# HARNESSING THE USER COMMUNITY FOR WEBSITE LOCATION AND EVALUATION

Tim Beal and Mimi Recker
Centre for Asia/Pacific Law and Business
and
University Teaching Development Centre
Victoria University of Wellington
PO Box 600, Wellington, New Zealand
Tim.Beal@vuw.ac.nz and Mimi.Recker@vuw.ac.nz

## ABSTRACT

The AsiaWeb Project is looking at the problem of accessing quality and relevant information on the Web. It is the product of collaboration between two disciplinary standpoints. Dr Mimi Recker is approaching the problem from a generic, systems viewpoint, seeing the information domain as a case study and Dr Tim Beal has an interest in the computer access and manipulation of information on Japan, and Asia generally. We are taking business information on Asia as the general subject and within this Japan is a priority area. However, it is considered that the techniques developed could be applied to any subject area. We are working from a number of premises:

- There is an explosive growth of Web information
- The Web lacks the validating and guidance structure of the traditional publishing/library environment
- There is therefore a problem in efficiently finding relevant, quality and validated information.

One attempted solution is to appoint domain specialists as 'virtual librarians' who validate and categorise World Wide Web (WWW) sites. The limitations of this are discussed. Our approach is to develop an infrastructure for supporting collaborative, distributed information filtering of Web resources. The filtering comes not from librarians and traditional guardians of information, though neither librarians nor library techniques are neglected, but the user community itself.

## INTRODUCTION

This paper is a description of a research project being undertaken by the authors. It is therefore a report on 'work-in-progress' rather than an attempt at a definitive solution to the problems addressed. Although it is based on joint research conducted by the two authors, and draws heavily on Mimi Recker's writings, much of this description was written by Tim Beal whilst attending conferences in Asia in June 1997, which meant that there was not the usual access to scholarly resources or even the Web itself during

that period. An earlier version of this paper was presented at the 5th International Fifth International Conference on Japanese Information in Science, Technology & Commerce, Library of Congress, Washington, DC, July 30 to August 1, 1997. Pressure of other commitments has meant that we have not been able to get back to the project since then and it has not been possible to revisit this description in any depth. We hope to do both over the summer and appreciate any comments and suggestions from Australian colleagues.

## THE WEB: EXHILARATING OPPORTUNITIES; BEWILDERING PROBLEMS

The conversion of Bill Gates[1] confirms that the Internet will be central to information technologies and that the WWW will be a prime source of information; perhaps in time it will become *the* prime source. Newly published material will go on it as a matter of course and there will be a partial transfer of existing material, of various media. To some degree the process has been started by CD-ROM publishing, but the Web will quickly take over from CD-ROM and greatly expand the range of electronically-stored and disseminated material. Many of the obvious classics have already been electronically published and the flow will continue. Shakespeare, the Bible, the standard literature texts, 'great' pictures from major collections, and so on are available on the Web, often having first been published on CD-ROM. What is interesting, and as yet quite uncertain, is how deep this process will go. How many forgotten authors or artists will have their works published on the Web, to be rediscovered by unknown readers in unknown countries? Presuming that the problem of language is mitigated, either by the growth of international languages such as English, or by developments in machine-translation, or both, then a potential global readership of hundreds of millions of people becomes conceivable. The combination of a vast audience, comprised of innumerable sectional interests, and a technology which allows the delivery, relatively inexpensively, of material from the existing corpus will have interesting and unpredictable consequences.

However, the conversion of printed material to electronic format faces barriers of cost and practicality. New material, created electronically, does not face those problems. Such material - this paper, a scholarly book, the daily newspaper, a patent, a new law - can all be published on the Web at negligible extra cost.

The result will be an explosion of material of all sorts - data and information, perhaps knowledge and inspiration. However, opportunities bring with them problems.

---

1 Gates (1996) ix-xii

## PROBLEMS

Associated problems concern access and locating information.

### Issues of access

Technical issues do not concern us here but financial and political issues, though peripheral to the main thrust of the paper, should be mentioned briefly.

