Barriers to Use?

  Weblogs are often open source, need no training for authoring" (Rajagopal et al, 2004)
- "Finding Relevant material issues, Quality Control?
  1. Favorite blogs act as filters and "bird dogs"
  2. Point at useful material in repositories building channels
  3. Bring links in from the internet (relevance)" (Rajagopal et al, 2004)

- "Enhanced learning by feedback and pointers?
    1. External links in weblogs illustrate thinking and make points
    2. Weblog feedback makes people react" (Rajagopal et al, 2004)
- "Credibility and Trust in information?
    1. Credibility built by Weblog (Posting) Frequency
    2. Word of mouth of other blogggers
    3. Blog links build credibility & relationships" (Rajagopal et al, 2004)

1.3 Learning Networks in Organizational Effort

Learning networks are social networks/environments that bring about sharing of not just the "know-what" but "know-who" and the "know-how" or contextual information as well. The benefits of such network enablement can be illustrated by use of a *vector-based model*. Let's assume that each person in an organization is represented by a vector, the length of the vector depicting the effort by the person and the direction indicating the direction in which the organization moved as a result of such efforts. (Valdemarin et al)

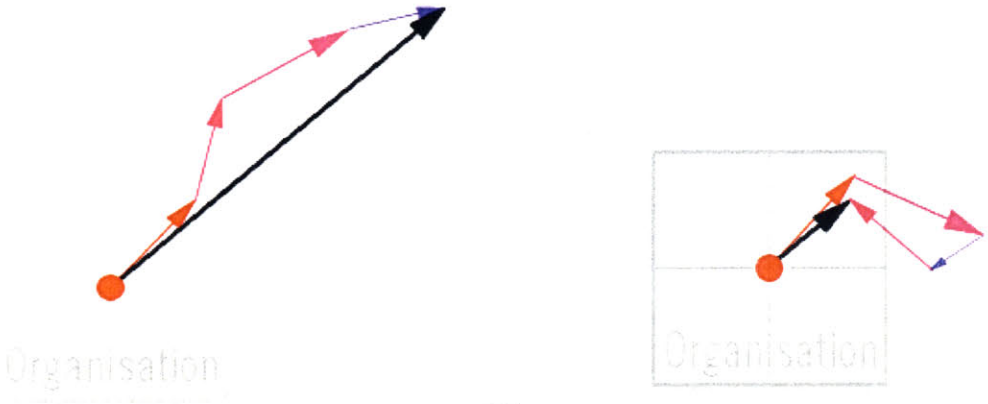
Figure 3: E-Vector for Organization (Valdemarin)

A group of people in the organization can be represented as a set of vectors, each person producing their own efforts and in their own direction.

OR

Figure 4: E-Vector for Organization (Valdemarin)

The progress of the organization can be traced out as a sum of the individual vectors of all the people in the organization. The lesson is that it does not just matter how much effort is expended by the people in the organization, the efforts need to be aligned with the strategic goals of the organization to realize the goals set out! So, the learning networks need to ensure that the output of such organizations or activities in such organizations are aligned with the strategic goals, for which the people need to learn and be informed in the best possible way.



VS.

Figure 5: E-Vector for Organization (Valdemarin)

1.4 Where do Weblogs come into the picture then?

What helps aligning efforts in any organization is shared knowledge and enhanced learning. The more easily and quickly knowledge is shared between people in a group, the more likely it is that every component of the group will move in the same direction as others. This is exactly the area where Weblogs could prove to be very useful. A **weblog** or **blog** is a web page that contains brief, discrete chunks of information called *Posts*. These posts are arranged in reverse-chronological order (most-recent posts come first).Each post can be identified by an anchor tag, and its marked with a permanent link that can be referred by others who wish to link to it. (Doctorow, 2002) Some blogs serve as *micro-portals* or *filters*, publishing commentary and links to other sites relating to a particular topic; whereas others lean more toward *online journals*, where the content focuses mainly on the thoughts and experiences of the author. (Lindahl, 2003) In any case, a blog usually takes on the character of the person or persons that contribute to it because it is so simple to update. This ease of use leads to frequent posting, which creates fluid, ongoing "conversations" with an audience that helps to bring out the nature of the person "behind the screen". (Stone, 2003) On the same note, publishing cycles have traditionally been slowest in case of books, faster in journal

publications, and fastest through conference papers. Some web publishing is done but there is no generally accepted channel or technology to drive this effort ahead. The practice of weblogging is a new dimension in the same spectrum, and promises to be the future of *online publishing*. (Rajagopal et al, 2004)

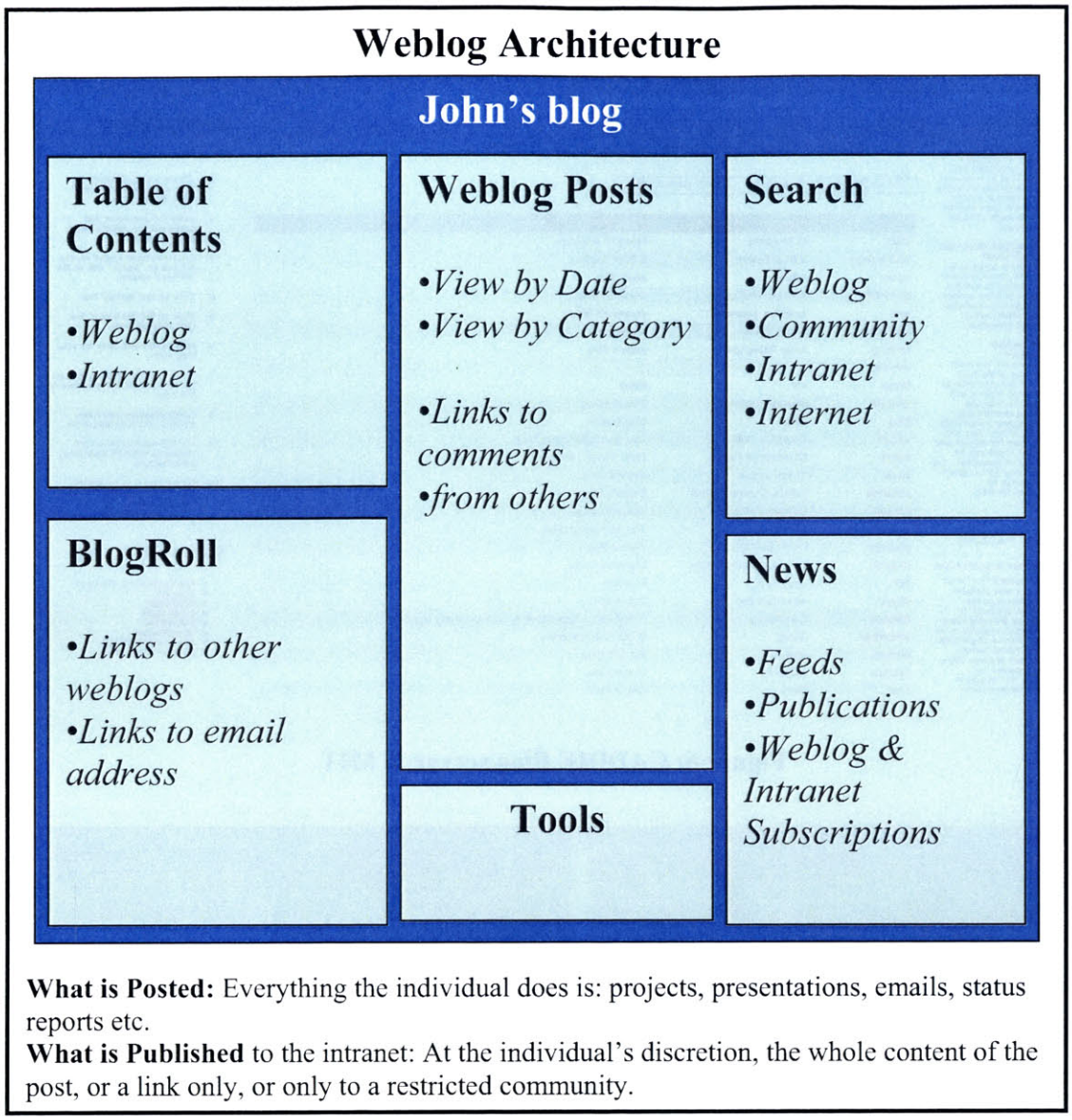
Weblogs are a very powerful tool for sharing knowledge in learning networks because they allow people to manage knowledge which is not easily shared by other common tools or approaches. Typically, people find knowledge in the form of a document or web page, but most knowledge really begins life as a granule of information, a piece of micro-knowledge which it may not be worth making into a whole document so that the knowledge disappears, to be re-discovered again later if needed. It is only when the cost of repeated rediscovery begins to bite that effort is finally made to turn it into a *formal* document. The corollary to this is that the benefits of knowing this information are not shared until quite late in its lifecycle. By contrast weblogs allow everybody to narrate, day by day, their experiences, their work, and their points of view. A weblog entry can range in length from a few words to whole essays. They also help people to quickly and effectively build relationships with others, which aids on collaborative learning activities. A good weblog is more than just the sum of its posts. It tells you something about the person behind it, about what is important to them. (Pollard, 2003)

1.5 Weblog Architecture and Implementations carried out

Weblogs aren't about new pages, but rather about new posts. In the age of ever-decreasing attention spans, reading a new paragraph or two each day was easy to expect of readers, and more likely than readers taking in a new 1000-word essay each day. While traditional web sites were based and organized around the **Page Paradigm**, weblogs were organized and built around the **Post Paradigm**. The Post Paradigm is a way of looking at chunks of information within a larger framework. (Bausch, 2002) This view of weblog content, in comparison to traditional web page-level content, is sometimes referred to as Micro-content. Micro-content is easy to read, easy to understand, and is small enough to enable flexible options for display. A post to a weblog could just as easily become the contents of a short email to someone, or to a group that is subscribed to daily updates via email instead of the web. A weblog post could also quite likely be sent in the entirety via instant message to someone. A post could even be sent to someone's phone, if that person has text messaging

support. Given the way weblogs are organized, each post, or chunk of micro-content, can have a permanent address on the web and allow others to point to specific ideas within posts.

In order to assay **Weblog Architectures** in organizations, blogs can be thought of as filing cabinets (Pollard, 2003). "The *filing cabinet* is more than just a place to store copies of documents. It is a representation of the way people think, learn and work. It is organized according to their personal mental model of how their jobs break down, so that two people doing the same job will often have completely different-looking filing cabinets." (Pollard, 2003) In the same spectrum of filing cabinets, weblogs allow each person to personally identify who *he or she* thinks actually belongs to and participates in his or her learning networks (using the *blogroll*), rather than who their management thinks should be in those networks. The blogroll consists entirely of *active links* to the blogs of the other community members, so knowledge is electronically and personally connected. Knowledge can be simply and flexibly indexed (and sorted or filtered) by date and category (using each individual's personal taxonomy or 'filing system', not some standard taxonomy system imposed by management). Instead of containing redundant copies of knowledge from other people like a filing cabinet, the blog simply hotlinks to the 'permalink' (the dynamically-generated URL for a particular piece of knowledge or 'knowledge object') in the other person's blog/filing cabinet. The knowledge is enriched by dynamic links to URLs of relevant news, bibliographies and other external resources used in its compilation, thus greatly increasingly its shelf life by allowing it to be more easily updated. The key external resources (journals, manuals etc.) that a person uses frequently can be stored in a 'resources roll', consisting of the URLs of these resources; by copying and using an expert's 'resources roll', an apprentice could discover and mimic the 'continuous learning' process of the expert. E-mails are the most valuable untapped codified knowledge resource in most organizations, and blogs allow knowledge to be simultaneously posted to one or more e-mail addresses *and* to the owner's indexed blog/filing cabinet. People can easily 'subscribe' to each other's entire blog by a process of syndication, followed by feed subscription (or an individual category/folder subset of it), so they are immediately notified about new knowledge or news that their work teammates or mentors deem valuable. Blogs do not require the learning of HTML or database management, though they perform both functions powerfully. They can easily be designed to either live within a company firewall or to transcend organizational boundaries, and to be accessible in whole or part to some or all other employees, as the trade off between security and value-of-sharing dictates.



Weblogs have grown at an exponential rate, from a handful in 1998 to over a million in 2003 (source - Technorati). This growth has been driven by the popularity of different blogging engines and by a process called as syndication. As part of the thesis, 2 separate forms of blog implementations were deployed and tested out by different types of users. On one hand, a **central blog server** (*CADDIE Blog server*) has been administered for the users at MIT. The server has attracted as many as 70 users – comprising of individuals, research labs, classes, academic programs such as System Design and Management and the MIT Admissions office. On the other hand, blogging systems – consisting of **group** and **individual** blogging systems have been deployed as part of *CADDIE .NET* – a web-service based content architecture, which has been distributed across many educational and non-profit institutions.

Figure 6: CADDIE Blog server at MIT

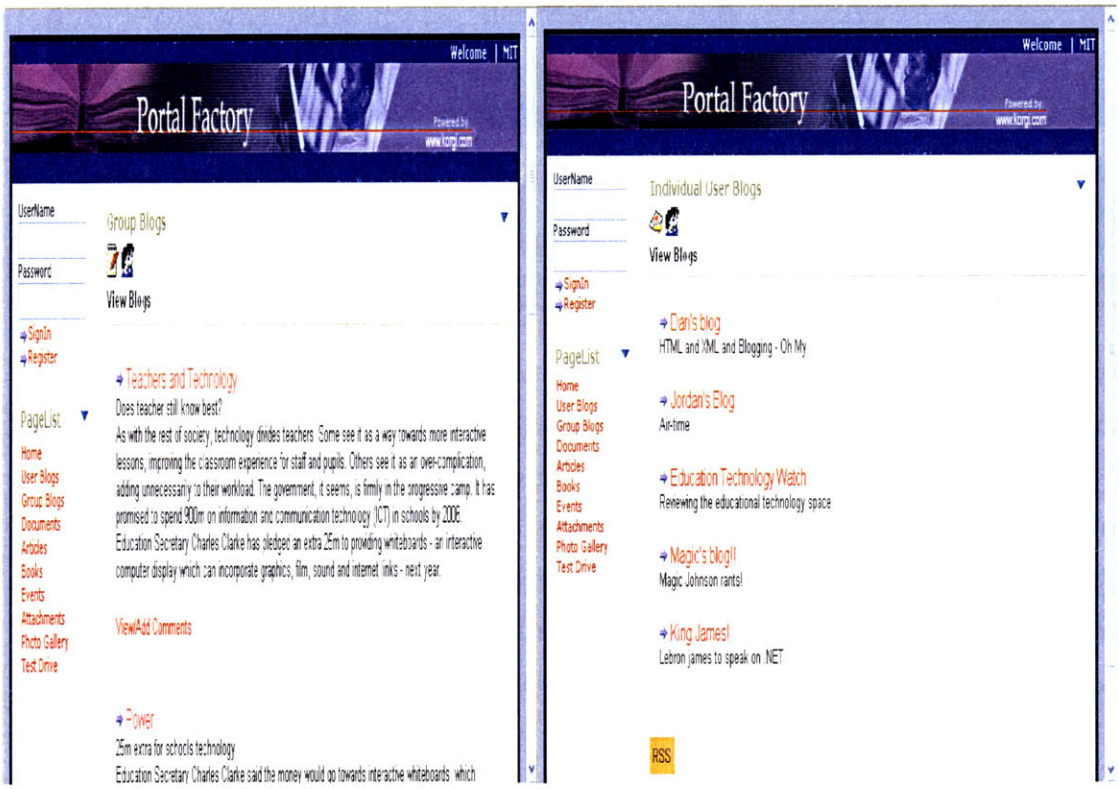


Figure 7: CADDIE .NET Blogging Systems – Group (left), Individual (Right)

1.6 Information Overload in Blogging – Role of Content Syndication

Bloggers or owners of weblogs are always looking for ways to make their content more popular, or attract more traffic. This is done by a process called **Content Syndication**. Content syndication makes part or all of a site's (in this case the weblogs) available for use by other services. (Stone, 2003) The syndicated content, or feed, can consist of both direct content itself and metadata – information about the content of the weblogs. The technology or data standards to do this range from the simple beginnings of **RSS 0.91**, through to the **RDF-based RSS 2.0**, all the way to industrial strength **NewsML, ICE, ATOM** etc. *Resource Description Framework Site Summary/Really Simple Syndication* or short for RSS is a dialect of XML. At the top level, a RSS document is a **<rss>** element, with a mandatory attribute called version that specifies the version of RSS that the document conforms to. If it conforms to this specification, the version attribute must be 2.0. Subordinate to the <rss> element is a single **<channel>** element, which contains information about the channel (metadata) and its contents. Each channel has 3 required elements, namely title (name of channel for users to refer to), link (HTML link to feed), and description (phrase describing the channel). A channel could also contain any number of **<item>** elements. Each item element represents a "story" -- much like a story in a newspaper or magazine; if so its description is a synopsis of the story, and the link points to the full story.

Figure 8: A Weblog – "Large Scale Computing" and its RSS Feed

The idea of content syndication through XML-based RSS/ATOM feeds seems to be robust, but how do users locate feeds? Information flow in this context is entirely dependent on location of appropriate feeds followed by subscription by the users in organizations. **Registries** such as *Syndic8.com* detail thousands of feeds which could be utilized by the users. At the same time, there are efficient **aggregators** available in the market today such as *SharpReader, Meerkat* Service etc. that add an additional layer of usability to RSS feeds. In addition to registries and aggregators, **search engines** such as snewp.com limit their indexing efforts solely to RSS feeds. (Hammersley, 2003) All these technologies together have enabled the adoption of RSS data formats in the corporate intranets to allow employees to track news sites for mentions of their organizations.

However, a comprehensive view of the blogging world reveals not just a deluge of content and subsequently RSS feeds, but also a *Power Law distribution* in weblogs and their feeds. (Colin, 2003) In this over-saturated market of blogs, "power laws" are inherent if "power" is equated to any of the following: internet traffic, inbound links, comments from readers or "near-top-of-list" keyword searches.

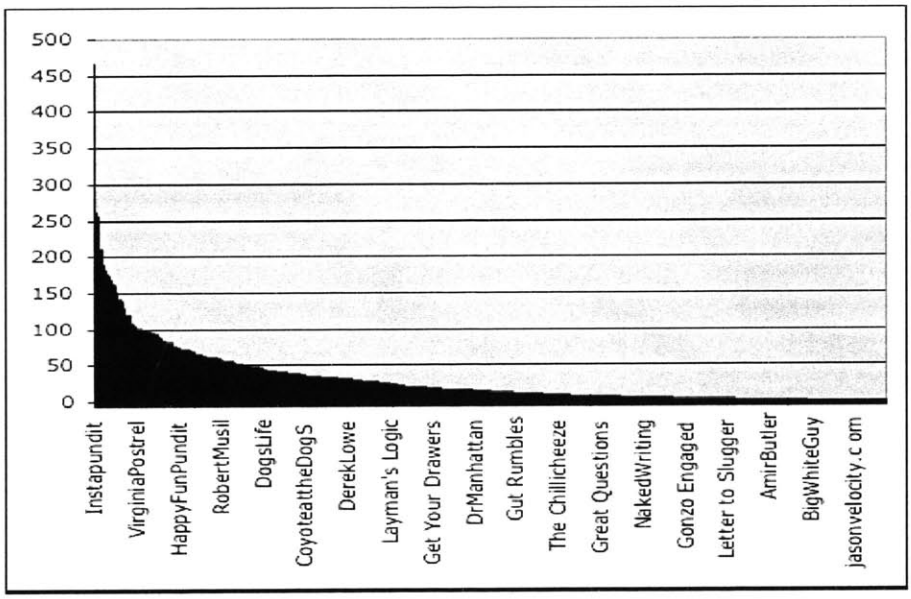


Figure 9: Shirky's Power Law Distribution (Colin, 2003)

The power law stated by Shirky leads to the following consequences: (Rajagopal et al, 2004)

1. Small subsets of bloggers attract disproportionate traffic and attention – "Head" bloggers.
2. Relatively harder for latecomers than early birds to become "blog stars".

