Preprint: final version available as:

BLANDFORD, A., ADAMS, A., ATTFIELD, S., BUCHANAN, G., GOW, J., MAKRI, S., RIMMER, J. & WARWICK, C. (2008) PRET A Rapporter: evaluating Digital Libraries alone and in context. *Information Processing & Management*. 44. 4-21. DOI http://dx.doi.org/10.1016/j.ipm.2007.01.021

There may be differences between this and the published version.

The PRET A Rapporter Framework: evaluating Digital Libraries from the perspective of information work

Ann Blandford[1], Anne Adams[1,3], Simon Attfield[1], George Buchanan[1,4], Jeremy Gow[1], Stephann Makri[1], Jon Rimmer[2] and Claire Warwick[2]

[1] UCL Interaction Centre, University College London, Remax House, 31-32 Alfred Place, London WC1E 7DP, UK.
[2] School of Libraries, Archives and Information Studies, University College London, Gower Street, London WC1E 6BT, UK

Corresponding author:
Ann Blandford: A.Blandford@ucl.ac.uk, tel. +44 20 7679 5288, fax +44 7679 5295

The strongest tradition of IR systems evaluation has focused on system effectiveness; more recently, there has been a growing interest in evaluation of Interactive IR systems, balancing system and user-oriented evaluation criteria. In this paper we shift the focus to considering how IR systems, and particularly digital libraries, can be evaluated to assess (and improve) their fit with users' broader work activities. Taking this focus, we answer a different set of evaluation questions that reveal more about the design of interfaces, user–system interactions and how systems may be deployed in the information working context. The planning and conduct of such evaluation studies share some features with the established methods for conducting IR evaluation studies, but come with a shift in emphasis; for example, a greater range of ethical considerations may be pertinent. We present the PRET A Rapporter framework for structuring user-centred evaluation studies and illustrate its application to three evaluation studies of digital library systems.

Keywords: digital library; usability evaluation; HCI; case study

[3] Present address: Institute of Educational Technology, The Open University, Milton Keynes, MK7 6AA

[4] Present address: Department of Computer Science, University of Swansea, Singleton Park, Swansea, SA2 8PP

The PRET A Rapporter Framework: evaluating Digital Libraries from the perspective of information work

1 Introduction

One of the priorities in setting up any evaluation project is to choose appropriate evaluation techniques and construct a plan of the evaluation. Within the Information Retrieval (IR) tradition, there are some well established approaches to evaluating the performance of retrieval algorithms (e.g. Tague-Sutcliffe, 1992) and, more recently, there has been an increasing focus on user-oriented evaluation criteria and methods for evaluating IR systems within the context of user–system interaction (Interactive Information Retrieval) (e.g. Borlund, 2003). Important as these evaluation criteria are, they continue to focus largely on algorithm evaluation; there are other criteria that need to be considered if IR systems are to be truly useful within the context of users' broader activities. People using IR systems are most commonly retrieving information in support of some larger task such as writing a news article or an essay, preparing a legal case or designing a novel device. Evaluating a system in terms of its fitness for purpose in this broader sense demands a different approach to evaluation from the methods that have become established within the IR tradition. In this paper, we present a framework for planning user-centred evaluation studies that set systems within the context of information work. We illustrate the application of the framework to evaluations of various digital libraries (DLs).

Digital libraries are coming into widespread use to support information work. While DLs are not simply "IR systems", they are an important class of systems within which IR algorithms are routinely implemented, and effective information retrieval is one essential feature of DLs. DLs typically bring together various subsystems to deliver information access and management facilities for users. There is no agreed definition of what a DL is; as Fox et al. (1995, p.24) note, "The phrase "digital library" evokes a different impression in each reader. To some it simply suggests computerization of traditional libraries. To others, who have studied library science, it calls for carrying out of the functions of libraries in a new way". What matters for the purpose of this paper is that DLs are systems that enable users to retrieve information, and that they can be evaluated in terms of how well they address users' needs.

Just as the term "digital library" is used in different ways by different people, so the term "evaluation" is interpreted in different ways by different communities. Within the IR community, evaluation is most commonly summative – that is, the outcome of an evaluation is summative measures (e.g. of precision and recall) of how "good" a system is. Within the Human–Computer Interaction (HCI) community, evaluation is more commonly formative – that is, the outcome of an evaluation is a description of how users interact with systems that highlights ways in which those systems could be improved. Formative evaluation can consider the "system" at various levels of granularity; as discussed more fully below, we take an inclusive view of evaluation as covering various aspects from details of implementation through to understanding how computer systems support work in context. The work reported here is based on the formative approach. The contrasts between these approaches are discussed below.

2 Background

To set the work on PRET A Rapporter in context, we consider evaluation from three different perspectives: the evaluation tradition within IR; the evaluation tradition within HCI; and approaches that have been taken to evaluating digital libraries. In doing this, we compare the evaluation cultures, to identify strengths and limitations of each, and show how they have influenced the evaluations of DLs.

2.1 An overview of IR evaluation

The classic approach to IR evaluation is the "Cranfield paradigm", within which as many variables as possible (including the database of documents over which retrieval is to be performed and the set of queries) are controlled in order to measure algorithm performance

on criteria such as precision and recall. Tague-Sutcliffe (1992) presents a detailed, and highly cited, methodology for conducting an evaluation study within this paradigm.

Tague-Sutcliffe (1992) highlights three criteria that any evaluation study must satisfy:

- Validity, which she defines (p.467) as "the extent to which the experiment actually determines what the experimenter wishes to determine". She highlights inappropriate measures (e.g. using a Likert scale to measure user satisfaction) and user populations (e.g. student users to represent professionals) as possible causes of low validity.
- Reliability, which she defines (p.467) as "the extent to which the experimental results can be replicated" – typically by another experimenter.
- Efficiency, or "the extent to which an experiment is effective" (p.467) relative to the resources consumed. This is an issue that is explored further by Toms et al. (2004).

She then presents a process for planning and conducting an evaluation that consists of the following decisions:

1. Decide whether or not to test (e.g. check that no-one else has already conducted the proposed study).
2. Determine what kind of test. She outlines four levels of rigour in testing, in which all four, three, two or one of the core components (users, databases, searchers and search constraints) are controlled, and discusses the relative merits of these different levels in terms of generalisability, level of realism, degree of focus and costs.
3. Decide how to operationalise the variables (i.e. determine what to vary and what to measure).

4-6. Select what database to use (step 4), where to source queries from (step 5) and how to process queries (step 6).

7. Decide how to assign treatments to experimental units (i.e. detailed experimental design).

8-9. Consider how to collect data (step 8) and then how to analyse it (step 9)

10. Plan how to present results. Here, she highlights the common experimental reporting structure of presenting the aims of the study, the background to it, the detailed methodology, the results and finally the conclusions.

In later sections, we compare Tague-Sutcliffe's (1992) presentation of the IR evaluation method with the PRET A Rapporter approach.