Bill Gates expressed surprise that "The level of investment in the Internet is amazing given that no one's making much profit yet"[2]. For academics, used to dwindling library budgets and the dearth of accessible information, the Web appears as a bonanza, showering free information to the ends of the earth. This will not last. There will be an increasingly commercialisation of quality information. Public relations (PR) and the self-published pages will remain free, other material will be sustained by advertising, but it is likely that much data will retreat behind financial barriers. However, these barriers will be, in general, much lower than those for traditional publications. With marginal cost virtually nil, it will be in the interests of information providers to keep prices low in order to garner a wider, global, audience. The main exception to this is where exclusiveness, real or perceived, is a major component of the value of the information. There is no value in knowing the winner of the next horse race if everyone else at the track knows as well.

However, even if information providers levy no or modest charges, there are other costs to be faced. Hardware and telecommunication charges[3] being the main ones, standard browser software, at the moment, being free. Restriction of access for political and social reasons is going to be a continuing and contentious issue. Statements by Joop Ave, Indonesian Minister for Tourism, Post and Telecommunications when launching a new Internet service in Jakarta are typical:

> "We are very much for the free flow of information, but it is quite obvious that there are some limits," Ave said.
> "If we talk about pornography, we say no. If we talk about things that will hamper or threaten national security, we will say no", Ave said ... Neighbouring Singapore licenses only three government-owned ISPs [Internet Service Providers], compared with Indonesia's 42 mostly private ones. The computers of all three use proxy servers capable of blocking banned sites[4].

---

2 Gates (1996): xii

3 To which we might the cost of telecommunications infrastructure, such as ISDN and fibre optic cables. This problem is most acute in the developing world.

4 The Nation, 19 June 1997:A8

Vietnam is about to license its first Internet provider and has delayed until now because "the government's prime concern was preventing subversive or other harmful material from being circulated in Viet Nam"[5].

### Locating information

This paper focuses on the difficulties of locating appropriate information on the Web. There are a number of such problems, often overlapping because they stem from the same causes. Currently, the World-Wide Web is characterised by a number of attributes including:

- Large number of sites, growing extremely rapidly
- Distributed, non-hierarchical structure
- Lack of accepted and coherent conceptual information structure
- Effervescence - Web documents are subject to constant change

It is well-recognised that the explosive increase in the number of World-Wide Web (Berners-Lee et al. 1994) resources has seen a corresponding growth in the problem of finding relevant, quality, and validated information. The Web lacks the structure and strong typing found in more closed database system (Pirolli, Pitkow, and Rao 1996). Moreover, its distributed nature and lack of accepted information structure precludes the implementation of filtering and reviewing conventions typically provided by libraries and publishers. There is no Web equivalent to the Library of Congress classification system. There is no mediating profession, such as has been provided by librarians in the past.

As a result the Web user faces serious difficulties in locating appropriate quality information. Firstly, the lack of structure leads to a sense of disorientation; it is no coincidence that 'navigation' is a favourite Web word. The Web has no end, so a search can never be exhaustive. Web documents are linked, but the nature of the linkage is uninformative. It is often difficult to gauge the authenticity and authoritativeness of Web documents because the traditional validating mechanisms are not available.

All this can be very liberating and it is one of the attractions of the Web. It can be argued that traditional classification and filtering systems both distorted the nature of information by imposing a typological straightjacket, and tended to limit publication to what was deemed acceptable. Nevertheless, being free is of limited attractiveness if one is lost at the same time, and new solutions must be sought.

## SOLUTIONS

The Web has quickly spawned a number of services to help users locate information. Bill Gates, with Panglossian optimism claims that "The interactive network's software will have to make it almost infallibly easy to

---

5 China Daily 25 June 1967:11

find information, to navigate, even when users don't know what they're looking for."[6]

His optimism is not entirely misplaced. Search engines can successfully locate required documents with incredible speed and ease. They are very good for information that belongs to a unique and compact data set, such as currency conversion, flight guides, web pages for specific organisations. They are less useful the more potential answers there are.

