Audience and Access: Evaluating Access Tools for Multimedia Collections

Douglas Tudhope,
Daniel Cunliffe
Hypermedia Research Unit
University of Glamorgan

Abstract

Major efforts are underway to digitise cultural heritage collections for the Internet and existing collections databases, previously the domain of the professional, are being unlocked to a new public. This raises the question of how people will try to access this new form of information and how we can assess the access points and tools we provide. While not claiming to provide a solution to these issues, this paper discusses relevant literature and its implications for cultural heritage information providers. Drawing on current research at University of Glamorgan, the problem of how to evaluate novel access tools is discussed and alternative approaches compared. The paper reviews evaluation methodology from a computing perspective, illustrating issues with a case study of Web evaluation and initial thinking on a planned evaluation of a thesaurus-based retrieval prototype.

Introduction

Major efforts are underway to digitise cultural heritage collections for the Internet and existing collections databases, previously the domain of the professional, are being unlocked to a new wider public. This is occurring not only in the museum and gallery domain, but in digital library research. It raises the question of how people will try to access this new form of information and how we can assess the access points and tools we provide. While not claiming to provide a solution to these issues, this paper discusses relevant computing literature and its implications for cultural heritage information providers. Drawing on current research at University of Glamorgan, the problem of how to evaluate novel access tools is discussed and alternative approaches compared. Some museums have invested significant effort in evaluation and are evolving traditional visitor techniques to deal with computer-based gallery interactives and Web sites. As a contribution to this effort, the paper reviews evaluation methodology from a computing perspective, illustrating issues with a case study of Web evaluation and initial thinking on a planned evaluation of a thesaurus-based retrieval prototype. We focus on the evaluation of an implemented computer application or prototype, as opposed to studies of user need or requirements (see McCorry and Morrison 1995; Morrison 1998; and the other papers in this session of the conference).

It would be convenient if rigorously following design guidelines obviated the need for evaluation. However, we are far from that situation - many computer horror stories involve a failure to evaluate a product until effectively too late in the lifecycle to make changes. Nor is there one prescribed method of evaluation. There are many different methods, each with its own set of advantages and disadvantages and suited to