Tague-Sutcliffe (1992) highlights the challenge that the more emphasis is put on user interaction with a system, the more difficult it becomes to achieve repeatability across studies. Nevertheless, there has been a growing recognition that the user perspective matters in assessing IR systems. As Borlund (2003) notes, users' information needs are individual and change over time, and the relevance of search results will be assessed against the need (rather than against the query, as is done in traditional IR evaluation). Borlund (2003) presents an IIR evaluation approach that considers three main factors: the components of an experimental setting designed to promote experimental validity (such as simulated work tasks); how to apply simulated work tasks within the experimental setting; and a richer set of performance measures (beyond precision and recall). Borlund's concern is to maintain a controlled experimental setting while making it as realistic as possible; an important mechanism for achieving realism is simulated work tasks (scenarios in which experiment participants are asked to imagine they have a particular task from which to derive their own information needs). She proposes two performance measures, relative relevance and ranked half-life, which aggregate over multiple interpretations of 'relevance' and can be used to assess the suitability of different results rankings respectively; the details of these measures are not important in the context of this paper beyond noting that, like the traditional measures, they are summative measures of system performance.

Blomgren et al. (2004) tested Borlund's approach, comparing the findings of her performance measures against subjective user assessments of the value of returned results; while the system performance measures were found to have some value, they did not conform well to participants' subjective assessments. For example, user satisfaction with results did not relate well to system measures of precision or ranked half-life.

Other researchers such as Koenemann and Belkin (1996) also argue for the importance of taking account of the user in measuring IIR performance. They present a comparative evaluation of three variants of an IR system – again, using well defined user tasks and basing the evaluation on quantitative performance measures.

Although most IR evaluation studies are based on classic experimental studies, this is not universally the case. For example, Nicholas et al. (2006) present a study based on transaction log analysis to understand how users work with scholarly resources. Bilal and Bachir (2007) illustrate the complementary use of both qualitative and quantitative data for conducting a richer evaluation of a system (in their case, the International Children's Digital Library). Belkin and Muresan (2004) discuss the relative merits of qualitative and quantitative measures for user-oriented evaluation in TREC (http://trec.nist.gov), noting that quantitative measures make comparisons easier whereas qualitative approaches typically have higher ecological validity.

2.2 An overview of HCI evaluation

Within Human–Computer Interaction, there is a higher preponderance of qualitative approaches to evaluating interactive systems, although quantitative approaches based on classic experiments (drawing on techniques developed in Psychology) are also common. Such experiments share many features in common with the IR approach outlined above (Tague-Sutcliffe, 1992), in that they involve the identification of independent and dependent variables and the control of confounds that might reduce the reliability of the study.

Just as Tague-Sutcliffe (1992) presents a set of three criteria that any study should satisfy (validity, reliability and efficiency), so within HCI a set of criteria have been identified. For example, Hartson et al. (2001) identify a set of comparison criteria including:

- Thoroughness: this is the usability equivalent of IR recall – i.e. the proportion of real problems found (as compared to the real problems that exist in the system).
- Validity: this is the usability equivalent of precision – i.e. the proportion of problems found that are real.
- Effectiveness: the product of thoroughness and validity.
- Reliability: defined in the same way as Tague-Sutcliffe (1992), as being a measure of consistency of results across different evaluators.
- Downstream utility: how well a method supports developers in generating re-designs.
- Cost effectiveness, which appears to be defined in the same way as Tague-Sutcliffe's "efficiency".

Wixon (2003) argues that downstream utility is the most important consideration, while Hertzum and Jacobsen (2001) focus their attention on the importance of reliability. All of these studies share the common view that evaluation is concerned primarily with identifying usability problems; while this is one concern for evaluating DLs, it is certainly not the only one. Other concerns include the IR evaluation measures discussed above and, more broadly, how useful the information retrieved is perceived as being, and how easily users can find and organise information (i.e. integrate IR into their information work). In the case studies presented in this paper, we consider a range of such user-focused evaluation issues.

Hartson et al.'s (2001) definitions (with the exception of downstream utility) are quantitative, being based on counts of real and identified problems. These criteria are less important if one takes a wider view of evaluation as focusing more on downstream utility, and hence as (qualitatively) identifying strengths and limitations of existing systems. With this broader view, validity is still important, but needs to be defined differently: can we have confidence in the

study findings, and on what basis? For example, user reports about their behaviour typically have low validity, so studies of user behaviour will ideally use methods that record behaviour, such as observations and think-aloud. Conversely, user perceptions – e.g. of how access to digital libraries has changed their working practices – are appropriately gathered through self-report techniques such as interviews.

One widely adopted approach to validating data and developing a deeper understanding of user evaluation issues is triangulating across data sources. Mackay and Fayard (1997) discuss this specifically in the context of evaluating interactive systems. Aula (2005) illustrates one approach to triangulation in her study of web use, in which she uses both interviews and observations as complementary data sources to develop a richer understanding of the challenges facing older users of search engines.

While there is an extensive literature on the strengths and limitations of particular evaluation techniques, there has been surprisingly little discussion on the process of planning an evaluation study within HCI. It seems that usability specialists are assumed to know how to design evaluation studies. To fill this gap, Preece et al. (2002) propose the DECIDE framework as a guiding structure for any evaluation; it involves the following stages:

1. Determine the overall goals the evaluation addresses
2. Explore the specific questions to be answered
3. Choose the evaluation paradigm and techniques to answer the question
4. Identify practical issues to address, such as selecting participants
5. Decide how to deal with ethical issues
6. Evaluate, interpret and present data

Stage 2 is a refinement of stage 1; Preece et al. (2002) give as an example the goal of understanding why people prefer paper airline tickets rather than e-tickets: this might be refined into specific questions such as what customers' attitudes are to e-tickets, whether they have adequate access to computer systems, etc.

Steps 3, 4 and 5 are interdependent: in particular, what evaluation paradigms are reasonable options will depend on practical and ethical considerations. The practical issues highlighted by Preece et al. are availability of suitable users, how to design appropriate tasks for testing, availability of equipment (computers for participants, data gathering and analysis systems, etc.), schedule, budget and what expertise is available in applying particular techniques.

Finally, step 6 includes assessing the reliability, validity and generalisability of findings.

As described below, the DECIDE framework was the basis for developing PRET A Rapporter.

2.3 Evaluation studies of digital libraries

In the digital libraries domain, as much as in any other, evaluation schemes are presented as having emerged obviously from the situation and the evaluation questions. Here, we briefly describe a selection of DL evaluation studies that illustrate contrasting evaluation questions and methods.

Many published studies are reports of evaluations of particular systems, involving either user testing or expert evaluation. For example, Hartson et al. (2004) report on an expert evaluation of the Networked Computer Science Technical Reference Library (http://www.ncstrl.org/). In their study, evaluators employed a co-discovery method, described by Hartson et al. as an approach in which two or more usability experts work together to perform a usability inspection of the system. The resulting verbal protocol forms the basis of the usability evaluation, which focused primarily on usability problems with the system interface.