Their search techniques vary. For instance, Yahoo[7] has its own classification system and so is more suitable for subject-based searching. AltaVista[8] indexes keywords from documents but does not attempt to categorise them. Nevertheless, for all their strengths they do not provide a total solution. In particular, they are virtually useless for intangible questions. Sometimes intangible questions can be broken down into tangible parts, and then searching can be done successfully on those parts. This is, after all, a standard academic procedure where data is brought together from different sources to make a composite picture.

Search engines generate lists of URLs. Each URL will have some accompanying text, usually taken from the metadata (information about the information in the file), but this is invariably not very informative, although this may be changing with the Dublin Core standards, as we discuss below. Although lists may be prioritised, with those sites best fitting the search criteria at the top, the search can yield some very bizarre results. Unless the search terms are very specific and limited the potential lists can be huge, giving tens and sometimes hundreds of thousands of URLs. Even if the search appears to have been successful, generating at the top of the list sites which from the information available seem to be appropriate, one can not be really sure until the site is visited. However, every visit imposes costs, perhaps of money and always of time. The 'worthless visit cost' is one of the main problems with the Web.

The inherent limitations of existing search engines have led to various attempts to build complementary mechanisms. Within information systems, there have been several promising approaches to the problem of labelling, categorising, and filtering information. Malone et al. (1987) describe three types of information filtering activities: cognitive, economic, and social. Cognitive filtering is based on indexing content (and is what most Web search engines do). Content-based filtering depends on a machine-readable and parseable format. Unfortunately, this can be difficult to implement in a multimedia environment. Economic filtering is based on a cost-benefit analysis of searching activities. While a powerful approach in large information repositories, it generally prevents serendipitous discovery of information. Social information filtering is based on word-of-

---

6 Gates 1966: 86
7 (http://www.yahoo.com)
8 (http://www.altavista.com)

mouth and recommendations from trusted sources (Maltz & Ehrlich, 1995; Shardanand & Maes, 1995).

Social information filtering is something which we are all, in practice, very familiar. Whether choosing a restaurant or buying a car or a computer we tend to ask friends, or look up guides of some sort. We turn to those whose judgement, for whatever reason and in varying degrees, we trust. This same premise applies to social information filtering and, as proposed by others (Hill, Stead, Rosenstein and Furnas 1995), holds promise for the Web.

**Dublin Core and PICS**

An important development was the convening, in March 1995, of a Metadata Workshop, sponsored by the Online Computer Library Centre (OCLC) and the National Centre for Supercomputing Applications (NCSA). This was held in Dublin, and brought together '52 selected researchers and professionals from librarianship, computer science, text encoding, and related areas, to advance the state of the art in the development of resource description (or metadata) records for networked electronic information objects.'[9]

This workshop and its successors have developed a 15-element metadata element set to describe electronic resources. Originally intended to be author generated, it also has value for librarians and similar information describers. It attempts to be simple so that it can be used by 'non-cataloguers', but also rigorous enough to be used over discipline boundaries by professionals; ' a catalogue card for electronic resources', 'a lingua franca for resource discovery on the Internet.'[10]

The Platform for Internet Content Selection (PICS) was originally designed as a tool for censorship, enabling parents and teachers to control what children access on the Internet. However, by developing an infrastructure for associating labels (metadata) with Internet content it provides the technological basis for identifying wanted material as well.[11]

PICS is attempting to devise a set of standards that facilitate the following:

**Self-rating:**
to enable content providers to voluntarily label the content they create and distribute.
**Third-party rating:**
to enable multiple, independent labelling services to associate additional labels with content created and distributed by others. Services may devise their own labelling systems, and the same content may receive different labels from different services.

---

9 http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html
10 http://purl.oclc.org/metadata/dublin_core/
11 http://www.w3.org/PICS/

**Ease-of-use:**
to enable parents and teachers to use ratings and labels from a diversity of sources to control the information that children under their supervision receive.[12]

PICS is quite similar in some ways to Dublin Core in its search for easy-to-use metalabels, but while Dublin Core aims for a sort of consensual impartiality, PICS embraces evaluation. In theory, one identifies a ratings service whose evaluations one accepts, and this leads to the filtering out (or filtering in) of web sites that match the evaluative criteria.

