Evaluating Digital Libraries: A User-Friendly Guide

NATIONAL SCIENCE DIGITAL LIBRARY - NSDL.ORG

Prepared by
Thomas C. Reeves
Xornam Apedoe
Young Hee Woo
The University of Georgia

Evaluating Digital Libraries: A User-Friendly Guide

Table of Contents

CHAPTER 7

CHAPTER 8

CHAPTER 9

CHAPTER 10

CHAPTER 11

CHAPTER 12

Chapter

1

Why evaluate?

> "So far, evaluation has not kept pace with efforts in digital libraries (or with digital libraries themselves), has not become part of their integral activity, and has not been even specified as to what it means, and how to do it."
>
> - Saracevic, 2000, p. 351

Evaluation is a term that many people fear. For others, it is not so much a matter of fear as a sense of being confronted with a concept that is shrouded in mystery and thus best left to someone else. The fear stems from the anxiety we all have concerning the possibility of being evaluated unfairly. The mystery stems from the fact that evaluation designs, instruments, and especially reports often appear to be overly complex and full of arcane statistics, daunting graphs, and verbose prose.

The purpose of this "User-Friendly Guide" is to allay any fears, dispel any mysteries, and, most importantly, help you become confident about evaluation. Using this guide, you will be able to design, implement, and report better evaluations within the context of developing, operating, and/or using digital libraries. This guide is written in an informal, easy-reading style intended to encourage you to make evaluation a routine aspect of your work with digital libraries.

It's all about decisions!

The first step in any evaluation is to establish a strong and clear rationale for evaluation. In other words, you should begin any evaluation by answering the question: "Why bother with evaluation?"

Evaluations are conducted for many reasons. Often, they are done to meet some sort of requirement established by a funding agency. They may also be done for political reasons, or simply because the people involved in an enterprise believe it is the right thing to do, akin to taking vitamins or engaging in vigorous exercise.

But, it is our belief that the only defensible rationale for evaluation is to inform decision making. Anyone involved with digital libraries (information scientist, collection specialist, manager, subject-matter expert, user, etc.) must make decisions on a regular basis. Some decisions are made on the basis of habit or tradition, others may be guided by politics or prejudice, a few may be guided by superstition or intuition, and far too many are founded on ignorance or best guesses. Ideally, decisions should be informed by timely, accurate information. That's where evaluation comes in. Evaluation should be conducted to provide decision makers with the information they need to make the best possible decisions. The better information provided by evaluation doesn't guarantee that the best decisions will be made. After all, traditions and politics are powerful forces in virtually every context. However, high quality information provided to decision makers in a timely manner certainly improves the likelihood that decision making will be enhanced.

Various kinds of professionals carry out evaluation activities to help them make decisions all the time. For example, physicians inquire about medical histories, conduct examinations, and run various medical tests before deciding upon a diagnosis and treatment. Attorneys interview clients, review documents, and conduct private investigations before deciding how to present their cases to judges and juries. Indeed, the reputation of any given doctor or lawyer is determined largely by his or her skill in conducting evaluative activities such as interviewing, examining, and testing. As someone involved in making decisions about digital libraries, you should and can become similarly skilled.

Fortunately, you don't need an advanced degree to be an effective evaluator. Although there are certainly many advanced topics within the realm of evaluation that may require graduate studies or professional development, anyone can learn to apply a simple, yet powerful, model to evaluate digital libraries.

The five basic steps of an evaluation

There are many complex evaluation models in textbooks and scholarly papers, and evaluation specialists often have graduate degrees in this area. However, the evaluation model promoted in this guide involves five basic steps:

1. Identify the decisions that you or others involved in your digital library enterprise must make.

2. Identify the questions that need to be addressed to inform the pending decisions.

3. Identify the evaluation methods and instruments that will be used to collect the information needed to address these questions.

4. Carry out the evaluation in a manner that is as effective and efficient as possible.

5. Report the evaluation results in an accurate and timely manner so that it can provide the information you and others need to make the best possible decisions.

Sounds simple, doesn't it? It makes you wonder why people haven't conducted evaluations of digital libraries more frequently in the past. We attribute at least part of this failure to evaluate to the fact that people rarely begin evaluation planning by identifying the decisions that an evaluation should inform. Instead, they usually begin an evaluation by struggling to identify the aspects of digital libraries to be evaluated (e.g., user interface or collection quality), the criteria for evaluation (e.g., accuracy or relevance), or the most appropriate methods (e.g., usability testing or online surveys). Beginning with these issues inevitably leads to disagreements and arguments that can stall or even halt the evaluation process. Although identifying decisions up front is not an easy task, the payoff is that all the other issues (questions, methods, instruments, criteria, standards, analysis, reporting, etc.) flow naturally from the specification of the decisions that the evaluation should inform. Chapter 2 in this Guide is devoted to planning an evaluation, including the critical process of identifying decisions.

Evaluation & research: What's the difference?

People sometimes use the terms evaluation and research synonymously, but this is a mistake. Evaluation differs greatly from research. The National Research Council (http://www.nas.edu/nrc/) has defined six key characteristics for scientifically-based research:

1. Pose significant questions, which are observable;

2. Link to relevant theory, in developing a hypothesis;

3. Use tools which are valid for addressing the questions;

4. Rule out counter explanations of observed evidence;

5. Replicate findings across groups and observers; and

6. Submit results and processes to scrutiny of colleagues and publics.

These characteristics can help clarify the differences between evaluation and research. With respect to characteristic number 1, the types of questions asked, both research and evaluation should be designed to address "significant" questions and involve the collection of observable data. But evaluation questions are much more closely linked to specific decisions that have a more localized, less "generalizable" scope. For example, an evaluation may be focused on decisions about the types of search delimiters that a particular audience would find most useful in a specific digital library whereas a research study may be focused on addressing issues related to the effects that digital libraries have on the scholarship within a field such as geology. Interestingly, although evaluation has a long history of being focused on decision-making, the shift to "evidenced-based decision making," long established in health and medical fields, is a relatively recent development in the context of educational research (Shavelson & Towne, 2002).

With respect to characteristic number 2, while there are evaluation models that emphasize theory, most evaluation plans do not have as strong a foundation in theory or research literature as research designs are expected to have. An evaluation focused on decisions about what types of metadata are most useful for K-12 teachers seeking educational resources might not be as informed by the theoretical underpinnings of metadata structures as would a research study focused on the mental models of metadata standards constructed by reference librarians. That said, there may well be great benefit to be gained by including relevant theoretical perspectives in evaluations (Chen, 1990).

With respect to characteristic number 3, both evaluation and research are concerned with the reliability and validity of the tools and instruments used in data collection. One difference is that within evaluation circles, there is more acceptance of the use of measures that may not be completely validated, whereas in research there is a much greater expectation that the reliability and validity of instruments be rigorously established before being used in a study. For example, an evaluation of community reactions to a new user interface for a digital library may utilize an original survey instrument that has not been completely validated, whereas a research study focused on the effects of digital libraries on plagiarism among undergraduate students would almost certainly require the use of a validated measure of academic honesty.

With respect to characteristic number 4, there are major differences between evaluation and research. Traditional quantitative educational research is generally designed on the basis of experimental or quasi-experimental designs intended to rule out or limit the plausibility of alternative explanations for results. Evaluations, on the other hand, are often designed to examine rival explanations from a variety of perspectives, and to provide decision makers with alternatives from which to choose. In that sense, many evaluations have more in common with qualitative research designs, and indeed, evaluations and qualitative studies may often look very similar in design, implementation, and even reporting. The capacity to rule out counter explanations of observed evidence has long raised many contentious issues within the educational research community (Lagemann, 2000), but evaluations can escape these issues by presenting the evidence for alternative explanations and allowing decision-makers to decide for themselves. We think it is better to think of evaluation as a process more akin to the judicial process. In a legal case, evidence is presented for the guilt or innocence of someone and a jury or judge decides, whereas, in the scientific process, findings are judged to be more or less warranted on the basis of peer review and replication. Of course, evaluations may also utilize quasi-experimental designs. For example, it might be feasible to roll out two different types of digital library search engines to randomly selected populations, and to collect data such as the number of return visits from the same domain names to judge user preferences. Although such an experimental approach might be feasible in an evaluation, we would not recommend it be used exclusively because of the difficulty of interpreting the results. Instead, we would advocate for a mixed methods approach that would include the collection of qualitative information to reveal why any preference patterns that emerged existed.

With respect to characteristic number 5, replication is much more common, and indeed, expected in research than in evaluation. In evaluation, the emphasis is on providing quality information to inform decisions in a timely manner. If evaluators have done their job well, the information they have provided has helped decision makers make better decisions, and hence future evaluations will be focused on different decisions. Suppose the National Science Foundation must decide funding priorities for collection development for the National Science Digital Library. If an evaluation has been done to collect usage data in K-12, undergraduate, scientific, and other communities, informed decisions can then be made. A subsequent evaluation might focus on the types of resources most highly valued in the communities targeted for increased funding.