Another approach that does not involve user participation is the use of transaction logs. For example, Mahoui and Cunningham (2000) present a study in which they compared the transaction logs for two collections of computer science technical reports to understand the differences in searching behaviour and relate that to the different designs of the two systems.

Mahoui and Cunningham argue that the value of transaction logs lies in the availability of data about a large number of transactions and users, making it possible to develop an understanding of behaviour with a particular system (though not of particular users).

Some studies use "surrogate users" – subject experts who can better assess features of a DL than the target user population, but who are not usability experts. One example is the work of McCown et al. (2005), who recruited eleven teachers to participate in a study comparing the effectiveness of NSDL and Google in terms of the quality of results returned for curriculum-related search expressions. In this case, the evaluation was not of the quality of the interaction or system design, but of the quality of the results returned in relation to the relevant school curriculum. Numerical ratings were given by the participants to grade the suitability of the results, and a statistical analysis was performed to yield quantitative measures for both information resources (which, in this case, indicated that Google in fact delivered more suitable results than NSDL for the areas of the curriculum investigated).

While some evaluation studies are based on quantitative analysis, many more are qualitative, based on techniques such as think-aloud, interview or observation. Such studies may focus on evaluating the interface or may be concerned with understanding the broader context within which users' information work takes place. For example, Blandford et al. (2001) conducted a study of how Computer Science researchers work with multiple digital libraries using a think-aloud protocol, while Kuhlthau and Tama (2001) used semi-structured interviews to study the information practices and needs of lawyers. Blandford et al. used an observational technique because what mattered was what people actually do rather than what they think they do; in contrast, Kuhlthau and Tama used interviews because the focus of their study was on lawyers' perceptions rather than the details of their information behaviour. In the latter case, general features of existing systems were considered, rather than the details of particular system designs. While this style of research does not fit the traditional IR meaning of 'evaluation', it fits a broader conception of 'evaluation' as assessing how well systems perform in the current work setting and, through this, identifying new design possibilities – delivering on the "downstream utility" evaluation criterion but not delivering on the problem count based criteria discussed above.

These examples illustrate the diversity of approaches it is possible to take when evaluating digital libraries, and the variety of possible evaluation questions. Further examples are presented in our case studies (section 4).

3 PRET A Rapporter

We found The DECIDE framework, described above, a helpful tool for thinking about evaluation studies but, in both applying it ourselves and teaching it on an MSc HCI course, we discovered limitations:

- The separate but interdependent questions of data gathering and analysis are split between steps 3 and 6.
- Data analysis is subsumed within a single stage with reporting, reducing its apparent importance within the planning process.
- Step 3 depends on steps 4 and 5, so seems to be out of order.
- The distinction between steps 1 and 2 is more of degree than being precisely defined: sometimes we have found that questions can be refined repeatedly.
- While the acronym 'DECIDE' has clear appeal in conveying the overall purpose (of deciding what to do and how to do it), the individual letters do not serve as good mnemonics for the steps involved.

The result of these reflections led us to adapt the DECIDE framework, resulting in PRET A Rapporter. The stages for designing an evaluation study using the PRET A Rapporter framework are as follows:

1. **P**urpose of evaluation: what are the goals of the study, or the detailed questions to be answered in the study?

2. **R**esources and Constraints: what resources are available for conducting the study and, conversely, what constraints must the study work within?
3. **E**thics: what ethical considerations need to be addressed?
4. **T**echniques for gathering data must be identified.
5. **A**nalysis techniques must be selected.
6. **R**eporting of findings is (usually) the final step.

PRET A Rapporter is not tailored specifically to the evaluation of digital libraries but, as illustrated in the case studies (section 4), can usefully be applied to evaluations of such systems. In two of the cases presented here, the application of PRET A Rapporter has been retrospective, as the framework had not been developed at the time of the studies; the ease with which it has been possible to do this retrospective fitting is one important check that the framework is well constructed.

We first outline each of the steps in the framework in more detail, then present three case studies described in terms of this framework. Compared to the IR and IIR methodologies presented by Tague-Sutcliffe (1992) and Borlund (2003), PRET A Rapporter takes a broader perspective. This is because it encompasses a wider range of possible approaches to evaluation, and to consider every decision in detail would take a text book rather than a paper; also, such text books already exist (though lacking the planning perspective that PRET A Rapporter provides).

3.1 Purpose of evaluation

In considering the purpose of an evaluation, an initial consideration is likely to be whether this evaluation is formative or summative – that is: whether the evaluation is to inform further design activity, or whether it is to provide a summary of design features (e.g. comparing performance measures for alternative system implementations).

In a research context, a related concern is whether the evaluation focuses on hypothesis testing or developing a deeper understanding of system use in context. For example, McCown et al. (2005) may have started with a hypothesis that NSDL supports the science curriculum better than Google, whereas Kuhlthau and Tama (2001) conducted their study in a largely exploratory way that helped identify requirements for future systems. Many of our digital library studies have been exploratory: aiming to understand a situation better in order to inform the design and deployment of future systems. In these situations, while there are initial themes to focus data gathering, which may be expressed in the form of particular questions (e.g. in an interview), that data gathering will be open to new possibilities around those themes.

In many situations, it is necessary to refine a theme into a set of questions that will provide a more detailed focus for data gathering. For example, in one study (Stelmaszewska and Blandford, 2002), where the overall goal was to better understand how naïve searchers formulated their queries (with a view to providing better support for this activity), one question was how often each user reformulated their query before adopting a different search strategy or giving up.

Perhaps surprisingly, identifying the purpose of a study is an issue that is downplayed by both Tague-Sutcliffe (1992), who subsumes this question within the first decision of whether or not to conduct a study at all, and Borlund (2003), whose work is based on the premise that the purpose of IIR evaluation is to measure system performance (from a user-oriented perspective). When the range of possible evaluation questions about a system is broadened, it is essential to consider purpose explicitly.

3.2 Resources and constraints

Any evaluation study has to work within what is practicable. In commercial contexts, many of the constraints are imposed by contractual considerations, such as the budget available, the timescale within which findings must be reported, and the form of report required. In research

settings (including all the case studies reported in section 4), these particular resources are of less immediate concern than others.

One consideration is what system representations are available. For many of the evaluations we have conducted, the only available representation has been the running system; in some cases, we have had access to developers, documentation or source code which facilitates better understanding of (for instance) how IR algorithms have been implemented, or the designers' reasons for providing particular features. Within Tague-Sutcliffe's framework, this consideration corresponds approximately to decisions 4 and 6 (what database to use and how to process queries).

Another central set of questions are where suitable participants can be recruited from, what tasks (if any) they should be given, and what environment they can be studied in. For evaluating digital libraries, it is usually important to work with participants who represent the intended user population; for example, it would rarely be appropriate to recruit undergraduates in computer science or information studies if the purpose of the study is to evaluate the effectiveness of a specialist medical or law library for supporting practitioners' work. The work of McCown et al. (2005) is a counter-example, in that it involved teachers rather than the students who are the target users for the systems. This corresponds approximately to Tague-Sutcliffe's fifth decision (where to get queries).