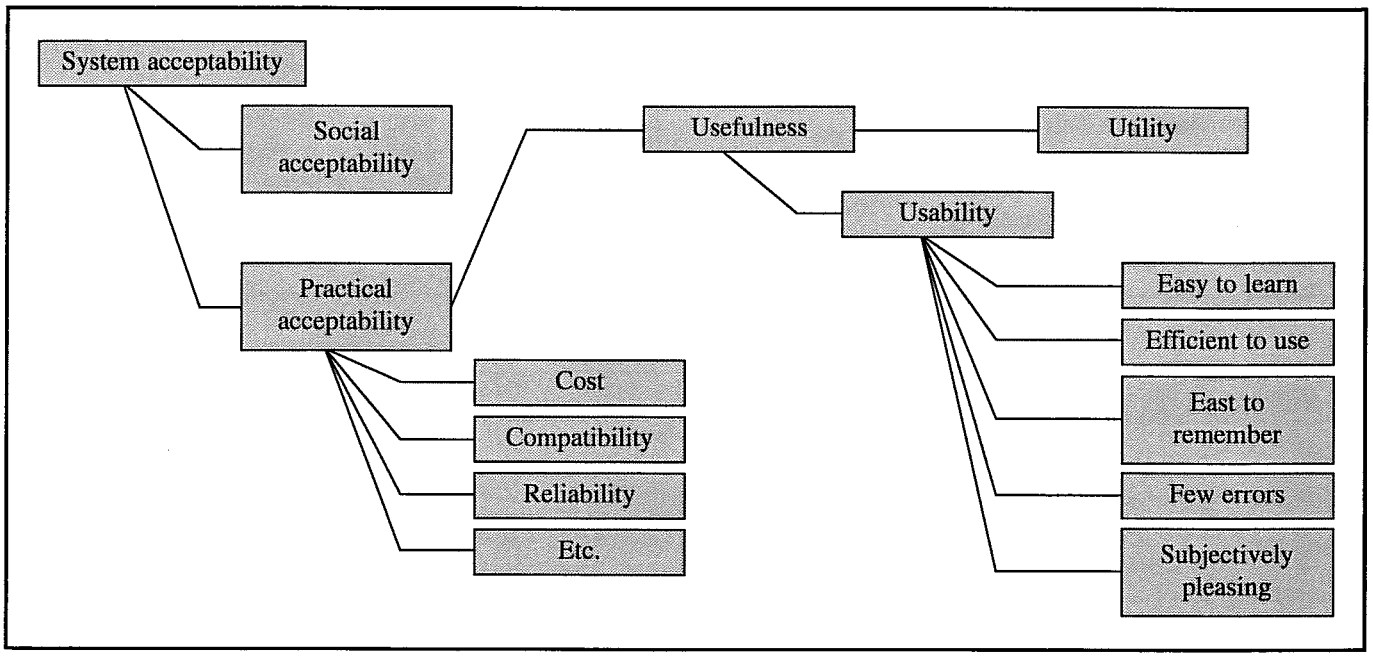

Figure 1: Aspects of system acceptability (from Nielsen 1993, p25)

different stages of the lifecycle. For detailed discussion, see HCI texts dealing with evaluation issues (e.g., Nielsen 1993; Shneiderman 1998). It is important to be clear about precisely what you wish to evaluate - usability is only one aspect of human factors issues (see Figure 1).

Evaluation can be formative - intended to help refine a design, or summative - a final test of the system's suitability. Measures can collect quantitative and/or qualitative data. Quantitative evaluation data can include details of time taken, errors noted, or affective measures, and lends itself to statistical processing. Qualitative descriptions of user behaviour are more difficult to analyse but can lead to richer understanding of the context and insights into reasons for mistakes and cognitive dimensions. The setting for the evaluation can vary along a continuum from a controlled usability lab equipped with one-way mirror, video recording and logging equipment to 'contextual' observations in a realistic workplace environment. The controlled setting may be suited for formally testing specific hypotheses and investigating causal relationships among experimental variables, such as the effect of different types of menu design or typography. A workplace setting may more faithfully reproduce typical situations, such as frequent interruption and task switching and may sometimes yield more validity when attempting to generalise to user populations beyond the trial subjects. Methods differ as to the personnel involved in an evaluation and characteristics identified as critical by researchers. Frequently this is subject to very practical constraints such as availability and expense, although the consequences may not be apparent until later in the development. A seemingly successful evaluation may be misleading if findings are subsequently generalised to a user population differing in key aspects. Even the very definition of success in an evaluation is not always obvious. For example, the common measure of the number of 'hits' on a Web site may not correspond in any direct manner to business or organisational objectives. Notwithstanding these complications, almost any evaluation will yield some insights and the inevitable limitations should not be used as an excuse to pass over the issue.

Evaluation Methods

Evaluation methods are typically subdivided into analytical, survey, inspection, and observational categories. Analytical evaluation is conducted early in the lifecycle on a (semi) formal specification, based upon a physical and cognitive model of the operations a user will perform. Single layer models consist of a linear sequence of micro physical and cognitive 'actions', such as the movement of the mouse or the typing of a key. Usually the measure is the time taken to perform some task. The 'keystroke-level' model is a well-known example of a single layer model. Multi-layer models (such as GOMS - Goals, Operators, Methods, Selection Rules) are hierarchical with tasks being divided into sub-tasks. No user testing is required with analytical evaluation. Such methods tend to focus on error-free performance by expert users and while particularly valuable in some application areas would typically be combined with another, more empirical method. Survey methods include the interview and questionnaire and are probably the most well-known to the lay population. There is an extensive literature on questionnaire design - a usual recommendation is to conduct a pilot before embarking on the main study. Design varies from fixed format to open-ended where the respondent is invited to give their own opinions or suggestions. Although response rate is always an issue, survey methods can be quite cost effective. It is important to remember that they are designed to elicit an opinion, usually post-hoc, and that there may sometimes be systematic reasons why respondents are unable or unwilling to remember the details sought. See Shneiderman (1998) for more details and further references on analytical and survey evaluation.