Finally, with respect to characteristic number 6, peer and public reviews are among the primary foundations of scientific research. In fact, pseudoscience would proliferate without high standards for reviews. By contrast, evaluations are rarely shared beyond the "stakeholders" (decision makers and other interested parties) who are part of any particular evaluation. The results of evaluations are sometimes presented at conferences, and a few even get published in journals, but most evaluation reports have very limited circulation. Frankly, we believe that the state-of-the-art of evaluations would be improved if there was more public sharing, and indeed there is a formal review process, called meta-evaluation (Cook & Gruder, 1978), that is essentially the evaluation of evaluations. As you get involved in evaluating digital libraries, you are strongly encouraged to share your methods and findings at conferences or through peer-reviewed publications. For example, Wildemuth, Marchionini, Yang, Geisler, Wilkens, Hughes, and Gruss (2003) from the University of North Carolina at Chapel Hill presented the results of their evaluation of video surrogates at the 2003 Joint Conference on Digital Libraries held in Houston, Texas. Mead and Gay (1995) published a paper about using concept mapping in digital library evaluations in the *ACM SIGOIS Bulletin*. For anyone in academe, evaluations of digital libraries can be a form of the "scholarship of teaching" (Shulman, 2000), i.e., systematic inquiry into the effects of various forms of instruction (e.g., web-based instruction) or instructional support (e.g., digital libraries).

Evaluation & assessment: What's the difference?

The terms evaluation and assessment are often used synonymously, but this leads to a great deal of confusion. Therefore, in this Guide, we will use these terms to mean two very different things. Both evaluation and assessment involve the collection of information to make decisions. However, evaluation is focused on things, e.g., programs, products, and projects. Assessment is focused on people, e.g., their aptitudes, attitudes, or achievement.

This Guide is primarily about evaluation, but assessment is an activity often used within an evaluation. For example, an evaluation intended to inform decisions about whether a digital library should contain only resources that have been scientifically validated might include an assessment of teachers' attitudes toward such resources. Thus, we view assessment as an activity that is often included in evaluations, but it is definitely not the same thing. Assessment is primarily an objective measurement function, whereas evaluation usually involves a much greater degree of subjectivity. The aforementioned assessment may find that teachers don't really care about whether resources have been scientifically validated, and prefer to choose resources based upon intuitive feelings regarding their utility in their classrooms. The decision makers who interpret the evaluation results, on the other hand, may decide that their digital library will only include validated resources because they believe that such a high standard for quality will distinguish it from other collections of resources.

Making the best use of this guide

This guide is designed to provide practical advice about planning, conducting, and reporting evaluations of digital libraries. This first chapter has introduced a decision-oriented perspective on evaluation and presented some important distinctions between evaluation, research, and assessment. Chapter 2 is focused on the evaluation planning process. Chapters 3-11 describe various types of evaluation methods that are feasible within the context of evaluating digital libraries. Chapter 12 is all about options for reporting evaluations so that they will have maximum influence on decision making.

This Guide is not a static document, and we fully expect to extend and enhance it based upon the feedback we receive from you and others involved in the development, implementation, and use of digital libraries. Revisions will also be informed by advancements in the area of evaluation as they evolve.

Chapter

2

Evaluation planning

> "Digital libraries serve communities of people and are created and maintained by and for people. People and their information needs are central to all libraries, digital or otherwise. All efforts to design, implement, and evaluate digital libraries must be rooted in the information needs, characteristics, and contexts of the people who will or may use those libraries."
>
> - Marchionini, Plaisant, & Komlodi, in press, p. 1

Every evaluation should begin with a well-crafted plan. Writing an evaluation plan is not just common sense. It is an essential roadmap to successful evaluation! An evaluation plan is essentially a contract between you (the evaluator) and the other "stakeholders" in the evaluation (e.g., clients who fund the evaluation, decision makers who will use the results, participants from whom data are collected, and any other interested parties).

Most evaluation plans go through several stages of revision before acceptance by the stakeholder community, and they are likely to be modified during their implementation. Negotiating an evaluation plan with stakeholders represents a major part of the effort you will invest in evaluating digital libraries. Indeed, once you have developed a well-designed plan and a set of reliable and valid evaluation instruments, it may even be possible to turn much of the actual data collection over to others. On the other hand, trying to evaluate without a carefully-vetted plan will almost always lead to major problems that otherwise could have been avoided.

An evaluation plan should reveal the what, how, when, where, and other technical and logistical requirements of an evaluation. It provides a means of keeping the decisions, questions, and methods involved in an evaluation open to review and enhancement by all stakeholders. In addition to enabling the support of stakeholders, the process of preparing a plan helps you understand the size and scope of an evaluation project. You need this understanding to establish a meaningful timeline and a reasonable budget for the evaluation. The first step in the planning process is identifying the decisions that the evaluation should inform.

Why kinds of decisions can you anticipate?

The concept of a digital library has been around for decades, at least since Vannevar Bush's description of the "Memex" in the July 1945 edition of *Atlantic Monthly*. Although enormous technical challenges remain, ideological struggles around the design and implementation of digital libraries are perhaps even more complex. For example, there are ongoing arguments about whether digital libraries should primarily retain the collection, cataloging, and service functions of traditional libraries or whether they should become something altogether different (Kahle, Prelinger, & Jackson, 2001). Underlying fundamental decisions about the nature of digital libraries are a host of smaller-scale, but extremely important decisions that confront digital library developers, patrons, and funding providers.

Trying to anticipate the decisions that can be influenced by an evaluation requires creativity and trust. Many stakeholders, especially those involved in funding or developing digital libraries, do not wish to anticipate negative outcomes for their efforts, but these too must be considered. In their short history, some digital libraries have been built under misguided assumptions about the existence of eager users who never materialized. For example, "Contentville.com, a commercial digital library enterprise into which millions of dollars were invested, failed after little more than a year of its release, and the URL is now for sale for a mere twenty thousand dollars. An evaluation focused on the existence of potential audiences might have avoided the loss of these enormous investments. Such an evaluation might have also revealed the fallacy that people would be willing to pay for resources from one website that could be found for free at others.

In most instances, you won't be able to create an exhaustive list of all the decisions that must be made about a digital library or its features. Nevertheless, although there will always be unanticipated decisions, the struggle to identify decisions up front is certainly worthwhile. Unless you strive to identify decisions in advance, your evaluation activities are simply not going to be as influential as they could be.

There are many different types of decisions surrounding digital libraries. Some involve the nature of collections, e.g., how stringently should they be reviewed? Others involve the basis for sustainable funding, e.g., should user fees be established? Still others involve the provision of service, e.g., should all service functions be automated or will human-to-human interactions be enabled?

Most decisions about digital libraries will not be at the scale or scope of the issues described above, but will involve more localized challenges. For example, myriad decisions will be required in the process of designing and refining the library's interface. Should icons be used in place of text buttons? How can the interface be designed to meet accessibility guidelines? How can an interface accommodate multiple audiences with varying levels of literacy? What fonts, colors, menus, and other graphical user interface (GUI) factors should be utilized?

Why kinds of questions should you address?

Once you begin to reveal important decisions that must be made about a digital library or some subcomponent of it, you can identify the questions that must be addressed to provide the information needed by the decision makers. The clearer and more detailed your evaluation questions are, the more likely that you will be able to provide reliable and valid information to your decision makers.

Suppose your evaluation is focused on decisions about the types of services that your digital library should include. Some of the questions that might be addressed are:

- What services are offered by other digital libraries?
- What services have patrons of your digital library requested?
- What services would your patrons be willing to pay for?
- What services would require external funding?

Obviously, this is only a partial list of the questions that might be addressed within an evaluation focused on decisions about services. One challenge in evaluation planning is limiting the questions to those that are the most relevant to the decisions that must be made without exceeding the time, money, and other resources allocated for evaluation. In most cases, there will be far more questions that could be asked in evaluating a digital library or its subcomponents than your resources will allow, and therefore some difficult choices must be made about which questions will actually be addressed. You should make these choices in collaboration with your stakeholders well in advance of any evaluation data collection activities.

Why kinds of methods should you use?

The delineation of important decisions and unambiguous questions is essential before deciding upon the methods of evaluation. Unfortunately, too many evaluations start with the specification of methods. People think: “We have to do an evaluation. Let’s begin by designing a survey.” Survey methods are just one option for evaluation methods, and this Guide has been written to provide you with a better understanding of how to choose the most appropriate methods for answering your evaluation questions and ultimately informing decision making.

It helps to think of evaluation methods as tools. Just as you would not select a carpentry tool (hammer, saw, or plane) before understanding the nature of the task you need to accomplish, you should not choose evaluation methods until you are as clear as possible about the questions you need to answer in order to inform the decision-making process. There are numerous evaluation methods (e.g., usability testing) and even more specific data collection strategies (e.g., keystroke analysis) that can be used within any given method. One key to successful evaluation is matching these options to the decisions and questions of your stakeholders while adhering to the budget and timeline limitations of your situation.

Most evaluations will demand multiple methods (Mark & Shotland, 1987). You will often need to "triangulate" your findings. You can triangulate findings by using more than one method to collect data related to an evaluation question. For example, suppose you are trying to decide whether the search functions in your digital library should only search across resources in the collections that are part of your library, or allow searching across the entire World Wide Web (WWW). A question you might address to inform this decision could be "What are middle school teachers' attitudes toward the use of open searching on the Internet?" An emailed questionnaire designed to elicit teachers' views about Internet searching would be one way of collecting that data, but most people, including teachers, are turned off by questionnaires and wary of sharing information via email. Thus, they may not provide you with sufficiently detailed information about their real opinions about this matter. A better strategy might be:

- conduct a series of focus groups with teachers, administrators, and parents about the pros and cons of open searching by middle school students,

- review the policies established by a representative sample of school districts concerning Internet access by students, and

- review the professional library literature concerning recommendations for Internet search policies and procedures.