Another issue is what facilities are available or appropriate for conducting studies and gathering and analysing data. In some of our studies, we have made use of a usability laboratory, with automatic key-press logging, screen capture and audio and video recording. In others, where higher ecological validity is needed, it has been necessary to visit participants in their work places, which limits what kinds of data gathering are possible. This issue is not explicitly discussed by Tague-Sutcliffe (1992).

Others who have reported on digital library evaluations (e.g. Hsieh Yee, 1993) have given their participants tasks tailored to the purpose of the study, typically including some tasks that have a single, well defined answer and others that require more sophisticated information seeking. Where it has been necessary to give participants tasks, we have favoured less well specified tasks that have higher ecological validity; for example, when our participants have been postgraduate students, we have often asked them to articulate an information need relevant to their current research project and to search for information to address that need. In some of our studies, where the focus has been on understanding information seeking in the context of work, the tasks are completely defined by the ongoing work of participants. Tague-Sutcliffe (1992) implicitly assumes the task of IR, whereas Borlund (2003) extends the notion of task to embed the IR within a scenario of use. A scenario attempts to artificially re-create an activity context. Implicit in this is that the scenario should specify those aspects of a context which will result in ecologically valid task performance. Assumptions need to be made about what those factors are, but it is never possible to guarantee that these assumptions are correct.

A third set of questions relate to expertise and support for data analysis. Often, the central question is what expertise is available in applying different techniques. For example, it would be unwise to plan a sophisticated quantitative study if there were insufficient expertise in experimental design and statistical analysis available. If new techniques need to be learnt in order to conduct an evaluation, it is necessary to consider what resources there are to learn those techniques. This issue is discussed by Tague-Sutcliffe (1992), but phrased the other way around: that it will be necessary to master some difficult quantitative analysis techniques in order to be able to complete any IR evaluation study.

In particular settings, there may be other constraints to consider; for example, in a study of user of DLs by clinicians (Adams, Blandford and Lunt, 2005), timing and location were important: one interview with a surgeon was held in the surgery ante-room between operations, and focus groups for nurses and allied health professionals (e.g. nutritionists, physiotherapists) were held at the end of meetings because they found it very difficult to find time to participate in the research individually.

Many of the explicit and implicit issues identified within Tague-Sutcliffe's (1992) methodology fall into this category of resources and constraints.

3.3 Ethics

In evaluation studies, it is important to consider ethical dimensions. Most professional bodies (e.g. ACM, 1999) publish codes of practice. Less formally, we have identified three important elements of ethical consideration:

- **V**ulnerable participants (young, old, etc.)
- **I**nformed consent
- **P**rivacy, confidentiality and maintaining trust

Few digital library studies involve studies with participants that might be regarded as vulnerable; counter-examples include the work of Theng et al. (2001), Druin (2005) and Bilal and Bachir (2007) on designing digital libraries with and for children, and Aula's (2005) work studying the information seeking of older users. Our studies of digital library use in clinical settings (Adams and Blandford, 2002; Adams, Blandford and Lunt, 2005) included observations of clinical encounters between doctors and patients, for whom privacy and confidentiality were paramount concerns. In this case, anonymisation of data, for both individuals and institutions as a whole, was imperative – both for protecting privacy and also to maintain individuals' and organisations' trust in the research procedure.

It is now recognised as good practice to ensure all participants in any study are informed of the purpose of the study (e.g. that it is the system that is being assessed and not them!) and of what will be done with the data. Also, participation should be voluntary, with no sense of coercion (e.g. by the exercise of a power relationship between evaluator and participants).

Data should normally be made as anonymous as possible, and individuals' privacy and confidentiality need to be respected. While immediate respect of individuals is reasonably obvious, less obvious is the need to continue to respect participants' privacy in future presentations of the work and to show similar respect to groups and organisations. Lipson (1997) discusses many of the less obvious pitfalls of publishing the findings of studies, such as participants feeling betrayed or embarrassed by descriptions of their behaviour or attitudes.

Mackay (1995) discusses 'best ethical practice' for researchers with personal participant data. She notes that professional ethics should ensure that multimedia data is used within acceptable boundaries. She also proposes that individuals' identities should be hidden, wherever possible, during recording. Adams (1999) highlights the importance of participants' awareness of who is seeing their information, in what context, and how they will use it.

Since traditional IR and IIR evaluation typically focus on quantitative data and focus more on system than user performance, many of the ethical considerations that are important in more contextualised studies are of minimal concern in traditional IR studies, so ethics has not featured as an explicit consideration in IR evaluation methodologies.

3.4 Techniques for data capture

Ethical considerations cover all aspects of a study, including data collection, analysis and reporting. While these steps may be interleaved (particularly in large studies), we consider them in order. Techniques for data collection cannot be addressed completely independently of intended analysis techniques; nevertheless, the purpose of the evaluation will inform what data collection techniques are likely to be appropriate.

Evaluation questions that involve counts of events or a test relating independent and dependent variables will clearly demand that appropriate quantitative data be gathered. Examples of such data within an IR context are presented in detail by Tague-Sutcliffe (1992) and Borlund (2003). Within studies that focus more on user behaviour, numerical data may include numbers of particular event types or user action types, or times to perform tasks.

Such data is most commonly captured using some form of computer logging (e.g. Nicholas et al., 2006).

In evaluating users' experiences of working with digital libraries, we have focused much more on qualitative data. Nevertheless, we have recruited a variety of data collection techniques including naturalistic observations, think-aloud protocols, in-depth interviews, access to server logs, and focus groups, as appropriate to the particular questions being addressed in each study. Such techniques are exemplified in the case studies presented in section 4; for detailed descriptions of such qualitative data collection approaches, see texts such as that by Kuniavsky (2003).

3.5 Analysing data

Quantitative data is typically analysed using statistical techniques (or simpler reports of numbers). Tague-Sutcliffe (1992) and Borlund (2003) present selected quantitative evaluation techniques in detail. For quantitative usability evaluation, standard psychology statistics texts such as Pagano (2001) are a useful resource.

Qualitative data analysis may take many forms, as described by Miles and Huberman (1994). For evaluation purposes, the analysis may vary from focused (e.g. what errors do users make with this system and how might they be avoided?) to exploratory (e.g. how do users work with information in this work setting, and how might their work be improved through system design?). The main data analysis technique we have employed is Grounded Theory (Glaser and Strauss, 1967). This is a social-science approach to theory building that can incorporate both qualitative (e.g. interviews, focus groups, observations, ethnographic studies) and quantitative (e.g. questionnaires, logs, experimental) data sets. The methodology combines systematic levels of abstraction into a framework about a phenomenon, which is iteratively verified and expanded throughout the study. As described below (section 4.3), we have also adapted the approach to relate it to relevant theoretical perspectives.