**Guides, recommender systems and virtual libraries**

Evaluation and selection is what guides are all about. There are innumerable guides to the Web, both in hardcopy and on the Web itself[13]. Some are commercial, especially the hardcopy ones, but it appears that most are not. Once pricing and payment technicalities are solved it is likely that there will be a great growth in commercial guides which will attempt to satisfy complex queries that standard search engines cannot handle satisfactorily. At the same time application software developers will attempt to improve information searching procedures[14].

Recommender systems or collaborative filtering complements impartial search engines by involving the preferences of the user. The staring point is that one of the best ways to find useful information is to find someone whose judgement you respect and ask for recommendations. Collaborative filtering is a way of mechanising this form of information search[15].

An alternative approach is, in effect, to use yourself, or what you have found useful in the past, as a recommender. An example is WebWatcher which is

> 'a "tour guide" agent for the world wide web. Once you tell it what kind of information you seek, it accompanies you from page to page as you browse the web, highlighting hyperlinks that it believes will be of interest. Its strategy for giving advice is learned from feedback from earlier tours.'[16]

Other examples of interactive, self-learning and recommender systems are given in the list of URLs at the end of this paper.

---

12 http://www.w3.org/PICS/principles.html
13 There is also a plethora of magazines some of which offer advice on Web sites, but usually in the form of 'the 100 best sites', rather than subject-specific guides.
14 Gates lists five techniques: queries, filters, spatial navigation, links and agents (Gates 1996; 88-95)
15 http://www.sims.berkeley.edu/resources/collab/conferences/berkeley96/collab-announce.html
16 http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/project-home.html

One Web guide that deserves special mention is the Asian Studies Virtual Library (VL)[17] started and maintained in part by Dr T. Matthew Ciolek of the Australian National University[18]. The Asian Studies VL is part of the wider World-Wide Web Virtual Library consortium.

The "virtual librarians" are generally subject specialists rather than professional librarians who volunteer to maintain a specific site[19]. Currently the Asian Studies VL is mainly divided on a geographical basis, with sites specialising in particular countries and regions rather than subject areas. This is perhaps a reflection of the historical divisions within Asian Studies which have traditionally tended to form around languages and thus be country-specific. A VL site is basically a set of links to sites which the maintainer judges to be relevant and of sufficient quality and substance. Some of the sites linked are themselves guides to other sites.

The virtual librarians add and delete links to appropriate sites. Some use sub-divisions with their sites but there is no standard pattern. The sites are ranked on a five-point scale from marginal to essential.

For anyone with a scholarly interest in Asia, or a part thereof, the Asian Studies VL makes an excellent starting point. Although fears of its decline are unfounded, it does face problems. It seems unlikely that a group of enthusiastic amateurs will be able to keep up with the growth of web resources and the competition from commercial services. There is a lack of management structure and inconsistency of formats and procedures.

There are currently attempts to set up an editorial board to address these issues. It is likely that there will be a shift towards professionalism with all that entails in terms of funding and managerial structure. There is a need to develop a more sophisticated classification structure; what was workable for a couple of hundred sites is no longer adequate when there are thousands. There will also be a need to provide more information about sites; again, where there are only a few sites to look at, it does not take too much time to visit them all. When there are numerous sites the user needs to have sufficient information to judge whether a visit is worthwhile. Minimisation of the 'worthless visit cost' must be a prime objective.

The AsiaWeb Project[20] faces many of the problems that the Asian Studies Virtual Library is encountering but it employs a different, and complementary, approach. Whereas the Virtual Library uses subject specialists to mediate between the user and subject, the AsiaWeb Project seeks ultimately to harness the input of user communities. Our approach is

---

17 http://coombs.anu.ed.au/WWWVL-AsiandStudies.html

18 Dr Ciolek also maintains Virtual Library sites for Aboriginal Studies, Buddhist Studies, Information Quality, Social Sciences and Tibetan Studies

19 Tim Beal is charged with maintaining the site for Asia as a whole- http://www.vuw.ac.nz/~caplab/asiavl/WWWVL-AsianCont.html)

20 The name 'AsiaWeb' is not unique to this project. It is also used by the Columbia Law School, and perhaps by others.

to develop an infrastructure for supporting collaborative, distributed information filtering of Web resources. The filtering comes not from librarians and traditional guardians of information but the user community itself.