Inspection methods involve one or more evaluators moving through an interface and assessing it for conformance to a pre-defined set of guidelines or heuristics. Inspection methods are a widely used form of evaluation for a number of reasons: some inspection methods require less formal training for evaluators than other evaluation methods, they can be used throughout the development process, they do not require the use of test users for experimentation, and they find a large number of potential usability problems (Sears 1997). Grey and Salzman (1998) characterise inspection methods according to the type of guidelines they use (level of abstraction and perspective) and whether or not the application of guidelines is guided by scenarios (scenarios focus the inspection on user tasks rather than being an exhaustive assessment):

	Scenario	
Guidelines	No	Yes
None	Expert review	Expert walkthrough
Short list	Heuristic evaluation	Heuristic walkthrough
Long list	Guidelines	Guidelines walkthrough
Information processing perspective	N/A	Cognitive walkthrough

Table 1: Grey and Salzman's characterisation of inspection methods

The number of evaluators needed to apply a particular inspection method and the expertise required by the evaluators varies. For example, Nielsen (1993, p156) suggests about five and at least three independent inspections are required for an effective heuristic evaluation. Nielsen also recommends that the evaluators should have experience in performing such evaluations, but also suggests that useful results can be achieved by non-experts. Different inspection methods use different numbers of guidelines, guidelines at different levels of abstraction and guidelines from different perspectives. For heuristic evaluation, for example, Nielsen recommends using around ten high level heuristics. The effective use of high level, abstract heuristics, such as "Speak the users' language" (Nielsen, 1993) requires a greater element of professional judgement, less experienced evaluators may find a larger, more detailed set of guidelines more appropriate. While HCI research has produced a number of sets of high level heuristics and also low level guidelines (e.g., Smith and Mosier 1986), the simple application of these guidelines to Web sites and multimedia products is not always appropriate nor sufficient. More specialist forms of inspection method have emerged for particular domains with faceted lists of guidelines. For example, the CIDOC Multimedia Working Group (Trant 1997) have developed multimedia evaluation criteria for kiosk/CD/Web applications in the museum domain, with 49 checkpoints arranged by categories: Content, Functionality, Interface, Implementation, Overall. In the CBL domain, Barker and King (1993) have developed an evaluation check list and supporting notes for interactive multimedia courseware. Typically such lists of heuristics require subjective assessment on the part of the evaluator, e.g. "Is the presentation consistent? Clear?" The scope of the guidelines is also an important consideration, some focus exclusively on usability issues, whilst others include consideration of wider issues such as the pricing of the product and the delivery platforms.

Recently, highly specific guidelines have been proposed based on an underlying model of the domain. Faraday and Sutcliffe (1997) use guidelines for timeline-based multimedia presentations based on a cognitive model of multimedia comprehension. They focus on the micro 'moment-by-moment' detailed design decisions on the choice and synchronisation of different media types (text, image, sound, animation, etc) and highlight the importance of 'contact points' for relating audio and visual channels. Garzotto and Matera (1997) are developing a systematic method for inspecting navigation issues in hypermedia systems (SUE). In the inspection phase SUE uses a set of specific abstract tasks which codify the inspection patterns and behaviours of experienced evaluators in a form that can be applied by novice evaluators. Whilst the philosophy of the approach is to support novice inspections, currently SUE involves a preparatory phase requiring the use of a formal hypermedia model, typically not appropriate for novice evaluators. However work is ongoing and the aim is to develop a method which yields high levels of inter-evaluator consistency.

Whilst the fact that inspection methods do not require test users has benefits in terms of costs and convenience, it is also a weakness. Concerns have been raised regarding the number of false usability problems identified by such methods (Bailey et al., 1992, cited in Grey and Salzman 1998), that is problems that real users performing real tasks would not encounter. This may also in part be due to the application of inappropriate guidelines or the inappropriate application of guidelines. The selection and application of guidelines can usefully be informed by consideration of users and their tasks. Another consequence of not involving users is that usability problems resulting from users behaving in ways unanticipated by the developer or evaluators will not be identified.

