Appropriate Policy Enforcement through the use of Network Information Brokers

CD Keen & JW Lamp
Department of Information Systems
University of Tasmania
Email: [Chris.Keen, John.Lamp]@infosys.utas.edu.au

Abstract

The rapid growth of the Internet and the ease with which information can be provided via such services as the World Wide Web has created a publication and information dissemination network with a highly chaotic component. It is questionable whether the current practices, services and tools will be sufficiently scalable to provide effective and efficient search and retrieval capabilities. This paper proposes the concept of the Network Information Broker (NIB) as a highly adaptable organisational unit, through which some order, quality assurance and scalability can be introduced into the Internet. The development of an NIB is part of a larger research project which will also consider the implementation and impact of the NIB.

1. Scalability of the Internet

Recent discussions on the development of the Internet have highlighted issues related to the scalability of current Internet practices, services and tools. Bowman (1994a) has made the following observations:

> *It is safe to say that at least 99% of the available data* [on the Internet] *is of no interest to at least 99% of the users.*
>
> *The Internet's rapidly growing data volume, user base, and data diversity will create difficult problems for the current set of resource discovery tools. Future tools must scale with the diversity of information systems, number of users, and size of the information space.*
>
> *As the volume of information continues to grow, organising and browsing data break down as primary means for supporting resource discovery. At this scale, discovery systems will need to support scalable content-based search mechanisms.*

The public media have embraced the prospects of the Information Superhighway with utopian enthusiasm. The Internet is portrayed as the forerunner of the Information Superhighway which will not only offer interactive, multimedia entertainment and resource discovery, but also free information providers from much of the costs and dependencies on publishers and producers which are intrinsic to the current technology and publishing industry. The utopia is that of instant access to information, entertainment and a vast array of electronic services. The recent announcement on the US government funding of Internet II has increased expectations of the benefits which will flow from the next generation of the Internet.

> *The explosion of the Internet has generated economic growth, high-wage jobs, and a dramatic increase in the number of high-tech start-ups. The Next Generation Internet initiative will strengthen America's technological leadership, and create new jobs and new market opportunities.* (USIS, 1996)

However, the current state of Internet and the information resources to which it provides access can be viewed as:

- heterogeneous, in its data formats, accessing services, user interfaces, etc;
- inconsistent: the data is dynamic, not coordinated, frequently replicated with little version control;
- incomplete and frequently ad hoc;
- unmanaged, in that the value of the data typically degrades rapidly with time and frequently no mechanisms are in place to maintain currency. An example is the various X.500 style "white pages" directories which were established in the early 1990s on the Internet. While these directories were presumably accurate when established, they have not been maintained;
- technology and service-oriented, not information or people oriented;
- tending towards anarchy in the style of information providers and the information content being provided;
- straining under the rapid increase in volume of data available, particularly in the World Wide Web.

This paper considers the extent to which the provision of information on the Internet is manageable, and, in particular, the appropriate realms of policy development and enforcement. This analysis is conducted within the context of the concept of a *network information broker* (Fikes, 1995).

2. Current Stage of Development of the Internet

The Nolan's Stages of Growth model (Gibson & Nolan, 1974; Nolan, 1979) proposed an evolutionary model of the introduction, adoption and management of information technology. The six stages of this evolutionary model are:

1. Initiation—the introduction of the new technology;
2. Contagion—rapid growth in the adoption of the technology as users demand more services;
3. Control—in response to rising costs management attempt to bring the spread of the technology under control;
4. Integration—diverse services and data sources are integrated;
5. Data administration—the development of new applications and services becomes driven by information requirements;
6. Maturity—further development is in line with business or strategic directions.

In terms of Nolan's Stages of Growth Model the Internet is passing through the second stage of contagion. The rapid, uncontrolled growth in Internet usage, availability of information and active subscribers is seeing the Internet expand in a fashion which few would have predicted just three years ago. However, with this contagious growth there have been few attempts to develop manageable practices, particularly with respect to the management and delivery of quality information.

