Finding Problems: When Digital Library Users Act as Usability Evaluators

Michael Khoo, Diana Kusunoki, and Craig MacDonald
The iSchool at Drexel University
3141 Chestnut Street, PA 19104 U.S.A
Email: {khoo, dsk38, cmm353}@drexel.edu

Abstract
Users in digital library usability evaluation typically participate as subjects in studies designed and conducted by usability experts and digital library researchers. What happens however when users take the role of the researchers, and with some basic HCI training, design and conduct their own evaluation of a digital library? For several years, teams of students in master’s level HCI classes at Drexel University were given the assignment of designing and carrying out heuristic evaluations of the interface of the Internet Public Library. Their final evaluation reports regularly focused on what, to a usability expert, would not be considered interface issues, such as problems with finding resources in the library. These outcomes contrasted with those of a parallel evaluation of the IPL carried out by doctoral students with a background in HCI, which found interface issues to be the main concerns. A post hoc comparison and analysis of these evaluations highlights differences between users’ and evaluators’ perceptions of usability, and has implications for the design of digital library evaluation and the roles of users and evaluators in such evaluation.

1. Introduction
Usability evaluation is an important component of digital library development [3, 4]. Data for usability evaluation is collected through a number of methods. Some of these methods involve users participating as subjects in studies designed and conducted by usability experts and digital library researchers. Four underlying assumptions of these studies are: (a) that developers are not typical digital library users (for instance, they are already experts in the design and use of the interface in question) and therefore need to consult users in order to identify requirements for their designs; (b) that users are in possession of useful opinions and knowledge about the usability of a digital library interface that are unknown to developers; (c) that these opinions and knowledge can be elicited from them through the medium and use of properly structured evaluation instruments; and (d) that such methods in general involve a ‘translation’ of users’ needs and knowledge (as revealed in the study) into useful knowledge for design requirements.

Usability methods are usually based on an underlying theory (cognitive, social, etc.) that explains how humans react to and interact with an interface, and the technical, social, and other factors that shape such interaction. It also assumes that expertise, gained in professional training and HCI work, is required to analyze the feedback from digital library users. This is not however the only possible configuration for digital library usability work. What happens for instance when digital library users take the role of the researchers, and with only limited HCI training, design and conduct their own evaluation of a digital library? How is the usability of a digital library measured, when users’ responses are not being shaped by instruments designed by usability experts? How does this affect the types of questions asked, and the data and results obtained? What insights might such evaluations tell us about a digital library, and its usability?

The following study reports a retrospective analysis of two sets of evaluation reports produced by two different groups of students at the iSchool at Drexel. Both groups evaluated the same interface, that of an online educational digital library, the Internet Public Library (‘IPL’: http://www.ipl.org/), which is hosted by the iSchool. The first set of evaluation reports were produced by three doctoral students with a background in HCI, who were members of the IPL project team. These students found a range of interface issues, such as the formatting and labeling of buttons, to be the main concerns. The second group consisted of teams of students in a master’s level HCI classes over several years, who were given the in-class assignment of designing and carrying out a heuristic evaluation of the IPL. The final evaluation reports of these masters’ students regularly focused on what, to a usability expert, would not be considered interface issues, such as the content (rather than the formatting and presentation) of the search results.

Below, it will be suggested that these differences arose because the master’s students saw digital library evaluation and usability in different terms than the
more experienced doctoral students. The latter were less familiar with the evaluation terminologies and methods used, and they had different interpretations and definitions of what usability meant, which led in turn to different understandings of evaluation. This raises the following general research questions:

- How does usability evaluation conducted by non-expert users differ from that conducted by experts?
- What are the advantages and disadvantages of having non-expert users carry out evaluation work?

This paper will present a preliminary exploration of these questions, and an analysis and discussion of the issues raised by the evaluation reports carried out by the master’s student teams.

2. Usability Evaluation for Digital Libraries

Usability is defined and assessed in different ways. Bødker [1] provides a three-stage model of HCI and usability research, based on three ‘waves’. In first wave work, attention focused on the direct relationship with the interface, and it differs from related concepts such as functionality (does the system work technically?). Cognitive and predictive models such as Fitts’ Law, GOMS, and keystroke analysis, calculate usability in terms of the placement and size of elements, task complexity, the number of actions required, and other factors [12]. In heuristic evaluations (one of the techniques described in this paper), experts analyze interfaces with structured checklists and heuristics, with requirements gathered from users being used to generate typical fictional user and use cases (personas and scenarios) [9]. Usability testing involved users (frequently in labs) interacting with an interface, following a structured task and guided by a usability researcher.