3. However, "head" would mainly be "broadcasters", as they have no time to correspond, converse or communicate.
4. The "tail" bloggers will be journal/notebook types writing for small audiences and engaging them in conversations.

In such a power law distribution, the problem of attaching value to blogs could prove to be very useful, since it would help users to better locate relevant feeds and find information in easier ways. Following the same thought, the thesis proposes a **Feed Location and Analysis Model**, which would comprise of the following:

1. *Feed Crawler*: use to crawl a URL/Site for possible RSS/ATOM feeds.
2. *Feed Analyzer*: analyze the feed for keywords, process text, comments, and trackbacks.

The Crawler, and Analyzer coupled together along with a mechanism with a traffic measuring/statistical functionality would provide an efficient **RSS Feed (Upstream) Generator** – essentially designed to serve "valued" and relevant feeds to the end user based on their keyword searches. A sequence diagram to depict the same is shown below:

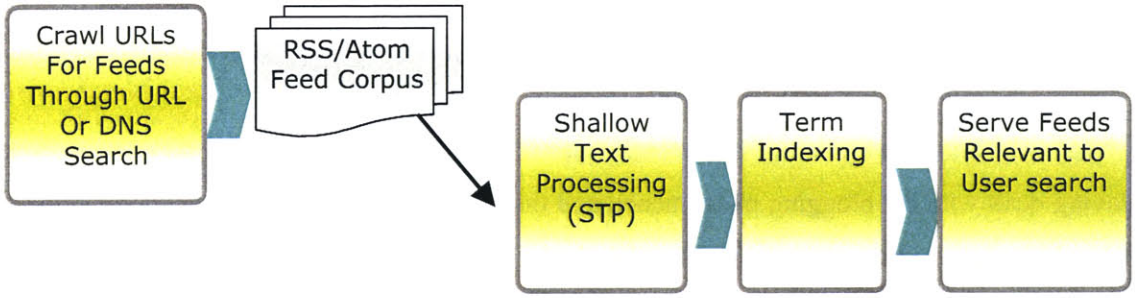


Figure 10: Sequence Diagram for Feed Location and Analysis Model

A brief description of key terms used:

Shallow Text Processing (STP): preliminary manipulation of text such as term extraction, word normalization, etc. (Chaiworawitkul, 2004)

Term Indexing: index terms from STP based on their frequency in a feed & overall corpus (Vector Space Model)

So, all the text in each of the feeds are analyzed using the vector space model and then their frequency of occurrence is found in each feed. Once the text or terms are analyzed, the user's keywords could be compared with the indexed terms/text and the feeds with the highest term

index/effective frequency would be served to the users as search results. However, the limitation with this approach is that the thesis focuses only on keyword search, while searching for cross-linking, comments and trackbacks along with traffic would be a futuristic goal.

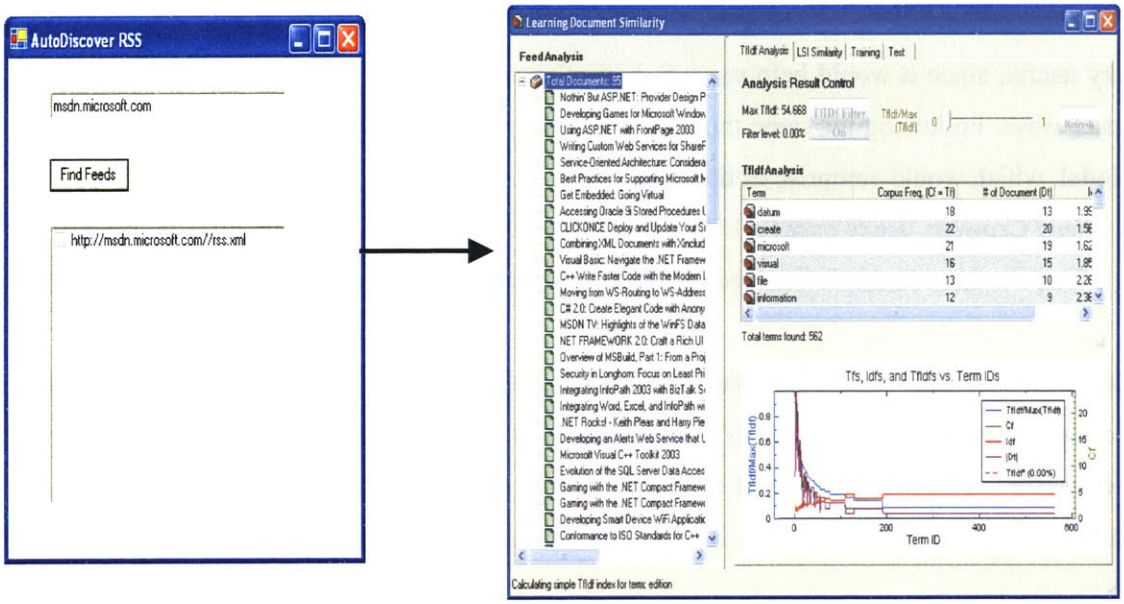


Figure 11: Snapshot of the Proposed RSS Upstream Feed Generator

1.7 Issues in Blog Services and Report Layout

Having dealt with the blogging phenomenon and the problems of information overload in the same, this thesis report also covers the ethical and legal issues in blogging services. While blogging can be applied to a variety of organizations such as the ones in academia such as schools and universities, and in industry, there remain a host of issues in the successful implementation of such systems. Some concerns are as follows:

- Internet addresses often change; weblogs can then have "**Dead Links**" hindering continuity in classroom materials (Ovarec, 2002)

- **Privacy** issues owing to surreptitious recoding of individual's travels on the Internet.

- Online activities and personal details often analyzed for marketing value, government officials without subjects' awareness or consent.

- Intellectual Property issues about "free voice" of employees blogging in firms, and also with researchers in R & D organizations at risks of losing their "original ideas".

This thesis is aimed at conducting a detailed study of a knowledge services roadmap (specifically weblogging) in enhancing online learning. The following chapters would address the following:

Chapter 2 goes into further details on the practice of weblogging, assaying the various applications and syndication standards, adoption in organizations both in academia and the industry.

Chapter 3 looks into the blogging implementations carried out as part of the thesis, including the centrally administered blog server and the CADDIE .NET blogging systems that have been distributed across many educational institutions.

Chapter 4 covers the areas of information overload in the weblogging spectrum, addressing issues such as Shirky's power law distributions and data mining efforts such as the RSS (Upstream) Feed Generator.

Chapter 5 deals with the typical applications and associated legal, ethical and major policy issues faced in the adoption of weblogging as a standard practice in the 21st century organizations.

2. Blogging Phenomenon: History, Existing Tools and Applications

Weblogs are pages consisting of several posts or distinct chunks of information per page, usually arranged in reverse chronology from the recent-most post at the top of the page to the oldest part at the bottom. Each post can be identified by an anchor tag, and it is marked with a permanent link that can be referred by others who wish to link to it. (Doctorow, 2002) Some blogs serve as *micro-portals* or *filters*, publishing commentary and links to other sites relating to a particular topic; whereas others lean more toward *online journals*, where the content focuses mainly on the thoughts and experiences of the author. (Lindahl, 2003) Weblogs are often created and maintained by one person, but they may also be done by small groups of people, and still others may involve large communities of many people on a single weblog. In any case, a blog usually takes on the character of the person or persons that contribute to it because it is so simple to update. This ease of use leads to frequent posting, which creates fluid, ongoing "conversations" with an audience that helps to bring out the nature of the person "behind the screen". (Stone, 2003) Blogs are more often non-commercial ventures done purely for enjoyment, although they are now being increasingly being added to commercial websites and being used in the workspace as a new form of business communication. Weblogs may be a small part of a bigger site, a small portion of a single page on the site, or they may be the entire website. This is due, in part, to the smaller changes and lower amount of effort required to add a new weblog post as opposed to adding an entire web page full of content to a typical site. While it is worthwhile to discuss about the weblog technologies and their applications, it is also important to analyze the evolution of weblogging since the early 1980s to the blogging phenomenon as it exists today.

2.1 Blog Evolution

The exact beginning of weblogs is an often-argued point in any discussion of weblog history. A page describing the new content available on a single site, or on the Internet as a whole, shares properties with weblogs. There's a chronological list of dates used to organize the page, and indicates small chunks of text peppered with links to specific locations described in the text. Some of the earliest pages on the web were "What's New" pages, covering new developments on single servers as well as new sites coming online. "One of the first graphical web browsers, National Center for Supercomputing Applications' (NCSA) Mosaic, was programmed by a group at the University of Illinois at Champaign-Urbana, which maintained a list of new sites on the web as early as 1983. Other pages that share properties with weblogs

were journals or diaries that were posted online. These were usually set up so each day's entry took up an entire page, and a journal spanning months or years could stretch for hundreds of pages in a site." (Bausch, 2002)

Since 1998, weblogs began taking up a new form that hadn't quite been seen before. A single page – the index page of the weblog would change slightly each day as small chunks of text or links would be "added" each day. In contrast, traditional weblogs or sites (before 1998) would have had new pages added for modifying the overall site or weblog. In other words, there has been a transition from a *Page Paradigm* in the good old days to the newer *Post Paradigm*, making small changes to the existing page's chunks or posts. They were easy to revisit and catch up on what one missed; one could simply go down to the last entry and read upwards. (Bausch, 2002) Such small modifications or changes made revisiting, reading and digesting weblogs easier compared to traditional sites, thereby increasing the weblog audience or "blogosphere" substantially. Early weblogs were also quite often link-driven, or loaded with links to interesting offline pages, enticing readers to revisit each day to find new links to obscure places. "With a steady stream of interesting content, updated daily or possibly multiple times per day, weblogs gained audiences that were likely to revisit the site often as well as tell others about it. In comparison to page-centric sites, weblogs that constantly churn over new material can be said to be dynamic. The index pages of weblogs don't stay static for very long, but instead are a hive of activity, and are bookmarked and regularly revisited by interested visitors." (Bausch, 2002) Weblog tools sprang up in mid 1999 to facilitate the creation of the sites. The tools varied in their specific support for features, but the major element shared by all of them was the decrease in friction. The systems were somewhat automated as they were web-based, meaning one could access them from anywhere. This took the tedium of having to ftp files back and forth from a desktop to a server, and making it a transparent background process. The ease of these web-based blog tools greatly reduced the friction imposed by publishing online and greatly helped spread weblogs to other sites. However, it's crucial to understand the value addition that people derive by reading blogs and creating/publishing knowledge through their blogs themselves.

2.2 Why do people read blogs?

Weblogs provide an alternative to the corporate-produced content found online, and offer their own Web-based versions of reality programming. Weblog posts, with all their

misspellings and typos, and unedited rush of emotion, resonate with readers searching for that authentic human knowledge creation experience online. (Bausch, 2002)

2.2.1 Filtering and finding relevant knowledge

"*Thematic weblogs* - those focused on a specific topic- and *link-based weblogs* provide a valuable means for readers to find news and information online." (Bausch, 2002) By linking to stories and articles related to specific interests, weblog authors filter the web to their readers in essence, pre-surfing the web and spotlighting the appropriate links, and hence the readers get to read from the "Know-Who". With the sheer size and amount of information available, weblogs provide a valuable service by helping readers locate stories of interest – essentially the "Know-What" with the "Know-How".

"Often, weblog authors are assisted by their readers - as readers become more familiar with the blog's content, and as their relationship with the writer develops, they begin to send along links and pointers to stories of interest. The author often posts these links, which in turn encourages the reader to send in more links, continuing the cycle. As a weblog gains more readers, an interesting knowledge loop emerges. Its readers send in more links of interest, creating a decentralized knowledge network that expands the blog's reach. Now the readers as well as the author are filtering the vast quantities of information online to find relevant pointers that address the blog's area of focus. This knowledge cycle ends only when each individual begins to recommend links beyond the scope of the author's interest. At that point, the knowledge system self-corrects: the author doesn't post the links, the reader notices his/her links are no longer posted; and stops sending them. Now, someone else quickly takes the reader's place and provides new links, thus leading to the notion of a never-ending weblogging knowledge cycle, with only the players changing with time." (Bausch, 2002)

2.2.2 Experts (Know-Who) offering opinions

Most weblog authors pepper their links with commentary. A compelling summary or opinion can help the reader decide whether or not to click on the accompanying link. Commentary also offers authors the opportunity to share their expertise on a subject. Because so many early webloggers worked in the IT industry, many weblogs are excellent technical resources within IT, for instance weblog authors talking about the latest Windows IIS web sever often have considerable experience with its installation and deployment.

2.2.3 Peer-to-peer journalism

A new genre of weblogs that has appeared recently is the "amateur" or "peer-to-peer" (p2p) journalism. This label is applied to weblogs that attempt to report on current events. Unlike traditional journalism, which is highly edited and written to be compensated for, p2p journalism is often written by people who experience the event first-hand. P2P journalism offers a powerful way to get news and information online for 2 unique reasons: Distributed content and Distributed Perspectives. In the world of weblogs, the content is distributed – there's no single location where one can find everything to read. Unlike going to the New York Times or BBC homepage for all the news, weblogs enable us to browse from one site to the next, following links as they create an inter-connected story across sites. In this manner, distributed content does away with issues of censorship of content, and bandwidth and traffic problems. On September 11, 2001, the value of this approach became apparent to surfers looking for information online swamped the websites of the major news organizations. Webloggers in New York were able to offer more information through the firsthand accounts published on their sites. Because each blog was located on a different web server, the traffic was balanced across multiple machines rather than targeted at one site. (Bausch, 2002) With authors all over the world, weblogs also offer readers distributed perspectives – opinions and viewpoints based on the author's location and experience.

2.3 Why do people blog?

"Examining how people use spoken language is useful in explaining why people blog, because blog writing frequently resembles conversation. Blog writing is naturally informal, undergoes limited editing, and is immediate." (Bausch, 2002) Some compelling reasons for people to blog are:

- Improve writing skills

  As blog posts are available to everyone, and by extension, everyone's commentary, blogging often forces authors to refine their analytical abilities. Statements without supporting facts or poorly reasoned opinions can lead to criticism from other bloggers. Hence, with continued practice, bloggers improve their writing skills.

- Share personal experiences and stories

  Dave Winer, a weblogger and the CEO of UserLand Software, observes "every human being observes things just by living". Blogs offer efficient means for people to share those observations with each other.

- Share expertise

  "Many of the most popular and useful weblogs tend to focus on the author's passions and areas of expertise. Weblogs provide a necessary outlet for these people to share their knowledge with the rest of the world, a process that can be empowering as well as potentially lucrative." (Bausch, 2002) Many bloggers have made names for themselves in their industry due to the caliber of their weblog posts, for instance Don Box's spoutlet (Microsoft) on XML Messaging and Distributed Computing.

- Assert individuality

  Weblogs can be thought of as the Speaker's corner, with their own choice of colors and fonts, stories and links to share with others.

Having looked into the reasons for people to read blogs and publish knowledge through their own weblogs, it can also be noted that loose patterns emerge within such weblogs that can be used to mentally organize and classify them into "**Weblog Genres**".

2.4 Weblog Genres – Taxonomy/Topology

Format has more to do with the way information is presented than the information itself. Some formats that are usually seen on the web are as follows:

- Short bursts of text

  This is the usual image of a weblog: short bursts of text separated by a space on the page, each group of text marked with a timestamp (the time it was posted). The posts are ordered chronologically from newest to oldest, with the number of posts per page limited by a time period chosen by the blogger, for e.g. current month posts. This format favors immediate thought-to-web publishing, and aids knowledge codification immensely.

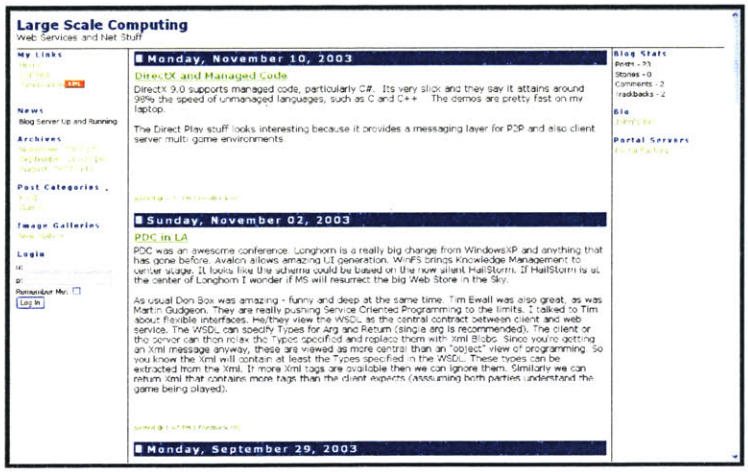


Figure 12: "Large Scale Computing" weblog – bursts of text

- Essay/Journal

  "An essay or journal site often organizes itself around the idea of pages. While bloggers writing short bursts of text see pages as simply a container - a way to group several days' worth of ideas together – essayists tend to see pages as a way to separate ideas." (Bausch, 2002) Journal sites usually consist of longer entries, several paragraphs at a time, with one entry per page, and this weblog format also favors more planning and organization. Instead of loosely organized thoughts on the fly, this format favors a record of thoughts after the author has had time to reflect and organize them.

Blog content genres differ with respect to the actual content shown. Hence, such genres are not easy to classify. Some regular content patterns are:

- Link-driven weblogs or Filters:

  Each post of a link-driven weblog contains a link to another site, usually with commentary about what's found on the linked page. Although the other genres may include links, they are used to give context or enhance the knowledge domain at hand.

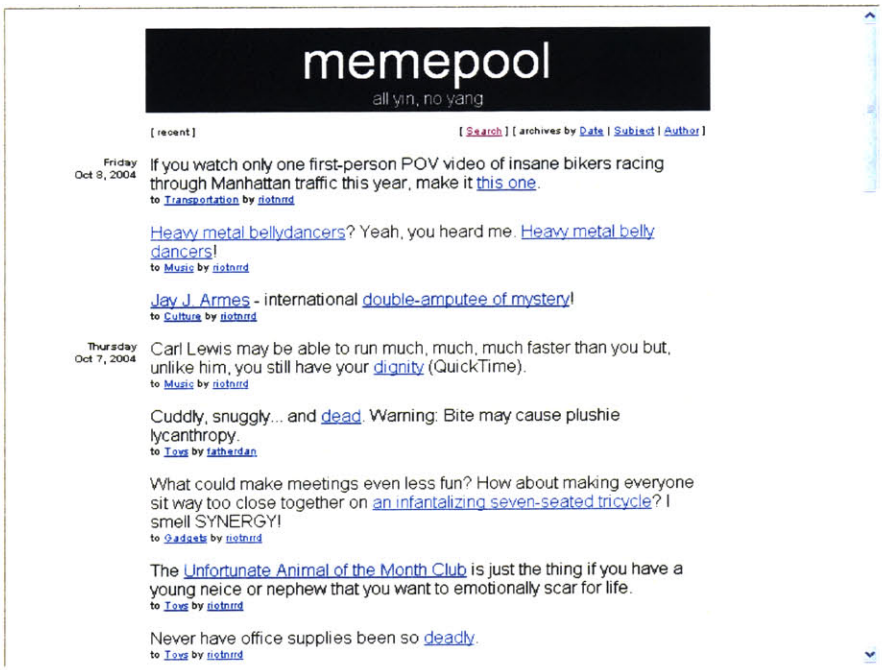


Figure 13: Memepool: A Link-Driven Group Weblog (Bausch, 2002)

- Single-topic weblogs

  Single-topic weblogs are weblogs with a single focus. While some blogs talk about any issue, single-topic blogs focus on one subject area and exclude everything else. More like

a special-interest magazine than a personal journal or newspaper, single-topic weblogs tend to build audiences around interests rather than personality.