## THE ASIAWEB PROJECT

The AsiaWeb Project is an attempt at utilising various Web filtering and labelling techniques within the context of a specific knowledge domain, in this case Asian business information broadly defined. It is an ongoing research project so this paper is a description of work in progress. The specific knowledge domain is a subject area of interest to one of the researchers (Tim Beal) but the project is considered as a pilot for generic information systems solutions by the other researcher (Mimi Recker).

In this research, we are developing an infrastructure for supporting collaborative, distributed information filtering of Web resources. This basically entails two things:

- The creation of a Web-based database of sites of interest to a defined domain-specific user community. This database, and its rating system, must be both valuable and useable, and the user must be encouraged to move beyond passive utilisation of the resources into active participation to develop it further.
- The creation of a rating service, which can dynamically serve labels from the database either independently or embedded into the meta tags of documents from participating web sites.

Such a system has several important requirements, which we briefly describe in turn (Malz and Ehrlich 1995).

First, the system must easily integrate into the existing Web infrastructure. A system imposing additional cognitive and technical overhead is much less likely to be used. To this end, we build on widespread Web technology by using a simple relational database that communicates with the user's Web client via a Common Gateway Interface (CGI) compliant script. The user is thus able to utilise the database to identify sites of interest and is also able, and encouraged, to add judgements on site content and characteristics, and on the database's description of the site.

Second, it must employ a useful yet simple rating scheme. The system must be detailed enough for users to be able to identify sites worth visiting, yet sufficiently uncomplicated for the user to move to the next step, that of participating in the evaluation by adding ratings. These contradictory objectives make this the most difficult part of all. We return to it in more detail below.

Third, the system must make it technically easy for users to add ratings of Web documents. Again, a simple Web-Forms interface embedded into a Web-Frame allows users to add their ratings as they view source documents.

Fourth, a critical mass of users must participate to ensure rating reliability. Naturally, Web availability ensures a large potential user pool, but turning sufficient numbers of passive users into active participants is the problem.

Fifth, it must be easy for information seekers to see and understand the ratings of other users. We are currently experimenting with several client-side displays of "community-relevant representations." For example, if the database has several relevant ratings, it presents a composite picture of the document, thereby capturing community knowledge. The definition and implementation of 'composite' raises a number of difficulties because it must convey the complexity and richness of varying (and perhaps contradictory) judgements whilst still presenting some sort of majority verdict, if that is possible.

Ratings are also augmented with contextual information, such as title of the document, author of the rating, and usage history (Hill et al.; 1995; Maltz & Ehrlich, 1995). Together, this meta and contextual information should help users evaluate the value and quality of particular Web resources.

## THE RATING SYSTEM

The rating system has two components - a set of attributes ('dimensions') that are applicable to Web resources generally, and a set which is more domain-specific. Virtually all these dimension are tentative at this stage and we are in the process of running pilot sessions with academics from both the discipline community (library and information sciences) and from the domain community (Asian Studies) to ascertain their usefulness and to generate ideas for new ones.

The dimensions are of two sorts, evaluative and descriptive. Evaluative dimensions are ranked on a scale of one to five. We are also working on a composite ranking which will attempt to convey the balance of evaluation as well as the spread of rankings.

General dimensions currently be experimented with are-outlined below. They are intended to be compatible with the Dublin Core as much as possible.

**URL**

**Name of Site**

Usually the title, but may be supplemented by an expanded title in the body of the document.

**Author/Creator**
**Publisher/Organisation Running the Site**
**Other Contributors**
**Nature of Organisation Running the Site**
Organisations are typed as commercial, governmental, academic and institutional. The rationale behind this dimension is that the nature of the organisation is quite a good indicator of the content of a site. In particular, a user who is unwilling or unable to pay may want to skip commercial sites. An academic may wish only to visit academic sites.

**Richness of Information**
This attempts to capture the degree to which the site gives substantial, meaningful information. Such measures are necessarily subjective and relative. It will probably be useful to give examples, at least for the top end of the scale i.e.