In second wave work, the research context widens to the social, organizational and other contexts of technology use. To understand users’ needs, it is necessary to observe them in ‘real life,’ and theories such as situated action, distributed cognition, activity theory, participatory design, contextual inquiry, and others, examine the social and organizational contexts of information technologies, and translate between the worlds of the user and the researcher.

Finally, Bødker identifies an emerging third wave of HCI work, in which ‘the use context and application types are broadened, and intermixed.’ Here, research has to address the design needs of a growing slew of information technologies that are ubiquitous, wirelessly connected, small, convergent, embedded, gesture/non-keyboard controlled, and supportive of distributed, fragmented, and mundane tasks.

This paper focuses on teams of students ostensibly practicing a ‘first wave’ usability evaluation method. According to Nielsen [9], usability is “a quality attribute that assesses how easy user interfaces are to use [and provides] methods for improving ease-of-use during the design process.” System usability is evaluated according to five overall quality dimensions: learnability (how easy is it for the users to accomplish basic tasks the first time they encounter the design?), efficiency (once users have learned the design, how quickly can they perform tasks?), memorability (when users return to the design after a period of not using it, how easily can they reestablish proficiency?), errors (how many errors do users make, how severe are these errors, and how easily can they recover from the errors?), and satisfaction (how pleasant is it to use the design?). For many reasons (relatively low cost, quick turn-around, no need for users, etc.), usability evaluation remains an important evaluation approach. Nielsen states forcefully:

On the Web, usability is a necessary condition for survival. If a website is difficult to use, people leave. If the homepage fails to clearly state what a company offers and what users can do on the site, people leave. If users get lost on a website, they leave. If a website’s information is hard to read or doesn’t answer users’ key questions, they leave. Note a pattern here? There’s no such thing as a user reading a website manual or otherwise spending much time trying to figure out an interface. There are plenty of other websites available; leaving is the first line of defense when users encounter a difficulty ... The first law of e-commerce is that if users cannot find the product, they cannot buy it either.

(http://www.useit.com/alertbox/20030825.html)

It may not be usual to think of digital libraries as e-commerce organizations with products and customers. However, the analogy is a useful one, and can be used to inform the evaluation of digital libraries. The aim of digital library usability is not just to provide a technically usable interface, but also to make sure that users do not leave without their goal being accomplished, and if they do leave, that they return again at some point in the future.

Digital library researchers have extensively addressed both first and second wave usability issues. Kling and Elliott [7] identify five forms of digital library usability: interface usability and direct interaction with a system; organizational usability and the fit with existing work practices; content usability, collections scope, and searchability; ‘design for usability’ (an early form of social informatics); and the cultural dimensions of system use. Jeng [3, 4]
describes how digital library usability has since followed a range of theories and methods, including “usability testing; usability inspection; card sort; category membership expectation; focus groups; questionnaires; think-aloud; analysis of site usage logs; cognitive walkthrough; heuristic evaluation; claims analysis; concept-based analysis of surface and structural misfits (CASSM); and paper prototyping.” Jeng notes that despite extensive work by a number of digital library researchers, usability research in general remains under-developed (a similar situation has been observed in the field of digital library evaluation) [13].

2.1 Usability work as translation work

As with usability research in general, digital library usability research makes a general commitment (a) to the researcher as someone who observes, (b) to the user as a source of data, (c) to the user as someone who is observed, and (d) to research methods as supporting the transfer of usability data from the user to the researcher (and then to designers, managers, clients, developers, and other stakeholders). Digital library usability data are seen as something to be obtained or derived from the thoughts, opinions and practices of users, through experiments, instruments and methods designed by researchers. There is an assumed data flow from users to researcher, facilitated by usability experiments. An abstract example of such a flow is shown in Figure 1.