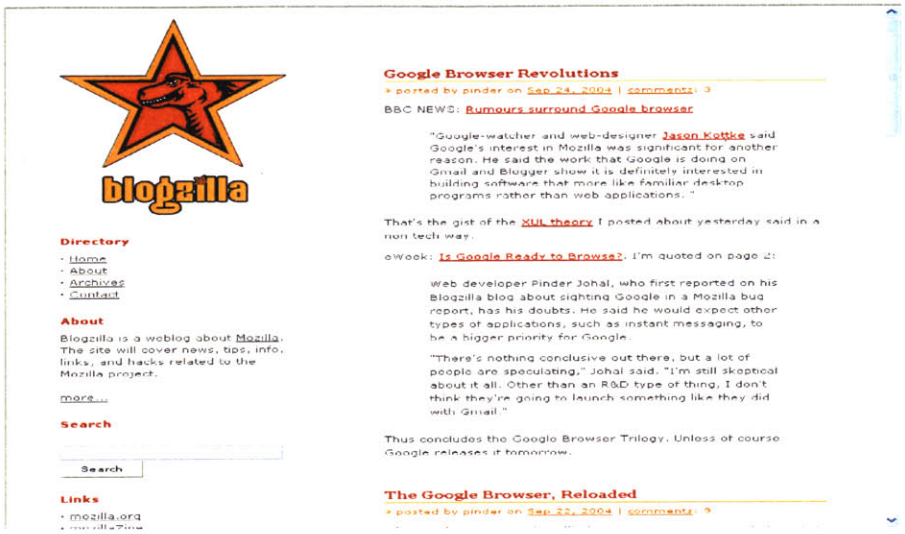


Figure 14: Blogzilla: A weblog on Mozilla web browser (Bausch, 2002)

- News opinion

  News opinion blogs combine elements of the single-topic and link-driven weblogs. They focus on the broad topics that are being discussed in the media, and usually include links to and commentary about mass media stories. (Bausch, 2002)

- Journal

  This is similar to a single-topic blog, but the focus is on the real-world life of the author. Because there are no common experiences to share with the reader, audiences are sustained by the author's voice and personality. (Bausch, 2002)

2.5 Anatomy of a weblog

Weblogs aren't about new pages, but rather about new posts. In the age of ever-decreasing attention spans, reading a new paragraph or two each day was easy to expect of readers, and more likely than readers taking in a new 1000-word essay each day. While traditional web sites were based and organized around the *Page Paradigm*, weblogs were organized and built around the *Post Paradigm*. The Post Paradigm is a way of looking at chunks of information within a larger framework. This view of weblog content, in comparison to traditional web page-level content, is sometimes referred to as *Micro-content*. (Bausch, 2002) The anatomy of a weblog in terms of the different containing pages, and weblog components such as posts, blogrolls, metadata, permalinks, and timestamps are discussed further.

Most weblogs consist of a handful of pages- of which there are 3 main types:

- Index Page

  The newest posts filled with the latest content are displayed on the index page (monthly or weekly index). Most weblogs put the index page first and foremost on their site, so visitors need only to remember a short web address to find or revisit it. (Bausch, 2002)

- Archive Pages

  Archive pages are ways to permanently store posts from the index page for safekeeping. When a new post is made, the post is usually copied to both the index and the latest current archive page.

- Additional Pages

  Aside from the index page and the automated archives, a weblog often consists of additional pages such as "About" or "Bio" pages that describe some information on the weblog's author. Other pages might include photo galleries or stories/articles written by the author. There might also be separate pages with links to other's blogs or contact forms to send an email to the weblog author. (Bausch, 2002)

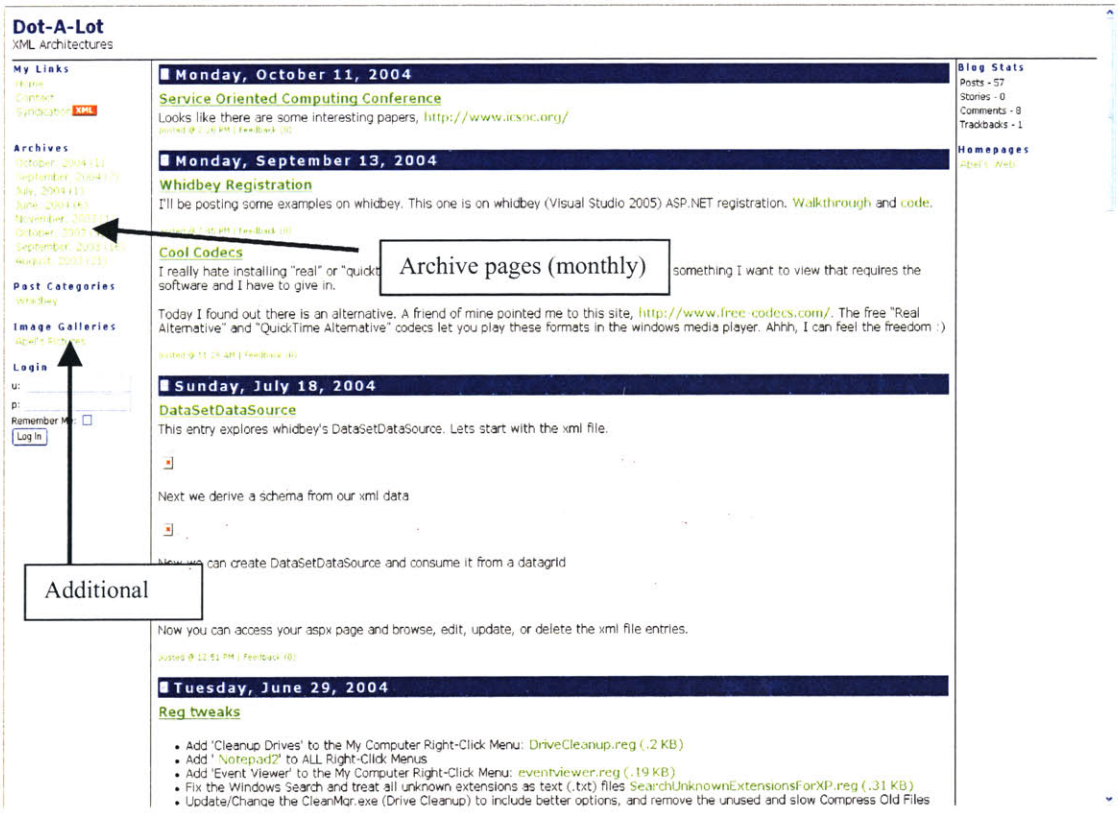


Figure 15: Anatomy of a Weblog (Rajagopal et al, 2004)

Blog Anatomy within the Pages:

- Posts

  "Posts are the fundamental chunks of content in the weblogs. A post is the primary building block of a weblog and what can be seen on the first (index) page when the blog is visited. Posts can range in sizes from a few words to several paragraphs in length, and can consist of not only text but images as well." (Bausch, 2002)

- Post Metadata

  "The first element on a weblog is the beginning of what's called as metadata, or information about a post. Metadata is supplemental to a post and gives more information about who made the post, what time and what day the post was made, offers links directly to that post, and what subject/category the post might fall under. Design-wise, the metadata is usually shown as different from the posts themselves so that visitors/readers know when a post had ended." (Bausch, 2002)

- Attribution

  The first thing that might be above or below a post is an attribution of the site's author. This is frequently done by following up a weblog post with a line that starts as "posted by John".

- Dates and Timestamps

  Weblogs are almost always organized by date, so the additional data for each post of what day it was made and at what time it was made is vital. Dates for most weblogs are shown as the month and day of the month, often followed by the year. Several posts may be made each day, but each will fall below a specific date to make that relationship between the posts clear. (Bausch, 2002)

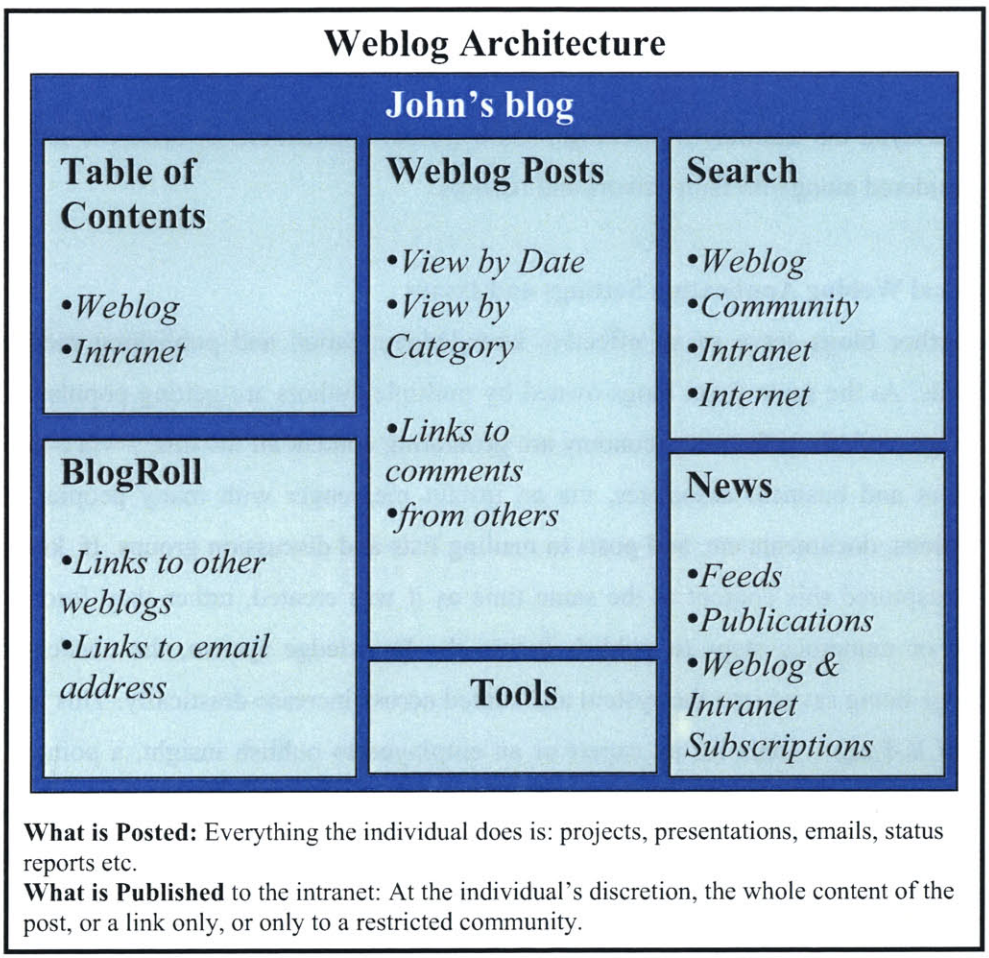
- Permalinks

  Permanent links, or permalinks, are ways of settling links to a specific post. This is an important concept and enables other weblog authors to link directly to a post, even if that post is on a long page of other posts. Permalinks are vital to understanding the importance of micro-content because every post on the site can have a unique and permanent address that anyone can use to point to the post. Permalinks are accomplished by way of an anchor tag, for instance:

  < a name="music"> This is a post about my favorite music. link to this (Bausch, 2002)

- Categorization/subjects/titles

  A common piece of metadata seen on weblogs is some sort of categorization, or some form of title or subject for each post. Depending on the weblog system used to manage a weblog, a subject or title can be added above a post, and organized posts into a number of categories. Such categories offer different ways to find and organize information in weblogs aside from simply by date. (Bausch, 2002)



- BlogRolls and External Resources like Feeds etc.

  Weblogs allow each person to personally identify who *he or she* thinks actually belongs to and participates in his or her networks (using the *blogroll*), rather than who their management thinks should be in those networks. The blogroll consists entirely of *active links* to the blogs of the other community members, so knowledge is electronically and personally connected. (Pollard, 2003) The key external resources (journals, manuals etc.) that a person uses frequently can be stored in a 'resources roll', consisting of the URLs of these resources; by copying and using an expert's 'resources roll', an apprentice could

discover and mimic the 'continuous learning' process of the expert. E-mails are the most valuable untapped codified knowledge resource in most organizations, and blogs allow knowledge to be simultaneously posted to one or more e-mail addresses *and* to the owner's indexed blog/filing cabinet. People can easily 'subscribe' to each other's entire blog by a process of syndication, followed by feed subscription (or an individual category/folder subset of it), so they are immediately notified about new knowledge or news that their work teammates or mentors deem valuable.

Having assayed the anatomy of weblogs, some typical applications in different settings are now considered alongwith their drivers and barriers.

2.6 Typical Weblog Application Settings and Issues

Single-author blogs serve as an effective knowledge-creation and publishing medium for individuals. At the same time, blogs owned by multiple authors are getting popular as well. Knowledge workers in the new economy are generating content all the time – via e-mail with co-workers and business associates, via an instant messenger with many people, through presentations, documents etc. and posts to mailing lists and discussion groups. If knowledge services captured this content at the same time as it was created, rather than forcing extra software or numerous steps to publish it into the knowledge system, the chances of the knowledge being saved into the system and shared across increase drastically. This leads to a notion of **K-Logs** – tools for an expert or an employee to publish insight, a point of view (POV), links to resources, important documents or emails, and other thinking to the intranet where it can be archived, searched and browsed. Unlike a weblog tool, K-Logs also include functionality for creating long-term connections (RSS news as knowledge streams for specific k-logs), and for fostering interchange between contributors. (Bausch, 2002)

Almost all knowledge systems target two key issues: document management or discussions. Both have serious flaws. Document management systems are complex and difficult to use. As a result, few people use them. K-Logs are easy-to-use for both the publisher and the reader. Additionally, K-Logs recognize the fact that knowledge is usually tied to a specific person. Knowledge is a continuous process of learning and analysis – it isn't static. By means of providing a means of publishing the process an expert goes through to build knowledge, K-Logs provide much more value than a relatively static, taxonomy-based document system. Furthermore, K-Logs provide organizations ways to identify, grow and show-case the Know-

Who (Experts). The problem with discussion groups is that they are totally dominated by a few individuals that drown out a lot of people who have the ability to contribute. K-logs change the dynamic by providing every contributor with a free voice, an open publishing medium so that everyone is included in the knowledge (management) system.

So, K-Logs can be useful for ensuring knowledge sharing and learning within a knowledge-intensive organization. Such services have been employed in a variety of research and technical institutions such as the MIT CADDIE Blog Server implementation, which will be discussed ahead in Chapter 4, or even blog services used for collaborative development efforts in Microsoft product development and research centers. Weblogs are tools not only for knowledge management within organizations, but also ensure ways for the entire enterprise to be more collaborative and connected. Such business blogs can be broken down into some categories:

- Workgroup blogs

  Most organizations have intranets that reflect a one-way publishing model – in which several departments or organizational units publish and disseminate knowledge to the rest of the organization. However, how does one ensure participation from everyone in the organization and have a KM system (or a federation of knowledge services) in place to work on a smaller scale, say in a workgroup? That's the spirit behind workgroup blogs. Imagine a 100-person firm with a engineering division of 30 people, the division being further broken down into 3 workgroups – software, hardware and support. By creating blogs for the most granular level (each workgroup), the focus of the blog has been narrowed in such a way that the knowledge created and disseminated is most useful to those at whom it's directed. The higher up the organization chart one goes, the more generic the information becomes. (Bausch, 2002) Workgroup blogs have benefits in enhancing team-work and reducing barriers between groups, as different voices can now been heard. At the same time, using the workgroup blogs as a document repository for workgroup documents cases versioning problems and provides a single point of access for workgroup-related knowledge.
- Project Blogs

  Project blogs are created specifically for projects and are open to all the project team-members. Such weblogs create a sense of identity for the project team, and provides a

single point of contact for all project-related communication, from meeting notes and timeline changes, to public acknowledgement of efforts and new ideas. Such weblogs help aligning the project teams around a common goal, and there's a sense of ownership for all the project-members as their voices are heard in an open, free publishing weblog format. Similar to workgroup blogs, project blogs also serve as document repositories, and hence provides a simple way to access the document management system and offers bare-bones knowledge management system as well. (Bausch, 2002)

- Extranet Blogs

  Such blogs are focused on improving relationships and communications between the external business parties, such as those transacted via an extranet. Extranets often provide a direct link between customers and suppliers, such as Widget Co. and the people that purchase its leading product, Widget 2.0. An extranet provides the communication channel by which customers can order additional widgets, find out version changes, and get technical support and so on. With use of extranet weblogs, effective conversations could be achieved between the customers and the manufacturers. (Bausch, 2002)

Weblogs do find various applications in organizational settings, but they have grown at an amazing rate since 1998. In fact, the growth has been at an exponential rate, from a handful in 1998 to over a million in 2003 (Technorati). This growth has been driven by the popularity of different blogging engines and by a process called as syndication.

2.7 Syndication and Data Formats

Bloggers or owners of weblogs are always looking for ways to make their content more popular, or attract more traffic. This is done by a process called Content Syndication. Content syndication makes part or all of a site's (in this case the weblogs) available for use by other services. (Stone, 2003) The syndicated content, or feed, can consist of both direct content itself and metadata – information about the content of the weblogs. The technology or data standards to do this range from the simple beginnings of RSS 0.91, through to the RDF-based RSS 2.0, all the way to industrial strength NewsML, ICE, ATOM etc.

At its most basic level, syndication is an agreement between a content producer and a content distributor. There are 2 requirements in content syndication: which content is available for distribution, and in what format will the content appear. For the content producer, syndication helps in making the weblogs more popular. The weblog's name and location could then be on

different sites, allowing different audiences to become potential audience for the specific weblog. For the distributor, providing weblog content to the readers has the advantage of increased traffic and prestige.

Because HTML is designed to only display information, it offers no clues about what that information is. Here's where XML-based formats come to the rescue. XML based formats do not give clues to display the information; instead they describe what the information is. However, simply using XML is not enough to facilitate syndication. Anyone using the document needs to know the rules as there would be differences in expressing the content, for instance one site might use <post> tags for weblog posts, while some other site might use <webpost> as the corresponding tag. If everyone agrees on the same format though, software can know the rules and make collecting and organizing data much easier.

2.7.1 Existing Syndication Formats

Netscape released its Rich Site Summary (RSS) format to allow independent news sources into its My.Netscape portal, and this looked like the beginning of a new standard. Any news source wanting to participate simply places an extra XML file in RSS format on its server (called a feed) and tells Netscape's portal where the feed is located, much like registering with a search engine. After a news source is registered, a bot visits the RSS files at regular intervals and stores the data locally.