Observational techniques can include various data gathering methods, ranging from simple direct observation by the researcher (taking fieldnotes), to taped 'think-aloud' sessions (concurrent or retrospective) or co-evaluation (two users working together with their conversation recorded), to videotape, interaction logging, and even eye tracking. These methods are sometimes classed as 'empirical' evaluation - some manner of representative user is observed operating on an implemented system or prototype. Dealing with the quantity of data gathered is often a practical problem for subsequent analysis and sometimes privacy concerns may be an issue. Another key issue is the extent and manner to which the users may be affected by the observation process itself (the 'Hawthorn effect'). This can arise in different guises with the different data gathering methods. A very practical concern is the need to re-assure users that they are not the focus of the evaluation but rather the system and that should they feel uncomfortable they may stop or ask questions at any time.

Museum Evaluation

Reports of museum evaluation studies can be found in relevant cultural heritage publications, including **mda** Conference Proceedings, the International Laboratory of Visitor Studies Review, Archives and Museum Informatics, Computers and the History of Art. Some computing literature has addressed museum evaluation issues specifically. One of the first studies was Hardman's (1989) evaluation of the early hypermedia system, Glasgow Online, which employed direct observation, think aloud and videotaping alongside an analysis of hypertext structure. A good example of a kiosk evaluation by an Apple HCI team, involving transaction logging and iterative design can be found in Salomon (1989). Garzotto and Matera's (1997) inspection method discussed above was applied to several cultural heritage CDs. Fidel (1997) discusses image retrieval from photographic archives, emphasising the importance of user tasks and context on the design of performance measures. Shneiderman et al (1989)

compared three museum evaluations and (among other issues) emphasised the critical issue of the initial instructions or training given to users. This point also arose in a trial evaluation we conducted of walk-up use of an early prototype exploring use of SHIC (the Social History and Industrial Classification) for presentation and access purposes (Tudhope et al 1994, 1998).

Web Evaluation

The Web poses a number of challenges for developers:

- The element of user control over the presentation of the Web page, the point of entry to the Web site and the pages visited.
- Lack of control over the users hardware and software platform.
- Lack of control over network and server response times. Speed of response has been identified as a major usability issue for users (GVU, 1998).
- The relative importance of content, visual appearance and usability can be both site and user specific.
- Usability must be considered at both site-level and page-level. Where a number of sites form a collaborative network, inter-site usability must also be considered.
- Although the current cultural orientation of the Web is towards the USA, there is a growing realisation of the need to consider the potential world wide audience and embrace a wide variety of cultural conventions, languages, and so on.
- Whilst it is generally accepted that the two main modes of user interaction are directed searching and exploratory browsing, users actually exhibit a range of behaviours which must be supported (Marchionini 1995; Smith et al. 1997).
- The need for definitions of site success which are more satisfactory than simple hit counts.
- The wide range and variable quality of guidance available, ignorance of existing research work, inappropriate application of research work, and the difficulty in applying theoretical research to practical developments.
- In large Web sites it may not be practical to evaluate each individual page due to resource constraints. A focus on user scenarios or on key pages can reduce the number of individual pages that must be evaluated.

A recent snapshot of HCI research on Web usability can be found in Shum and McKnight (1997), for example Shneiderman (1997). Bevan (1998) lists a good set of practical Web-specific guidelines while Smith (1996) provides a more theoretical perspective on hypertext evaluation and discussion of an experiment measuring 'lostness' in hyperspace. To illustrate practical issues involved in Web evaluation, we now consider a recent 'low-budget' evaluation conducted at Glamorgan.

Case Study

The Web site developed provides information about The New Review of Multimedia and Hypermedia, an annual review journal (NRHM 1999). It is a relatively small site, consisting of approximately one hundred and sixty pages, mainly paper abstracts. No formal development model was used though Bevan's paper (Bevan 1998) was used as an informal guide. The development of the site was completed without any formal user input, relying on the intuitions of the information providers and competitive analysis of existing Web sites to provide user needs models.