As part of this data flow, the researcher interacts with users, for instance by asking them questions about the digital library interface being evaluated. It is preferable (and usually assumed, because what would be the point otherwise) that the usability test subjects can understand the usability questions being asked of them. However, this is not always necessarily the case. Information systems are almost inevitably described in terms of webs of interrelated technical terms, and while this terminology may be clear to developers and practitioners, it is often opaque to users, even when simplified for usability testing. An ongoing meta-study of usability testing of bricks-and-mortar websites by Kupersmith [6] has analyzed 51 usability studies between 2002 and 2011. Included in this meta-study is a report and analysis of the terms that individual studies have reported as being ‘misunderstood, not understood, or not preferred.’ These bricks-and-mortar library related terms included: periodical, database, collection, electronic journal, reference shelf, web guides, archives, FAQs, browse versus keyword, basic search versus advanced search, reference desk, inter-library loan, circulation (desk), keyword, and many others, as well as many local acronyms for libraries and collections. In other words, many of the terms are commonly used in library and information science when talking about library computer and web systems, and whose definitions are often taken for granted, are mysterious to the actual users.

No equivalent meta-analysis has been carried out with regard to digital library studies (although such an analysis would be useful). It would not be unreasonable to expect that digital library users involved in usability testing (and digital library users in general) likewise sometimes do not understand many of the terms and concepts regularly involved in discussion about digital libraries. For instance, usability work carried out by the first author with the National Science Digital Library showed that test subjects did not understand terms such as browsing, topic, subject, collection, catalog record, etc.

2.2. Eliciting usability data

Usability data do not emerge unbidden from users; they have to be coaxed out through experiments, tests, and forms; collected and aggregated, and compared and analyzed; and written into usability reports to be distributed to wider stakeholders. Incorporated into a usability workflow are usually, therefore, a series of what we refer to as ‘translations.’ Here, a translation involves the conversion of data from one form or representation into another, as part of the process of creating the final usability report. Some examples of translation in HCI work are: a user’s comments in a ‘think-aloud’ exercise are translated into an observer’s notes; values from Likert scales are translated into charts; and evaluation spreadsheets, notes, and charts are translated into a final report. Significant translation work is required between the raw data and initial findings, and between initial findings and the final report(s) that will be distributed to various stakeholders (see the arrows in Figure 1).

There are a number of phases to this translation work. First, there is the translation of the data in the users’ heads into the test instruments, the translation from the gathered data to the initial analyses, and the translation from the raw findings into the polished usability report. Users in usability testing are unlikely to express themselves in terms of the findings written
up in final reports, and usability experts have to collect, sift, analyze, interpret these data, and ‘write up’ the results based on their interests and expertise. Data have to be translated into charts and tables, and charts and tables have to be translated into reports.

Second, there is translation between the researcher and other digital library stakeholders, such as developers, designers, users, managers, and funders. Each of these stakeholders can have different understandings or ‘technological frames’ [11] of digital libraries, “assumptions, expectations, and knowledge [organizational members] use to understand technology in organizations.” The presence of multiple groups and frames in the same setting, can lead to problems with communicating about a technology between these different groups [2, 5, 6].

Digital library usability evaluators will translate data and findings between various stakeholders. For instance, the users may be K-12 teachers, and the stakeholders may be managers and funders. Translation work between these groups and their technological frames is handled by the usability researcher, who designs the experiment(s), collects and analyzes the data, and reports research findings. This is however just one possible configuration of workflow and translation in usability work. Other configurations are possible, even if they are less frequently encountered. In the case described below, it is digital library users—students in a graduate level HCI class—who act as evaluators, and who collect and analyze usability data, then write it up into reports. A summary of their workflow is presented in Figure 2.

![Figure 2. Usability translation (II)](image)

Here, the users also take the role of usability researchers who have received basic training in heuristic evaluation techniques. It is the users (and not researchers) who translated evaluation data into a series of reports, which they sent to the first author, who then compiled the individual student reports into a summary report, for presentation to the other members of the IPL.

An important difference between Figure 1 and Figure 2 is that the translation from data to findings is carried out by different groups with different technological frames. In Figure 1 the translation is carried out by the usability researcher, within the context of her HCI technological frame. In Figure 2, the translation from data to findings is carried out by the users, within the context of those users’ technological frames. In both cases usability reports are produced. However the translation from data to results to reports takes place at different stages.