The development of the World Wide Web is a typical case in the introduction of a new service on to the Internet. The concept of a distributed encyclopedia, accessible via a simple graphical user interface, is very appealing in its simplicity. The technical development of such a service proceeded fairly rapidly, with a relatively simple minded approach. The documents were structured using the established standard SGML, but based on a simple document type definition for the markup language, HTML. The access protocol, HTTP, is equally simple and effective at fetching whole documents. Extensions have been required to permit more complex searching and the handling of forms. The referencing scheme was based on Universal Resource Locators (URLs) which typically

contain site specific information. A proposal to address the issue of site independence through the use of Uniform Resource Names (URNs), based on SGML, is contained in a recent IETF Internet Draft (Internet Engineering Task Force, 1995).

The World Wide Web was developed rapidly for experimentation purposes, but has not advanced significantly beyond its initial design. In particular, it has lacked the following:

- a richer document structuring language, such as more complete SGML Document Type Definitions (DTDs)'
- a more complex search and retrieval language, including provision for in-text searching;
- a site independent referencing scheme
- enforceable guidelines in such fundamental information resource management areas as referential integrity, version control, synchronisation of replicated data fragments;
- display independence, which permit the markup language to describe only the structure of documents, and allow the viewer to specify appropriate formatting and display schemes.

Unfortunately the World Wide Web appears to be suffering from a similar fate to other Internet services, such as mail, news, FTP, Archie, in which a great idea has not been fully developed, but has simply progressed to the level of technical sufficiency, without reaching the level of sophistication needed by the majority of users.

In a recent paper (Bowman, 1994a) the members of the Internet Research Task Force on Resource Discovery and Directory Service, characterised the problems associated with resource discovery and information dissemination over the Internet at three conceptual levels:

1. The *Information interface* layer at which the user perceives the available information. This layer needs to support scalable means of information organising, browsing and searching.
2. The *Information dispersion* layer which needs to support information object replicating, distribution and caching.
3. The *Information gathering* layer which needs to be responsible for collecting and correlating information from many incomplete, inconsistent and heterogeneous sources.

At present the World Wide Web service presents all of these levels directly to the user.

Some automated tools have been developed to assist in the *Information interface* layer to permit the user to search and browse pages which may contain relevant information. Notable among these developments is the Harvest project (Bowman, 1994b) which has produced software to build indexes using Internet gleaning systems, and then provide a relative simple search and retrieval interface to these indexes. The results of such searches typically have a high recall factor and very low precision.

There is relatively little development in the *Information dispersion* layer, with users being able to duplicate and install versions of information with very few restrictions, including some disregard for the principles of intellectual property rights. As a consequence the searcher is left with multiple copies of very similar documents, and unable to verify which is the latest or authorised version. This is most notable with FTP sites, where mirroring was seen as a useful caching mechanism, but lacked tight version control.

The *Information gathering* layer is beginning to be addressed, but there is virtually no attempt made to provide common information models, through which data from heterogeneous sources can be compared or merged.

The Internet Research Task Force on Resource Discovery and Directory Service (Bowman, 1994a) identified a number of research problems in the development of

scalable information services which can handle the increasing bulk and complexity of information, and the need to provide a high quality of service to customers, both information providers and information retrievers.

These problems include:

- client customisation of their own interfaces to support a wide variety of searching, browsing and retrieval styles;
- the provision of indexing schemes which scale up to very large, amorphous collection of information;
- the development of tools which will automate some of the tasks of maintenance of information quality, and of information discovery and indexing, without the need for intensive human guidance.

3. Appropriate place of policy

While one can easily point to the deficiencies of the Internet, it is unrealistic and indeed inappropriate to suggest that widespread adoption of policies on information management and retrieval can solve these problems.

Development of policy with respect to a heterogeneous network can be characterised as being applicable in three distinct, but not independent, areas:

1. IT infra-structure – the technology that constitutes the network

   This structure includes the physical network, the protocols and interchange conventions, interconnected data stores and the inter-information system messaging schemes.

2. Service delivery

   This area includes the types of services, the quality, relevance and usefulness of those services, the modes of interaction, the sensory media employed and the associated costs of provision of the services.

3. Information management

   This area covers the manner in which the information accessible via the network is managed, and includes the degree of dispersion and replication of information, version control, access control, integrity control, synchronisation and coordination of duplicated data fragments, and the legal issues associated with security and intellectual property rights.