Most carpentry jobs require multiple tools, and similarly, most evaluations require multiple methods. There are many examples of library evaluations that have employed multiple methods. For example, Norlin (2000) employed surveys, unobtrusive observations, and focus groups to evaluate user services in a university library.

How should an evaluation plan be organized?

Planning an evaluation requires political savvy and astute negotiation skills. Just as politicians must engage in persuasion and negotiation to get anything accomplished within legislative bodies, evaluators often find themselves in the position of having to persuade their stakeholders of the value of anticipating difficult decisions and asking hard questions in an evaluation. Unwilling to confront the complexities involved in most evaluations, stakeholders in a digital library may demand direct and simple answers to complex questions. However, simple answers to complex questions are extremely rare, and "it depends" and other conditional statements are inherent in even the best evaluations.

A sound evaluation plan will expose as many of these conditionals as possible up front, but the trick is doing so without having the clients decide to abandon evaluation altogether. Therefore, an evaluation plan should be presented in a straightforward, easy-to-read manner. You can organize your evaluation plan using the following sections:

- Introduction
- Background
- Stakeholders
- Purposes
- Decisions
- Questions
- Methods
- Sample
- Instrumentation
- Limitations
- Logistics
- Budget

Introduction

The Introduction section introduces the major sections of the plan as well as the primary people involved in preparing the plan. It informs the reader about the type and amount of information upon which evaluation planning has been based, both in terms of human input and review of other materials. Here is a brief example of an Introduction section:

> INTRODUCTION:
> This document describes the background, purposes, stakeholders, decisions, questions, methods, sample, instrumentation, limitations, logistics, and budget for the evaluation of the Digital Library for Engineering Education (DLEE) being developed by the North American Association of Engineering Professors (NAAEP) with funding from the American Science Foundation (ASF). The design, methods, and instrumentation included in this plan are based on three day-long meetings between members of the DLEE development team (Jane Jones, Sam Smith, and Wanda Watson from DiglibRUS.com) and the evaluation team (Bill Biggs and Tracey Toliver from North Island University), as well as a review of the original DLEE funding proposal and numerous draft design documents accessible on the DLEE development website. This is the first draft of an evaluation plan that is being released for review by members of the DLEE stakeholder community.

The Introduction section of an evaluation plan should answer the following questions:

- What is included in this plan?
- Who prepared it?
- What information was accessed in the planning process?
- What is the status of the plan, e.g., preliminary draft or final?

Background

The Background section of the evaluation plan should describe the information needed to provide stakeholders with an understanding of the background of the digital library or its subcomponents being evaluated. This section should provide enough information to convey the nature of whatever is being evaluated, but not so much detail as to overwhelm readers. Explain any jargon used in describing the digital library (e.g., metadata), especially if the plan will be read by stakeholders unfamiliar with technical terms. Although most evaluation plans make for dry reading, it does not have to be that way. Your evaluation plan can tell an interesting story, and you can include screen images from the digital library or its components to clarify its nature. If lengthy background materials are needed, consider putting them in Appendices. Here is a much abbreviated example of a Background section:

> BACKGROUND:
> DLEE is a digital library that provides engineering faculty with free access to high quality interactive learning resources. With a planning grant of 2.3 million dollars from the ASF that commenced in November 2002, the first beta version of DLEE.net was released in September 2003. DLEE currently contains education resources primarily related to undergraduate engineering education, but future plans include the provision of interactive learning resources for both graduate engineering education and continuing professional development. All of the resources currently available via DLEE have been validated by the members of the North American Association of Engineering Professors (NAAEP). In addition, most of the resources in DLEE contain reviews by practicing engineering faculty members who have previously used the resources in their teaching. Appendix A contains screen captures from the beta version of DLEE illustrating the interface and interactive features of this digital library.

The Background section of an evaluation plan should answer the following questions:

- What digital library is the focus of this evaluation?
- What is the current status of the digital library?
- Who uses this digital library and why?

Stakeholders

The Stakeholders section of the evaluation plan describes the primary and secondary audiences or consumers of the evaluation. Patton (1997) recommends the use of the term "stakeholders" to designate evaluation audiences, and we have adopted it within this Guide. Patton wrote "…stakeholders typically have diverse and often competing interests" (p. 42). Competing interests in an evaluation should not be obscured, and therefore, you are advised to share information about an evaluation with as many stakeholders as is technically possible and politically feasible.

Primary stakeholders include the people most directly involved in or affected by an evaluation, e.g., representatives of the agencies that funded the digital library, its developers, and its intended patrons. Secondary stakeholders encompass any people who may have an interest in the evaluation or who have a right to know about its methods and results, e.g., students who might be expected to learn using resources located through a digital library. Which stakeholders will receive evaluation plans and reports may even become a major focus for negotiation between you and your clients, i.e., the people paying for or commissioning the evaluation. This is especially likely when the results of an evaluation are expected to inform hard decisions about how resources should be allocated. Here is a brief example of a Stakeholders section:

> STAKEHOLDERS:
> The primary stakeholders in this evaluation are the members of the DLEE design and development team at DiglibRUS.com, the project officers at ASF, and executives at NAAEP. Important secondary audiences include all external consultants involved in the DLEE development effort (e.g., content experts and metadata specialists) as well as the larger engineering education community. The designers and implementers of this evaluation are Bill Biggs and Tracey Toliver, evaluation specialists from North Island University.

The Stakeholders section should answer the following questions:

- Who are the primary stakeholders for this evaluation?
- Who are the secondary stakeholders for this evaluation?
- Who is responsible for evaluation planning and implementation?

Purposes

The Purposes section of the evaluation plan thoroughly describes the rationale and goals of the evaluation. An evaluation can address a variety of purposes, but all must be delineated clearly. Because evaluation is inevitably a political process, all stakeholders should seek consensus about its purposes if it is to succeed. According to evaluation experts, there are two primary types of purposes, formative and summative. An evaluation with formative purposes is primarily aimed at providing information to inform decisions about how to improve whatever is being evaluated, e.g., how can the graphical user interface of a digital library be made more user-friendly? An evaluation with summative purposes, on the other hand, is primarily aimed at informing decisions related to the worth or merit of whatever is being evaluated, e.g., should another year of funding be extended to a digital library initiative? Many evaluations will have both formative and summative purposes. Here is a brief example of a Purposes section:

> PURPOSES:
> The overall purpose of this evaluation is to provide decision makers at the North American Association of Engineering Professors (NAAEP) with the timely, accurate information required to support decisions regarding the enhancement, expansion, and promotion of the beta version of DLEE. A list of anticipated decisions is presented in a separate section below. As a result of this formative evaluation and the decisions and actions stemming from it, DLEE should be ready for Version 1.0 release in the third quarter of 2004.

The Purposes section of an evaluation plan should answer the following questions:

- Why is this evaluation being done?
- Is this evaluation primarily formative (to improve), summative (to judge merit or worth), or a blend of both formative and summative goals?

Decisions

The Decisions section of the evaluation plan is usually the most difficult part of a plan to prepare, but it must be included if the evaluation is to have a sufficient impact on decision making. This Guide is based upon what may seem like a simplistic premise: **Decisions informed by sound evaluation are better than those based on habit, ignorance, intuition, prejudice, or guesswork**. Although the history of digital libraries is still young, experience indicates that far too often poor decisions are being made about the design and implementation of digital libraries because critical decision makers lack pertinent information when they most need it. For example, "Build it and they will come" appears to have been an underlying assumption in failed digital libraries such as Contentville.com. If more effort had been made up front to expose such assumptions and collect information related to decisions about audience identification and income revenues, millions of dollars might have been invested more wisely. Here is a brief example of a Decisions section:

> DECISIONS:
> If this evaluation is to provide timely, accurate information to inform decision making regarding the improvement of DLEE, we must anticipate decisions that will be made. It is important to remember that most of these decisions have to be made regardless of the quantity and quality of information available to the decision makers at NAAEP and DiglibRUS.com. Therefore, it is essential that evaluation be conducted efficiently so that decisions are informed in a timely manner. The following decisions are anticipated:

1. The beta version of the graphic user interface for DLEE must be enhanced so that all members of the intended user community are empowered to access engineering education resources with ease.
2. The basis for sustained funding for DLEE after the ASF support is exhausted must be identified.
3. Mechanisms for recognizing the volunteers who validate DLEE resources must be identified so that they receive proper acknowledgement within their academic departments.
4. Whether to include the DLEE collection within larger scale digital libraries must be decided.

The Decisions section of an evaluation plan should answer the following questions:

- What decisions are pending regarding the digital library being evaluated?
- Who are the primary decision makers?