"The oldest and most established RSS standard still in use, **RSS 0.91** was originally released by Netscape's team, in July 1999. It was later refined and further documented by Netscape, with Userland Software's Dave Winer." (Hammersley, 2003) RSS 0.91 is XML-based, consisting of one channel with a maximum of 15 items. Each item has a title, a description and a URL. The metadata set is restricted to the channel. RSS 0.91 is also a pull-based format, as the user must request the feed. This RSS version was designed to describe news web sites. News sites typically have a large number of articles with an index page linking to various articles. To describe this index of articles, each <item> in the RSS file included a <title> to describe the headline and a <link> to give its location. A typical news headline would look as follows:

<item>

<title> John's Blog</title> <link>http://blogs.mit.edu/jrw</link>

</item>

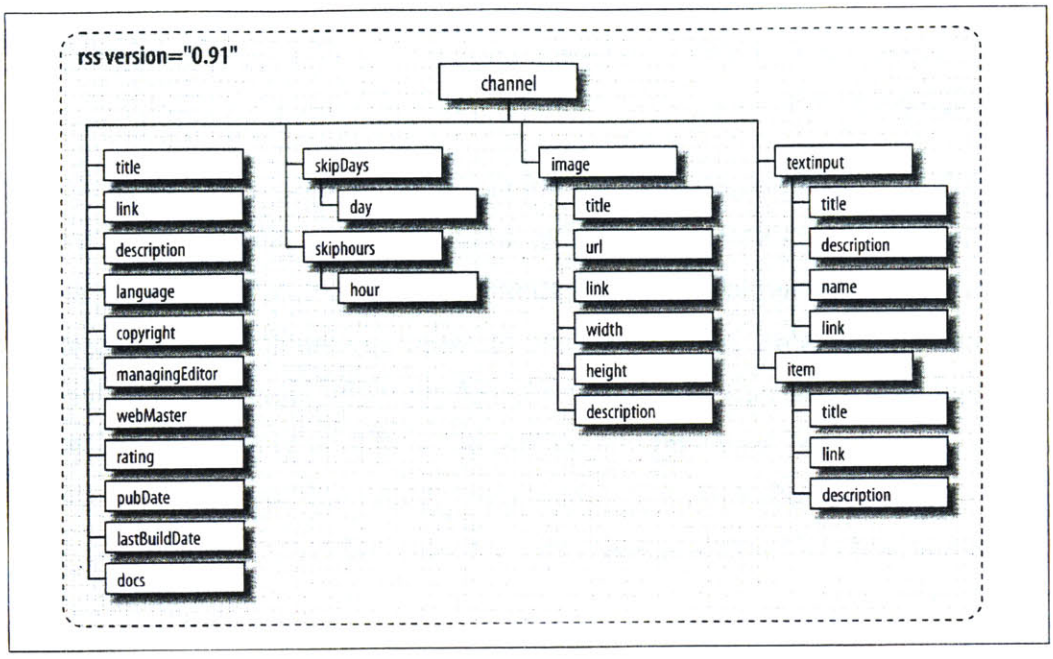


Figure 16: RSS 0.91 Format Tree Representation (Hammersley, 2003)

"**RSS 0.92** arrived on Christmas Day 2000. Written by Userland Software's Dave Winer, it expanded on 0.91 with five additional elements and a rethink of various restrictions placed on string length." (Hammersley, 2003) In this format, there could be one channel with an unlimited number of items. Each item may have a title, description, and URL, as well as a source, category, and enclosure. The metadata was not pertaining to the channel, but the items as well. This format was also pull-based like RSS 0.91, but it gives facilities to enable Publish and Subscribe.

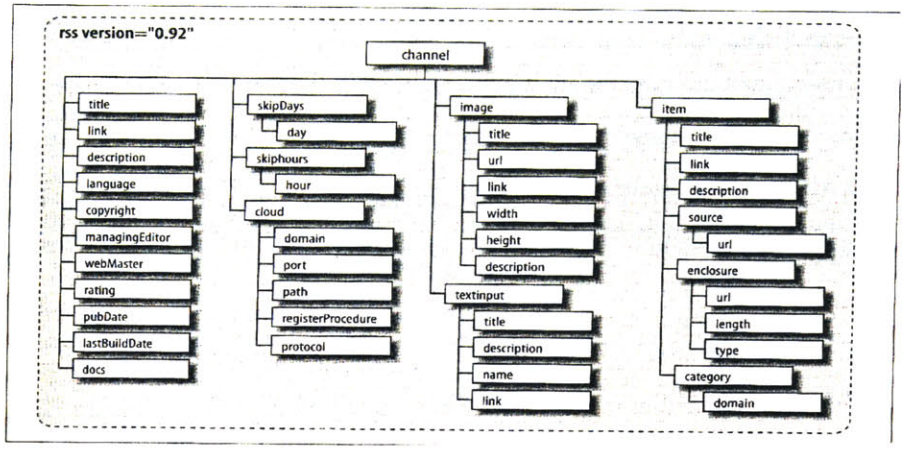


Figure 17: RSS 0.92 Format Tree Representation (Hammersley, 2003)

The development of the RSS standard has been forked split into 2 different groups. The heart of the conflict is extensibility vs. simplicity. In December 2000, the RSS-DEV working group, a collection of developers, released RSS 1.0. Around the same time, Userland Software released RSS 0.92 and higher versions such as RSS 2.0. Version 1.0 changed some of the underlying format technology, allowing greater flexibility. The 0.92 format on the other hand, keeps the simplicity of the earlier versions, while adding some new features for more compatibility with weblogs. (Hammersley, 2003)

With **RSS 2.0**, Dave Winer and Userland Software declared the simpler strand of the RSS specification frozen. RSS 2.0 can be extended by the use of modules. This format was far more complex compared to previous versions. It offers a modularized format, providing massive extensibility but also additional complexity. (Hammersley, 2003)

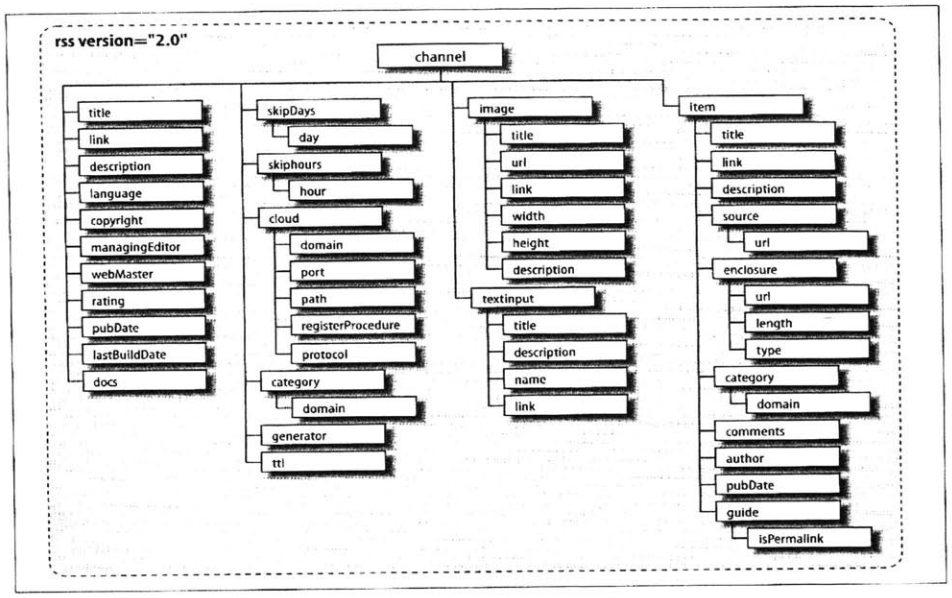


Figure 18: RSS 2.0 Format Tree Representation (Hammersley, 2003)

2.7.2 Issues in the Syndication Standards

The industry is also looking for RSS alternatives currently, although giant corporations such as Microsoft are increasingly embracing the RSS content syndication specification for a variety of community-oriented tasks. RSS has established (in its various forms) a substantial foothold on the web; it's used for everything from news feeds from major media organizations to alerts from social networking sites.

But rather than moving to incorporate RSS 2.0 itself into the web standards process, the two major Internet protocol standards bodies have instead been focusing on another syndication protocol, ATOM. Tim Bray of Sun Microsystems, and Sam Ruby of IBM, co-chairs of the Atom Project, have been leading the effort to turn the project into an Internet Engineering Task Force (IETF) working group. There have also been efforts by the World Wide Web Consortium (W3C) to adopt ATOM as a syndication standard. (Gallagher, 2004)

There's a good deal of disagreement between the two specification camps; while Atom is rooted in RSS, the specification in its current form is not backwardly compatible with any of the previous RSS versions. Atom is incompatible with the RSS standard, for a number of reasons; for instance, it creates a new XML data format for syndicated content to allow for more data types to be built into an Atom feed. The publishing API is being brought into alignment with other web development APIs and approaches such as SOAP and REST. Atom will also link into enterprise security models like WS-Security, through WSSE. (Gallagher, 2004)

The question that arises quite often is as follows: Are the changes that Atom implements really necessary for the majority of syndication applications-particularly those outside of the blog universe, like media content delivery? It may not matter much to users in the long run, as long as software developers incorporate the ability to handle both types of syndication feeds in their software, or ways to bridge between one and the other. But with the lion's share of existing feeds based on the RSS format--and Atom still in its early stages of development (the current version is 0.3)--there's a lot of momentum already behind RSS. Blogger users who already had RSS feeds still can publish using that syndication format (though Blogger supports the 1.0 version), and third-party services such as Feedburner will allow those Blogger users stuck with the default Atom feed to generate RSS feeds. For the foreseeable future, developers will end up getting stuck supporting either formats one way or another, because users will just want syndication feeds to work. (Gallagher, 2004)

A detailed discussion on syndication standards and the key policy issues within the same is provided in Chapter 5 of this report.

3. Approaches to Weblog Services and Implementations

The previous chapter focused on the general aspects of weblogging: how and why it is done; the blog anatomy and topology; and the syndication phenomenon driving weblogging ahead. This chapter moves towards a more concrete discussion of different weblogging implementations that could be adopted either for personal learning and use or for an enterprise learning system as well. First of all, weblog services can be classified as locally installed, remotely installed or centrally hosted services.

3.1 Installed Weblog Services

Locally installed services

Some weblog services are installed locally – that is, on our computer. All the files needed to run the service sit in directories on our computer's hard drive, including Word documents etc. An example of a locally installed weblog tool is Radio UserLand. (Bausch, 2002)

Remotely installed services

Some installed services reside remotely, meaning the files required to run the application reside on a different computer from the one we are using, for instance a remotely installed weblog service may be located on a web server that hosts our website, or may just be installed on another computer. An example of a remotely installed weblog service is Movable Type (www.movabletype.com). (Bausch, 2002)

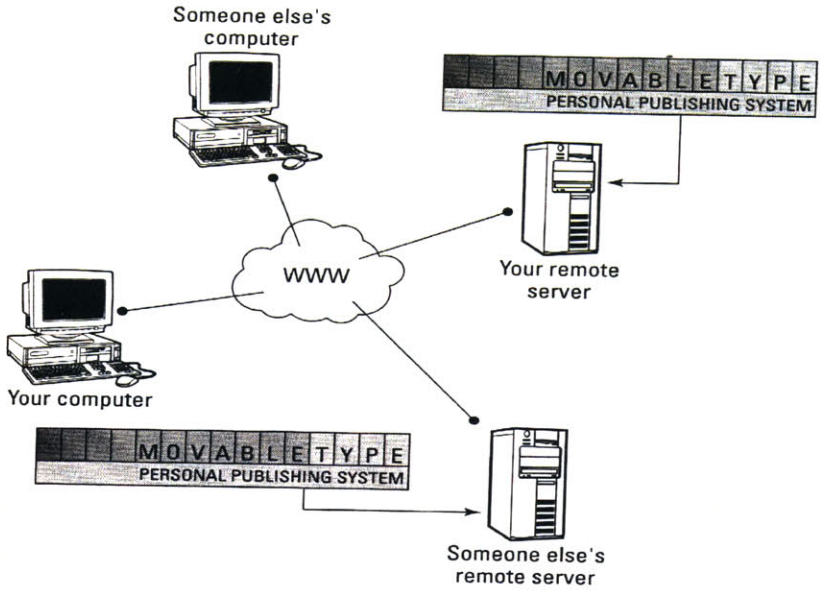


Figure 19: Remotely Installed Services – MovableType (Bausch, 2002)

Installed services offer the following benefits:

- *Flexibility and reliability*: With all the necessary files located on a desktop computer or server, the installer (user) manages or controls the service. Some services also allow for modification of their source code, enabling users with programming experience to customize and enhance the application feature set. (Bausch, 2002)
- *Content management and control*: Since the user has the server and the weblog service, he/she owns the content as well. One could also write custom scripts to retrieve data entries in specific formats. (Bausch, 2002)

However, these weblog services exhibit the following drawbacks:

- *Technical Barrier*: Perhaps the biggest challenge to using an installed weblog service is the installation process itself. Some services require installation of a client application, while others require software installation on a web server. Experience with directory permissions, open source Web servers, programming languages, IIS, Apache, VBScript etc. can be required for installation of some weblog services. (Bausch, 2002)
- *Hardware requirements*: Like any installed service, a certain amount of disk space is needed to install the weblog service. If the user runs a web server, he/she has to ensure that the server is up and running for access, and this might need some level of system administration competence and understanding.

3.2 Centrally Hosted Weblog Services

While some weblog services are installed services, others are offered as centrally hosted services. A hosted service is different from an installed service since there are no files for us to install.

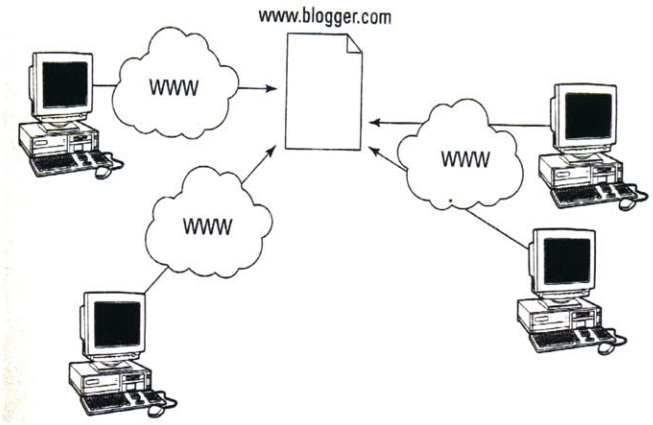


Figure 20: Centrally Hosted Service-Blogger.com (Bausch, 2002)

The people that offer the service control the service functionality and application in a central location. Some examples of centrally hosted weblog services are LiveJournal (www.livejournal.com) and Blogger (www.blogger.com). (Bausch, 2002)

Hosted weblog services have the following benefits:

- *Ease of use*: Since there are no files to install usually, one can setup a weblog almost immediately. One does not need to have knowledge about running a server or installing software to blog in this case. (Bausch, 2002)
- *Latest code*: The code base for a centrally hosted service is centrally located, which means that whenever one goes to the service's website (say www.blogger.com or www.livejournal.com), one can access the latest version of the service with additional features for trouble-shooting and technical support. (Bausch, 2002)

However, centrally hosted services exhibit the following drawbacks:

- *At other's mercy*: While using centrally hosted blog services, the user is at the mercy of the service provider. If the service is down for upgrades or experiencing poor performance, the user has no recourse, and cannot access the blog content until the service gets restored. (Bausch, 2002)
- *Trapped content*: The entity managing the service also manages the weblog content. This means that a user can't switch his/her content across services, and risks losing all the content if the weblog service security is compromised. (Bausch, 2002)

3.3 Implementations (Pilot projects at MIT)

3.3.1 MIT Caddie Blog Server

The Caddie Blog Server (CBS) is a centrally hosted service and is part of the Portal Factory - a platform based architecture for information products. It allows many different kinds of information systems to be deployed from a single code base. Some features of the portal factory include: administrators can manage 100's of portals across an enterprise; each portal having the advantage of the central Web Services, such as Content Stores and Registration.

The Blog server is open to all members of the MIT community and has more than 100 users currently– mostly from classes/research labs/academic programs at MIT. The results of such a pilot blog server implementation have revealed the tremendous diversity of bloggers in an

academic institution such as MIT. The blog server has been used by individuals; research teams such as IESL (Intelligent Engineering Systems Laboratory); Classes (such as 1.124); Academic Programs such as Systems Design and Management (SDM) to communicate the latest alumni events across SDM lists; and also the MIT Admissions Office to educate prospective students about admissions and financial aid processes.

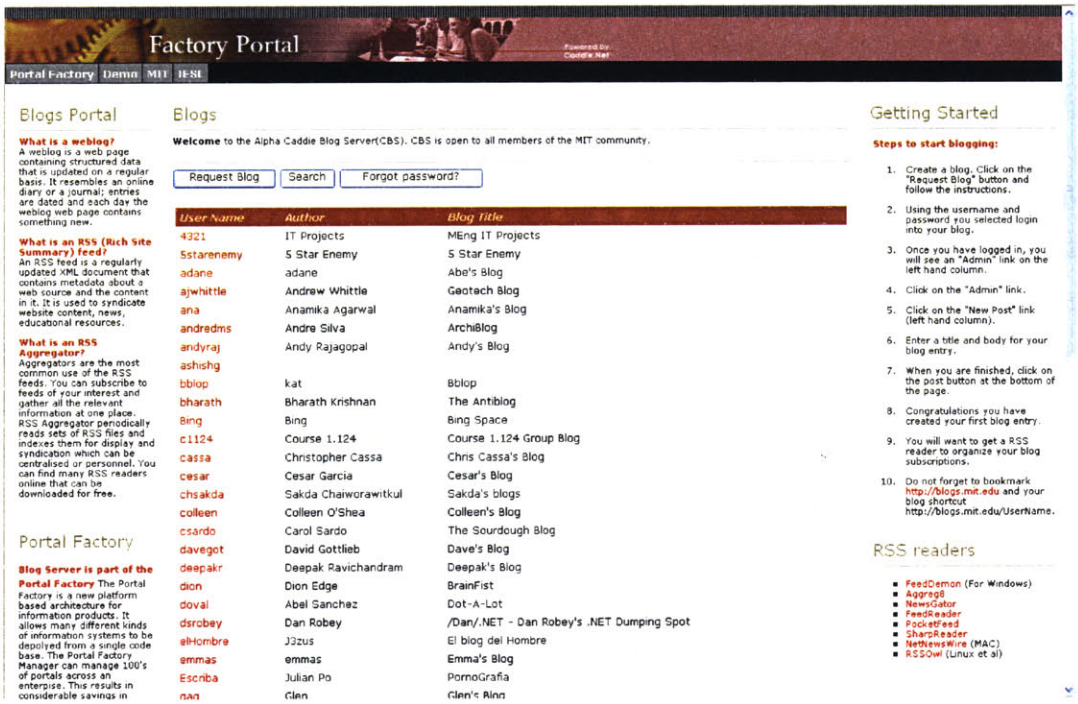


Figure 21: MIT Caddie Blog Server (http://blogs.mit.edu)

3.3.2 CADDIE .NET Portal Factory (Distributed Installed Systems)

The CADDIE .NET portal factory is an installed weblog application/service. It is based on a new Web Service oriented architecture for educational institutions and allows many different kinds of portals to be deployed from a single code base. The Portal Factory Manager can manage 100's of portals across educational institutions. This results in considerable savings in maintaining portals. The platform has its own development SDK that allows the end user to extend both the Web Services and the Web Component library that allows dynamic rendering of the User Interface. Templates are provided for different types of portal ranging from Course Management to facilities management.