The summative evaluation of the site employed a combination of methods, selected according to the following considerations:

- The evaluation had to be completed in approximately two weeks.
- The intended users of the Web site were spread across the globe.
- The perceived importance of testing common user tasks.
- Reported methods used in similar evaluations.
- The reported advantages and disadvantages of the methods.
- The evaluation team consisted of one inexperienced person, and there was no access to specialist HCI testing facilities or equipment.

Based on these considerations, four techniques were selected: direct observation of usability tests; heuristic evaluation; on-line questionnaire; and transaction log analysis.

Direct Observation

Within the constraints of the evaluation, the gathering of a test group of real users proved problematic. In order to perform an evaluation with the proxy users that were available, a set of six task scenarios was constructed, each task scenario presented a specific task (Kritou 1998, p37):

Assume that you are an author who wants to submit a paper and looks for an appropriate journal. You have found the NRHM Web-Site. Find out what topics can be submitted in the next issue and the guidelines for submitting a paper to this journal.

Evaluations were performed in the subjects offices to provide a natural context. During the evaluation subjects were encouraged to 'think aloud', a video camera was used to record the users utterances and on-screen behaviour. While the 'think aloud' protocol helps in understanding why users behave the way they do, most of the subjects reported that they were uncomfortable with it and that it interfered with their execution of the tasks. The evaluator also completed an observers notebook during the evaluation. This was used to record the time taken for each task, whether the task was completed, whether the subject followed the optimum path (the path involving the fewest 'clicks') and the subjects affective state. In practice the evaluator found it difficult

to record all the information required and some of the details gathered were of little practical worth. The evaluation was followed by a short interview during which the subjects were encouraged to comment generally on their experience and to express their subjective satisfaction.

Heuristic Evaluation

The evaluator reviewed a number of well-established principles of usability and Web site design in order to derive a set of eight high level heuristics. The evaluator then performed a page by page inspection of the entire site. Each heuristic was described by a heading (shown below) and a short paragraph highlighting key concerns.

- Consistency and conformance to standards.
- Recognition and predictability.
- Web pages should stand alone.
- Flexibility and efficiency of use.
- Effectiveness.
- Readability.
- Every page should express one topic or concept.
- Consider the global audience.

Log Analysis

The evaluation collected Web access logs over a period of sixteen days. In addition to simple 'hits per page' information the evaluator was interested in trying to identify particular patterns of use. The evaluator encountered several difficulties with log analysis:

- IP addresses are not unique identifiers for users, therefore identifying users and tracking their behaviour over several interactive sessions is problematical.
- The log of accesses may not contain a full record of interactions if pages are cached.
- It is difficult to identify specific usability problems from the analysis of Web logs alone, and this analysis may be highly subjective.

The importance of capturing real user interactions should be stressed even though we may lack effective tools for gathering and analysing it. One technique which may prove effective for sites with search facilities is the capture and analysis of search terms (Bevan 1998; Rosenfeld and Morville 1998, p173).

On-line Questionnaire

An on-line questionnaire was created containing questions to gather three types of information: demographic information, technical information, and visit information. The questionnaire was linked to prominently from the NRHM homepage and e-mail was sent to various mailing lists and individuals.

Demographic information included occupation, age, locality and Internet experience. Technical information included the type of browser used and the speed of their Internet connection. As the intended audience were expected to be technically literate, this type of question was appropriate. Where the users are technically naive they may not know the answers to them potentially leading to non-completion of the questionnaire, guessed answers, or missing data. Visit information included the number of times the user had visited the site, their purpose in visiting the site, the page they entered the site on, pages they visited, how useful they found the pages they visited and which pages they bookmarked. These questions were intended to build up a more detailed picture of user behaviour. Users were also asked to rate their general satisfaction with the site and were offered the opportunity make any general comments.

The two major concerns with on-line questionnaires and similar feedback mechanisms are the self selecting nature of the sample and the response rate required to draw reliable conclusions. The response rate for the questionnaire was between 5.9% and 2.3% depending on how the number of visitors is determined. Although this compares reasonably to the 2.0% rate reported in the evaluation of the Science Museum Web site (Thomas and Paterson, 1998), the actual number of responses was low and it was impossible to draw reliable conclusions from the results.