It is clear that the formal and informal adoption of policies has been most successful in the IT Infrastructure area. Most notably the RFC (Request for Comment) process of proposal and adoption of protocols ensured that the Internet developed new services rapidly, without a lengthy delay in the ratification of standards.

In the area of service delivery, the adoption of policies, has been replaced by the rapid dispersion of new software and new experimental services. In most cases, such as email, FTP, archie and gopher, the simple experimental services have survived years of use, without significant enhancement from the users' viewpoint. While these services are familiar and easy to use for the technically aware, they present a major stumbling block to new or casual users who do not understand the mechanisms and are dissuaded by the primitive interfaces.

The dissemination and adoption of policy has been least successful in the area of information management. The widespread enforcement of information policies is considered inappropriate in a largely experimental venue in which a high degree of serendipity is seen as a desirable means of encouraging new ideas and trialing new services.

4. Models of communities in which policies can be successful

Although one should not regard Nolan's Stages of Growth mode as normative or deterministic, when considering the development of the Internet it is reasonable to expect greater movement towards the third, fourth and fifth stages of Control, Integration and Information management.

The chaos which is currently the Internet produces a level of frustration of the form: "the data I want must be out there somewhere, but how can I describe, find, retrieve and efficiently utilise it?"

Further, the rapid expansion of the Internet will cause greater scrutiny from government and public network providers as they see the need to finance this technology, and the opportunity to gain revenue from its usage. This will in turn, force the Internet to become less of an experimental and more of a production tool. The increasing adoption of electronic commerce will provide the Internet with the financial incentive to maintain a reliable, secure IT infrastructure.

To examine how the Internet may develop, we can consider a similar chaotic system, such as the financial markets, and in particular the stock exchange system. We have in this example an inherently unstable system which is continually in a state of change, the operations of which are largely incomprehensible to the naive or casual user. The incentives for external users to utilise the stock exchanges and other financial markets are fairly obvious. However, naive handling of these systems frequently results in financial loss. Each stock exchange system is regulated internally by its board, and its players are scrutinised by government security commissions. The provision of high quality information and honest transactions to the naive investor is provided through brokers. The brokers are presumed to have a knowledge of the operation of the markets, be honest and abide by a code of ethics, and are closely monitored by the security commissions.

The stock exchange system provides an example in which the widespread adoption of policies is generally undesirable in that there is no wish to inhibit the free operation of market forces. Policies are only enforced to ensure a level of fair trading and to protect the investor against illegal or unfair practices. The introduction of brokers imposes some order on the stock market operations, at least from the viewpoint of the investors who use brokerages. Further, brokers provide an interface between the users or investors, and the complex trading system.

The Internet is not based on the same financial model as the stock exchange. It is dealing with the less tangible commodity of information. That information has value, at least to the provider and potentially to the retriever, and hence we have the basis of a trading system.

The increasing commercialisation of the Internet emphasises the significance of this last point. The ability to provide a means by which searches can be made with a high degree of precision, recall and repeatability will attract paying clients.

5. The Network Information Broker

The concept of an information broker is widely used in the information science and library fields to refer to a middle agent who deals in information as a commodity, enabling customers to gain more efficient access to quality data. The role of this middle agent is described as “information retrieval and information organisation” (Rugge & Glossbrenner, 1995)

To address the scalability of network services and to reduce the chaos of heterogeneous protocols, the judicious use of Network Information Brokers (NIB) is proposed.

A NIB could be an organisational unit, a network site, a range of specific services, a source of information and a negotiator between customer–supplier parties. A NIB is a focal point through many suppliers can channel information to many customers.

Cameron and Clarke (1996) identify five dimensions which can affect the adoption of such electronic commerce technologies:

- the nature of the participating organisations

  The NIB would largely involve small or medium sized enterprises (SMEs), as larger organisations could probably rely on their size and reputation to attract customers.

- the cardinality of the linkages

  The NIB would be a many-to-many system, many customers would be simultaneously accessing many information suppliers. The advantage to the customers is that the multiple suppliers they would be querying, and the protocols that they were using would be hidden from them.

- the degree of competitive versus collaborative orientation

  The suppliers, especially those who were SMEs, would gain significant advantage through the collaborative nature of the NIB. This collaboration would also advantage customers.

- the longevity of the associations among the participating organisations

  One of the features of the Internet is the rapid change and growth it experiences. Provision of a stable NIB would hide changes in name, location and technological levels and facilities from the customer.