Do these two configurations produce different usability results? If so, what differences are there, what is their nature, and why do they occur? On the one hand, it could be argued that while the chain of translations has been reconfigured between different groups in Figure 2, the overall process (from users to stakeholders) remains roughly equivalent, and the final reports will be similar. On the other hand, the presence of different groups with different technological frames at different stages of the overall usability workflow could ‘filter’ the usability data in different ways, resulting in different sets of results being presented in the final report. We have already noted that this latter scenario was what indeed seemed to have happened in the case of the masters’ students. To understand this process further, the next section will analyze the content of these masters’ student reports.

3. Case Study – The IPL

![Figure 3. IPL homepage: http://www.ipl.org](image)

The IPL is a largely volunteer maintained project that provides online reference services, reviewed collections, and other services. It has 40,000 online resources, 12,500 web pages, and a virtual reference service that has answered over 95,000 questions. For the 12 months to May 2011, the library received approximately 765,000 unique visits and 2.12 million page views per month. In 2008, work began on a
redesign of the web site. The work took place within a number of constraints, including limited time, and staff turnover (students cycling in and out of the project). Due to these constraints and the largely voluntary nature of IPL, the interface evaluation work was often incorporated into the existing research and teaching interests of the faculty and student members of the project. Two types of evaluations, generating two evaluation data sets, were carried out. Both of the evaluations were carried out using available student expertise. In the first set of evaluations, doctoral students (second and third authors, led by the third author) tested the IPL website with student volunteers. In the second set of evaluations, students in a masters’ level HCI class taught by the first author acted both as users and heuristic evaluation experts. (It is important to note that these evaluations were not set up as a formal experiment; they were intended to provide support for the development of the IPL website. It was only after separate evaluations had been carried out that the different outcomes were noted, and it was decided to investigate these differences, with the aim of gaining a better understanding of the individual evaluation processes.)

The first type of evaluation involved usability testing produced by doctoral students working with the IPL project. These students had interest, background and training in HCI evaluation. The evaluation was carried out in three phases. In the first phase, eight participants were recruited to complete a series of tasks using a paper prototype of the newly designed IPL homepage and its six major sub-pages; in the second phase, five graduate students performed a heuristic evaluation on a beta version of the IPL website; in the third phase, five participants were recruited to provide answers to ten simple questions using the fully functioning IPL website. The main findings of the evaluation focused on specific usability issues with the site, such as button labeling.

The second type of evaluation consisted of a series of heuristic evaluations carried out by teams of students in master’s level ‘Introduction to HCI’ classes at the iSchool. These students received an introduction to HCI concepts in the ten-week course, including the work of Norman [10] and Nielsen [8] (see below), but generally, they had no other prior background in HCI. They conducted the evaluation exercise in order to gain ‘hands-on’ experience with applying evaluation techniques in a realistic setting with a realistic client (the IPL) and design brief. Heuristic evaluation was chosen as a technique because it could be learned and applied relatively simply in the context of a class group project. The students were expected to work as a project team, and to carry out individual evaluations before meeting as a group to prioritize the evaluation findings, and to produce an extensive written report that described these findings and design recommendations. This latter teamwork was included in order to introduce students to working on an evaluation team in an organizational setting.

In an unexpected contrast to the findings of the doctoral students, the evaluation reports of the master’s students often appeared to focus on what, to a usability expert, would not be considered interface issues. In particular, they focused on the difficulties they had with efficiently discovering resources on the site. This focus appeared consistent across all teams and classes. The contrast between the two sets of results (the doctoral students and the master’s students) was intriguing.

It was hypothesized initially that this contrast arose not just because of the different methods that were used by the two groups, but also because of the differences in HCI experience between the two groups, combined with differences in technological frames, and definitions and understandings of usability and digital libraries. In terms of the ‘translation’ hypothesis advanced in this paper, the doctoral students’ workflow was similar to that portrayed in Figure 1, where researchers collect data from users and present it to stakeholders. The master’s students’ workflow was similar to that portrayed in Figure 2, in which the users also analyze the data before passing it onto the main researcher. In the case of the doctoral students, the initial data were filtered through their experience as HCI experts; in the case of the master’s students, the initial data were filtered through their experience as users (with some introduction to HCI concepts).