Comparison

A total of twenty-two potential usability problems were identified by the methods used (Table 2). Ten problems were identified solely by heuristic evaluation, nine were identified solely by direct observation, one was identified by heuristic evaluation and direct observation, one was identified by heuristic evaluation, direct observation and log analysis, and one was identified by all four methods.

The usability problem identified by the online questionnaire was contained as a general comment rather than in answer to a specific question. The two usability problems identified by log analysis, relied heavily on the evaluators subjective interpretation. These methods are best suited to capturing information about users and their behaviour, rather than identifying specific usability problems. Questionnaires can also be useful for determining if a site is meeting the general needs of its users. The potential usability problems identified by direct observation include some probable false positives resulting from using test subjects who were proxy users. For instance 'The difference between Editors and Editorial Board is not explained' is unlikely to be experienced by real users. Where real users are observed, such false positives should not occur, providing the users are representative of the user population as a whole. The heuristic evaluation also identified potential false positives, such as 'There is no Help facility'. The direct observation did not identify the need for a help facility (though this could be due to the use of proxy users). This reflects the difficulty in selecting and applying appropriate heuristics. Heuristic evaluation identified the largest number of potential usability

Description of the problem	HE	DO	LA	OQ
Search facilities are not efficient enough	✗	✗	✗	✗
Index is not consistent across the Web site	✗	✗	✗	
'Return to top' is not used consistently	✗			
There are four broken links in the Web site	✗			
There is no Help facility	✗			
Search facilities not visible instantly/Grouping of links not effective		✗		
Instructions to Authors link not available next to each theme for submission		✗		
A Volume page in Hypermedia Journal doesn't link back to the Volume	✗			
Pages don't provide the creator, date of creation, update and copyright	✗			
The title (in <TITLE> tag) is not always representative of the Web site	✗			
In a small screen the Home Page is too long	✗			
As the Web site gets larger the Index will be very long		✗		
The layout is too simple and not inviting		✗		
Hypermedia Journal not visible in Volume Contents in a small screen	✗			
The distinction between the two journals is not emphasised enough	✗	✗		
The author's address is not available when clicking on his name		✗		
The difference between Editors and Editorial Board is not explained		✗		
Subscription information is insufficient		✗		
The list of papers under an authors name is not numbered		✗		
The purpose of the Web site is not stated	✗			
The 'no abstract available' message causes confusion	✗			
There are no instructions on how to get a full paper		✗		

Table 2: NRHM Comparison of methods
(HE = Heuristic Evaluation, DO = Direct Observation, LA = Log Analysis, OQ = Online Questionnaire)

problems, but some of these, such as 'The no abstract available message causes confusion' were only found because the inspection was comprehensive. There are only a very small number of papers which have no abstract, so if a scenario based inspection had been performed it is unlikely that this would have been discovered.

The purpose in identifying usability problems is that they can then be rectified. However in many cases it may not be cost effective to rectify all the problems that have been identified. In order to make an informed choice, some form of severity ranking is necessary (Nielsen 1999). Often this involves ranking by, and agreement between expert evaluators. Methods for placing this on a systematic footing so that non-expert evaluators can perform this activity effectively have yet to be developed.

This case study used a variety of complementary evaluation methods and suggests that a combination of methods is effective in assessing usability from a number of perspectives. Direct observation and inspection methods proved useful for identifying specific usability problems. The questionnaire and log analysis provided useful information on actual user behaviour rather than behaviour under evaluation conditions, and could be useful for tracking changes in user behaviour over time. Whilst each of the methods used can potentially produce useful information, it is important to recognise that each method also has limitations. The use of an inappropriate method, or of a method under inappropriate conditions or assumptions is likely to result in misleading or incomplete results.