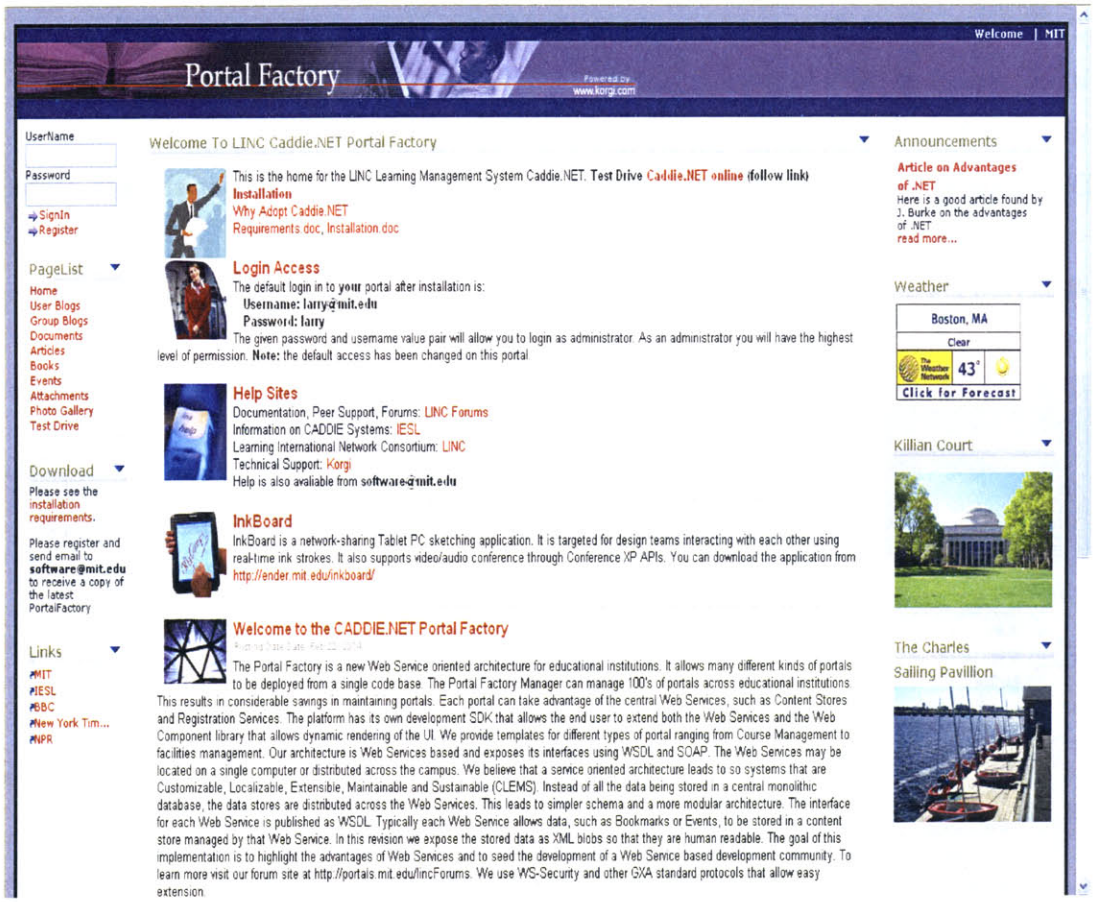


Figure 22: Installed Weblog Service–CADDIE .NET Portal Factory (http://iesl.mit.edu)

The architecture is Web Services based and exposes its interfaces using WSDL (Web Services Description Language) and SOAP (Simple Object Access Protocol). The Web Services may be located on a single computer or distributed across the campus. Such a service oriented architecture leads to so systems that are Customizable, Localizable, Extensible, Maintainable and Sustainable (CLEMS). Instead of all the data being stored in a

central monolithic database, the data stores are distributed across the Web Services. This leads to simpler schema and a more modular architecture. The interface for each Web Service is published as WSDL. Typically each Web Service allows data, such as Articles or Events or Blog Content, to be stored in a content store managed by that Web Service. In this portal factory revision, the stored data is exposed as XML blobs so that they are human readable. The goal of this implementation is to highlight the advantages of Web Services and seed the development of a Web Service based development community.

Installed Blog Services in the CADDIE .NET Portal Factory: (Rajagopal et al, 2004)

- 2 versions: Group Blog spaces and Individual Blog Spaces. While group blog spaces allow groups of users to blog (post) together into the same space, individual blog spaces allow each user to post to his/her own space.
- This blog service was distributed to hundreds of educational institutions and non-profit organizations as part of the CADDIE .NET Portal Factory.
- Received good feedback about usage in academia and knowledge-intensive organizations (http://biztalk.mit.edu/AspNetForums/Default.aspx).

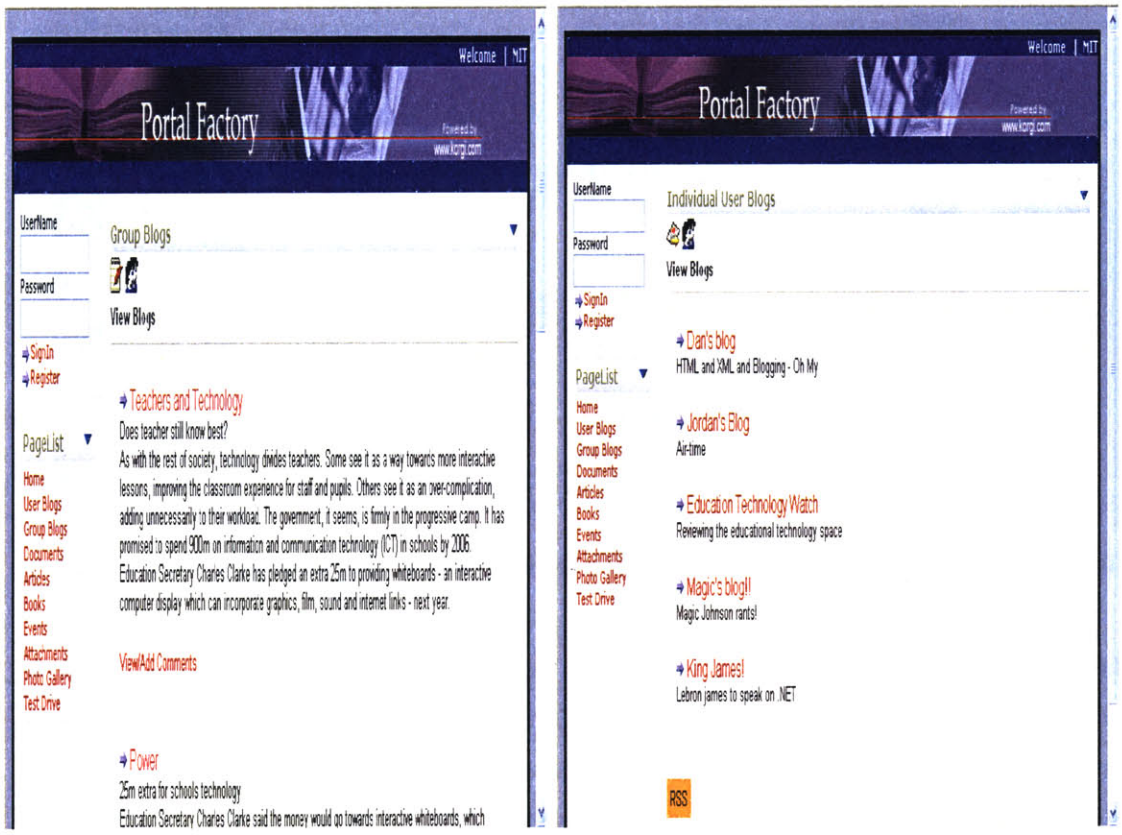


Figure 23: Group (left) and Individual Blog Services – CADDIE .NET Portal Factory

3.4 Approaches to Enterprise Blog Services

This chapter dealt with different types of blog services and implementations so far. Enterprises are now integrating blog services as part of their knowledge architecture or information ecology. Following are approaches in which blog services can be integrated with other information aggregators and publishing tools within enterprises. This discussion goes a step beyond Sections 3.1 and 3.2 where only the foundations of blog services were assayed.

3.4.1 Centralized Approach

In this approach, there is only one enterprise weblogging and news feed aggregation application, consisting of the following components: (Angeles, 2004)

- Integrated information/news aggregator
- Integrated classification / controlled vocabularies (subject, author. etc.)
- Integrated publishing

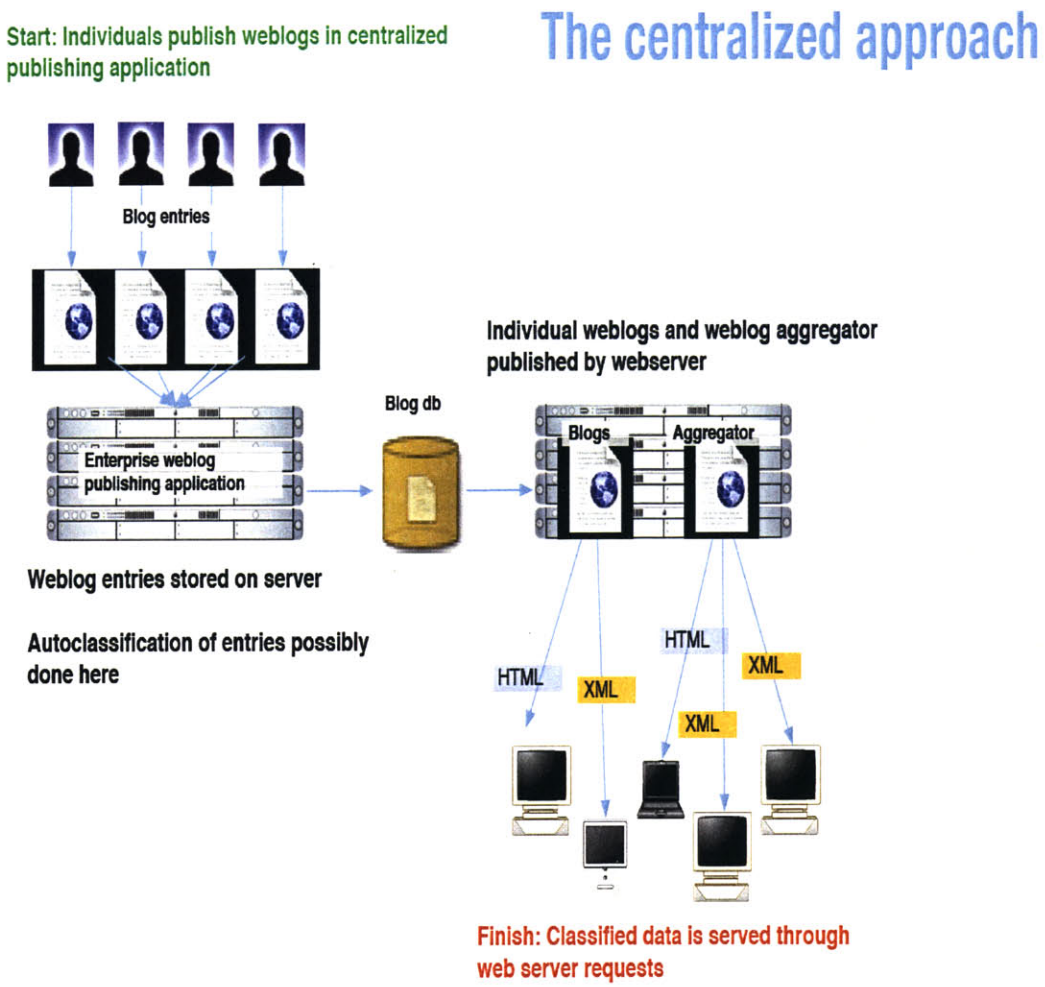


Figure 24: Centralized Weblog Services Approach (Angeles, 2004)

Issues/Benefits of Centralized Approach:

- Demands standard processes, ways of working
  - Might work well in small organizations.
  - Tough communicating value of centralization in large organizations.
- Needs user research, communication and training
- Ability to use more complex application features e.g. email based blogging, contextual commenting which are of great use in learning contexts.
- High returns in ease of administration

3.4.2 Decentralized Approach

This approach is the reverse of the centralized approach, wherein many diverse / disparate publishing services including weblogs co-exist. Such applications have a common XML output format and the enterprise application gets facilitated by the XML feed aggregators and search service with possible auto-classifier.

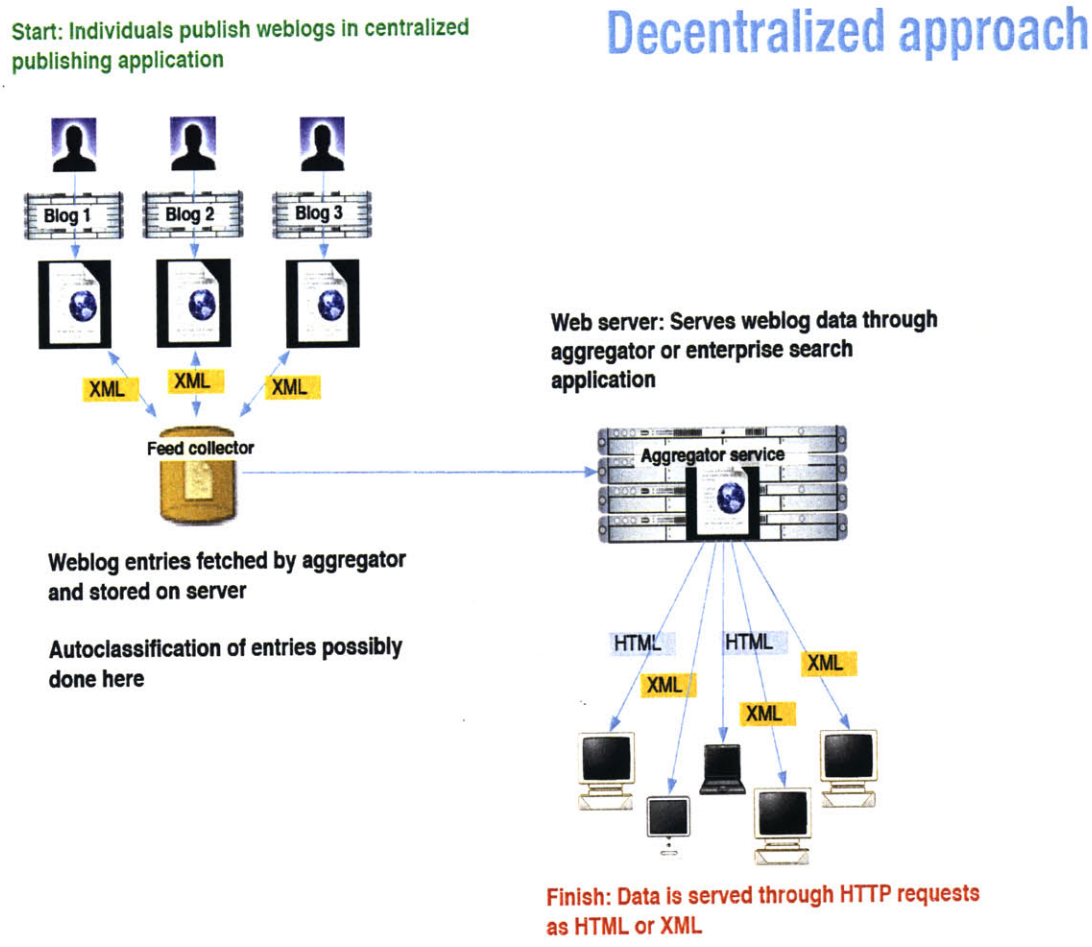


Figure 25: Decentralized Weblog Services Approach (Angeles, 2004)

Issues/Benefits of Decentralized Approach

- Perhaps the easiest to execute, especially in learning environments.
- Supports diversity; gives users the freedom to choose tools that match their processes.
- Least amount of human resources and effort on the system side.
- Requires considerable financial investment.
- May not provide most relevant results by machine classification.

Choice between Centralized and Decentralized Services Approaches

The following factors would influence the choices made by enterprises while adopting a particular approach towards incorporating weblog services into their knowledge architectures. (Angeles, 2004)

- **Size**
  - Small groups may not need large knowledge networks. A centralized approach would work to be better then.
  - Large organizations seem to benefit most from decentralized and mediated approaches (between centralized and decentralized).
- **Resources**
  - Does the enterprise in question have the financial and human resources to pull off any of these solutions? The Centralized approach would be preferred if resources are scarce, but the decentralized approach would be preferred if resources are not constrained.
- **Culture**
  - Do personal home pages/bloggers exist now?
  - Does the enterprise have an attitude of transparency regarding knowledge sharing? The centralized approach would be followed if transparency is accorded more priority.

4. Information Overload in Weblogging: Mine the Weblog?

The previous chapters focused on the reasons behind the weblogging phenomenon, the blog anatomy and topology; the syndication phenomenon driving weblogging ahead, as well as the approaches to blog services for different kinds of organizations dedicated to enhanced learning and knowledge management.

This chapter deals with a detailed discussion of 'information overload' in weblogging, and mentions the services available to mine/search relevant blog information in the web today. An independent weblog-ranking mechanism proposed as part of the thesis is also assayed in this paper, along with a discussion of ways to enhance weblog (information) retrieval in the future.

4.1 Weblog Overload and Power Laws

There has been tremendous growth in the weblogging phenomenon: from a handful in 1998 to over a million in 1993 (Technorati). A few facts about weblogs as a knowledge service today: (Technorati and Pew Internet Study)

- Around 11%, or about 50 million, of Internet users are regular blog readers.
- A new weblog is created every 5.8 seconds or ~ 15,000 new blogs a day.
- Bloggers — people who write weblogs — update their weblogs regularly; there are about 275,000 posts daily, or about 10,800 blog updates an hour.

While users in online learning environments have a tremendous quantity of weblog content to draw upon for learning, they do face the problems of finding relevant information (mining weblogs), or becoming powerful (blog) writers themselves. A comprehensive view of the weblogging world reveals not just a deluge of content and subsequently RSS or other feeds, but also a *Power Law distribution* in weblogs and their feeds. (Colin, 2003)

A study by Clay Shirky addresses power law distributions in blogging by linking to choices and preferences among diverse (knowledge) options or sources. Shirky notes that an "A-list" of extremely high profile bloggers has emerged and to, a large extent, consolidated, for instance Don Box, Scott Water, Dave Winer and the like in the technical areas of distributed computing, XML based architectures, content syndication formats and web services. This concerns the fact that a relative few blogs account for the majority of traffic or "hits".