If this were to be true, this could have interesting implications for the design of digital library evaluation instruments, and the roles of users and designers in digital library evaluation, as different groups with different frames could generate different (but always useful) sets of evaluation data. It was therefore decided to analyze the master’s students’ team heuristic evaluations in further detail.

3.1. Data sources

The data in the analysis consists of reports generated by students over several years in a master’s level ‘Introduction to HCI’ class. A total of 141 master’s students in 24 groups carried out evaluations over a period of three years. (The reports were de-identified in line with IRB policies. An ‘honest broker’ unaffiliated with the project was provided with copies of the student reports; non-relevant pages were removed, and all occurrences of student names were blacked out. The resulting documents were scanned, and printouts of the scans were used in the analysis. This produced a corpus of approximately 450 pages.)
All versions of this class were conducted online. The overall aim of the class is to provide an introduction to ‘first wave’ HCI concepts, focusing on cognitive models and direct interactions with an interface, web site, etc. The students were introduced to Norman’s *Design of Everyday Things* [10] as a way of thinking about a cognitive theoretical perspective to the analysis of the usability of interfaces. They studied topics such as task analysis, personas and scenarios, and heuristic evaluation. In order to experience and apply these concepts first hand, in an example of problem-based learning, the students carried out a variety of ‘real-life’ evaluation exercises, both individually and in groups.

One of these exercises was a heuristic evaluation of the IPL website. Depending on class enrollment, the students were split into groups of five to seven students. They used an online course management system to coordinate their work. They were provided with some background reading on digital libraries, digital libraries, heuristic evaluation, and the IPL. They were then asked to carry out individual heuristic evaluations of the IPL, using previously developed personas and scenarios. They recorded data on a 10-point evaluation instrument based on Nielsen’s evaluation heuristics (i.e., 1, visibility of system status; 2, match between system and the real world; 3, user control and freedom; 4, consistency and standards; 5, error prevention; 6, recognition rather than recall; 7, flexibility and efficiency of use; 8, aesthetic and minimalist design; 9, help users recognize, diagnose, and recover from errors; 10, help and documentation).

They were asked to pay particular attention to interface issues that violated the various heuristics; to rate the severity of these violations; and to identify and describe overall usability themes that emerged from the heuristic evaluations. They then had to write a final report describing the main usability issues that they had found in IPL, prioritize these issues, and support their findings with concepts from the course. The groups were provided with a copy of an exemplary HCI student report, with an executive summary, description of methods, description of results, recommendations for design, etc. It was suggested that they follow this exemplary report in terms of formatting, although not all teams did so.

3.2. Analysis

Each student group had been asked to provide an executive summary with a concise prioritization of their evaluation results. These summarized findings were assumed to represent the main findings of the group. In some cases, a group followed its own report formatting, and the overall findings had to be located in the main body of each report. Once the main findings in each report were identified, they were marked up. Brief notes of each finding were written on index cards, and each card was provided with a reference number to link it back to the report and passage where it occurred. The first and second authors each made their own set of index cards. The first author prepared 93 such cards, and the second author prepared 91 cards.

Six high-level categories common across all the index cards were then inductively identified, through qualitative coding. These six categories were:

- **Search/browse** issues related to problems with the search engine, difficulties with browsing, confusing subject categories, poor search results (empty results, etc.), lack of advanced search, lack of search results refinement, etc.
- **Navigation** issues related to broken links within the site or to external resources, confusion when leaving the site to visit an external resource, lack of navigation back to the home page, and poor quality ‘breadcrumb’ trails.
- **Interface** issues related to a site design that was at times considered to be cluttered, inconsistent, and confusing.
- **Help** issues related to a general lack of online help, documentation, and FAQs on how to use the site.
- **Ask a Librarian** specific issues related to the online reference section of the site.
- **Content** issues related to the quality of the IPL content.

These six categories are independent of Nielsen’s ten heuristics, although there are some overlaps, for example in terms of the provision of help information and error prevention.

Next, a collaborative card sort was used to allocate the individual index cards back into the high-level categories. There was often discussion as to which category was appropriate, and there was some overlap between the categories. For instance, a problem with navigation could also be a problem with the underlying browsing metadata that supported the navigation in the browsing pages. Again, the category ‘broken links’ could refer both to internal navigation and also to links to external resources, where the link is stored in the metadata. This ambiguity reflects the general interdependence of metadata and navigation in the IPL website, and also the ways in which these were reported in the original student reports.