Very rich e.g. site aaa
Quite rich e.g. site bbb
Average
Less than average
Poor

There is no guarantee that a participant will visit these sites to clarify what is meant by the measure so it would be preferable to quote well-known sites as benchmarks. However, identifying which sites are well-known to the particular user-community is a problem. Moreover, sites change, getting better or worse. In addition, the standard may change; what was a rich site yesterday may be perceived as an average one today and a poor one tomorrow.

An associated, and perhaps incorporated measure, is quantity. This is a standard measure for books and presents no substantial problems in definition. It can be very useful. We respect author X as an authority in the field. She has published two books at about the same time. Book A is 25 pages long and Book B is 400 pages. If we are looking for an authoritative overview of the subject we will go for book A, and we want a lot of detail we go for book B.

However, with Web sites, quantity is a difficult measure. There are external hyperlinks of course, and we use a separate dimension for that. Here we are thinking in terms of internal pages, but since a 'page' has no fixed standard length there is no simple arithmetical solution. Counting the number of words is often an acceptable solution for text, but not of much help in a multimedia environment.

It is also often difficult to determine when one site ends and another begins. Which part of the URL is taken as the root, and which part the branches? It should be noted that this measure is, in theory at least, distinct to the more domain-specific judgement as to authoritativeness. The domain expert may accept that a site is information rich but consider that information to be of

poor quality. Alternatively, domain experts may concur that a site is information rich, but differ as to its quality. This sort of information is useful for the user to decide whether to visit a site. A site which has a low ranking in information richness and variable rankings for quality may be skipped, but a site with the same mix of quality ratings, but a higher richness rating may be considered worth visiting.

**Links**
The number of links, probably divided into three categories (>20,5-20,<5) so that it is easy to make an estimate without adding them up. Links to other sites are often the most valuable attribute of a site.

**Time Stamping and Currency**
Since the Web is so volatile and effervescent it is important to know when a site was last edited (time stamped). It is also relevant to time stamp the review, for two reasons. If we are looking at a review in September which gives the last edit date in March and the review date in April, we do not know if the site has been updated after April. However, a large gap between the two indicates that the site is not frequently updated. As the number of reviews increase it may be feasible to calculate some measure of currency from the relations between these two sets of dates.

**Web Quality**
This is a minor measure, and one difficult to quantify. It may be broken down into various components, such as navigation, innovativeness, attention grabbing and legibility. It is somewhat analogous to book layout and print quality. If we need to choose between two books with equivalent information quality but differing standards of layout and printing then we will chose the one with better layout. Layout is an important component of comprehending information[21]. However, in general for the type of users we are thinking of, layout (and navigation) are minor considerations compared to information quality[22]. For this particular domain it is the cream on the coffee and it may be unwise to devote too much space to it. For other domains, such as Web marketing, it assumes much greater importance and would be expanded and emphasised.

**Graphical Reliance**
This is another minor measure which we may delete. Whilst of importance in some contexts where the graphical element is vital, in most cases graphics (for our users) are a minor consideration.

**Text Alternative**
This is useful for users whose browsers do not support graphics, frames etc.

---

21 Its importance varies with the nature of the information. Poorly laid-out text may be a nuisance but poorly laid-out tables may be incomprehensible.

22 Navigation will be considered more important by computer/information science professionals.

**Language Options**
The Web is currently mainly in English but the dominance is decreasing. This is important information for users who may be able to read English but would prefer to use their mother tongue if available. For some subject domains, such as Asia business, this information has added relevance.

**Domain specific dimensions**

**Authoritativeness and Information Quality**
As mentioned above this measure complements, and is conceptuality distinct from, 'richness of information'. For some domain groups which are popular and where the level of expertise is low, the distinction may be too fine to be of practical use and it might be necessary to amalgamate the two. However, for the particular user community we are targeting, the level of expertise can be expected to be relatively high, as will be the value placed on authoritative information. Whether this prediction is valid will only be confirmed after extensive testing. If it turns out that there is a strong correlation between sites which are ranked information rich and those which are ranked authoritative then the distinction will have to be abandoned.