3.6 Reporting findings

The final step is reporting findings. Since our evaluations have all been elements of research projects, the main means of reporting has been through academic publications. The common practice is, as described by Tague-Sutcliffe (1992), to present the aims (or purpose) of a study; the background; the method or methods applied; the results; and conclusions (usually including a discussion of implications for design). For quantitative experiments, it is customary to provide information at the level of detail that would enable other to replicate the study to establish its reliability. For qualitative studies, it is rarely possible to replicate the conditions of a study closely enough to expect this degree of reliability, so the focus is usually more on presenting the method, analysis and findings in sufficient detail to enable the reader to assess their validity. It is not yet common practice to present all the constraints under which the study was conducted or to present the ethical procedures followed in detail, although this might change in the near future with a growing focus on ethical practices.

In interacting with developers, such as the developers of the Greenstone Digital Library system (Witten et al., 2001) and policy makers, less formal reporting channels have been appropriate. These have included executive summaries that focus on problems found and possible design solutions proposed.

4 Case studies

The case studies we present address a range of evaluation questions relating to the usability and user experience of working with digital libraries, from the local (evaluating a single system or understanding the user's experience of one technology) through the interactive (understanding use of multiple digital libraries) to the larger scale questions of how digital libraries are perceived and used in the context of professionals' every day work. We present the studies as a series moving from the local to the highly contextualised (this ordering is not chronological):

- A formative evaluation of a single system

- A comparative evaluation of two interfaces to the same underlying databases
- A study of system use in context

In each case study, we focus on the elements of particular interest, rather than aiming to provide complete accounts.

4.1 Formative evaluation of a single system

Our first example is an evaluation of a single system, Garnet, that integrates a digital library (based on the Greenstone architecture) with a spatial hypertext system, so that users can not just gather documents but also organise them in a way that is personally meaningful. Spatial hypertexts (Shipman et al., 1995) provide an opportunity to create both formal organisations such as hierarchical and linear structures, and also informal organisations such as piles and offset positions.

Prior to the development of Garnet, the organising tools of spatial hypertexts had not been integrated with information finding resources such as digital libraries. In offices and libraries, users had been seen to flow from information organisation to seeking and back again in a fluid and unpredictable pattern, in response to the found material. The new integrated tool provided an opportunity to establish whether this same fluidity could be observed in an electronic environment, and to identify whether implicit information in the organisation of documents could be used to glean data on the user's structuring decisions or their information goals.

4.1.1 Purpose

Existing research in spatial hypertexts (Shipman, 2001) indicated that the integration of digital library and spatial hypertext may not be straightforward. The interesting evaluation question for Garnet was whether such a system was comprehensible to and of potential benefit to authors of research papers. This rather abstract question was refined into three related questions (Buchanan et al., 2004a):

- Can spatial hypertext provide an effective interface to a digital library?
- Will users demonstrate the interleaving of information seeking and structuring reported in physical environments?
- Can the information structuring performed by the user in the spatial hypertext be used to support information seeking? Particularly, can the theme of a document group be determined from the text of its members?

These questions are still quite general – for example, the word "effective" is open to multiple interpretations. Also, the third question relates more to technical issues than user perceptions or behaviour. Clearly, these questions bear little relationship to traditional IR goals of precision and recall, and the study goals are not focussed upon query performance. However, we discerned a relationship between the user's information structuring and the detection of topical themes and the established IR tool of document clustering. Within that limited scope, therefore, we could adapt and apply existing measures and evaluation approaches.

4.1.2 Resources and Constraints

In this case, since this was a 'home grown' system we had full access to the system and a deep understanding of its implementation.

We also had access to all the Greenstone collections, so could build a collection that was suitable for use by the most conveniently available participants – i.e. local students recruited from our MSc course in HCI with Ergonomics. We also had access to a usability lab that supported screen capture and audio recording. Given that we were interested in evaluating the system, rather than users' more general information seeking behaviour, it was most appropriate to make use of the usability lab facilities and give participants reasonably well defined tasks to perform.

In an ideal world, it might have been possible to give users system access over a long period of time (e.g. several weeks), but given the immature state of development this was considered impractical: it was unlikely that participants would actually choose to use the system in its present form over an extended period.

4.1.3 Ethics

There were relatively few traditional ethical issues to consider during this study; informed consent was obtained from participants and although the participants were students, there was no power relationship between participants and the main investigator.

4.1.4 Techniques for data collection

The three separate questions for investigation required different sets of data to be collected. In the case of usability problems when the spatial hypertext interface was used to access DL features, the users' use of the system was recorded to video tape, and attitudinal questions asked at the end of the study, followed by a semi-structured interview to elicit specific problems. The flow of activity between information seeking and information structuring was also captured through the video recording of the computer screen during the experiment, and specific questions were asked during post-experimental interviews to better understand the decision-making processes that led to particular courses of action. Finally, the system continually recorded extensive logs of the user's activity. This log included details of the user's grouping of documents, changes in these, and the searches performed by the user. After the interactive session, representative text profiles were created and stored for each group of documents created during the user's session.

4.1.5 Analysis

The experimental data was analysed using a variety of techniques from human-computer interaction and information retrieval.

A qualitative analysis was applied to the participants' responses during the post-experimental interview, and this was compared with the data gathered through pre- and post-experimental questionnaires to ensure that a balanced view was achieved. Trails of the participants' sequences of actions were tabulated and compared against each other, the screen capture videos and the computed IR scores for groups of documents created by the users in the spatial hypertext workspace. This data was used to triangulate over the behaviour reported by the users during the interviews and the behaviour observed during the experiment.

As noted above, the focus of the study was not the act of retrieval itself – but rather the organisation subsequent to retrieval. One particular focus was the textual consistency of the document groups created by the user. To evaluate these, we took the document set selected by each user, and applied two established clustering algorithms: Grouper (Zamir et al., 1997) and Scatter/Gather (Cutting et al., 1992). In contrast, we computed the scores for the same algorithms applied to the user's chosen organisation. We further identified the key word terms for each group using the Grouper algorithm, which is used in Grouper to measure the match between any two document groups of one or more documents each, and also serves as a descriptor for a group in the Grouper interface.

During the post-experimental interviews, participants' intentions in organisation were elicited, and these were compared with the characteristic keyphrases selected by the clustering algorithms applied to the same groups. This issue was further probed by the participants' use of an advanced feature that matched documents in a search result set against the user's own organisation. The participants' responses to this feature were measured by a Likert scale in the questionnaire, and also correlated with the IR scores for the document groups involved.

The analysis revealed that human organisation of documents was similar to that created by clustering algorithms, with the exception that humans created specific 'miscellaneous' groups or tolerated singleton documents that the clustering algorithms avoided. This meant that the human-created document groups were more topically focussed and had higher individual scores for textual consistency, but the overall organisation of humans scored slightly lower

when the entire set of groups was computed, as singleton documents and specific "miscellaneous" piles are seen as undesirable in clustering.

4.1.6 Reporting

The results have been reported in conference papers Buchanan et al. (2004a; 2004b) using a standard experimental structure of reporting based on aims, methods, results and conclusions supplemented by detailed description of the system design so that the evaluation is meaningful to the reader. Since the system is 'home grown', it has not been necessary to formally report back to the developers.