Following the card sort, **search** and **navigation** emerged as the most important categories identified in the student reports. The average number of cards allocated per category is shown in Table 2. The
average number of categories identified per report was 4.0.

<table>
<thead>
<tr>
<th>Category</th>
<th>A</th>
<th>B</th>
<th>Ave</th>
<th>% (n/92)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search/browse</td>
<td>37</td>
<td>33</td>
<td>35</td>
<td>38%</td>
</tr>
<tr>
<td>Navigation</td>
<td>29</td>
<td>20</td>
<td>24.5</td>
<td>27%</td>
</tr>
<tr>
<td>Interface</td>
<td>13</td>
<td>20</td>
<td>16.5</td>
<td>18%</td>
</tr>
<tr>
<td>Help</td>
<td>7</td>
<td>12</td>
<td>9.5</td>
<td>10%</td>
</tr>
<tr>
<td>Ask</td>
<td>3</td>
<td>4</td>
<td>3.5</td>
<td>4%</td>
</tr>
<tr>
<td>Content</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3%</td>
</tr>
</tbody>
</table>

Table 2. High-level usability categories

Search and navigation together comprised 65% of all identified issues. These two categories could frequently be confused, particularly in relation to search, as finding one’s way through the IPL in many cases could either be seen as an issue with search and browse, or an issue with navigation. The remaining issues identified—Interface issues, help and supporting documentation, the ‘Ask a Librarian’ online service, and resource content—reported usability issues in more conventional and unambiguous terms.

The search and navigation issues were framed in three main ways. The first way was in terms of specific search interface issues. For example, elements of the search interface were found to be confusing, including multiple links to different features of the search box itself, a lack of easy-to-read help and error recovery documentation, and a lack of ability to retain and refine queries from previous searches. The second way was in terms of navigation interface issues. For instance, the use of different page styles within the site made it difficult to understand whether or not one had moved to a new site or not; the fact that the IPL had both ‘home-grown’ and third-party resources in its catalog made it hard to understand when one had left the site and wished to return to it; it was hard to navigate to the home page; and the descriptions of different collections in navigation buttons were hard to understand. Both of these findings are to be expected in a heuristic analysis.

A more unexpected finding of the analysis was that the student groups used a third way to frame search and navigation issues, in terms of more general search functionality issues. These search issues were generally associated with locating resources while searching or browsing the site. They seemed to lie somewhere between search and navigation (and were placed into both categories during the card sort), but at the same time they did not fit precisely into either category. Rather, they seemed to refer to issues that were related to the quality of the underlying metadata (something that the students had not been asked to evaluate). A sample of ways in which back-end/metadata issues were raised in reports that should have addressed front-end issues is as follows:

The most highly prioritized usability problems ... concerned the presentation of results retrieved from Search or by drilling through successive categories [i.e. browsing].

The IPL2 site could offer services to more advanced users that allow them to access the information they need faster and more efficiently. These services include the ability to sort results sets and to search within the site's five high-level categories.

Unable to refine search ...: A number of team members were frustrated because they were unable to refine their search to eliminate inaccurate search results.

Search problems included anything that was encountered while directly attempting to search for information - several students had difficulty finding different types of information on the website. Categorization and classification included problems encountered when trying to find information. Some information was inconsistently classified, causing multiple lists with same options, or nearly the same options with only slight differences.

The website is broken down into 5 main categories (see below); however, it was challenging to utilize these categories for finding information … Information within the selected categories is not always available where it should be.

The … website includes a search functionality to help users of the website find the information they need. Almost all Internet users today are familiar with searching on the Internet using search engines, and have grown to expect certain behavior from a search feature on web sites … Participants seemed to have high expectations for the usefulness of the search functionality. In reality, most of the participants encountered one or more significant issues when using search, and frequently resorted to alternate ways of finding the information.

The first and most major usability issue that we found related to the uniqueness of results using the search function located on each page ... Some search terms, particularly those which return a page within the IPL, will have the top two results be the exact same page, but with different descriptions.

Searching for an item in one area of the website often produces a "No Results Found" message, yet the item might be available in another area of the website.