**Geographical Focus**
This is clearly of importance to this particular domain and is relatively straightforward. We are using a set of geographical descriptors ranging from the general (global), through the regional (Asia), the sub-regional (Southeast Asia) to the country (Thailand) and then, as required, to cities (Thailand, Bangkok) or divisions within countries (China, Guangdong). Work on a set of geographical descriptors for an Asian Studies directory[23] identified a number of issues. The distinction between geographical and subject descriptor can be difficult; 'the overseas Chinese', 'ASEAN', 'Japanese investment in China' are examples where editorial decisions are required. The first two are considered subject descriptors but since users may initially consider them as geographical descriptors a 'see reference' should appear in the geographical list.

An associated issue is that terms commonly used by scholars, and others, do not necessarily coincide with 'official' ones[24]. Tibetan Studies scholars, for instance, would use 'Tibet' rather than 'China, Tibet' and many, of course, would oppose the inclusion of Tibet within China. However, the problem does not quite stop there. When such scholars use the word 'Tibet' they probably mean 'those places where Tibetans live', which in the case mean not merely the province of Tibet, but also adjoining (Chinese) provinces and parts of India, etc. The solution is to use terms which users employ, giving a 'see reference' from the official term.

---

23 Beal 1966: Geographical Index

24 It is clear that terms should be as politically neutral as possible, but there will be occasions, such as dealing with disputed territories, when this may be difficult.

Nevertheless, whilst there can be a certain amount of fuzziness about geographical descriptors, the issues are quite minor.

**Subject Focus**

This is a far more difficult area. Standard classifications of information, such as the Library of Congress Classification system, are not appropriate and there are no commonly accepted thesauri or classification systems for Asian Studies.

In the case of the NZASIA Directory of Asian Studies (Beal 1996) it was necessary to allow respondents to generate their own keywords. After a certain amount of editorial intervention[25] this appears to have produced a reasonably satisfactory set of descriptors, though there has not been the opportunity to verify that by testing user search strategies.

At the moment we are looking at the possibility of drawing on the Yahoo classification system. That has the advantage that is it relevant to the Web. However, it is, of course, not focussed on the subject domain, so as a classification system in is not appropriate, though individual terms might be utilised.

There is also the problem of how many terms to allow. Since participants will be amateurs rather than professional librarians, and they will have limited time and patience, we cannot expect them to select terms from a long list. However, if the list is too short it does not allow sites to be identified with any precision.

Our approach will be to set up an initial list of subject descriptors, probably drawn from Yahoo, but allow participants to add terms. The resulting list would need to be periodically culled and rectified. Much work remains to be done on the issue, which is probably the key to the success of the rating system and hence of the project.

**Testing and Refining the System**

We are currently conducting pilot evaluations of our system involving several groups of users from within Victoria University. These studies examine the usability and usefulness of the approach, from both human-computer interface (HCI) and social information filtering perspectives. The next stage will be to take the evaluation further by inviting general participation from Asianists in New Zealand and elsewhere known to the authors. We also hope to involve a government-funded commercial organisation in a tailored pilot study.

After each stage of evaluation we will reassess the project, particularly the rating system. No doubt many problems will surface during the course of these evaluations. Some can be anticipated, though the solutions are as yet unclear. For instance, defining the user community raises a host of problems. Should there be barriers to entry, and if so, what? The validity of

25 Beal 1996: Guide to the Keyword Indexes

the judgements of participating users will vary. Some will know more than others. Some will give more thought to entering ratings that others. Some will be flippant or mischievous. Some may be malevolent. Do these things matter and, if so, how do we tackle them?

If we can surmount these difficulties we hope to construct a rigorous collaborative filtering infrastructure that is of sufficient value to users to produce self-sustaining growth. The techniques developed could then be transferred to other subject-domains and we might see the formation of a consortium of user groups, each using a rating system tailored to their specific subject areas, but all sharing techniques and experiences. None of this would replace other search mechanisms, such as search engines and virtual libraries, but we would hope that it would provide a complementary tool by which Web resources can be located and utilised.