- the extent to which the project is revolutionary versus evolutionary in nature

  The NIB does not, of itself, constitute a revolutionary change in information search and retrieval. The use of networked information provision and electronic publishing, and the acceptance of this within the academic community in particular, could be considered a revolutionary change, and, to the extent to which the NIB contributes to the acceptance of electronic publishing, it could be seen as more significant.

The NIB would seem to have many of the characteristics which Cameron and Clarke consider contribute to a successful electronic commerce initiative.

The services which an NIB can provide include:

- facilitation of the delivery of goods (ie information) to the customers;
- value enhancement of the information provided;
- adherence to a code of conduct and so strengthen the honesty and reduce the chaos of network services;
- acting as a guarantor of standards of information integrity and quality of information services;
- representation of the supplier to the customer and vice versa;
- provision of new information by integrating sources from many suppliers;
- acting as a revenue gatherer for suppliers;
- advertisement of suppliers' information and services;

An NIB can determine policy at each of the levels of a network system:

1. Information infra-structure

   eg. protocol confusion – the NIB will communicate with end users in their preferred protocol and hide any conversion from differing protocols used by information providers.

2. Service delivery

   eg. levels of service – levels of service provided by information providers through the NIB would be the subject of a formal agreement between the NIB and the provider, with a minimum level required before access through the NIB.

3. Information management

   eg. integrity of information provided, such as validity of URL references – an information provider's commitment to maintaining the integrity and currency of their information would be an aspect of the agreement between the information provider and the NIB.

However, an NIB can only determine and enforce policy to the extent that:

- it does not diminish the quality of the service provided;
- it does not dissuade wanted customers or suppliers;
- it does not extend beyond the agreed realm of influence of the NIB;
- it does not inhibit the creativeness of the suppliers or customers;
- it respects the intellectual property, privacy and integrity of suppliers and consumers.

6. Quality aspects

Traditional paper publishing, the theatre and the circus, all required a large central investment and minimal costs of promotion and distribution. Quality was enforced by the need to protect this investment. The various media developed in the twentieth century have increasingly spread this investment outwards (Smith, 1980). The development of the Internet has taken this to an extreme, where there is no identifiable central investment. The development of TCP/IP based networking software for common personal computer platforms has resulted in a situation where it is possible to connect to the Internet for an investment of less than $2,000.

The rapid growth and decentralised nature of the Internet has left us with no potential to implement the sorts of quality controls with which centralised media has made us familiar. At the same time, there is a large impetus to move towards electronic publishing. To a large degree, this has been contributed to by the increasing costs of traditional publishing and therefore many proposals are being driven solely by market forces (Okerson, 1995).

Paul Ginsparg's High Energy Physics (HEP) preprint distribution (http://xxx.lanl.gov/) system is one of the more successful developments in the field of electronic publishing. It has over 20,000 users and handles 35,000 transactions per day (in Okerson, 1995). Within a year of being established, the HEP server had become the standard information distribution in that area of academic endeavour (Odlyzko in Okerson 1995). However, online access to these papers has also been criticised as resulting in a drop in the quality of the papers because the value added by peer review has disappeared (Quinn in Okerson, 1995).

Apart from sentimental attachment to the patina of paper publishing, Stephen Harnad, who has been publishing an electronic journal for some years, believes that the only factor holding back electronic publishing is the matter of quality control (Okerson, 1995). For a successful electronic journal, Harnad specifies:

- rapid peer review;
- rapid copy editing, proofing and publication of accepted articles;
- rapid interactive peer commentary; and
- a permanent, universally accessible, searchable and retrievable electronic archive (Okerson, 1995)

The rapid development of electronic publishing and the entry into the field of people with no background or experience in paper publishing, has lead to a situation where there is a confusion of roles in the area—librarian, publisher, serials agent, document delivery service, writers, editors, users (Naylor in Okerson, 1995).

The introduction of an NIB could significantly reduce this role confusion and provide a mechanism for defining and regulating levels of quality and access. In the absence of a NIB, a user has to deal directly with information providers. Apart from the difficulty of locating these information providers, there is no way, other than by further information gathering, for a user to obtain verification of the quality of the information being provided. This could be a contributing factor to the extremely low rate of citation of articles in electronic publications. Harter (1996) found that the great majority of scholarly, peer-reviewed electronic journals have had essentially no impact on scholarly communication in their respective fields.