Several of the pages we evaluated consisted of lists of links, with no apparent organization. Students expect results to be ranked by relevancy, similar to the major search engines they use.
This is just a sample of references in the student reports to search, navigation, and metadata issues, which were described in different and overlapping ways. A summary refrain repeated in several of the reports was that, in the end, commercial search engines offered a potentially better service than IPL. The comment “The [IPL] search engine must be improved so that users can not only find relevant information, but also so that they will choose ipl2 over other popular resources such as Google and Yahoo” is typical of this form of evaluation comment.

As can be seen from these examples (and there were many more direct and indirect examples in the reports), a significant factor for the master’s students in assessing the usability of the IPL site was that of the ease of accomplishing what was presumed to be the main task, locating and retrieving relevant resources. However, this was often framed not in terms of whether the interface supported search, but rather in terms of whether the ‘right’ results were found. These results reflect previous findings from the evaluation of the National Science Digital Library (NSDL) by the first author, which suggested that users wanted to spend as little time as possible on the site itself, preferring to identify relevant resources as quickly as possible and then leave the site.

3.3. Discussion

Why were the masters students discussing non-usability issues when they were specifically asked to focus on usability? According to the assignment instructions, the master’s students should have focused their analysis on the properties and features of the interface itself (buttons, links, colors, layout, etc.). However, their observations on the usability of IPL often described usability in terms of finding (or not finding) what they wanted. Many of the issues that the groups identified were in fact attributable not to the interface design, but to the performance of the underlying information architecture, specifically the back end database that contained the metadata and search engine (although it was apparent from the way that these issues were described, the students themselves were often not aware that these were related to metadata).

For example, the lack of an advanced search interface was an issue. However, this was not just a matter of creating a new ‘advanced search page’ with appropriate buttons and drop-down menus; such a function would also depend on having metadata of a high enough quality to support various advanced search features. The comments regarding the navigation and link issues associated with the task of locating and browsing to resources in the IPL’s web site were also metadata related. While the user, when browsing, appeared to be clicking links on a series of static web pages of ever more refined subject categories, in reality the links on each page were being generated ‘on-the-fly’ by the underlying web code that in turn drew on specific browsing metadata in the catalog record for each resource, in order to display the browsing categories which that particular resource was in. Reported problems with browsing were therefore reflecting problems with the specific IPL browsing and subject metadata, rather than any hard-coded elements of the interface itself. As the IPL interface was basically a front end for the back end catalog and metadata repository, with both the front end and back end being closely integrated, these issues could often be conflated.

Figure 4: Search, navigation and findability

One way to visualize the ways in which these issues overlapped is to understand that the process of finding something in the IPL site could be seen and analyzed by the students in multiple ways. These include: (a) finding something in the search results (an interface issue); (b) finding something on an appropriate page in the IPL site (a navigation issue); and (c) finding something in the catalog (a metadata issue). As used in the student reports, these three senses of finding were ambiguous and varied between these three definitions. This situation is illustrated conceptually in Figure 4.

One thing to be stressed about this figure is that while the students’ conclusion that ‘finding things in IPL was difficult’ was not situated in line with the requirements of the assignment, it was in fact a very useful evaluation outcome. In terms of Nielsen’s criteria for usability ([9], see Section 2), poor findability (however defined) will result in users leaving the site; and in the end, some of the search results were just not satisfactory for the evaluators.

3.4. Implications for the evaluation of digital libraries
The master’s students’ emphasis on being able (or not able) to locate the resources they wanted quickly and with little effort, is a dimension of usability that depends on the quality of the underlying metadata. What therefore drew the master’s students to define usability in terms of ‘findability’ rather than specific interface issues? To consider the wider implications of this question for digital library evaluation, we return to the concepts introduced at the start of this paper—technological frames, the three waves of HCI, and translation.