Shirky's study draws conclusions from a project by N.Z.Bear (www.myelin.co.nz/ecosystem) also quoted in Pew Internet studies (2003), arranged more than 400 weblogs in rank order by the number of inbound links they have, i.e. number of weblogs that link to them found via aggregation services such as blogdex.com, daypop.com, popdex.com etc. "Of the 433 blogs, the top 2 ranked for 5 % of inbound links, the top 12 for 20 percent and the top 50 (12 % of the total sample) accounted for 50 % of the inbound links." (Colin, 2003) The graph for inbound links tracked in 2003 is shown below:

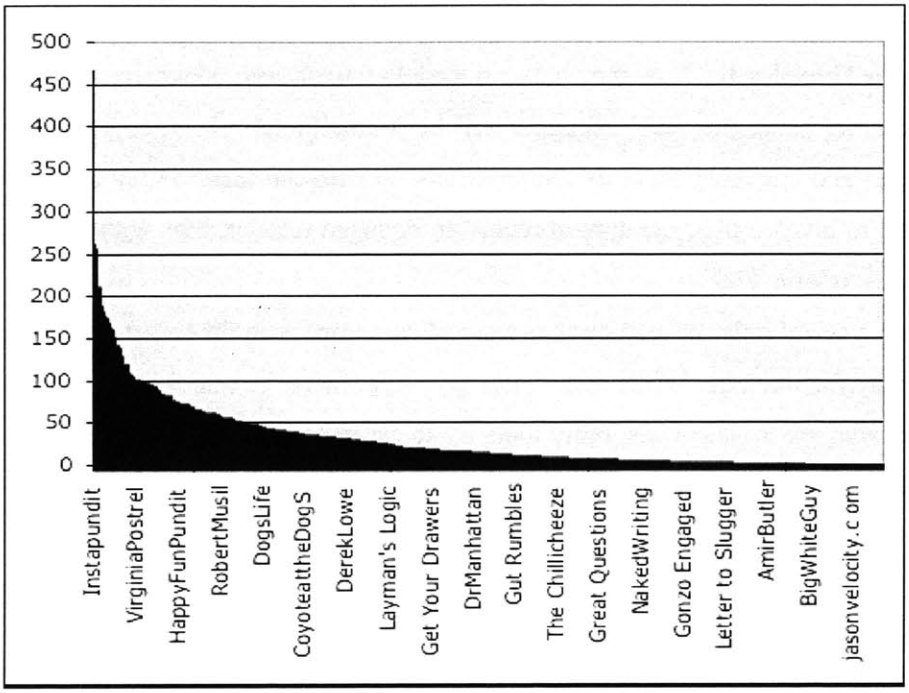


Figure 26: Shirky's Power Law Distribution (Colin, 2003)

This can be seen as a "power law distribution" since a weblog knowledge system with a diversity of (weblog) choices for users to learn from can witness a small subset of providers (bloggers) attract a disproportionately large amount of choices exercised. "The very act of choosing spread widely enough and free enough, creates a power law distribution". (Colin, 2003)

"In this over-saturated knowledge market of blogs, "power laws" are inherent if "power" is equated to any of the following metrics: internet traffic, inbound links, comments from readers or "near-top-of-list" keyword searches." (Colin, 2003) Consequently, being a powerful blogger might simply be equated to scoring well on such metrics. Such a power law distribution leads to the following consequences:

1. Small subsets of bloggers attract disproportionate traffic and attention – A-list or "head" bloggers. Hence, a social order of bloggers gets created, i.e. an order of the "head" and "tail" bloggers/blogs. (Rajagopal et al, 2004)
2. It is relatively harder for latecomers than early birds to become "blog stars". This is because of the fact that in addition to being noticed and linked to by others, bloggers have to actually work against preference premiums of other users (learners) that are already in place. (Rajagopal et al, 2004)
3. However, the power law distribution also reveals different types of activity or forms of blogging with such a social order of bloggers. On one hand, those at the "head" would mainly be "broadcasters", as they have no time to correspond, converse or communicate. On the other hand, the "tail" bloggers will be journal/notebook types writing for small audiences and engaging them in conversations. In between these extremes will be blogs intended to involve their creators in relatively engaged relationships with moderate sized audiences. (Colin, 2003)

In fact, such a social order of weblogs has a significant impact on the search and mining tools existing today for weblogs. While some tools give importance to searching for inbound links or bloggers who are broadcasters, other tools try to estimate the ability of weblogs to engage users/readers in conversations for learning or for other purposes such as review etc. Such estimates would depend on the commenting/track back done on the blogs.

4.2 Weblog-specific Search/Indexing Services

- **Google** (with Blogger): Google favors blogs in terms of delivering up links when users search the web for relevant information. This is favored by the blog anatomy since blogs inherently contain links with the commentary or contextual knowledge provided by the blogger (writer). (Stone, 2003) Google treats such links as votes, i.e. the service interprets a link from blog A to blog B as a vote, by blog A, for blog B. But, emphasis is not only given to the sheer volume of such links or votes, but also to the blog that casts the vote. Votes or links cast by blogs that are themselves popular or important weigh more heavily and help to make other blogs important or 'valued' as well. Basically, this means that if (popular) blog A links to blog B, then blog B increases in popularity and climbs the ranks among Google's search results pages. (Stone, 2003) Such popularity schemes help the user view relevant online information in easier ways since the 'value' or 'popularity' tags get attached to each weblog – a source of knowledge (K-Log).

- **Eatonweb** portal: This was possibly the first directory dedicated to weblogs, which are simply listed by name, category, language or country (http://portal.eatonweb.com) (Bausch, 2002)

- **Daypop**: Calling itself a current events search engine, Daypop indexes sites that are updated on a regular basis, including online newspapers, magazines and weblogs. Similar to any search engine, the user can enter a term and search for matching content. The difference between this service and Google is that it's only looking at updated content. Pages that aren't frequently updated and other static documents aren't in its index. Another feature of Daypop is a list of the top 40 URLs posted to weblogs, offering a way to get a quick glance at popular topics in discussion. (www.daypop.com) (Bausch, 2002)

- **Blogdex:** The MIT Media Laboratory has created a service called Blogdex that tracks links users are posting to their weblogs. After a weblog is added to this service, the service bot visits the site once a day to see the links that have been posted. "It awards points to the links found in weblogs, and lists the most popular ones. If a blogger posts a link to a news article about distributed computing, and ten other bloggers link to the same article, then the link would be rated higher." (Bausch, 2002)

  Blogdex then creates charts for the newest popular links as well as all-time most-popular links. Its primary purpose isn't to drive traffic to weblogs, but one of its features includes listing the weblog resources for each popular link. Such a service greatly enhances the choice for any user trying to acquire knowledge in a certain (scientific) topic and carries the process of online learning ahead. A user visiting Blogdex who is interested in a certain news item may want to visit each of the weblogs that has linked to the story to see what has been said about it. This concept of choice goes back to the notions of tacit vs. explicit knowledge explained in the previous sections. While the link or the story is knowledge in explicit form, each blogger's perspective or contextual interpretation may be different and that's the implicit knowledge sought by different online users.

- **Bloghop**: This site is part-directory, part popularity contest. After adding a weblog to this site, Bloghop visitors or users can rate a weblog on a scale of one to five. The highest rated weblogs are listed first. (Bausch, 2002)

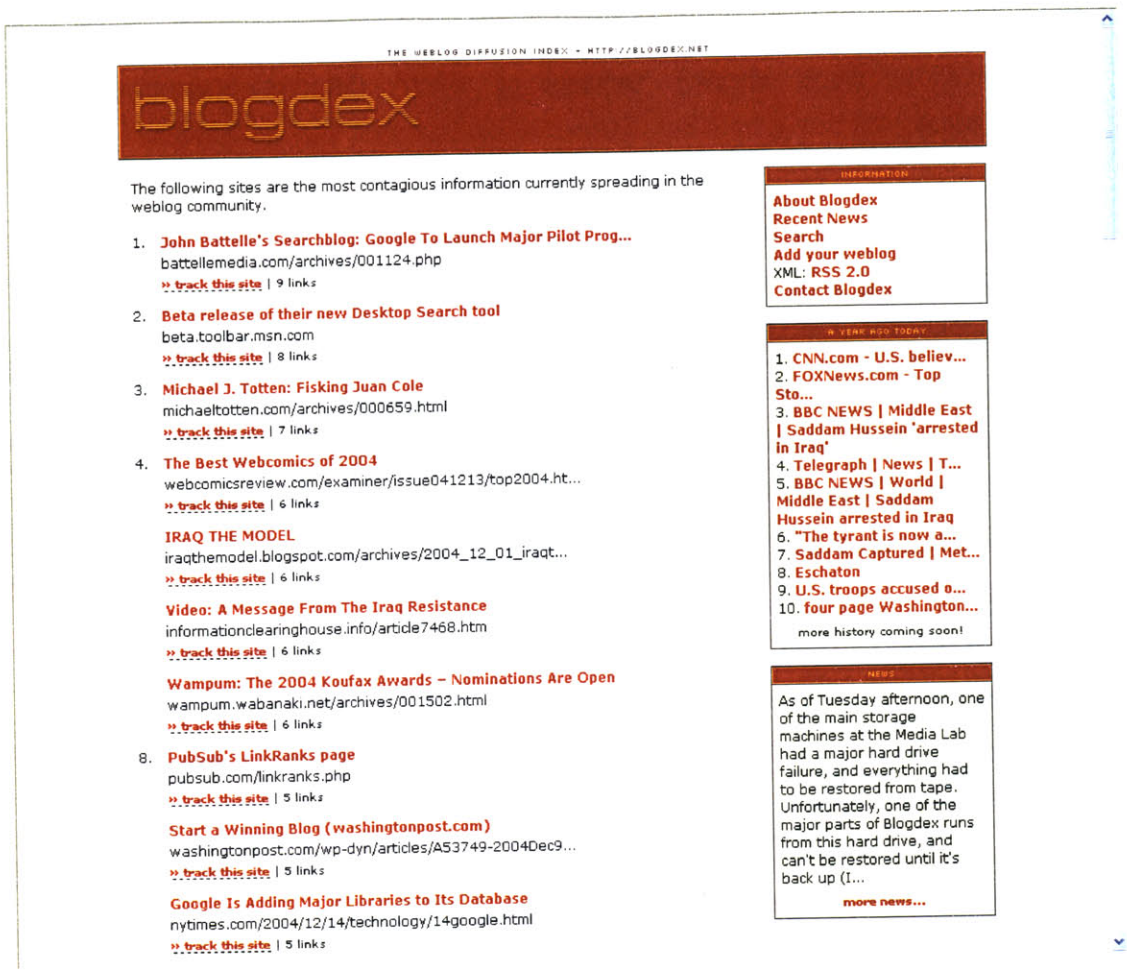


Figure 27: Blogdex Linking Service (MIT Media Labs)

- **Webrings**: A webring is a group of sites, usually focused on a single topic, that are connected to each other through links. A search on a service such as Google for a topic should point to all existing webrings that the user could join and learn from. Some webrings that are popular today include Diet and Exercise Bloggers, Christian Bloggers, etc. (Bausch, 2002)

- **Geographic Directories**: "Some weblog authors have also started organizing themselves by their physical location. Weblog groups are starting to appear around the world. These groups are often more than a collection of links on a web page; they are social groups where the members meet in real life. The topics about which each blogger post may be different from the other, but they all share one common trait between them: location.

The geographic groupings range in scope. While Gblogs is open to any UK blogger, DFWBlogs is open to bloggers ONLY from the Dallas, Texas area. Bloggers in the New York City area came up with an interesting way to see where their fellow blogger live. A site called nyc bloggers (www.nycbloggers.com) allows people to register their weblog, noting which subway stop they live closest to. Using a subway map, they show bloggers with their corresponding stop." (Bausch, 2002)

4.3 Pilot Search/Mining Proposal

As discussed earlier in section 4.1, weblog 'power laws' are inherent and blog value/power is equated to any or a combination of the following metrics: 'near-top-of-list' keyword searches, internet traffic, inbound links, or comments from readers. While different services existing today give more importance to different metrics, a **Feed Location and Analysis Model** is proposed here which would serve to value blogs and other syndicated web content (in the form of feeds) by giving appropriate weight to each of the above metrics. Weblogs get syndicated into 'Feeds' using different formats such as RSS, ATOM etc. These feeds get crawled for and further analyzed to mine relevant content from the weblogs. On one hand, some users interested in learning about a specific topic would be more interested in blogs that are on the 'A-list' blogs or have been linked the most (maximum number of inbound links), for e.g. Don Box's Spoutlet. On the other hand, other users might be more interested in viewing blogs that engage Internet users in conversations and commenting, not necessarily linked to the most.

The Feed Location and Analysis Model would comprise of the following components:
Feed Crawler: used to crawl a URL/Site for possible RSS/ATOM feeds.
Feed Analyzer: analyze the feed for keywords, process text, comments, and trackbacks.

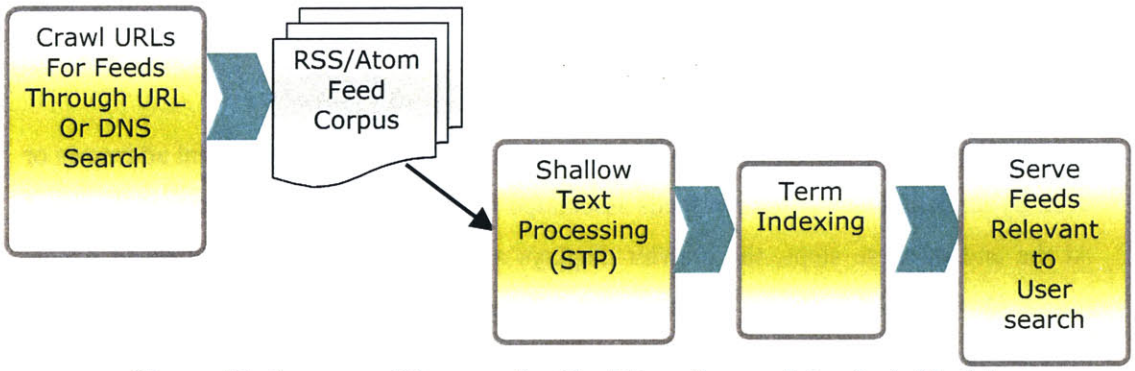


Figure 28: Sequence Diagram for Feed Location and Analysis Model

The Crawler, and Analyzer coupled together along with a mechanism with a traffic measuring/statistical functionality would provide an efficient *RSS Feed (Upstream) Generator* – essentially designed to serve "valued" and relevant feeds to the end user based on their keyword searches. (Rajagopal et al, 2004)

4.3.1 Feed Crawler

The crawler to search for RSS feeds given the URL (s) of the page/weblogs has been developed using C# .NET language and is based on Mark Piligrim's RSS Auto discovery algorithm adopted in different forms by other feed aggregator services as well. A brief description of the algorithm used in the feed crawler is outlined below: (Pilgrim, 2002)

- The crawler takes the address of the feed directly if the feed's URL is supplied. This is the base case (do nothing).

- However, if the address of the main web page is given, the crawler downloads the page and looks for LINK tags that point to feeds. If such feeds are found, the crawler uses them. An important point to note here is that the main page's URL need not be prefaced with http://.
- If the main site does not support RSS auto-discovery, the site might have a regular link to its feed. The crawler scans all such links on the page, and makes a best guess to point to a feed.
- Links to addresses on the same server that end in **.rss**, **.rdf**, **.xml**, or **.aspx** are prime candidates for being feeds. The crawler gathers these links up and investigates which ones actually are RSS feeds. (by downloading them and analyzing them)

- If feeds can't be found, the crawler looks for links to addresses on the same server that contain rss, rdf, xml or aspx in the address.

- If feeds are still not found, the search is scaled up to include links to external addresses or even Syndic8 that keeps track of feeds associated with many different addresses.
- At the end of these steps, the crawler displays a failure message if feeds could not be found. However, if feeds are indeed found, all such feeds get displayed in a list and the user has the further option of analyzing them for specific terms to enhance information retrieval.

4.3.2 Feed Analyzer

The analyzer part consists of the following segments: feed corpus shallow text processing, term indexing and ranking feeds by term queried for.

A brief description of key terms used:

Shallow Text Processing (STP): preliminary manipulation of text such as term extraction, word normalization, etc. (Chaiworawitkul, 2004)

- Tokenize: find appropriate boundary of a word
- Normalize: return normal form of a word, e.g. "taught - teach", "geese - goose"
- The text processor used here is named 'Monty Lingua' (Hugo Liu), with implementation being Java based on Python. This implementation has been wrapped up and exposed the interface through the following web service:
  http://biztalk1.mit.edu:8080/axis/MontyLinguaService.jws?wsdl
- Note that Monty Lingua does NOT do the followings:
  1. Compound analysis, e.g. database ◊ data base
  2. Name recognizer, e.g. B2B ◊ B+2+B and therefore, is not captured
  3. Word sense disambiguation, e.g. "I can swim" and "Juice can"

Term Indexing: index terms from STP based on their frequency in a feed & overall corpus using the Vector Space Model. (Chaiworawitkul, 2004)

- *Term frequency* (Tf) can reflect its degree of importance and we want to use only important terms to efficiently represent a document
- However, term occurrence by itself is not sufficiently guarantee its importance, e.g. "with" can appear several times over a document and corpus but it is not necessary
- Through observation, it is well-known that terms with high occurrence in both documents and corpus are not importance. A good index to measure this is the Inverted Term Frequency (Idf):

$$\mathbf{Idf = \log(D / df)}$$

- df (document frequency): number of documents in corpus *D* that a term occurs
- Terms that appear in many documents will quickly be damped by this index. So we use Idf to balance the Tf, i.e.

$$\mathbf{Tf\ Idf = Tf \log (D / df)}$$

- Tf Idf is used in filtering unimportant terms out and is efficiently used for term indexing

So, all the text in each of the feeds are analyzed using the vector space model and then their frequency of occurrence is found in each feed. Once the text or terms are analyzed, the user's keywords could be compared with the indexed terms/text and the feeds with the highest term index/effective frequency would be served to the users as search results. However, with this Crawling and Analysis Model, the thesis focuses only on keyword search, while searching for cross-linking, comments and trackbacks along with traffic would be a futuristic goal.

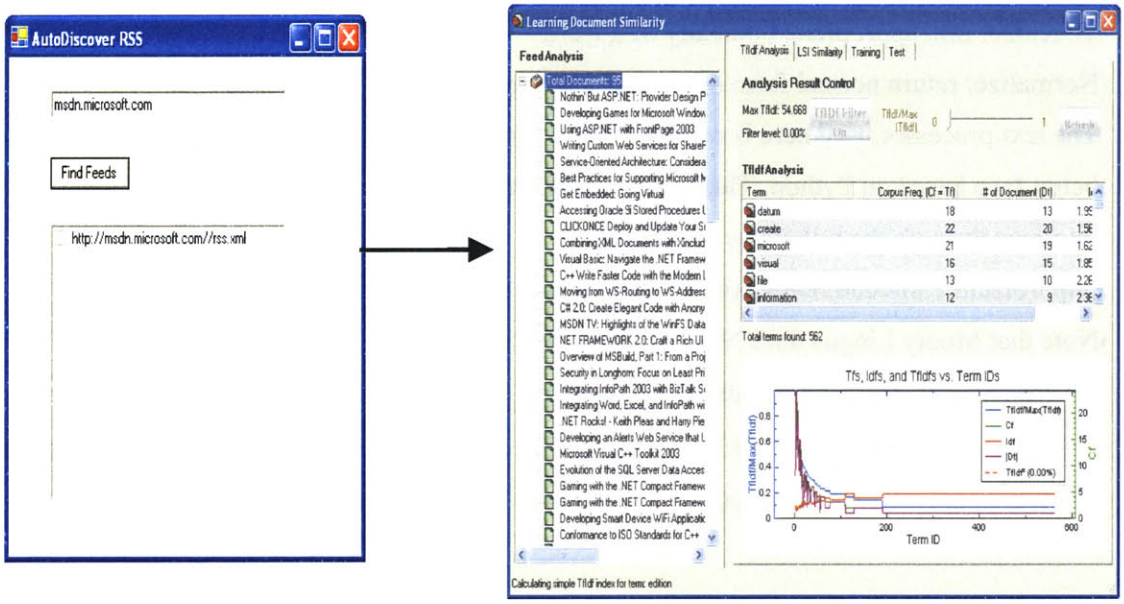


Figure 29: Snapshot of the Proposed RSS Upstream Feed Generator

5. Weblogging and Issues in its Adoption as a Knowledge Service

This chapter investigates the typical applications of weblogs as a knowledge service and the associated issues that emerge in such applications – ranging from intellectual property rights, privacy of information of bloggers, and other issues involved in standardization of content (weblog content) syndication formats such as RSS, ATOM etc.