Questions

The Questions section of the evaluation plan should flow naturally from the Decisions section. For each decision that the evaluation should inform, there will be one or more (usually several) questions that the evaluation must address. The answers to these questions provide the essential information the decision makers need to make their decisions in a timely manner. Questions will rarely be posed in a form that can be answered with a simple Yes or No response. The issues involved in digital libraries are usually too complex for simple questions. Instead of asking a question such as "Is there a real audience for this digital library?", an evaluation question might be posed in this way, "What evidence can be provided that describes the nature and size of the likely audience for a digital library focused on engineering education?" Here is a brief example of a Questions section:

QUESTIONS:
The following questions will be addressed to inform decisions related to the redesign of the graphical user interface of DLEE:
a. How do digital library experts judge the effectiveness and efficiency of the DLEE GUI?
b. How do graphical design experts judge the effectiveness and efficiency of the DLEE GUI?
c. How do members of the user populations of engineering education faculty judge the effectiveness and efficiency of the DLEE GUI?
d. What enhancements are recommended for the DLEE GUI?
e. What costs are associated with the feasible enhancements to the DLEE GUI?

The Questions section of an evaluation plan should answer the following questions:

- What questions must be addressed to answer all of the decisions that the evaluation is intended to inform?

- How do the various evaluation questions align with the anticipated decisions?

- What priorities, if any, can be established for addressing the various questions?

Methods

The Methods section of the evaluation plan spells out the overall evaluation design and data collection strategies to be employed. There are scores of designs and many more data collection strategies that can be used. Unfortunately, traditional evaluation textbooks do not provide sufficient practical guidance in the area of methodology because the examples they commonly include are based upon the assumption that one design will suffice (e.g., a quasi-experimental design that could be used to compare a digital library with a traditional one). If money and time were unlimited, it might be possible to carry out large-scale experimental evaluations, but this is rarely the case. Instead, you will be lucky if you can use several different smaller-scale methods such as usability testing, expert review, and user surveys to collect the information needed to answer your evaluation questions and ultimately inform the decision-making process. Here is a brief example of a Methods section:

> METHODS:
> No single evaluation design can encompass the major questions specified for the evaluation of DLEE. Therefore, a variety of evaluation designs and methods will be utilized to collect the information required to address these questions. The data collection methods include:
> a. heuristic evaluation
> b. usability testing
> c. expert reviews
> d. user focus groups
> e. keystroke tracking
> f. user questionnaires

The Methods section of an evaluation plan should answer the following questions:

- What methods will be used in the evaluation?

- How are methods aligned with the evaluation questions that are, in turn, aligned with the decisions the evaluation must inform?

One way to illustrate how your evaluation methods align with your questions is to use a matrix. The matrix below illustrates the relationship between specific questions and the data collection methods used in the evaluation of the graphical user interface (GUI) of a hypothetical digital library. On one axis of the matrix are listed the abbreviated versions of the questions to be addressed by the evaluation. Listed on the other axis are the appropriate data collection methods (i.e., reliable, valid, and feasible) for this particular evaluation. An advantage of using a matrix is that you, your colleagues, your clients, and other stakeholders can review the alignment between the evaluation questions and the proposed methods of collecting data. It also allows you to ensure that each question is addressed by one or more data collection methods. Although it is not always feasible in every evaluation, it is desirable to triangulate most questions with more than one evaluation method.

Methods **Questions**	Heuristic Evaluation	Usability Testing	Expert Reviews	User Focus Groups	Keystroke Analysis	User Questionnaire
a. DL Expert Perspectives?			X			
b. GUI Expert Perspectives?	X		X			
c. User Perspectives?		X		X	X	X
d. Recommended Enhancements?	X	X	X	X	X	X
e. Enhancement Costs	X		X			

This sample matrix is by no means an exhaustive list of all the evaluation data collection methods that could be employed in such an evaluation. Other chapters in this Guide provide examples of additional methods.

Sample

The Sample section of the evaluation plan specifies the digital library users, information scientists, subject matter experts, and other people from whom data will be collected. They are also called the evaluation participants. Except in rare situations, it is not possible to collect information from everyone in any given population of potential participants. Therefore some sort of sampling is required whereby a subset of the population is selected to represent the information that would be collected from everyone if that was feasible. One way of sampling is to use some sort of random selection process, but this is impractical in most real world evaluation contexts. You will want to put some serious thought into your sampling plan. Involving people in an evaluation should not be done carelessly because you are asking for their valuable time and energy.

The nature of your sampling strategies will vary considerably depending upon the methods selected and the status of the digital library being evaluated. For example, early in the stages of development of a digital library, fewer participants will be involved for longer and more intensive evaluation sessions. On the other hand, when a digital library is ready for beta testing, it can be shared with large numbers of reviewers who might try it out and complete a pop-up questionnaire about it. Here is a brief example of a Sample section:

SAMPLE:
The participants in this evaluation will include a:
- non-random sample of people who log into DLEE during first quarter 2004
- panel of digital library experts identified by the editor of D-LIB Magazine
- panel of engineering education experts identified by the NAAEP Board
- usability testing expert from Usability Gurus, Inc.

The Sample section should answer the following questions:

- Who will participate in the evaluation?
- How will the participants be identified and recruited?

Instrumentation

The Instrumentation section of the evaluation plan describes the measurement tools to be used in the evaluation. Copies of the instruments can be included in appendices for review by your clients or others. The descriptions in this section should provide enough information to permit readers to judge the various purposes and uses of instruments such as questionnaires, interview protocols, and observation recording tools. Some digital library evaluations will require the development of new instruments, in which case the plan may only include an outline of how the instruments will be developed. Here is a brief example of an Instrumentation section:

> INSTRUMENTATION:
> 1. A User Questionnaire will pop up on the screen after someone has interacted with DLEE for more than 30 minutes. A copy of the questionnaire can be found in the appendices. This type of pop-up questionnaire has been utilized in previous evaluations of digital libraries conducted by Bill Biggs and Tracey Toliver. Previous evaluations have yielded acceptable support for the reliability and validity of pop-up questionnaires. As illustrated in the appendices, the pop-up survey is very brief with only four questions, but it includes an invitation to link to a longer questionnaire. Previous evaluations have found that 45% of the users complete the pop-up survey, and that 35% of those go on to complete the longer survey.
> 2. Usability testing of DLEE will be conducted using a sample of engineering education faculty from North Island University. The protocol for the usability testing can be found in the appendices along with reliability and validity data. The actual testing will be conducted by two experts from Usability Gurus, Inc.

The Instrumentation section of an evaluation plan should answer the following questions:

- What measurement instruments will be used to collect data for the evaluation?
- What is the reliability and validity of these instruments?

Regardless of the types of instruments you use, issues of reliability and validity are important. The reliability and validity of instruments must be considered in light of the purposes of the evaluation (Patton, 1997). Reliability deals with the consistency of measurement of an instrument. For example, a bathroom scale that provides the same weight if you step on it ten times in a row can be said to be reliable. Validity is about the degree to which an instrument achieves its aims. For example, if you want an accurate report of your weight, the reliable bathroom scale must be calibrated with another scale of recognized accuracy. It could be giving you the same weight ten times in a row, but be off by five pounds. Any evaluator can learn the fundamentals of establishing the reliability and validity of evaluation instruments, but it may be necessary to hire measurement specialists to provide expert consultation in this area, especially when new instrumentation is being developed.

Limitations

The Limitations section of the evaluation plan describes any known limits on the implementation, analysis, interpretation, and application of the evaluation. Every evaluation has limitations, and there is often an arguable basis for alternative explanations of even the most robust findings. The Limitations section of your plan should also describe potential threats to the reliability and validity of the evaluation design and instrumentation. Here is a brief example of a Limitations section:

> LIMITATIONS:
> Two constraints on this evaluation should be clarified. First, all resources in DLEE during this evaluation should be regarded as a small sample of the resources and collections that will eventually be available. In fact, additional resources and perhaps whole new collections will be added to DLEE during the evaluation. The "moving target" nature of the DLEE should be kept in mind when interpreting the results of the evaluation. The second constraint has to do with the different perspectives of the participants in this evaluation. Some of the participants will be fulltime engineering educators who have an immediate need to access and use DLEE resources. Other visitors to DLEE are likely to be engineers seeking continuing professional development resources or engineering graduate students seeking help with their courses. These distinctive perspectives must be kept in mind when the results of the evaluation are considered.

The Limitations section of an evaluation plan should answer the following questions:

- What constraints or limitations exist that may influence data collection, analysis, interpretation, and use of the evaluation findings?

- How are constraints or limitations being handled?

Logistics

The Logistics section of the evaluation plan describes who will be responsible for evaluation implementation, analysis, and reporting. It usually includes some sort of timeline that illustrates the logical dependencies among various evaluation activities. Evaluation data are often time-sensitive. Keeping track of when, where, and how various data need to be collected requires strong project management skills, and a large-scale evaluation team may even include someone whose sole function is the management of the logistical arrangements for an evaluation.

One advantage of implementing an evaluation of digital libraries is that some data can be collected online. However, if data are collected online, it is important to make users aware of such strategies. You should also keep any evaluation data separated from any information about the user's online activity not needed by the evaluation. Here is a brief example of a Logistics section:

> LOGISTICS:
> Bill Biggs and Tracey Toliver will coordinate the implementation of this evaluation plan, including scheduling, data collection, and data handling, with the DLEE development staff. The primary point of contact will be Jane Jones, who is the DLEE project manager. All data will be processed, analyzed, interpreted, and reported by Bill Biggs and Tracey Toliver. All reports will be provided to the DLEE project manager and members of the development team. Further dissemination of the evaluation findings to stakeholders at NAAEP, ASF, and beyond will be determined by the project manager. Additional details about the logistics, including due dates for deliverables, can be found in the timeline presented in Appendix D.