To recap, a technological frame is “the assumptions, expectations, and knowledge [organizational members] used to understand technology in organizations. This includes not only the nature and role of the technology itself, but the specific conditions, applications, and consequences of that technology in particular contexts” [11]. In the context of the IPL evaluation work discussed in this paper, and Bødker’s ‘3 wave’ model, it is suggested that doctoral students were mainly concerned with evaluation from a first wave perspective, and with a detailed evaluation of the IPL interface itself. The master’s students’ heuristic evaluations, on the other hand, while intended to be examples of first wave evaluation, often addressed second wave criteria, such as the overall interaction experience of using IPL and the frustration with not being able to discover the resources that they wanted in the context of the personal, organizational, and social goals they had outlined in the personas and scenarios they used in the evaluation. Their evaluations also seemed to address some third wave issues, in that some of them advanced an evaluation benchmark—the use of search engines—that is increasingly based on a ubiquitous, wireless, fragmented, and mundane task (that is, search engines are used everywhere, all the time, for a wide variety of tasks; their use has become embedded in common computational device use). Of course, this lack of focus on first wave issues could also possibly be explained in terms of the masters’ students being inexperienced evaluators and receiving inadequate classroom preparation.

The two groups of students also had different technological frames of digital libraries. The doctoral students had a more technical understanding of IPL as a digital library, acquired during the time that they had spent working on the project. This frame was expressed in their analysis in terms of the attention to detail in the HCI work carried out, and in the granularity of the elements of IPL that were evaluated. The master’s students, on the other hand—many of who had not thought in depth about digital libraries before—had a frame that often assumed that the IPL was some kind of a search engine. It should again be emphasized that neither frame was more nor less valuable than the other; they were just different, and both frames supported the generation of useful usability evaluation results. What is important to remember though is that in terms of translations (Figures 1 and 2), each group of students translated the data collected from their test subjects into a report that reflected their technological frames with regard to HCI work and digital libraries. In the end, the different frames, and the different translations between different groups in the evaluation process, gave rise to different evaluation findings. When acting as heuristic evaluators, the masters’ students translated what they considered to be problems into a language that the client (in this case, the IPL) could then analyze further. Working with users through this more direct channel thus has the potential to identify further dimensions of what users care about in a more standardized and efficient fashion.

3.5. Lost in translation

This prompts the question as to what evaluation data may be ‘lost in translation’ between researchers’ reporting and analysis of usability data, and users’ reporting and analysis of the same data. On the one hand, the masters’ student reports contained useful information about the IPL that were not contained in the doctoral students’ report; such information would be hard for experienced HCI researchers to collect, as they will have forgotten what it is like to approach a digital libraries and usability work in a naïve fashion. On the other hand, however, if users can find it hard to understand basic HCI and digital library concepts and vocabulary, how should questions regarding the usability of digital libraries be presented to them, for instance in the form of usability instruments?

The usability and design literature already recognizes that ambiguity and misunderstanding can occur in user studies. What questions are asked of users, can determine what sorts of data are gathered; but what are the right questions in the first place? There are ways and methods to approach this (pilot studies, definition of tasks with user involvement, etc.) These techniques can be time and resource consuming but they can play an important role in a successful study. One possible solution is to support users to carry out evaluation, for instance by developing HCI instruments that support users to collect evaluation data on their own terms, that would also be useful for researchers. One way to do this would be to begin an evaluation by observing and understanding how users think about evaluation. Kupersmith [6] recommends that when studying users in usability tests and focus groups, researchers capture the users’ terminology and vocabulary for technical concepts; in the case of the
digital library evaluation reported here, the concept of 'findability' has been abstracted from the master's student reports. Kupersmith also notes that such exercises may capture "opinions not behavior" and that it may be difficult to “separate terminology-related problems from other design issues.” Nevertheless, we believe that there is room here to develop more user-centered HCI instruments that support users to carry out evaluation, and which address users’ central evaluation concerns (such as findability). In evaluation work, such instruments would act as ‘boundary objects’ [14] between HCI researchers and users, allowing users to participate more fully in the design process itself and bring their technological frames to the actual research process.

4. Conclusion

This paper has presented a post hoc analysis and comparison of two sets evaluations of a digital library, carried out by two different groups of students with differing levels of HCI experience. The paper has highlighted the differences between these evaluations. Our analysis of these differences, based on the concepts of technological frames and translation, has shown how the evaluation reports were shaped by the different technological frames of the evaluators, and by how and at what stage these frames were translated between different groups. We identified findability as being a significant non-interface concern for the master’s students, and conclude that metadata is concretely linked to user satisfaction in digital library usability. Not only does resource retrieval depend on metadata, but so too do wider perceptions of the usability of digital libraries in general. Both of these findings have implications for the design of digital library evaluation and the roles of users and designers in such evaluation.

5. References