4.2 Comparing two versions of a system

Moving from the evaluation of a single system, our second case study compares two versions of a system. This case study is drawn from an ongoing investigation of Humanities use of digital libraries, and has not been completed yet. In this project, we are investigating interfaces to digital resources and how their design can support these users' broader work activities. When planning a study, we seem to face a choice between focusing on specific interface details or on the context in which users use these systems, and then retrospectively mapping between the two. A system-focused study may provide incremental refinements of the given designs, but is unlikely to lead to significant redesigns which better support users' work. Conversely, a work context-focused study may increase our understanding of the context of use, but is disconnected from the details of interface design. To address this dilemma, we are planning a system-focused study in which two systems are contrasted. The aim is to encourage users to think more generally about the interaction possibilities and relate these to their own experiences. We hope this will provide us with more than incremental design improvements.

4.2.1 Purpose

The purpose of the study is to allow participants, specifically Humanities scholars, to contrast two different interfaces to the same digital collection. This will facilitate more general reflection about design possibilities and we will encourage them to relate this to their own research activities. Another important motivation for the study is to provide design feedback to ProQuest, our commercial research collaborator, to 'pay our way' for access to content they have provided. Because the study is part of a larger effort to understand Humanities scholars' use of digital libraries, we want a naturalistic investigation of scholars which reflects their work practices as far as possible.

4.2.2 Resources and Constraints

We have been granted access to ProQuest's "Eighteenth Century Fiction", a collection of ninety-six English prose texts from 1700-1780, and "English Poetry", a collection of over four thousand works of English poetry from the 15th to 19th centuries. These collections are available by subscription as part of a much larger collection via the "Literature Online" (LION) website (http://lion.chadwyck.co.uk). An alternative interface to the collections has been developed using the Greenstone digital library, which the study will contrast with the original (see Figure 1). Having developed it ourselves, we have full access to the Greenstone implementation, but only online access to LION. Both collections can be accessed from any web browser, so observations could feasibly take place in participants' actual working environment.

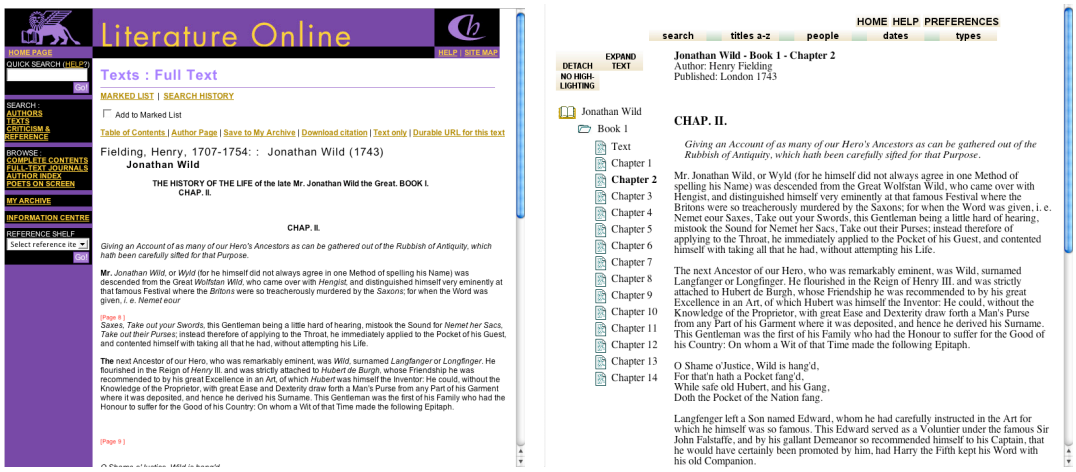

Figure 1. LION (left) and Greenstone (right) interfaces to the ECF collection.

Humanities scholars are a highly distinct and relatively uncommon user group, so we anticipate that finding participants will be challenging. A significant resource is the location of the main research team in London, where it has access to a large number of Humanities research institutions. However, by restricting ourselves to specific collections we have placed even heavier constraints on potential participants as we require them to already use LION or similar systems for accessing English literature material.

Interviews have already been conducted with Humanities scholars from a variety of institutions in the UK, as well as from Thailand, New Zealand and Australia. The interviews have been transcribed and thematic discourse analysis applied so that commonalities and exceptions could be drawn out. This has provided insights into the information seeking habits of Humanities scholars and details of the resources (both physical and digital) that they use, as well as their experiences of, and attitudes towards, them. Typical tasks have been identified to form the backbone of our task list for the test of the two system interfaces we are using.

4.2.3 Ethics

There are no particular ethical considerations in conducting this usability study other than ensuring that both systems are treated fairly and equally. Care may be needed in reporting findings in a way that is acceptable to all stakeholders.

4.2.4 Techniques for data collection

The study will take place in participants' working environments in order to encourage recruitment: busy scholars may be less keen to take the time to come to a lab. This setting may give us greater insight into their normal working practices, although it should be noted that the tasks themselves are artificial. The setting also constrains the type of data collected, e.g. screen capture would be inappropriate. A think aloud protocol can be audio recorded, along with a video recording in order to help later interpretation (but not to record low level interaction). After piloting with an English postgraduate researcher, it is proposed to run 20 participants in one hour test sessions.

4.2.5 Analysis

After the user test sessions are complete, the think aloud protocols of the participants' interactions will be transcribed and coded for breakdowns in the users' task-flow and times of confusion or error. Two experimenters will independently code the transcriptions to improve inter-observer reliability. In addition, clips of incidents will be edited from the screen capture videos for deeper analysis.

4.2.6 Reporting

As well as being written up in academic form, the results of this study will be reported back to the ProQuest and Greenstone developers through an executive summary highlighting strengths and weaknesses of the existing interfaces, and possible design changes to improve system usability.

4.3 Understanding professionals' use of digital libraries to support their work

The final study we present exploits more exploratory qualitative approaches to understand professional information users' use of digital information resources that were available in the context of their everyday research and writing tasks, with a view to exploring how systems might be designed to better support that work. Here, we focus on the work of journalists.

4.3.1 Purpose of evaluation

The study was motivated by the idea that information retrieval research has often treated finding information as an endpoint, to the exclusion of how it might then be used (Wilson, 1999; Kuhlthau and Tama 2001). However, seeking, manipulating and using information are not unrelated but (ideally) form a flow of interdependent activity, each modifying and conditioning the other. By reducing the separation between tools, opportunities may arise for better enabling continuity.

To explore this, the study looked at the information seeking and use of newspaper journalists. It aimed to discover and explain journalists' information behaviours in relation to electronic news cuttings services and, in this light, to evolve possible significant redesigns.