The Logistics section of an evaluation plan should answer the following questions:

- Who is responsible for the logistical aspects of the evaluation such as scheduling, data collection, processing, analysis, and so forth?
- Who will receive the evaluation reports generated by the evaluators?
- How will further dissemination of the evaluation reports be controlled?
- What timelines have been established for implementation, analysis, and reporting of the evaluation?

Budget

The Budget section of the plan describes the finances for the evaluation. Evaluation is usually a people-intensive process, and therefore, most of the money spent on evaluation usually will be for dedicated evaluation personnel and/or external consultant costs. If specialized equipment and facilities such as a software usability laboratory are used, additional costs will be incurred. For example, one round of usability testing in a professional usability testing lab can easily cost $3,000 - $8,000 or more. Budgeting for evaluation is challenging because most people are reluctant to spend money for evaluation in the first place. When things get tight during a digital library development project, people often look at cutting the evaluation budget first.

Unfortunately, many digital library initiatives have not included sufficient funding for evaluation. What should an evaluation cost? One rule of thumb is to budget 5-10 % of an overall digital library development budget to evaluation. Evaluation consultants often cost $800 to $2,000 per day depending upon their expertise and experience. It is sometimes feasible to hire graduate students from nearby universities to carry out many of the data collection duties that might otherwise be done by a higher paid consultant. Here is a very simple example of a Budget section:

BUDGET:

ITEM	RATE	AMOUNT	COSTS
Biggs & Toliver Consulting	$1,000 per day	10 days	$10,000
Expert Review Honorarium	$1,000 per expert	5 experts	$5,000
Usability Lab Rental	$5,000 per day	2 days	$10,000
Travel & Per Diem	$1,000 per trip	2 trips	$2,000
Printing, Communications, etc.	$250 per month	12 months	$3,000
TOTAL			$ 30,000

The Budget section of an evaluation plan should answer the following questions:

- What items are included in the budget?

- What are unit costs per item?

- What are the total costs?

What's next?

This chapter has stressed the importance of identifying the decisions to be affected by a digital library evaluation up front, and then aligning evaluation questions and methods with those anticipated decisions. A model of how to prepare an evaluation plan has also been presented.

Now it is time to go into detail about various approaches to digital library evaluation such as service evaluation, usability testing, and information retrieval. Careful review of these approaches will provide the guidance you need to plan, implement, and report your evaluation activities in such a way that they impact important decisions in a timely manner.

Chapter

3

Service evaluation

> "In an ideal world, with unlimited resources, it would be possible to provide a full range of digital library services to all users. In reality, resource constraints require a consideration of priorities. Consequently, it would be useful to evaluate potential benefits, as determined by patrons and end users, regarding digital library services."
>
> - Choudhury, Hobbs, Lorie, & Flores, 2002

Service is what people appreciate most from a library, virtual or otherwise. Reference librarians are one of the primary points of contact most people have with library collections, and the experience can range from very disappointing to incredibly rewarding. Unobtrusive evaluation of reference librarian services in physical libraries has indicated a search satisfaction rate of less than 50% (Dilevko & Dolan, 1999). Can digital libraries do better, especially given that they depend largely on automated reference services? This chapter provides guidance for evaluating reference and other services provided by digital libraries.

What is service evaluation?

Service evaluation within a traditional library system is focused on evaluating how effective a library is in carrying out its mission (Marchionini, 2000). For a digital library, the focus is very much the same: How is the digital library carrying out its mission and providing service to its users? Anyone involved in digital libraries knows that this sounds much simpler than it is. Whereas librarians working in a physical library can see users, request feedback, and observe interactions with various service functions, people who work with digital libraries lack these opportunities. Nonetheless, as described below, there are effective strategies that can be used for service evaluation in the virtual environment of digital libraries.

Evaluating how well a library is meeting its service goals inevitably requires obtaining user feedback. Depending on the decisions to be made and the resources available for evaluation, a service evaluation can take many forms. One promising methodology is called the multi-attribute, stated-preference economic model (Choudhury, Hobbs, Lorie, & Flores, 2002). This technique involves the use of surveys that are based on the idea of providing users with "choice experiments" in which they get to state which alternatives they prefer (for services or features). It is important that the alternatives that are offered have multiple attributes and are realistic and credible so that users can make meaningful choices. The multi-attribute, stated-preference economic model has been used widely in marketing research focused on predicting the demand for new products. More information about this technique is provided below.

How do you do a service evaluation?

So how would you actually go about conducting a multi-attribute, stated-preference survey? Well, let's explore a simple example of what it might entail. Suppose you are considering providing additional services to support your digital library. For example, you wish to decide on the viability of having a reference librarian available to provide "live" help for users. There are a number of factors that would be associated with providing this additional service to your library, and a good way to gather information about how this service would be viewed by typical users is to use a multi-attribute stated-preference survey. Here are the steps involved in this approach:

Step 1: Identify the different service attributes for which you want user input.

For our example some attributes might include:

- Number of hours reference librarian is available for live help
- Method of communication between librarian and patron (i.e., online chat, email, or telephone?)
- Price users are willing to pay to have a live reference librarian

Step 2: Identify levels of attributes that you want to explore.

One good way to do this is to create a chart:

Choice Attributes and Levels	
Attribute	**Range of Levels**
Reference librarian availability	4, 8, 12, 24 hours
Communication mode	Email, telephone, online chat, WebX
Price for service	Per use/monthly/3 months/6 months/yearly subscriptions (You could also include actual prices to see what people are willing to pay.)

Step 3: Design the survey

Surveys typically begin with easy questions, to help focus the participant on the subject at hand, before moving into the more "thought provoking" questions. (Guidelines for constructing surveys can be found in Chapter 8 of this Guide.) You may want to begin your survey with questions that gather demographic information (e.g., age, position, etc.) or information about the respondent's knowledge of or use of the digital library. In the next section of the survey, you can begin presenting your choice experiment. You'll want to be sure that the options you list in your questions are ones that would be within the range of possibilities. You certainly don't wish to create expectations for a new service that is simply not feasible.

Here is an example of what a question might look like using our reference librarian example:

Example question:

Which of the following systems do you prefer?

(a) Existing system, no reference librarian, no extra costs

(b) Reference librarian available 8 hours/day; available via email; pay fixed price per use

(c) Reference librarian available 8 hours/day; available via online chat; pay fixed price per use

(d) Reference librarian available 8 hours/day; available via telephone; pay fixed price per use

This question specifically explores what *type* of contact (if any) users would prefer to have with a reference librarian given that it would only be available for 8 hours/day and would have a fixed price. A good multi-attribute, stated-preference survey would explore all of the possible combinations of hours available (4 levels), methods of communication (3 levels) and price for use (5 levels). Because there are so many available combinations (4x3x5 = 60), you would probably not want to present all 60 choices to every user, because by the end of the survey all the choices would begin to look the same! But by assigning a set number of questions (perhaps 20) for each user to provide feedback, you could get the information you want, provided the options are logically presented across the range of different surveys.

Step 4: Administer survey, analyze results, and present findings.

The final steps in conducting a multi-attribute, stated-preference survey would be to administer it to the relevant group of people, analyze the results, and present your findings to the decision makers in a timely manner. Today, many surveys are administered over the WWW. This seems particularly appropriate if you are surveying typical users of a digital library. However, if you are seeking input from non-users, having the survey available via the Internet alone may introduce an unacceptable level of bias because people who are not online won't be able to access your survey. You'll need to consider alternative methods of collecting the surveys such as regular mail, telephone, or person-to-person. Recommendations about analyzing and presenting your findings can be found in Chapter 10.

When attempting to influence decisions about the types of services to be offered in a library, digital or otherwise, it only makes sense to consult with the users. Using the multi-attribute, stated-preference technique is an insightful way to gather information from actual users of the digital library about the services they value and the ones they would like to have.

Other common techniques for evaluating the services of a digital library include the use of interviews, focus groups, and observations to determine patron satisfaction with various library services. One important thing to note when doing user evaluation of services is to take into consideration the difference between use value and option value (Choudhury et al., 2002). Use value refers to the value attributed to a service or feature by actual users of the service. Option value is the value attributed to a service or feature by individuals who might use the service in the future but do not currently use it. Often individuals may place a high value on a service even if they currently do not use it.

If influencing decisions related to services provided by the digital library is a goal, then focusing your evaluation on service performance based on user feedback is the way to go. Establishing the degree of user satisfaction with existing services and revealing user desires for alternative services through evaluative methods should guide future decisions concerning the services your digital library will provide. Of course, you don't want to ask your users to respond to surveys too frequently or they'll ignore them. Only do a service evaluation when important decisions need to be informed.

Service evaluation case study

An example of an evaluation that was done using the multi-attribute, stated-preference technique is the evaluation of the Comprehensive Access to Printed Materials (CAPM) project at John Hopkins University (Choudhury et al., 2002). The CAPM project involves a robotic retrieval system that provides users with the ability to view, full-text search, and scan materials that are shelved off-site. The CAPM system allows a user to control a robot at an off-site shelving location to retrieve materials, and bring the materials to a scanning station where the user can actually browse the materials to determine whether to request the materials be delivered or returned to the shelf.