## ACKNOWLEDGEMENTS

Collaborators on this research are Tony McCrae, Samuel Ng, and David TenHave. We are grateful to the Internal Grants Committee, Victoria University of Wellington, for financial assistance and to the US Department of Commerce for a travel grant to enable Tim Beal to attend the Washington conference.

## REFERENCES

### Books and scholarly papers

Beal, T, 1966. *NZASIA Directory of Asian Studies and Expertise,* Wellington, NZ Asian Studies Society; also at http://www.vuw.ac.nz/~caplab/nzasia.htm

Gates, B. , Myhrvold, N. and Rinearson, P. 1996 (2nd Edition). *The Road Ahead,* Harmondsworth, Middlesex: Penguin

Hill, W., Stead, L., Rosenstein, M., and Furnas, G. 1995. *Recommending and evaluating choices in a virtual community of use.* ACM Conference on Human Factors in Computing Systems, 194-201. New York, NY: ACM.

Malone, T., Grant, K., Turbak, F., Brobst, S., and Cohen, M. 1987. *Intelligent information sharing systems.* Communications of the ACM, 30(5).

Maltz, D. & Ehrlich, K. 1995. *Pointing the way: Active collaborative filtering.* ACM Conference on Human Factors in Computing Systems, 202-209. New York, NY: ACM.

Miller, J., Resnick P., and Singer, D. 1996. *Rating services and rating systems (and their machine-readable descriptions),* PICS-1.1: August 8 1966, W3C Reports

Shardanand, U. & Maes, P. 1995. *Social information filtering: Algorithms for automating word-of-mouth.* ACM Conference on Human Factors in Computing Systems 210-215. New York, NY: ACM.

**Newspapers**

*The Bangkok Post,* 23 June 1967, Business section p. 5.
"How to be noticed and net a profit".

*The Nation,* Bangkok
"Jakarta to restrict Internet access", 19 June 1997, p. A8
"China rebels launch drive in cyberspace", 19 June 1997, p. A11

*China Daily,* Beijing
"VietNam to licence Internet provider", 25 June, 1997, p. 11

**URLS**

AsiaWeb
http://aorangi.vuw.ac.nz/asiaweb/

Dublin Core
The Dublin Core Metadata Element Set, Home Page
http://www.oclc.org:5046/research/dublin_core/
http://purl.org/metadata/dublin_core_elements

PICS
Platform for Internet Content Selection (PICS)
http://www.w3.org/pub/WWW

Recommender systems
GroupLens
http://www.cs.umn.edu/Research/GroupLens/

Fab
http://fab.stanford.edu/questions/what.shtml

ReferralWeb
http://www.research.att.com/~kautz/referralweb/index.html

PHOAKS
http://www.phoaks.com/phoaks/

**Web guides**

Daiwa Foundation Bridge to Japan
http://www.Daiwa-foundation.org.uk/

**Web guides**

Daiwa Foundation Bridge to Japan
http://www.Daiwa-foundation.org.uk/

Asian Studies Virtual Library
http://coombs.anu.edu.au/WWWVL-AsianStudies.html
http://www.vuw.ac.nz/~caplab/asiavl/WWWVL-AsianCont.html

**Others**

Infomine (UC librarians)
http://lib-www.ucr.edu/pubs/postlcsh.html

The World-Wide Web Virtual Library, Evaluation of information sources
http://www.vuw.ac.nz/~agsmith/evaln/evaln.htm

NZ Asian Studies Society, *Directory of Asian Studies and Expertise*
http://www.vuw.ac.nz/~caplab/directy.htm

Collaborative Filtering Workshop (Berkeley)
http://www.sims.berkeley.edu/resources/collab/collab-report.html

Communications of the ACM; March 1997,
Special section on recommender systems
http://www.acm.org/cacm/MAR97/marchtoc.html

Information Filtering Resources
http://www.enee.umd.edu/medlab/filter/filter.html

Carnegie Mellon University Webwatcher Project
http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/project-home.html

Journal of Economic Literature Classification System
http://www.epas.utoronto.ca:8080/ecipa/JEL.html

Australia: The Resource Discovery Unit
http://www.dstc.edu.au/RDU/