4.3.2 Resources and Constraints

Organisations are busy places and hurdles must be jumped to gain the confidence and 'buy in' of the host. They may want to help, but they also want to know that your activity will not upset their employees or their work. Personal contact is essential for navigating resource and constraint issues. For the study, contact was delegated to a 'Systems Editor'. Within the newspaper, his role was to ensure smooth running of newsroom systems and to explore new opportunities. Consequently, he had excellent knowledge of the work and personalities around the company and he made a knowledgeable and sympathetic guide to the culture of the organization.

Newspaper newsrooms have daily cycles and we learned to organize data gathering around this. Before mid-morning editorial meetings, many journalists had not been assigned tasks and the newsroom was quiet. This created a good opportunity for interviewing. Later, it was usually possible to spot people furiously working to deadlines by their posture and general demeanour, and we learned to avoid them. Also, journalists can get called out with little notice and we needed to be flexible. We knew that a 3pm interview with a home news correspondent was cancelled when we spotted him on the television standing outside a court room at 2.50pm!

4.3.3 Ethics

Whilst the participants could not be described as 'vulnerable', there were sensitivities to be observed. The aims of the study and commitment required were clearly explained and permissions obtained from participants and line managers. This not only reassured people that data gathering did not present a threat, but helped participants interpret the questions asked.

Publishing results from an organizational study requires sensitivity too. It may be difficult for a researcher to predict how material may play into the hands or a competitor, for example. Part of the agreement for this study was to allow the organization to review resulting papers and bring potential problems to our attention. Some issues were identified and these could be addressed without affecting the conclusions we were able to draw.

4.3.4 Techniques for data collection

Data were gathered primarily through semi-structured interviews. Twenty-five journalists were interviewed. Interviews lasted 20 to 40 minutes and were conducted at participants' desks. This allowed them to explain what they did, and also demonstrate key artefacts and show how they might be used. Interviews began with the researcher prompting for a description of the assignment process with reference to recent work. The researcher then steered the interview towards information seeking and use issues.

After 14 interviews, the recordings were transcribed and analysed. This allowed specific questions to be formed that could become the focus for further interviews. A questionnaire was produced to act as an interview script to ensure coverage of specifics. However, the script became a barrier to data collection: it was disconcerting for participants, so the interviews returned to a more conversational style. Issue coverage depended on the researcher exploiting and directing the flow of conversation.

4.3.5 Analysis

Grounded Theory was used for the analysis, but adapted slightly to incorporate concepts from Rasmussen, Pejtersen and Goodstein's (1994) Cognitive Systems Engineering (CSE).

CSE aims to model socio-technical work systems in order to predict how people would behave in response to engineered changes—to ask, 'what could be done differently and better?' CSE recommends understanding behaviour in relation to the active constraints and available resources in a situation (not to be confused with constraints and resources within the PRET A Rapporter framework): that the analyst should build a picture of a person's 'action alternatives' and that this is bounded (and consequently determined) by a set of constraints and resources.

Using this perspective to guide the analysis may be seen to compromise the openness advocated for Grounded Theory. Rasmussen et al. argue that a well-defined point of view supports rapid convergence in the analysis. The question of using a framework often involves a trade-off between openness and efficiency, and this needs to be informed by a judgement about the value of the framework. In this case, it was considered a valuable filter through which to analyse the data.

The study motivated a number of requirements, including the need to have source material readily available during writing, even if its use had not been anticipated. A tool supporting this was subsequently built and evaluated with journalist users with very positive results.

4.3.6 Reporting

The work has been reported in both academic papers (Attfield and Dowell, 2003) and a confidential report back to the organisation. In particular, to 'pay our way' for access to the study context, we performed and reported on a comparative usability evaluation of two digital libraries that were used within the organization.

5 Conclusion

By illustrating the application of the PRET A Rapporter framework for planning and conducting evaluation studies, we have conducted one kind of test of it as a framework – i.e. it is helpful for planning (or even describing) studies, and has been used effectively by the several authors of this paper. (An earlier version of this paper included six case studies; three were removed for space reasons.) This is, of course, a weak form of validation of an approach, and further work needs to be done on testing its utility with a wider spectrum of users.

Compared to the evaluation methodologies presented by Tague-Sutcliffe (1992) and Borlund (2003), PRET A Rapporter is presented at an abstract level. This is because it is intended for more generic application: the concern is not only with the evaluation of IR or IIR systems, but with evaluation questions pertaining to the use of systems in the broader work setting.

In the presentation of the framework, we have noted how the various decisions discussed by Tague-Sutcliffe (1992) fit within it: apart from the explicit consideration of ethics, similar

criteria are considered, albeit at different levels of detail and with different emphases. For example, there is an emphasis in PRET A Rapporter on resources and constraints because such practical considerations often limit what is possible in naturalistic evaluation studies, whereas the emphasis in Tague-Sutcliffe's methodology is on decisions.

This shift in emphasis reflects a focus that is less concerned with how IR algorithms perform and more with understanding how users experience and work with systems that include IR at their core. The three case studies we have presented illustrate a range of possible evaluation questions pertaining to the detailed design of an interface, the design of the interface within a broader work setting and the fully contextualised use of systems to support information work. With these broader questions, new strategies (typically involving triangulation) have to be devised for assessing study validity. Reliability becomes less important than it is for more focused IR studies; conversely, downstream utility becomes more important.

PRET A Rapporter has been based on several years' experience of planning and conducting evaluations such as those described in this paper, and also of teaching evaluation methods. As noted in the Call for Papers for this Special Issue, user-centred evaluations are essential for understanding the effectiveness of IR systems and user–search interactions, and hence for designing systems that are usable, useful and used. By presenting the PRET A Rapporter framework and selected case studies, we have illustrated one approach to planning such evaluations and highlighted many of the issues that need to be considered if evaluations are to be ethically conducted, valid and informative.

Acknowledgements

This research has been supported by EPSRC grants GR/S84798 and EP/D056268, and by ESRC grant RES-335-25-0032. We are grateful to ProQuest, the Greenstone developers and all other organisations and individuals who have supported our work, including all study participants and anonymous referees of an earlier version of this paper.

References

ACM (1999). Software Engineering Code of Ethics and Professional Practice. http://www.acm.org/serving/se/code.htm (accessed 31/3/06)

Adams, A. (1999). Users' perceptions of privacy in multimedia communication. Proceedings of CHI' 99, Pittsburgh. ACM Press. (pp 53-54).

Adams, A. & Blandford, A. (2002). Acceptability of medical digital libraries. Health informatics Journal. Sheffield Academic Press. Vol 8(2). 58 – 66.

Adams, A., Blandford, A & Lunt, P. (2005). Social empowerment and exclusion: a case study on digital libraries. ACM Transactions on Computer–Human Interaction. ACM Press. 12(2). 174-200.

Attfield S., & Dowell J. (2003). Information seeking and use by newspaper journalists. Journal of Documentation. 59(2). 187-204.

Aula, A. (2005) User study on older adults' use of the Web and search engines. Universal Access in the Information Society, 4. 67-81.

Bilal, D. and Bachir, I. (2007) Children's interaction with cross-cultural and multilingual digital libraries: I. Understanding interface design representations, Information Processing & Management. 43(1). 47-64.