5.1 Typical Applications

As mentioned earlier in Chapter 2, single-author blogs serve as an effective knowledge-creation and publishing medium for individuals. At the same time, blogs owned by multiple authors are getting popular as well. Knowledge workers in the new economy are generating content all the time – via e-mail with co-workers and business associates, via an instant messenger with many people, through presentations, documents etc. and posts to mailing lists and discussion groups. If knowledge services captured this content at the same time as it was created, rather than forcing extra software or numerous steps to publish it into the knowledge system, the chances of the knowledge being saved into the system and shared across increase drastically. This leads to a notion of **K-Logs** – tools for an expert or an employee to publish insight, a point of view (POV), links to resources, important documents or emails, and other thinking to the intranet where it can be archived, searched and browsed. Unlike a weblog tool, K-Logs also include functionality for creating long-term connections (RSS news as knowledge streams for specific k-logs), and for fostering interchange between contributors. . (Bausch, 2002)

Almost all knowledge systems target two key issues: document management or discussions. Both have serious flaws. Document management systems are complex and difficult to use. As a result, few people use them. K-Logs are easy-to-use for both the publisher and the reader. Additionally, K-Logs recognize the fact that knowledge is usually tied to a specific person. Knowledge is a continuous process of learning and analysis – it isn't static. By means of providing a means of publishing the process an expert goes through to build knowledge, K-Logs provide much more value than a relatively static, taxonomy-based document system. Furthermore, K-Logs provide organizations ways to identify, grow and show-case the Know-Who (Experts). The problem with discussion groups is that they are totally dominated by a few individuals that drown out a lot of people who have the ability to contribute. K-logs

change the dynamic by providing every contributor with a free voice, an open publishing medium so that everyone is included in the knowledge (management) system.

So, K-Logs can be useful for ensuring knowledge sharing and learning within a knowledge-intensive organization. Such services have been employed in a variety of research and technical institutions such as the MIT CADDIE Blog Server implementation as discussed in Chapter 4, or even blog services used for collaborative development efforts in Microsoft product development and research centers.

In addition, External blogs or Internet blogs are also being used by businesses to inform their customers/consumers. The elements of interactivity, community and collaboration in weblogging are really crucial as growing businesses adopt blogs for customer relations, advertising, and promotional initiatives. (Bausch, 2002) One well-known business example is software company Macro-media's use of blogging so that its customers could learn about recent product updates and be notified of any new product launches. (Carroll, 2002)

Macromedia, the developer of Flash, Shockwave, Dreamweaver and other digital creative tools, recently established several Weblogs pertaining to its products. Staff members maintain a blog that features a running commentary with news, tips on bug fixes, hints, and links to sites that have been built using the product, customer feedback and so on. There are blogs on ColdFusion, Dreamweaver and Macromedia MX. Some Weblogs are official; others clearly "personal". But in essence, all of them offer a smattering of information aimed at someone rather important - the customer. (Carroll, 2002)

Other firms in the technological industry are also taking note of the applications of weblogging in consumer applications. For instance, some of the Microsoft's initiatives aimed at weblogging include: (Foley, 2003)

- Microsoft's FrontPage 2003 product can be used as a front-end blogging tool, as can InfoPath.
- Microsoft's OneNote note-taking application (with its one-button "publish to Web" feature) can be a blog builder.
- Microsoft's ASP .Net team's recently released Community Starter Kit template can be used to develop blogs.

- Microsoft released a blogging plug-in for its Windows XP Media Fun Pack, allowing bloggers to annotate their music selections.
- Blog tools such as BlogX and .Text released by Microsoft developers are being used widely now.
- Other Microsoft developers are experimenting with related technologies, such as RSS, an XML format for syndication, for instance Microsoft's WebData team member Dare Obasanjo built a desktop news aggregator called "RSSBandit" that allows interested parties to read streaming Web-site content and headlines on their desktops.

While blogs offer an easy publishing medium for individuals, educational and research institutions, and industrial organizations; there are some issues involved in the adoption of weblogging as a standard knowledge service. Not only do the blogs existing today provide an insight into the emotional/professional lives of the young bloggers, but the ubiquitous presence of demographic details, 'consumer-talk' and trend analysis allows for a much more comprehensive vision of their everyday lives. Despite the wide-ranging educational campaigns warning children and teens of the dangers of revealing specific demographic data (name, address, etc.), the continuity offered through blogging allows for small revelations to be connected and a large amount of very precise personal data to be extrapolated. In addition to the enhanced risk of being tracked (and found) by information predators, revelatory information such as this could facilitate highly targeted marketing campaigns, by providing a continuous recording of the bloggers' 'real life' experiences (where they go and when) and social behaviors. (Grimes, 2003) This argument forms the basis for issues raised in adopting weblogging as a standard online knowledge service for learning purposes.

5.2 Issues in Adoption of Weblogging as a Knowledge Service

There has been a considerable tendency for bloggers to engage in various types of 'consumer-talk': from brief movie reviews, to listings of favorite bands etc. Often included in the retelling of the events of the day, this emphasis on brand names, media content, and new product discoveries reveals a rich body of information on consumer preferences and purchase behaviors. Bloggers often reveal the movies they see (sometimes immediately after viewing), the television shows they watch, the new videos they like (and hate), the clothing lines they're wearing, and perhaps more importantly what they would like to be wearing. (Grimes, 2003)

Due to the temporal element of blogging, this consumer-talk is not only immediate (feedback on new product launches and film releases) but also measurable. "New movies, for example, were often discussed in relation to their release dates, making it possible to determine which nights the teens were most likely to attend new theatrical releases, or how long they would wait to buy home videos." (Grimes, 2003) Thus, there's a deluge of consumer information in the blogs today that can be a source of competitive advantage for industry players that invest in activities such as information monitoring and predation, and more specifically 'trend-watching' on the Internet.

However, there are some critical ethical issues concerning both children's rights and intellectual property on the Internet today. (Grimes, 2003) Young people's blogs or websites are actually being data-mined for personal details and information, the opportunity for exploitation is nonetheless substantial. When one considers the proprietary claims that cool hunters and youth marketers have made in the past over the data collected from their young subjects, the problem becomes compounded with questions of intellectual property rights and the unauthorized commodification of children's self-produced culture.

"When considering cases that involve children and teens, the issue becomes more complicated by the ambiguity that surrounds the legitimacy of their claim to the right to freedom of speech. Children's rights to freedom of expression reside in somewhat of a legal gray area: even the United Nations Convention of the Rights of the Child does not clearly address children's ability to formulate opinions, nor does it suggest that any differentiation should be made in regards to children's age and maturity level (unfairly placing infants on the same legal footing as pre-teens) (Smith, 2002). As the very concept of freedom of speech on the Internet is currently being reconsidered, as seen in recent legal proceedings in Canada and abroad (such as the 'Zundelsite' case in Canada, or the Yahoo! France ban against the on-line auctioning of Nazi paraphernalia), and given the strong tendency in many cultures to create policies aimed at *protecting* children, rather than *empowering* them, it remains to be seen whether children's adoption of Internet technologies will even be interpreted as a civic right that should be defended." (Grimes, 2003)

Intellectual Property Rights (IPR) is a catchall term used to describe the legal status and protection that can be claimed for information and knowledge. (Casey) It is fair to say that the law is lagging behind the digital technology that is changing the way that the creation,

publication and access to the products of intellectual activity now happen, as is specifically relevant to the adoption of weblogging as a standard knowledge service. (Casey) The law for IPR is fairly clear in terms of the principles and guidelines that it embodies. However, the knowledge/information sector is currently characterized by low levels of awareness and understanding of IPR law and how it is applied to (online) learning activities using services such as weblogging.

Although the weblog intellectual property is not physically tangible, it can nevertheless be owned, sold, rented and otherwise exploited by those with a legal right to do so. Unlike physical property it may also exist in more than one place at once i.e. it may be copied. "The legal right governing who may copy a piece of intellectual property, called copyright, is one of the most important laws affecting learning activities. Some of the primary laws that govern the IPR issues include Copyright, Moral Rights, Performers Rights, Database Right, Patents, Confidentiality, Know How, Trademarks, Design Right etc." (Casey) However, the areas of IPR law that most affect (online) learning content development especially through weblog services, are those of copyright and moral rights.

"The owner of the copyright in an intellectual work enjoys the right to grant or withhold the right to others to make copies of the work; copyright is often described as a restrictive right because it is concerned with stopping others doing something with the work." (Casey) Copyright exists immediately for the creator of a work as soon as it is fixed or recorded in some material form such as in writing or on film or video etc. Moral rights are rights the original author has automatically as the creator of the work. One of the reasons these rights are called 'moral' rights is that they are not economic in nature - they cannot be sold or bought. These rights stay with the author even when the copyright to the work has been sold or given to someone else; they also can be passed on to others after the author has died. However they can be waived. (Casey)

A basic question often gets raised when investigating the Internet Intellectual property rights. Does an Internet content provider have the right to the expectation of privacy in relation to the subject matter and information she or he broadcasts on a mass (and therefore public) medium? (Grimes, 2003) Once content is released into the public domain it is assumed that it becomes, at least to some degree, a part of our shared culture. From this perspective, it would

seem that ethnographic research into children's blogs, websites, or chatrooms is in many ways equivalent to content analysis of any other publicly released media.

"Ever since the web-bubble burst at the turn of the millennium, the race to establish a viable economic model on the Internet has instigated dramatic changes in society's conception of what the Internet should be. Often referred to as the shift from free-to-fee, corporations and advertisers are desperately trying to make a profit from web tools and capabilities. Legal proceedings against file-sharing networks and copyright infringement are becoming increasingly common, and more powerful encryption codes (such as DRM) are slowly being introduced into the new media environment. The debates surrounding fair dealing (or fair use in the U.S.), and the enclosure of the commons (Boyle, 2002) versus the possible demise of copyright (Vaidhyanathan, 2001), are emerging issues that will take center stage in the years to come. When it comes to children and teens, however, the situation is again complicated by their ambiguous status as underage citizens. The question of whether minors can legitimately claim ownership over content is one that will need to be considered if we are to fully recognize their contribution to the Internet landscape." (Grimes, 2003)

In the meantime, marketers are able to take advantage of the current confusion over children's status and the lack of regulation concerning copyright. The corporate appropriation and commodification of young people's information and cultural contributions remains a sizeable problem that is insufficiently tackled in international law or national policies. Although youth marketing strategies can be described as ethnographic studies, the end result is in effect a commercial transaction. Marketers such as SMG (Student Marketing Group) compile vast databases of the information they collect, selling access to a variety of corporations and manufacturers. This invites the application of what the Canadian Copyright Act calls "moral rights", described in the *Guide to Copyrights* (2001) as an author's right "that no one, including the person who owns the copyright, (be) allowed to distort, mutilate or otherwise modify your work in a way that is prejudicial to your honor or reputation." (Grimes, 2003) The guide gives the example of a composer who has sold the copyright of a song to a publisher. An infringement of the moral right would occur if "the publisher converts your music into a commercial jingle without your permission." (Grimes, 2003) What this suggests is that the unauthorized transactions that occur when the information hunters sell their findings could be construed as not only a breach of the authors' intellectual property rights, but also of their moral right to control how that property is used. (Grimes, 2003)

5.3 Issues in Content Syndication Standards

As mentioned earlier in Chapters 2 and 4, Content syndication makes part or all of a weblog's or website's content available for use by other services. The syndicated content, or feed, can consist of both the direct content itself and metadata – information about the content. Such feeds are generated from the original blogs/sites using any of the syndication data standards or formats such as the RSS, ATOM, NewsML, OPML etc.

Currently, the two dominant standards/formats in the syndication domain are RSS (Rich/RDF Site Summary) and ATOM. With the lion's share of existing feeds based on the RSS format, and ATOM still in its early stages of development (the current version is 0.3), there's a lot of momentum already behind RSS. However, there is a competitive struggle between different industry stakeholders regarding the syndication format, as is seen between the RSS-backers (Microsoft, Yahoo) and the ATOM-backers (IBM, Google). Such competitive tussles in the market only get worsened by the competition between the standard setting organizations – the IETF (Internet Engineering Task Force) and the W3C (World Wide Web Consortium), since IETF supports the ATOM standard while the W3C is approached by the RSS-backers to adopt RSS as an industry-wide standard. The result of such market conditions is the uncertainty in coming up with a common, industry-wide and extensible content syndication standard, which severely impacts the ability of different bloggers or individuals/firms to share information between each other. This is crucial to weblog services-enabled learning initiatives since sharing of knowledge (through content syndication) is of critical importance in such contexts. (Mallett, 1998) Therefore, there is a need for the standards setting organizations to work together with the different stakeholders and develop a newer standard such as ATOM in ways that do not affect a concentrated portion of the industry/market severely (concentrated costs of transforming existing systems to comply with the new standard) or confer the advantage to a select few within the industry (such as IBM and Google in case of ATOM). (Olson)

Currently, a majority of the knowledge services and businesses in the IT industry support for RSS syndication. This enables such businesses/services to share information in an easily-understood and interoperable manner. The selling point for RSS format is that it is simple, easy to use and offers a way for information to be exchanged in a way that is platform-agnostic (Windows, Linux etc.) and easily interpreted by machines (interoperability argument). However, the development of RSS is concentrated and has one gatekeeper – the

Berkman Center for Internet and Society at Harvard Law School and Userland Software. While the RSS standard efforts are concentrated, the benefits of all the developmental efforts are diffuse (Olson) – shared by all the organizations in the industry which are willing to syndicate their information to be available for use by other knowledge services. This can be perceived as free-riding, but some organizations also derive competitive advantage by supporting their (knowledge) services and content with RSS syndication format, for e.g. a majority of Microsoft community-based software applications can easily share information between each other (interacting services) using RSS syndication. At the same time, another RSS backer - Yahoo plans to test the standard for its personalization tools, giving people the ability to automatically receive news and information feeds from third parties onto MyYahoo pages. Yahoo already started using RSS for its Yahoo News service, allowing other sites to automatically "scrape" Yahoo's top stories daily. RSS would also let MyYahoo users transport feeds from third-party content sites onto their personal pages, intermingling outside links with tailored news, video and financial information from Yahoo. The outside links then direct the reader to content on third-party pages. (Hicks, 2004) Such a layout would be a first for Yahoo, and would distinguish it from other competitors, hence the RSS-induced competitive advantage for Yahoo.

However, the RSS format has been plagued with the difficulties faced in meeting different challenges in the IT industry today, for instance scaling and extending the format ahead to support different types of data and support for web services and security enhancements. On one hand, the RSS gatekeepers want to keep the format simple for anyone to use. On the other hand, the demand for enhancements in the content syndication domain has led to an open-source, vendor-neutral initiative to develop a newer, easily-extensible standard called ATOM. RSS gatekeepers face the *trade-off between retaining simplicity and making the standard more extensible by adding complexity*. This trade-off can be perceived as *"satisficing"-a notion between "satisfying" and "sufficing"*-coined by Herbert Simon in 1957. Simon argues that people are only 'rational enough', and in fact relax their rationality when it is no longer required. (Simon) Applied to the RSS community, satisficing meant that the RSS development would achieve at least some (acceptable) level of information exchange with an easy and simple standard (such as RSS), but which does not strive to achieve its possible value – make RSS capable of supporting the ever-changing, different needs of the Internet community in terms of supporting different data types or extensibility needs.

This is indeed the domain for the standards tussle between the RSS-backers and the ATOM-backers in the IT industry today. An open standard such as ATOM will likely enhance efficiency in the industry by supporting all forms of data and communication exchanges. Using the notion of "satisficing" again, the ATOM project strives for the best possible outcome in the area of content syndication and is supported by all standards bodies (W3C, IETF) as a consequence. However, the ATOM format is *not backward-compatible with the different versions of the RSS format* used today. (Gallagher, 2004) If ATOM is established as the open, industry-wide standard, there could be a severe impact on the organizations and existing knowledge services already built on the RSS standard. To start with, there could be inertia to switch to the new standard –ATOM, in the form of increased costs to the firm - in terms of time, money and human effort required to do so. Also, the competitors of the RSS standard fueled the development of the ATOM standard– namely Google and IBM, which derive competitive advantage by being the first-movers in this domain and supporting their existing applications to share information using the ATOM standard. On the reverse side is a firm such as Yahoo which has built up personalized content services such as MyYahoo based on the RSS format. In the event of a transition to the ATOM standard, the switching costs to Yahoo services would be quite high based on their current information architectures.

So, the syndication standards tussle is best seen between the major competitors in the industry, with big industry players such as Microsoft and Yahoo increasingly embrace RSS for community-based tasks, while IBM and Sun Microsystems have been the front-runners in the development of the ATOM standard. The two major Internet protocol standards bodies – the IETF (Internet Engineering Task Force) and the W3C (World Wide Web Consortium) have been focusing on development and adoption of ATOM as an industry-wide standard. Tim Bray of Sun Microsystems, and Sam Ruby of IBM, co-chairs of the Atom Project, have been leading the effort to turn the project into an IETF working group. However, the representatives of the W3C approached the ATOM folks, suggesting that they were a better fit than the IETF. So far, the consensus within the ATOM community seems to be still leaning toward the IETF, if only because they had momentum toward the IETF route already. (Gallagher, 2004)

On the other hand, the RSS backers such as Dave Winer (Harvard Law) and Microsoft have offered to *bring RSS 2.0 to the W3C if ATOM decided to go with the IETF*, thus creating a standards tussle not only between *industry stakeholders* but also between the two *primary*

standards bodies, namely the IETF and the W3C. However, industry experts like Tim Berners Lee of W3C, Sam Ruby of IBM and the ATOM project believe that given the limited resources of the two standards bodies, it wouldn't make sense for them to take on opposing standards specifications. Needless to say, there's a good deal of disagreement between the two specification camps; while ATOM is rooted in RSS, the specification in its current form is not backwardly compatible with any of the previous RSS versions. (Kaulins, 2004)

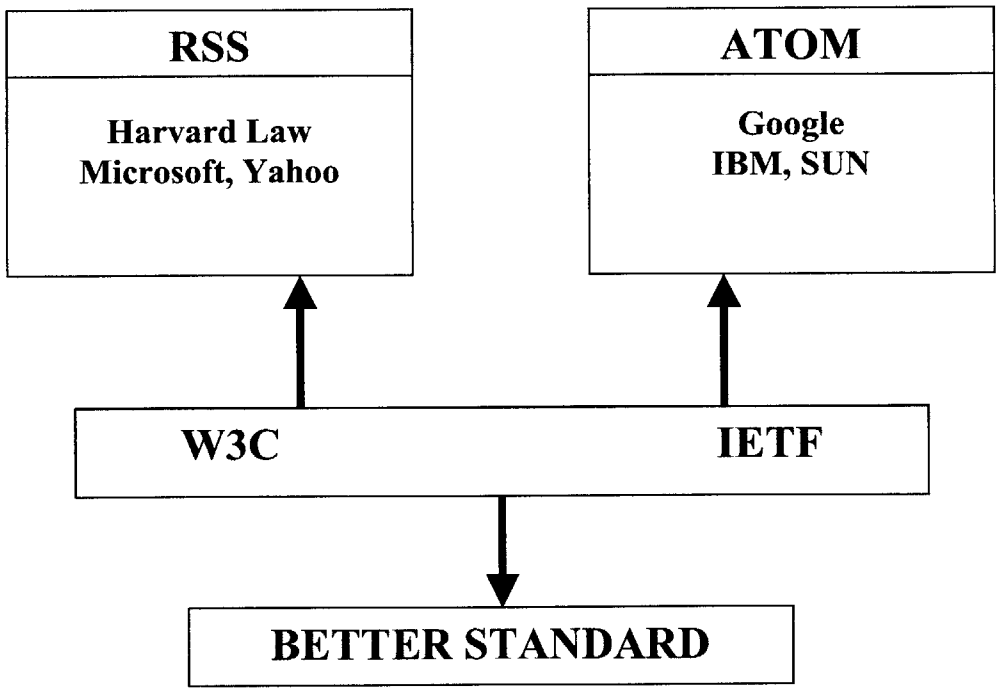


Figure 30: RSS vs. ATOM Standards Tussle

5.4 Tacking the Issues in Content Syndication Standards

The two major Internet protocol standards bodies – the IETF and the W3C should be focusing on working together on the development and adoption of ATOM or a common merged standard (between ATOM and RSS) as an industry-wide standard. Such an industry-wide standard would enable the different organizations to support their applications/services with the capability of sharing information between them. It is critical to note that such information sharing can happen within an organization or even between different organizations, thus driving the need for a well-accepted, industry-wide standard.