The evaluation team used the multi-attribute, stated-preference technique to determine users' preferred combination of service level and price from 36 possible options. Factors considered in these options included: presence or absence of digital images; presence or absence of full-text search; delivery time to receive materials from off-site location; and price per semester for using CAPM service. A Web-based choice survey was constructed and refined. A total of 2,000 randomly selected John Hopkins faculty, students, and staff were invited to participate. An incentive for participation was offered (a chance to win a $500 travel certificate), and eventually 603 people responded, reflecting a 30% response rate.

From the survey results, the evaluation team was able to determine an approximate amount of money users were hypothetically willing to pay for services provided by CAPM (Choudhury et al., 2002). This information was used to inform the decision to continue development of the CAPM robotic system at John Hopkins University. More information about the project is available at: http://dkc.mse.jhu.edu/CAPM/.

Print references

Although there is not an extensive literature to be found on the topic of service evaluation in libraries, one useful text reference is *Library Evaluation: A Casebook and Can-Do Guide* edited by Danny P. Wallace and Connie Van Fleet (2000). Published examples of service evaluations include Pettigrew and Durrance (2001) and Norlin (2000).

Online references

The best available reference for service evaluation in the digital context is *A Framework for Evaluating Digital Library Services* (Choudhury et al., 2002) available online at: http://www.dlib.org/dlib/july02/choudhury/07choudhury.html. It provided much of the material used in this chapter.

Another useful online reference is *Emerging Tools for Evaluating Digital Library Services: Conceptual Adaptations of LibQual+ and CAPM* (Heath, Kyrillidou, Webster, Choudhury, Hobbs, Lorie, & Flores, 2003) available online at: http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Heath/.

Online instruments, tools, guidelines, etc.

Some useful service evaluation tools can be found at: http://www.si.umich.edu/libhelp/toolkit/index.html

Although intended primarily for traditional libraries, other tools can be found at: http://www.libqual.org/index.

Chapter

4

Usability evaluation

> "To discover which designs work best, watch users as they attempt to perform tasks with the user interface. This method is so simple that many people overlook it, assuming that there must be something more to usability testing. Of course, there are many ways to watch and many tricks to running an optimal user test or field study. But ultimately, the way to get user data boils down to the basic rules of usability:
> - Watch what people actually do.
> - Do not believe what people say they do.
> - Definitely don't believe what people predict they may do in the future."
>
> - Nielsen, 2001

Usability, in the context of digital libraries, can be defined as the effectiveness, efficiency, and personal satisfaction with which people are able to access and make productive use of the resources in a digital library. People interact with a digital library (and most other computers programs) through some sort of human-computer interface. The importance of the interface and the functionality it enables cannot be overemphasized. William Y. Arms (2000), author of a definitive text on digital libraries wrote: "A digital library is only as good as its interface" (p. 160). Thus, the usability of a digital library's interface and functionality should be evaluated and enhanced to the greatest degree possible.

Shneiderman (1987) maintains that the usability of any type of computer program is determined by a combination of five user-oriented characteristics: (1) ease of learning, (2) high speed of user task performance, (3) low user error rate, (4) subjective user satisfaction, and (5) user retention over time. With reference to digital libraries, this means that (1) learning how to access the resources in a digital library should be intuitive or easy-to-learn, (2) finding a desirable resource should take minimal time, (3) errors of omission (not finding what the user wants) or commission (finding the wrong things) should be rare, (4) searching should be a pleasant and rewarding experience, and (5) returning to the digital library within a reasonable time should not require learning the user interface all over again.

Usability is about much more than the "look and feel" of the digital library. The interface of a digital library should communicate its functions and navigational structure to new users with a minimum of "cognitive overload." In other words, novice users should be able to devote most of their thinking to the task at hand (e.g., reviewing various educational resources to find the best one to meet specific instructional needs) rather than to the task of figuring out how to search within the collection of resources in the first place. There are many challenges that novices are bound to face, including:

- unfamiliarity with functionality of computers in general,
- lack of information-age skills such as effective search strategies, and
- unfamiliarity with the interface and functionality of the library being used.

Usability is not just a concern for new users. Frequent patrons of a digital library expect to pick up where they left off. If new features have appeared since their last visit, these must be communicated without distracting from the patrons' reasons for accessing the library. Arms (2000) maintains that digital libraries frequently change their interfaces. Obviously, this is not done for malicious reasons, but with the intention of enhancing users' experience, effectiveness, and efficiency. However, design changes are far from infallible, and thus any substantive modifications should be carefully evaluated.

What is usability evaluation?

There are numerous methods that can be used to evaluate the usability of a digital library. Usability evaluation methods can be classified as belonging to one of three categories: inspection, testing, and inquiry.

Usability "inspection" refers to a number of processes whereby experts systematically review the usability of a digital library and recommend improvements. Two such processes are described below: heuristic evaluation and cognitive walkthrough.

Usability "testing" refers to evaluative processes whereby the interface that enables human-computer interactions are systematically tested and enhanced. Usability testing can be done in a professional usability laboratory, locally using a portable usability lab, or even with standard video equipment. Typically, usability testing involves having people follow predetermined protocols so that specific aspects of a digital library's usability can be evaluated. The "think aloud" approach to usability testing is described below in more detail.

Usability "inquiry" refers to processes that are somewhat like usability testing, except that in usability inquiry evaluators observe users working with digital libraries while doing real work rather than evaluator-assigned tasks. There are a number of evaluative methods that fit within the usability inquiry framework including: field observations, focus groups, interviews, use logs, and questionnaires.

How do you do usability evaluation?

Usability Inspection

There are two common approaches to usability inspection: heuristic evaluation and cognitive walkthrough. As conceived by Jakob Nielsen (1993), arguably the world's most famous usability expert, the heuristic evaluation method employs a set of principles (termed heuristics) which have been defined prior to the evaluation. Heuristic evaluation is usually done with experts such as human-computer interface design specialists, digital library designers, or graphic artists. The experts independently examine the product and judge its compliance with a set of heuristic principles. Here are the original ten heuristics listed at Neilsen's website (http://useit.com):

Visibility of system status: The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
Match between system and the real world: The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
User control and freedom: Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
Consistency and standards: Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
Error prevention: Even better than good error messages is a careful design which prevents a problem from occurring in the first place.
Recognition rather than recall: Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
Flexibility and efficiency of use: Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
Aesthetic and minimalist design: Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
Help users recognize, diagnose, and recover from errors: Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
Help and documentation: Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the users' task, list concrete steps to be carried out, and not be too large.

As each expert spends time interacting with the digital library, usually two to four hours depending on the complexity of the library and its functions, he or she make notes of the features of the library interface and functionality that violate one or more of the heuristics on a predetermined list. The expert may also identify usability flaws that do not obviously match one of the predefined heuristics.

After reviewing the system, each expert usually goes back through all of the problems identified to rate each one according to its frequency and severity. Subsequently, the various experts may be brought together for a debriefing in which they compare the problems found and attempt to come to an overall recommendation concerning each problem area. The consensus of the experts might be guided by using a final rating scale as represented below.

Usability Problem Rating Scale

0: This is not a usability problem.
1: This is a cosmetic usability problem only, and it need not be fixed unless extra time is available on project.
2: This is a minor usability problem and fixing this should be given low priority.
3: This is a major usability problem, and it is important to fix so it should be given high priority.
4: This is a usability catastrophe, and it is imperative to fix this before the digital library can be released.

Another method used in usability inspection is the cognitive walkthrough. In a cognitive walkthrough, a group of expert evaluators (e.g., composed of graphic user interface specialists, software developers, etc.) work through a paper mock-up, prototype, or full version of the digital library, with the goal of completing a set of realistic tasks while evaluating the library's ease of learning, user-friendliness, and understandability. Before beginning the walkthrough, evaluators are informed of important issues such as:

- Who will be the users of the system?
- What types of tasks will typically be done?
- What is the status of the digital library (e.g., early prototype versus beta version about to go public)?

While completing the walkthrough, evaluators ask themselves questions such as:

- Will the user know what to do here?
- Will the user associate the correct action with the desired effect?
- If the correct action is chosen, will the user know he or she is on the right path?

In doing a cognitive walkthrough, evaluators should put themselves in the user's "shoes" so to speak. The goal of the walkthrough is to have the evaluators look at the digital library through a user's eyes, trying to create scenarios of where and why a user might be successful in completing a task, and also scenarios of where and why a user might experience difficulty in completing a task. Evaluators usually speak their feedback aloud as they go through the digital library. What they say can be recorded for later transcription or another evaluator may take notes as the expert is doing the cognitive walkthrough.

There are several other methods of conducting usability inspections. These include pluralistic walkthroughs, feature inspections, perspective-based inspections, and claims analysis.

Pluralistic walkthroughs (Bias, 1994) are similar to cognitive walkthroughs. But instead of just using experts, the walkthroughs are conducted with a mix of typical users, subject matter experts, and usability experts who are expected to discuss and hash out their different reactions to the program.

Feature inspections (Kahn & Prail, 1994) are expert or user reviews that are focused on specific features of a system. For example, if a digital library is being redesigned to include a much richer set of search delimiters, different approaches to enabling these functions might be targeted for inspection and feedback.

Perspective-based inspections (Zhang, Basili, & Shneiderman, 1999) involve reviewing the interface design and functionality from a variety of different viewpoints. For example, viewpoints could be from the perspective of the novice visitor, the frequent patron, or a content expert.