The NIB can undertake to provide a number of value added services:

- the NIB can, by selection of, and feedback to information providers, enforce standards of quality of information, both in terms of content, and accessibility;
- the NIB can take on the roles of publishing, distribution and document delivery from the information provider;
- by having a number of NIBs covering different subject areas, indexing of material, instead of being general, could be tailored to the subject area of the particular NIB;
- the NIB can offer the user a limited number of sites which could guarantee quality, accessibility, and a high degree of precision and low recall in their information searching; and
- the NIB could handle accession of information stored in different formats and accessed by differing protocols, but present these to the user in a consistent format using the protocol with which the user is familiar.

Because the broker would act as an additional, rather than a replacement method of accessing electronic information, it would not limit access in absolute terms.

7. Conclusion

The current explosive growth in the usage and volume of information provided by the Internet raises questions about the extent to which the current services can be scaled up to provide meaningful and efficient access to, and publication of, information. If one considers the global Internet community as a single organisation, then one may reasonably expect attempts to be made to control the spread of the technology on which the Internet is based, and attempts to integrate the currently diverse services and sources of information. While it is infeasible and undesirable to attempt to impose global policies of information management on the Internet, it is realistic and desirable to create realms of information and service provision in which policies can be determined on publication, access and quality of service delivery.

The concept of the NIB is seen as a flexible structure, based on both technology and an organisational structure that can deliver quality of service in selected realms. However, the authors do not envisage that the NIB should be the dominant or exclusive model of information provision on the Internet. Advancement of the Internet, or its successor, will be achieved through experimentation , ad hoc service creation and serendipity.

A trial NIB is being developed, based on the adoption of international standards, such as SGML and Z39.50. This broker will be used as an experimental apparatus to evaluate the broker concept and determine appropriate policies and mechanisms to achieve a high quality of service.

References

Bowman CM, Danzig PB, Manber U & Schwartz MF (1994a) “Scalable Internet Resource Discovery: Research Problem and Approaches”, *Comm of the ACM*, 37, 8, Aug ‘94.

Bowman CM, Danzig PB, Hardy DR, Manber U & Schwartz MF (1994b) “The Harvest Information Discovery and Access System”, *Proceedings of the Second International World Wide Web* pp. 763-771, Chicago, Illinios, Oct 1994.

Cameron J & Clarke R (1996) “Towards a Theoretical Framework for Collaborative Electronic Commerce Projects Involving Small and Medium-sized Enterprises” *Proceedings of the Ninth International Conference on EDI-IOS* Bled, Slovenia, pp142-160

Fikes R, Engelmore R, Farquhar A, & Pratt W (1995) *Network-based Information Brokers*, Technical report KSL-95-13, Knowledge Systems Laboratory, Dept of Computer Science, Stanford University, Jan 1995

Gibson CF & Nolan RF (1974) “Managing the four stages of EDP growth”, *Harvard Business Review*, Jan 1974.

Harter SP (1996) “The Impact of Electronic Journals on Scholarly Communication: A Citation Analysis” *The Public-Access Computer Systems Review* 7(5) [Online] Available URL: http://info.lib.uh.edu/pr/v7/n5/hart7n5.html

Internet Engineering Task Force (1995) *An SGML-based URC Service* [Online] Available URL: ftp://ds.internic.net/internet-draft/draft-ietf-uri-urc-sgml-00.txt, June 1995.

Nolan RF (1979) “Managing The Crises In Data Processing”, *Harvard Business Review*, Jan 1979.

Okerson A & O'Donnell JJ (1995) *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing*, Association of Research Libraries, Washington DC, USA

Rugge, S and Glossbrenner, A (1995) *The Information Broker’s Handbook*, McGraw-Hill, USA

Smith, A (1980) *Goodbye Gutenberg: The Newspaper Revolution of the 1980's*, Oxford University Press, UK

United States Information Service (USIS) (1996) *On New Internet Initiative* [Online] Available Email: Hassan Bin Hassan <paocoord@po.pacific.net.sg>, 16 Oct 1996, Email to usis-super@spice.com