Future evaluation studies at Glamorgan

We are currently considering the design of future evaluations of research prototypes of thesaurus-based retrieval systems. The University of Glamorgan Hypermedia Research Unit

has a longstanding interest in the question of how the semantic structure underlying information can be used to enhance browsing and search tools, particularly in the cultural heritage domain. A classification system, or thesaurus, embodies a semantic network of relationships between terms. Thus it has some inherent notion of distance between terms, their 'semantic closeness'. Distance measurements between terms can be exploited to provide more advanced navigation tools than relying on fixed, embedded links. The algorithm is based on a traversal function over the underlying semantic net (Tudhope and Taylor 1997), a function of the steps to move from one term in the index space to another term - exact match of terms is not required. Each traversal diminishes the semantic closeness by a cost factor which varies with the type of semantic relationship connecting the two terms. Previous research employed a small testbed, largely archival photographs from the Pontypridd Historical and Cultural Centre, indexed by the Social History and Industrial Classification (SHIC 1995). Advanced navigation options included query expansion when a query fails to return results and requesting information similar to the current item (Cunliffe et al. 1997).

One current research project, emphasising spatial access to cultural heritage collections, investigates the combination of various distance measures to augment the retrieval capabilities of online gazetteers or geographical thesauri (for example, the TGN - Getty Thesaurus of Geographic Names). Another EPSRC funded collaborative project with the National Museum of Science and Industry (NMSI) will make use of the Getty Art and Architecture Thesaurus (AAT) as the terminology system. We are currently considering evaluation strategies for these projects. As discussed above, previous museum setting evaluation of prototype information exploration tools raised issues to consider before inviting members of the general public to trial any system (Tudhope et al. 1994). It proved difficult to separate access and navigation issues from general user interface issues in the evaluation of the prototype, which had not involved surface details of the interface as a prime concern. Based on this experience, we intend to distinguish evaluation and refinement of the distance measures from empirical, operational studies of online use with representative users. Evaluation, more narrowly focused on the distance measures will need to occur before more contextual evaluations of actual use which will include consideration of broader HCI issues and visualisation of results. The initial evaluation of distance measures is better suited to expert users in the subject domain with some knowledge of the use of controlled vocabularies in indexing and retrieval. Key issues include tuning of controlling parameters for the semantic distance measures and feedback on application scenarios.

The basic assumption that there is a cognitive basis for a semantic distance effect over terms in thesauri has been investigated by Brooks (1995) in a series of experiments exploring the relevance relationships between bibliographic records and topical subject descriptors. Subjects are essentially asked to assess the similarity of two texts (the record and the descriptor). Analysis found a semantic distance effect, with an inverse correlation between semantic distance and relevance assessment, modified by various factors. Rada has also conducted experiments on semantic distance measures using the MEdical Subject Headings (MESH) thesaurus and the Excerpta Medica (EMTREE). In these experiments, expert users (physicians familiar with information retrieval systems) were asked to judge the semantic closeness of the same set of citations and the query (Rada and Barlow 1991). Results were ranked and the experts' evaluation compared with different versions of the algorithm. These experiments suggest a starting point for initial evaluation of measures of semantic distance measures. Expert subjects will be given an initial information need, expressed as a set of thesaurus terms and asked to compare that with a sample result set of information items and their index terms. Subjects will be asked to rank (or score) items and then compare that with retrieval tool results. Providing information seeking scenarios will probably be necessary to assess the use of semantic distance measures in retrieval in a realistic manner.

References

Barker P., King T. 1993. Evaluating interactive multimedia courseware - a methodology. *Computers in Education*, 21(4), 307-319.

Bevan N. 1998. Usability issues in web site design (version 3, April 98), http://www.usability.serco.com/netscape/index.html, also in Proceedings UPA '98.

Brooks T. 1995. People, Words, and Perceptions: A Phenomenological Investigation of Textuality. *Journal of the American Society for Information Science*, 46(2), 103-115.

Cunliffe D., Taylor C., Tudhope D. 1997. Query-based navigation in semantically indexed hypermedia. *Proceedings 8th ACM Conference on Hypertext* (Hypertext'97), 87-95.