Claims analysis (Keith, Blandford, Fields, & Theng, 2002) focuses on identifying the positive and negative effects of a feature that may influence the usability of a digital library. The goal of this variant of usability inspection is to be able to describe the benefits and disadvantages of features, and then consider and propose alternatives that could improve the design.

Usability Testing

Usability testing involves having a number of individuals, who are considered representative of typical users, complete routine and/or special tasks in the digital library while evaluators observe and collect results to see how the interface supports the users in completing the tasks. Usability testing can occur in formal settings such as a professional usability laboratory or it can occur in virtually any other setting using portable usability equipment or only a regular video camera.

Professional usability labs generally consist of two rooms separated by a one-way glass window. In one room, a computer user sits at a desk and interacts with the application being evaluated, e.g., a digital library. Two or three video cameras mounted in the room are focused on the user from different perspectives. For example, one camera might be focused on the user's hands whereas another might be recording the user's facial expressions. In the other room, usability evaluators sit at control panels where they can simultaneously observe the user in the room through the one-way glass or any of the video screens displaying selected aspects. The user may be instructed to "think aloud" as he or she uses the digital library, e.g., talk about why certain choices are made or describe any confusion about the library's interface. Alternatively, the evaluators may question the user via headsets or speakers about why he or she has acted in certain ways.

In setting up a usability test, users are informed that they will be observed, and they have the right to discontinue a test at any time for any reason. Typically, these sessions are videotaped for later analysis and documentation. Some professional usability labs actually have a third room where clients can observe through a one-way glass the usability testing as it is being conducted

A portable usability lab is much simpler than commercial usability laboratories, and it has the advantage of allowing users to stay in their own environment rather than forcing them to come to a lab and test the digital library in an artificial environment. This may increase the validity of the usability test.

You can even do usability testing with only a single video camera, especially if you cannot afford to rent a professional laboratory or buy a portable usability lab. For example, a single video camera can be used to record two users while one explains to the other how to use a digital library. This has some unique advantages. The two people have to talk as one explains how the digital library works and the other asks questions. Reviewing the video record of these interactions can be very informative. You can interpret the adequacy of the mental model of the library held by the person doing the instruction and also estimate how hard the interface is to learn by the questions asked by the other user.

Regardless of what type of lab is used (professional or portable), usability testing enables evaluators to collect both quantitative and qualitative data related to issues such as user interface, mental models, navigation, ease-of-learning, documentation utility, effectiveness, and efficiency. There are a variety of protocols used in usability testing. One of the most common usability testing methods is the *think aloud protocol.*

The think aloud protocol is useful in gathering qualitative data about users' mental models of a digital library, as well as their general impressions and feelings about library factors such as layout, navigation, and design. Below are guidelines for implementing the think aloud protocol:

1. Begin by finding participants who are representative of a typical user. Typically, a small number (3 – 5) of participants will be sufficient to gather the type of information needed to start informing decisions about improving the interface of the digital library.

2. Create a set of tasks (called a script) for your test users to work on during the usability testing These tasks should be typical of the kinds of tasks you expect real users to complete while using your digital library system.

3. Ask test users to think aloud while they complete the tasks. The more vocal a test user is, the better chance you have of gaining deeper insight into how other users will approach and interact with your library. Most users become tired after about an hour of testing.

4. Although an evaluator is usually present and taking notes during the testing, the testing session is often videotaped as well. This way you will have a permanent record to return to, allowing for a more in-depth analysis after the session. Notes can be taken freehand or with the aid of special usability testing software such as UsabilityWare™ 4.0 from UsabilitySystems.com.

5. After a user has completed a think aloud protocol session, there will often be a debriefing session during which the results are reviewed. At that time, adjustments to the protocol may be made before doing another test with the next user.

You may want to ask questions to users during observations, but asking questions during a usability test can change what the user would naturally do. An alternative is a delayed think aloud approach whereby you video the user, and later play the tape back to the user. During the playback, you ask the user to state what he or she was thinking while interacting with the digital library. You can ask specific questions such as "Why did you decide to use those search terms?" The tape assists the user in recalling the recorded session. Later, the same tape can be shown to other experts for their advice and interpretations. Alternatively, a focus group of designers can review the videotapes of users to stimulate new ideas about enhancing the interface of the digital library.

There is a lot more to usability testing than can be described in this brief Guide. If you are going to engage in serious usability testing, we recommend going to a workshop on the subject and perhaps hiring a consultant your first time out. If you get even more serious about usability testing, you should join the Usability Professionals Association (http://www.upassoc.org/) and attend one of their annual conferences.

Usability Inquiry

Usability inquiry is much like usability testing, except that in the former the evaluator observes users working with the digital library while trying to complete their own real work rather than tasks defined by the evaluator. You will want to use usability inquiry methods if you are interested in gathering information about users' likes, dislikes, needs, and understanding of a digital library. Indeed, some experts maintain that usability inspection methods are better than usability testing approaches because only the former allow you to evaluate the user, the tasks, and the working environment at the same time (Hackos & Redish, 1998). There are a number of evaluative methods that fit within the usability inquiry framework including: field observations, use logs, focus groups, interviews, and questionnaires.

Field observations require you to schedule visits with real or potential users in the workplace or home where they would normally access a digital library. For example, if you want to understand how teachers would access a digital library, it would be useful to observe them in their schools. You may want to begin your observations of each teacher with a brief "get-acquainted" interview, spend some time observing the teacher using the digital library, and then conduct a debriefing interview.

User logging is a method that involves having computers record what people actually do when they are using a digital library. It is relatively easy to collect user statistics with user logs, but much more difficult to make sense of the data so that actual enhancements to an interface can be made. Some statistics should be routinely logged such as error messages or time patrons spend using specific resources. When and where people go for online help can also be useful data.

Focus groups are useful when you want to collect information about a digital library's usability from a group of people who have already been using it for a while. This normally requires at least two people, one to moderate the discussion and the other to take notes. The session may also be recorded. The focus group method is useful in terms of getting users' reactions to an interface as they use it over time. But it has the disadvantage of only collecting information about what users say they do, and not what they actually do.

Interviews are similar to focus groups in terms of utility and limitations, except that instead of interviewing a group, interviews are normally with one user at a time. An interviewer questions the user, the user replies, and the interviewer records those responses using either written notes or a recording that is later transcribed. Interviewing can be relatively unstructured, although for usability inquiry, structured interviews are more common.

Questionnaires are probably the most common form of evaluation instrument used in evaluations in general, but they have limited utility in usability evaluations. One form of questionnaire that can be useful sometimes is a "pop-up" questionnaire that is programmed to appear whenever the user does something unexpected in a digital library. Alternatively, a brief on-screen questionnaire about usability issues may be initiated after a user has been using the library for a certain period of time or upon exiting the library. Questionnaires may also be emailed to users of a digital library if accessing the library requires some sort of identification protocol that would give you the email addresses of users. Regardless of how they are presented, questionnaires employed in usability inquiry should be brief and clear if you expect many people to respond. Consider including an incentive to respond such as entry into a drawing or a coupon good for online shopping.

Usability evaluation case study

Few evaluations of digital libraries have employed a sufficient blend of usability evaluation methods to reap the many benefits of these various approaches. One notable exception is the evaluation of the Virtual Data Center (VDC), a collaborative project between researchers at Harvard-MIT Data Center at Harvard University and the University of Michigan's School of Information and College of Engineering (http://TheData.org) (Hovater, Krot, Kiskis, Holland, & Altman, 2002). The VDC is an open-source digital library that is designed to facilitate the management and dissemination of social science research data.

In the evaluation of this digital library, evaluators adopted not only usability inquiry approaches such as focus groups and user surveys, but they also employed usability inspection methods such as cognitive walkthroughs and think aloud usability testing methods. The VDC evaluation team used focus groups and user surveys to gain preliminary feedback on the usability of the digital library. The goals for the focus groups method were to identify: (a) how users were conducting research, and (b) how the VDC could help them better conduct research. The goals for the user survey method were to identify: (a) current patterns of use, (b) qualities of other sites that users found helpful, (c) traits of other users seeking data, and (d) usability issues within the library (as it existed then).

After the usability inquiry, the VDC evaluation team used the cognitive walkthrough protocol to inspect the usability of the system. Finally, the VDC evaluation team employed the usability testing think aloud method to test real users' abilities to navigate their digital library system. Through the systematic use of multiple usability evaluation methods, the VDC team has been able to address many usability issues to create a more user-friendly digital library.

Print references

The classic print reference about usability evaluation is Jakob Nielsen's (1993) book, *Usability Engineering*. Nielsen has written several other texts since then, but this book remains the standard guide.

Two print references from the American Library Association are more closely related to usability evaluation of digital libraries: Nicole Campbell's (2001) *Usability Assessment of Library-related Web Sites: Methods and Case Studies* and Elaina Norlin and CM! Winters' (2002) *Usability Testing for Library Web Sites: A Hands-On Guide.*

Online references

The most popular resource available online related to usability evaluation methods is Jakob Nielsen's usability website: http://www.useit.com/.

Several papers related to usability evaluation are available at the web associated with the workshop on the usability of digital libraries held at the 2002 Joint Conference on Digital Libraries: http://www.uclic.ucl.ac.uk/annb/DLUsability/JCDL02.html.

Online instruments, tools, guidelines, etc.