Belkin, N. & Muresan, G. (2004). Measuring web search effectiveness: Rutgers at Interactive TREC. In Proc. Workshop on Measuring web search effectiveness: the user perspective. At WWW 2004 Conference. Available from http://www.scils.rutgers.edu/~muresan/Publications/wwwWsBelkin2004.pdf (accessed 10th August 2006)

Blandford, A., Stelmaszewska, H. & Bryan-Kinns, N. (2001) Use of multiple digital libraries: a case study. Proc. JCDL 2001. 179-188. ACM Press.

Blomgren, L., Vallo, H. and Byström, K. (2004) Evaluation of an information system in an information seeking process. In R. Heery & L. Lyon (Eds.) Proc. ECDL 2004. LNCS 3232. 57-68.

Borlund, P. (2003). The IIR Evaluation Model: a Framework for Evaluation of Interactive Information Retrieval Systems. Information Research, vol. 8, no. 3, paper no. 152.

Buchanan, G., Blandford, A., Jones, M. & Thimbleby, H. (2004a). Integrating information seeking and structuring: exploring the role of spatial hypertexts in a Digital Library. Proc. ACM Hypertext & Hypermedia 2004. 225-234.

Buchanan, G., Blandford, A., Thimbleby, H. & Jones, M. (2004b) Supporting Information Structuring in a Digital Library. In R. Heery & L. Lyon (Eds.) Proc. ECDL 2004. LNCS 3232. 464-475.

Cutting D., Karger D., Pedersen J., and Tukey, J. W. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, Proceedings of the 15th Annual International ACM/SIGIR Conference, Copenhagen. pp. 318-329.

Druin, A. (2005). What children can teach us: Developing digital libraries for children. Library Quarterly , 75(1), 20-41.

Fox, E. A., Akscyn, R. M., Furuta, R. K., and Leggett, J. J. (1995) Digital libraries. Commun. ACM 38, 4 (Apr. 1995), 22-28.

Glaser, B. & Strauss, A. (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago, Aldine

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Evaluating usability evaluation methods. International Journal of Human-Computer Interaction, 13 (4), 373-410.

Hartson, H. R., Shivakumar, P. & Pérez-Quiñones, M. A. (2004). Usability inspection of digital libraries: a case study. International Journal of Digital Libraries. 4.2. 108-123.

Hertzum, M., and Jacobsen, N.E. (2001). The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. International Journal of Human-Computer Interaction, 13(4), 421-443.

Hsieh-Yee, I. (1993). Effects of Search Experience and Subject Knowledge on the Search Tactics of Novice and Experienced Searchers. Journal of the American Society for Information Science. 44.3. 161-174.

Koenemann, J. & Belkin, N.J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. Proc. CHI'96. Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 205-212.

Kuhlthau, C.C. & Tama, S.L. (2001). Information Search Process of Lawyers: A Call for 'Just for me' information services. Journal of Documentation, 57(1), pp. 25-43.

Kuniavsky, M. (2003). Observing the user experience: a practitioner's guide to user research, Morgan Kaufmann, San Francisco.

Lipson, J. G. (1997). The politics of publishing: protecting participants' confidentiality. In J. Morse (Ed.) Completing a qualitative project: details and dialogue. Sage Publications.

Mackay, W.E. (1995). Ethics, lies and videotape... . Proceedings of the ACM conference on Human Factors in Computing Systems (CHI '95), ACM Press, pp.138-145.

Mackay, W. E. & Fayard, A.-L. (1997), HCI, Natural Science and Design: A Framework for Triangulation Across Disciplines. Proc. ACM DIS'97. pp. 223-234.

Mahoui, M. & Cunningham, S. J. (2000). A comparative transaction log analysis of two computing collections. In proceedings of ECDL'00. Heidelberg: Springer. 418-423.

McCown, F., Bollen, J. & Nelson, M.L. (2005). Evaluation of the NSDL and Google for Obtaining Pedagogical Resources. Proc. ECDL 2005. Lecture Notes in Computer Science, Volume 3652, Pages 344 - 355

Miles, M. B. and Huberman, A. M. (1994) Qualitative Data Analysis: An Expanded Sourcebook. Sage Publishers.

Nicholas, D., Huntington, P., Jamali, H. R. and Watkinson, A. (2006) The information seeking behaviour of the users of digital scholarly journals, Information Processing & Management, Volume 42, Issue 5, Pages 1345-1365.

Pagano, R. (2001) Understanding Statistics in the Behavioral Sciences, 6th edn, Wadsworth.

Preece, J. Rogers, Y. & Sharp, H. (2002). Interaction Design. Wiley, New York

Rasmussen, J. Pejtersen, A.M. & Goodstein, L.P. (1994). Cognitive Systems Engineering. New York, Wiley.

Shipman, F., Marshall, C., and Moran T. (1995). Finding and Using Implicit Structure in Human-Organized Spatial Layouts of Information. Proceedings of Human Factors in Computing Systems (ACM CHI), ACM Press, pp. 346-353.

Shipman, F. (2001). Seven Directions for Spatial Hypertext Research. First International Workshop on Spatial Hypertext, ACM Hypertext Conference 2001, Aarhus, Denmark. Online at: http://www.csdl.tamu.edu/~shipman/SpatialHypertext/SH1/shipman.pdf (accessed 2/4/6)

Stelmaszewska, H. & Blandford, A. (2002). Patterns of interactions: user behaviour in response to search results. In A. Blandford & G. Buchanan (eds.) Proc. Workshop on Usability of Digital Libraries at JCDL'02. 29-32 Available from http://www.uclic.ucl.ac.uk/annb/DLUsability/JCDL02.html

Tague-Sutcliffe, J. (1992) The pragmatics of Information Retrieval Experimentation, Revisited. Information Processing and Management. Vol 28 no 4, pp 467-490.

Theng, Y. L., Mohd-Nasir, N., Buchanan, G., Fields, B., Thimbleby, H. and Cassidy, N. (2001). Dynamic Digital Libraries for Children. First ACM and IEEE Joint Conference in Digital Libraries, Roanoke (Virginia), pp. 406 - 415.

Toms, E., Freund, L. & Li C. (2004) WiIRE: the Web interactive information retrieval experimentation system prototype. Information Processing & Management, 40(4). 655-675.

Wilson, T. (1999). Models of information behaviour research. Journal of Documentation, 55(3), pp. 249-270.

Witten, I. H., Bainbridge, D. & Boddie, S. J (2001.) Greenstone: Open-source digital library software with end-user collection building. Online Information Review, 25 (5), 288-298.

Wixon, D. (2003). Evaluating Usability Methods; Why the Current Literature Fails the Practitioner. Interactions July and August 2003

Zamir, O., Etzioni, O., Mandani, O. and Karp, R. M. (1997). Fast and Intuitive Clustering of Web Documents. Third International Conference on Knowledge Discovery and Data Mining, August 14-17, Newport Beach, California,. AAAI Press, Menlo Park, California. pp. 287-290.