Faraday P., Sutcliffe A. 1997. *Evaluating multimedia presentations. New Review of Hypermedia and Multimedia*, 3, 7-37.

Fidel R. 1997. The image retrieval task: implications for the design and evaluation of image databases. *New Review of Hypermedia and Multimedia*, 3, 181-199.

Garzotto F., Matera M. 1997. A systematic method for hypermedia usability evaluation. *New Review of Hypermedia and Multimedia*, 3, 39-65.

Gray W., Salzman M. 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, 13, 203-261.

GVU 1998. GVU's 10th WWW User Survey. http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/

Hardman L. 1989. Evaluating the usability of the Glasgow Online hypertext. *Hypermedia*, 1(1), 34-63.

Kritou E. 1998. A case study of web usability evaluation. MSc thesis, School of Computing, University of Glamorgan.

Marchionini G. 1995. *Information seeking in electronic environments*. Cambridge University Press.

McCorry H., Morrison I. 1995. *The Catechism Project*. National Museums of Scotland: Edinburgh.

Morrison I. 1998. SCRAN and its users. *New Review of Hypermedia and Multimedia*, 4, 255-260.

Nielsen J. 1993. *Usability Engineering*. Academic Press: Boston.

Nielsen J. 1999. *Severity ratings in heuristic evaluation*. http://www.useit.com/papers/heuristic/severityrating.html

NRHM 1999. http://www.comp.glam.ac.uk/~NRHM/

Pejtersen A. 1989. The Book House: Modelling user's needs and search strategies as a basis for system design. Riso-M-2794 Technical Report, Riso National Laboratory, Denmark.

Rada R., Barlow J. 1991. Document ranking using an enriched thesaurus. *Journal of Documentation*, 47(3), 240-253.

Rosenfeld L., Morville P. 1998. *Information architecture for the World Wide Web*, O'Reilly.

Salomon G. 1990. Designing casual-use hypertext: the CHI-89 InfoBooth. *Proceedings ACM Conference on Computer-Human Interaction*.

Sears A. 1997. Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, 9 (3), 213-234.

SHIC 1993. *Social History and Industrial Classification: A subject classification for museum collections*. Museum Documentation Association, Cambridge, UK.

Shneiderman B. 1997. Designing information-abundant web sites: Issues and recommendations. *International Journal of Human-Computer Studies*, 47, 5-29.

Shneiderman B. 1998. *Designing the user interface*. Addison Wesley:Reading, Mass.

Shneiderman B., Brethauer D., Plaisant C., Potter R. 1989. Evaluating Three Museum Installations of a Hypertext System. *Journal of the American Society for Information Science* 40(3), 172-182.

Shum S., McKnight C. (eds.) 1997. Special issue on World Wide Web usability. *International Journal of Computer Studies*, 47.

Smith P. 1996. Towards a practical measure of hypertext usability. *Interacting with Computers*, 8(4), 365-381.

Smith S., Mosier J. 1986. *Guidelines For Designing User Interface Software*, Mitre Corporation Report MTR-9420, Mitre Corporation.

Smith P., Newman I., Parks, L. 1997, Virtual hierarchies and virtual networks: Some lessons from hypermedia usability research applied to the World Wide Web. *International Journal of Human-Computer Studies*, 47, 67-95.

Thomas N., Paterson I. 1998, *Science Museum Web site assessment*. Research Report by Solomon Business Research, http://www.nmsi.ac.uk/eval/

Trant J. 1997. *Multimedia evaluation criteria: revised draft*. ICOM CIDOC Multimedia Working Group. http://www.cidoc.icom.org/

Tudhope D., Beynon-Davies P., Taylor C., Jones C. 1994. Virtual architecture based on a binary relational model: A museum hypermedia application, *Hypermedia*, 6(3), 174-192.

Tudhope D., Taylor C. 1997. Navigation via Similarity: automatic linking based on semantic closeness. *Information Processing and Management*, 33(2), 233-242.

Tudhope D., Taylor C. 1998. Terminology as a search aid: Using SHIC for public access. ***mda*** *Information*, 3(1), 79-84. Museum Documentation Association, Station Road, Cambridge.