ATOM is perceived as a "vendor neutral standard, implemented by everybody, freely extensible by anybody, and cleanly and thoroughly specified". (Kaulins, 2004) At the same time, the standards bodies also need to make sure that existing firms in the industry already using RSS versions do not have to bear the burden of transforming their existing systems to comply with the newer standard – ATOM. If this is not ensured, the costs of such standards' development efforts would be concentrated and confer competitive advantage to a select few (Google, IBM in this case). Needless to say, the benefits of such efforts would be diffuse and costs would be concentrated on a few such as Microsoft and Yahoo (switching costs mainly to revamp their information systems and now support the newer standard). The standards bodies could adopt the following policies to enhance the standards setting process to achieve competitive market conditions:

- Work together on the development and adoption of ATOM or a common open, merged standard (between ATOM and RSS) as an industry-wide standard.
- Allow standards to co-exist with one another, for e.g. ATOM and RSS, so that the firms would have the final choice regarding the standard.
- Ensure backward compatibility of ATOM with the earlier versions of the RSS in order to maintain the competitive market conditions and also avoid conferring the competitive advantage to a select few – IBM, SUN, Google in this case.
- Most of the firms in the WWW domain follow the standards bodies such as IETF and W3C working together with one or more of the big players such as Microsoft, IBM, Yahoo, SUN, Google etc. However, there is always a problem of free riding here. Other firms have to be forced to engage in the collective action of establishment of standards. Hence, standards bodies such as the IETF and the W3C have to create conditions (incentives) for all firms to represent themselves and contribute to the collaborative development of standards in the WWW syndication and publication domain.

6. Discussion: The Road Ahead

This thesis report addressed the need for efficient knowledge infrastructure in organizations today. This need is even more evident currently as we live in a world characterized by "Information Overload". With advances in the publishing media (online and offline), our ability to generate information has far exceeded our abilities to find, review and understand it. In particular, the volume of information on the Internet has exceeded the ability of most people to find the information they need, thus giving rise to the concept of "Information Overload". The term - Information Overload, refers to the inability to extract needed knowledge from an immense quantity of information for one of many reasons. Wurman (1989) explains that information overload can occur when there is:

- Inability (on part of the person) to understand available information.
- Overwhelming volume of available information.
- People do not know if certain information exists.
- People do not know where to find information.
- People know where to find information, but do not have the key to access it.

These issues could be addressed by having efficient **Knowledge Management Systems/Knowledge Services**, so that people can create and understand available information, and have efficient services to help them learn effectively and make better decisions.

As explained in this thesis, a *Knowledge management system* or a federation of interacting *Knowledge Services* is *information ecology*, comprising of the interacting components of people (knowledge workers, managers etc.), technologies, and knowledge itself. However, a big problem with Knowledge Management (KM) itself is that the term has come to mean many different things to different people, and hence nothing at all. (Pollard, 2003) To tackle the new information needs, there is a need to have a *diverse set of knowledge services that allow personalized, customizable and independent usage.* In other words, the capture, organization, recall and dissemination of documents, messages and other personal knowledge has to be in an intuitive, transparent, automatic, personally customizable and simple manner.

New tools such as **Weblogs**, wikis and discussion group forums fulfill this requirement to a great extent. So, organizations now use a combination of technologies to address the issues in

Knowledge sharing and management. Such systems allow people to participate locally rather than following enforced technology-based KM Policies. The use of technologies such as **Weblog-Enabled Knowledge Services** (or just **Weblog Services)** offer opportunities for decentralized knowledge creation especially in learning and business applications (industry) and dissemination as such tools put the authors/users in charge of knowledge creation process without any administration-enforced policies.

Learning environments are typically characterized by challenges such as barriers to use, quality control and relevance issues, or issues of credibility of information. These issues are effectively tackled by weblog services since weblogs are often open source and need no training for authoring. In addition, favorite blogs act as information filters or "bird dogs" and point at useful information. Feedback incorporated in weblog services makes people react and learn "interactively" and also enhances credibility and trust in information. (Rajagopal et al, 2004)

Weblogs are like web pages that contain brief, discrete chunks of information called *Posts*. Such weblogs mark a shift from the Page Paradigm (of traditional web sites) to the new Post Paradigm (of weblogs and other services). These posts are arranged in reverse-chronological order (most-recent posts come first).Each post can be identified by an anchor tag, and its marked with a permanent link that can be referred by others who wish to link to it. (Doctorow, 2002) Some blogs serve as *micro-portals* or *filters*, publishing commentary and links to other sites relating to a particular topic; whereas others lean more toward *online journals*, where the content focuses mainly on the thoughts and experiences of the author. (Lindahl, 2003) In any case, a blog usually takes on the character of the person or persons that contribute to it because it is so simple to update. This ease of use leads to frequent posting, which creates fluid, ongoing "conversations" with an audience that helps to bring out the nature of the person "behind the screen". (Stone, 2003) On the same note, publishing cycles have traditionally been slowest in case of books, faster in journal publications, and fastest through conference papers. Some web publishing is done but there is no generally accepted channel or technology to drive this effort ahead. The practice of weblogging is a new dimension in the same spectrum, and promises to be the future of *online publishing*. (Rajagopal et al, 2004)

Weblogs have grown at an exponential rate, from a handful in 1998 to over a million in 2003 (source - Technorati). This growth has been driven by the popularity of different blogging engines and by a process called as syndication. As part of the thesis, 2 separate forms of blog implementations were deployed and tested out by different types of users. On one hand, a central blog server (*CADDIE Blog server*) has been administered for the users at MIT. The server has attracted as many as 70 users – comprising of individuals, research labs, classes, academic programs such as System Design and Management and the MIT Admissions office. On the other hand, blogging systems – consisting of **group** and **individual** blogging systems have been deployed as part of *CADDIE .NET* – a web-service based content architecture, which has been distributed across many educational and non-profit institutions. The encouraging aspect of weblog services is that they have been adopted in different ways in different organizational contexts – centralized, decentralized or combination approaches as were discussed in this thesis report.

Bloggers or owners of weblogs can make their weblog content more popular, or attract more traffic by a process called **Content Syndication**. Content syndication makes part or all of a site's (in this case the weblogs) available for use by other services. (Stone, 2003) The syndicated content, or feed, can consist of both direct content itself and metadata – information about the content of the weblogs. The technology or data standards to do this range from the simple beginnings of RSS 0.91, through to the RDF-based RSS 2.0, all the way to industrial strength NewsML, ICE, ATOM etc. *Resource Description Framework Site Summary/Really Simple Syndication* or short for RSS is a dialect of XML. The idea of content syndication through XML-based RSS/ATOM feeds seems to be robust, but how do users locate feeds? Information flow in this context is entirely dependent on location of appropriate feeds followed by subscription by the users in organizations. **Registries** such as *Syndic8.com* detail thousands of feeds which could be utilized by the users. At the same time, there are efficient **aggregators** available in the market today such as *SharpReader, Meerkat* Service etc. that add an additional layer of usability to RSS feeds. In addition to registries and aggregators, **search engines** such as snewp.com limit their indexing efforts solely to RSS feeds. (Hammersley, 2003) All these technologies together have enabled the adoption of RSS data formats in the corporate intranets to allow employees to track news sites for mentions of their organizations. Technological enhancements in the domain of registries, search engines as is applicable to weblog Feeds is crucial to enhancing adoption of weblog services in learning environments everywhere – be it the academic world or the corporate knowledge

infrastructures. This thesis also proposed a Feed Location and Analysis Model that would comprise of the following:

1. *Feed Crawler*: use to crawl a URL/Site for possible RSS/ATOM feeds.
2. *Feed Analyzer*: analyze the feed for keywords, process text, comments, and trackbacks.

The Crawler, and Analyzer coupled together along with a mechanism with a traffic measuring/statistical functionality would provide an efficient **RSS Feed (Upstream) Generator** – essentially designed to serve "valued" feeds to the end user based on their keyword searches. Through this model, all the text in each of the feeds is analyzed using the vector space model and then the frequency of occurrence is found in each feed. The user's keywords could then be compared with the indexed terms/text and the feeds with the highest term index/effective frequency would be served to the users as search results. However, the thesis focuses currently only on keyword search, while searching for cross-linking, comments and trackbacks along with traffic would be a futuristic goal. Such enhancements would drastically impact the accuracy of weblog feeds searched for. There's a lot of technological research going on currently for improving feed location and mining mechanisms, for instance the **Google** search mechanism coupled with blog services of **Blogger**.

While weblog services can be applied to a variety of organizations in academia and the industry, there remain a host of issues in the successful implementation of such systems. Some concerns that have been mentioned in the thesis report include:

- Internet addresses often change; weblogs can then have "**Dead Links**" hindering continuity in learning materials (Ovarec, 2002)
- **Privacy** issues owing to surreptitious recoding of individual's travels on the Internet. Online activities and personal details often analyzed for marketing value, government officials without subjects' awareness or consent.
- Organizational tussles in the content syndication standards domain between RSS and ATOM backers, including industrial organizations such as Microsoft and Yahoo on the side of RSS and IBM and Google backing ATOM; and standards-setting bodies such as the IETF and the W3C. The two major Internet protocol standards bodies – the IETF and the W3C should be working together on the development and adoption of ATOM or a common merged standard (between ATOM and RSS) as an industry-wide standard. Such an industry-wide standard would enable the different organizations to support their applications/services with the capability of sharing information between them.

References

- Anderson, T., (2004), "Toward A Theory Of Online Learning", ISBN: 0-919737-59-5.
- Angeles, M., (2004), "Supporting enterprise knowledge management with weblogs A weblog services roadmap", Lucent Technologies, Computers in Libraries 2004 conference, Washington D.C., http://urlgreyhot.com/files/cil-presentation/
- Bausch, P., Haughey, M., Hourihan, M., (2002), "WeBlog Publishing Online with Weblogs", Wiley Publishing, Inc.
- Breivik, P., (1985), "Putting libraries back in the information society", American Libraries. 16, 11.
- Carroll, J., (2002), "Blogging Alone", Internet Magazine, Sept-Oct, http://www.jimcarroll.com/articles/mktg22.htm
- Casey, J., "Intellectual Property Rights (IPR) in Networked E-Learning A Beginners Guide for Content Developers", JISC Legal Information Service, http://www.jisclegal.ac.uk/publications/johncasey_1.htm
- Chaiworawitkul, S., (2004), "Document Similarity, Searching and Learning", IESL Presentaton Reports
- Colin, L., Michele, K., (2003) "Do-It-Yourself Broadcasting: Writing Weblogs in a Knowledge-Society", Research Paper Presentation, Annual Meeting of the American Educational Research Association, Chicago, April 2003.
- Curtice, R., Rosenberg, V., (1965), "Optimizing Retrieval Results with Man-Machine Interaction", Center for the Information Sciences, Lehigh University, Bethlehem, PA.
- Davenport, T.H., and Prusak, L., (1998), "Working Knowledge:How Organizations Manage What They Know", Boston, MA, Harvard Business School Press.
- Davenport, T.H., De Long, D.W., and Beers, M.C., (1998), "Successful Knowledge Management Projects", Sloan Management Review, Winter, 39, 43-57.
- Doctorow, C., Dornfest, R., Johnson, J.S, Powers, S., Trott, B., Trott, M., (2002), "Essential Blogging", O'Reilly & Associates, Inc.
- Fine, S., "Information Technology: Critical Choices for Library Decision-Makers". ed.
- Foley, M.J., (2003), "Microsoft's Blogging On the Brain", Microsoft Watch
- Gallagher, S., (2004), "RSS, ATOM and Syndication Standards Dance", http://blog.ziffdavis.com/gallagher/archive/2004/06/05/1208.aspx

- Gallupe, R.B., (2000), "Knowledge Management Systems: Surveying the Landscape", Framework Paper 00-04, Queen's Management Research Centre for Knowledge-Based Enterprises, http://www.business.queensu.ca/kbe
- Grimes, S.M., (2003), "All About the Blog: Young People's Adoption of Internet" Technologies and the Marketers Who Love Them, ACM SIGCAS Computers and Society, Vol. 33, Issue 1
- Hammersley, B., (2003), "Content Syndication with RSS", O'Reilly & Associates, Inc.
- Herring, S.C., Scheidt, L.A., Bonus, S., Wright, E., (2004), "Bridging the Gap: A Genre Analysis of Weblogs", Proceedings of the 37th Hawaii International Conference on System Sciences
- Hicks, M., (2004), "RSS Backer Seeks Merged Syndication Format", http://www.eweek.com/article2/0,1759,1546077,00.asp
- Horton, F., (1983), "Information literacy vs. computer literacy", Bulletin of the American Society for Information Science. 9, 4.
- Kassop, M., (2003), "Ten Ways Online Education Matches, or Surpasses, Face-to-Face Learning"
- Kaulins, A., (2004) "ATOM vs. RSS – Advantages for ATOM – The Monopolists Lose", http://www.lawpundit.com/blog/2004/04/atom-vs-rss-advantages-of-atom.htm
- Lindahl, C., Blount, E., (2003), "Weblogs: Simplifying Web Publishing", Computer (IEEE Proceedings), http://computer.org/tcbb
- Mallett, R.L., "Why Standards Matter", Issues in Sci Tech Policy, Winter 1998, http://www.nap.edu/issues/15.2/mallett,htm
- Murray, H. Jr., (1966), "Methods for Satisfying the Needs of the Scientist and the Engineer for Scientific and Technical Communication in a press release", Washington, D.C..
- Nelson, M.R., "We Have the Information You Want, But Getting It Will Cost You: Being Held Hostage by Information Overload", ACM Crossroads Feature Articles, www.acm.org/crossroads/xrds1-1/mnelson.html
- Olson, M., "The Rise and Decline of Nations", Yale University Press
- Ovarec, J.A., (2002), "Bookmarking the world: Weblog applications in education", Journal of Adolescent and Adult Literacy, Vol. 45, Iss.7, pg. 616.
- Pilgrim, M., (2002), "Ultra-liberal RSS Locator, dive into mark blog", http://diveintomark.org/archives/2002/08/15/ultraliberal_rss_locator

- Polanyi, M., (1966), "The Tacit Dimension", London, U.K., Routledge & Kegan Paul
- Pollard, D., (2003), Social Networking, Social Software And The Future Of Knowledge Management, How to Save the World Blog, http://blogs.salon.com/0002007/2003/05/28.html#a251
- Pollard, D., (2003), "The Future Of Knowledge Management", How to Save the World Blog, http://blogs.salon.com/0002007/2003/10/29.html#a496
- Pollard, D., (2003), "Blogs In Business: The Weblogs As Filing Cabinet, How to Save the World Blog", http://blogs.salon.com/0002007/2003/03/03.html#a101 Pollard, D., (2003), A Weblog-Based Content Architecture For Business, How to Save the World Blog, http://blogs.salon.com/0002007/2003/03/23.html#a133
- Rajagopal, A., Williams, J.R., Sanchez, A., (2004), "Enhancing Online Learning And Knowledge Management: What Can Weblogging Do?", The 10th Sloan-C International Conference on Asynchronous Learning Networks, Nov 12-14 2004, http://www.aln.ucf.edu/
- Simon, H., "Decision Making and Problem Solving", http://www.dieoff.org/page163.htm
- Stone, B., (2003), "BLOGGING Genius Strategies for Instant Web Content", New Riders Publishing
- Udell, J., (2003), "Trends bode well for KM", Infoworld.com, 03.17.03
- Valdemarin, P., Mower M., "Weblogs for Knowledge Management", evectors, http://www.evectors.com/itkcollector/story$num=4&sec=1&data=kcollector
- Wurman, R.S., (1989) "Information Anxiety". New York: Doubleday.

Acknowledgements

I could not have completed this thesis without the help of a number of people. Here, I take the opportunity to thank them.

My deepest and most sincere thanks go to my advisor-Prof. John Williams. His contributions to this work, and to my entire learning experience while at M.I.T., have been invaluable. I will always appreciate the time and effort that he gave to me during this process. Thanks are also due to Dr. Abel Sanchez, Dr. Ning Hai, and Mr. Sakda Chaiworawitkul for their guidance in completing this thesis report along with the required implementations. They always answered my questions promptly and patiently, and as a result of our discussions, my knowledge in the domain of knowledge services and distributed computing has substantially improved. I would also like to express my gratitude to Ms. Joan McCusker for her constant support and guidance during my stay at the Intelligent Engineering Systems Laboratory. On the same note, special thanks are also due to Ms. Cynthia Stewart in the CEE Office and Ms. Sydney Miller in the TPP office.

During my time at my research laboratory – the Intelligent Engineering Systems Laboratory, I benefited greatly from discussions with fellow graduate students, including Scott Johnson, Stefan D'Heedene, Ching-Huei Tsou, Sakda, Hariharan Lakshmanan, Anamika Agarwal, Deepak Ravichandran, Naz Majidi, Bharath Krishnan, Steve, Dan Robey, and Sudarshan. These conversations were on range of topics, from IT to holiday deals to future career plans, and I learned a lot and laughed a lot as a result of them.

My experience in Boston was not limited to my lab, and I want to acknowledge a few of the many people who helped me enjoy the other side of life. Thanks to my friends-Mahesh, Vipin, Vlad, Gustavo, Ashish, Aadel, Ayanna, Manish, Rupa, Charu and Chang who were always there to do fun stuff like gym workouts or partying with me. Special thanks to all these guys and also my friends from MIT Bhangra and MIT Casino Rueda who helped me learn dancing and add considerably to my experience at MIT.

Finally, my deepest thanks go to my mother, father, and my sister – Amrita, for their support throughout this time. Without their love and encouragement, I would not be where I am today.