Some usability evaluation tools can be found at the following site:
http://jthom.best.vwh.net/usability/.

A resource providing guidelines and resources such as example consent forms and task scripts can be found at:
http://www.infodesign.com.au/usabilityresources/evaluation/usabilitytesting.asp.

Another resource focused on usability in the context of digital libraries can be found at:
http://dkc.mse.jhu.edu/dkc_usability.html.

A resource comparing usability evaluation methods can be found at:
http://www.userdesign.com/usability_uem.html.

Another resource providing information and guidelines for conducting three different types of usability evaluation methods can be found at:
http://www.pages.drexel.edu/~zwz22/UsabilityHome.html.

Although primarily aimed at software engineers, some useful usability tools are at:
http://www.otal.umd.edu/guse/.

An example of a report of a usability evaluation can be found at:
http://eprints.cs.vt.edu:8000/archive/00000619/01/iLumina.pdf.

Chapter 5

Information retrieval

> "Users look for information for many different reasons, and they use many different strategies to seek for information. Sometimes they are looking for specific facts; sometimes they are exploring a topic. Only rarely are they faced with the standard problem of information retrieval: to find every item relevant to a well-defined topic, with the minimal number of extraneous items. With interactive computing, users do not carry out a single search. They iterate through a series of steps, combining searching, browsing, interpretation, and filtering of results. The effectiveness of information discovery depends on the users' objectives and how well the digital library meets them."
>
> \- Arms, 2000, p. 205

Information retrieval in the context of digital library evaluation is defined as finding the information (e.g., a text document, a media object, or a fact) that a user is seeking. People seek for information with many different purposes in mind. The scholar may be doing a "comprehensive search" of the American Memory historical collection (http://memory.loc.gov/) for all the existing material about the connection between President Theodore Roosevelt and teddy bears. Another scholar may conduct a "known-item search" of the Library of Congress online catalog (http://catalog.loc.gov/) to find a specific edition of George Orwell's book, *Animal Farm*. A middle school student may go to a digital library seeking a specific fact such as the surface temperature of Mars. The same student's teacher might conduct of general search of the Digital Library for Earth System Education (http://www.dlese.org/) for lessons that would help her students learn about the differences in the geology of Earth and Mars. Perhaps the most common form of searching is "browsing" whereby a user may enter a digital library with a vague information need in mind and just wants to spend time "looking around."

Information retrieval within the context of any single digital library collection is very complex, involving concepts such as metadata, cataloging, indexing and so forth. The complexity increases dramatically when a digital library is designed so that a user can search across multiple collections that may use different metadata systems for cataloging, indexing, and other functions. However, most users care little about these complexities and their inherent challenges. They just want to find the information they desire in the most effective and efficient manner possible.

What is information retrieval evaluation?

Information retrieval evaluation is twofold, user-oriented and systems-oriented. From the user perspective, information retrieval evaluation is appropriately focused on evaluating how effectively and efficiently a user's search for information meets his or her needs or interests. From the systems perspective, information retrieval evaluation is focused on evaluating the effectiveness and efficiency of the retrieval system that is at the core of any digital library.

If you have to focus on one form of information retrieval evaluation over another, the user perspective is recommended. In user-orientated evaluation, emphasis is not on the user's ability to conduct "good" searches, but rather on the user's experiences with the information retrieval tools presented by a digital library. No matter how comprehensive a digital library is in terms of the quality of its collections nor how sophisticated its underlying technology may be, a digital library is of little value if users cannot find needed information in an effective and efficient manner.

Information retrieval evaluation is also not about evaluating the technical functionality of the information retrieval system, e.g., MARC-based catalogs versus automatically generated indexes. There are numerous information retrieval systems being employed in digital libraries and a great deal of research and development is focused on issues such as the utility of metadata. While these issues are incredibly important, they are beyond the focus on information retrieval evaluation as described in this Guide. What is important for our evaluation purposes is how well the information retrieval system is performing with respect to the needs, interests, and expectations of our users.

As noted above, one of the key challenges for digital libraries is the storage, organization, and retrieval of its contents. Most digital libraries aspire to have an information retrieval system that allows users to locate items of interest in the most efficient and cost-efficient ways possible. Evaluating the information retrieval capabilities of your digital library can provide useful information for making future decisions about the theoretical and technical components of your digital library in ways that maximize the effectiveness of user searches.

How do you do information retrieval evaluation?

Several information retrieval evaluation methodologies and measures have been developed (Harter & Hert, 1997). But one of the first challenges in evaluating information retrieval is deciding upon the criteria to be used to judge success. Historically, the two primary criteria that have been used are "precision" and "recall" (Arms, 2000).

Precision is the proportion of documents retrieved that are relevant to the information an individual is seeking (i.e., meets the requirements of their search). If a scholar is searching for information related to the relationship between President Theodore Roosevelt and teddy bears, precision would be high if most of the documents retrieved are directly related to this topic, and not to President Roosevelt's role in the development of the Panama Canal.

Recall is simply the proportion of relevant documents that are retrieved from the collection of all relevant documents. Recall is much more difficult to estimate than precision because few digital libraries are cataloged in such a way that all the possibly relevant documents are known or can be identified.

In any case, Arms (2000) points out that the precision and recall criteria were originally defined in terms of evaluating a single search. Few users search that way any more because one of the major benefits of digital libraries is that they enable "interactive searching" whereby a user employs multiple iterative strategies (e.g., searching for a specific topic combined with delimiters followed by browsing followed by a new search with more specific search terms). Hence, evaluating information retrieval is much more difficult because it will usually involve evaluating interactive search sessions rather than simple, one-time searches.

So how should you proceed with respect to conducting an information retrieval evaluation of your digital library? The first issue you should consider is what kinds of decisions you hope to inform and then what kinds of information the evaluation must provide to influence those decisions. Suppose you and your colleagues are facing a decision concerning whether the metadata approach you are utilizing in your digital library should be replaced with another one or abandoned altogether. You could consider conducting an information retrieval evaluation focused on the performance of your specific system using either external standards or data from real users. Alternatively, you can do a comparison of your system with other information retrieval systems.

If you decide to focus on the performance of your system based on its use by real users, you will need to identify criteria and standards. The limits of traditional criteria such as precision and recall are obvious when evaluating today's complex interactive searches, but new criteria are evolving, e.g., search cost. Search cost is calculated in terms of the time and money that a user expends before reaching a satisfactory conclusion to an information retrieval session. Another important criterion could be relevance, i.e., how relevant does a user regard the results of using the information retrieval system? User satisfaction is another possible criterion upon which to base your evaluation of information retrieval. In this case, user satisfaction would be viewed as equivalent to effectiveness. Methods of collecting data would encompass qualitative methods such as observations, interviews, open-ended questionnaires, and even think aloud protocols similar to those used in usability testing. Qualitative approaches are necessary when the criteria are as subjective and situational as relevance and satisfaction are.

On the other hand, if you wish to compare the information retrieval of your digital library with external standards, the Text REtrieval Conference (TREC) (http://trec.nist.gov/) has several databases available for use in comparison evaluations. Indeed, TREC is one of the main venues for discovering the latest information about research in the information retrieval community. Two of TREC's main goals are: (a) increasing the speed of transfer of technology from research into commercial products demonstrating substantial improvements in retrieval methodologies on real world problems, and (b) increasing the availability of appropriate evaluation techniques for use by industry and academia. TREC is a unique evaluation community in that it has developed test collections (a set of documents, topics, query questions, and corresponding relevance judgments) and evaluation software that is available to the research community and other organizations so that any developers can evaluate their own retrieval systems at any time.

If you choose to do a comparison study, you will need to decide which system to use for comparison purposes. Do you wish to compare the information retrieval effectiveness of your digital library with another digital library or with some sort of standardized database? In the first instance, you might wish to compare the information retrieval results of your digital library with the search results obtained using a popular search engine such as Google.

Instead of focusing only on qualitative user-based evaluations or standards-based performance evaluations, Wu and Sonnenwald (1999) promote the concept of multiple-methods approaches that would blend together user-studies with systems-oriented studies. This might involve extending the criteria beyond the relevance and satisfaction perceptions of individual users to external factors such as the effects on research, productivity, and decision-making (Saracevic, 1995).

Information retrieval case study

An evaluation conducted by Berenci, Carpineto, Giannini, and Mizzaro (2000) is a notable example of an information retrieval evaluation. Berenci et al. examined how visual displays could be used to increase the effectiveness of using ranked-output retrieval systems, which are commonly used in web-based search engines such as AltaVista. They developed a system called VIEWER (VIEwing WEb Results) which acts as an interface to any selected search engine.

VIEWER acts to provide a graphical representation of the search results along side the ranked search results provided by the search engine. The visualization of the data displays red horizontal bars that each represents the number of "hits" for each sub-query (which are formed by combining the number of query terms). Users are then able to click on any of the bars to select the associated documents. The advantage of having such a system is that users do not have to search through all retrieved documents to find the information they are looking for. Rather, the VIEWER system allows users to immediately select documents that have the relevant combination of terms.

Berenci et al. (2000) decided to do a comparative study of their VIEWER system and a typical web-based search engine (in this case AltaVista was chosen). Their goal was to evaluate the effectiveness of VIEWER in comparison with AltaVista in a realistic search situation. They hypothesized that the VIEWER system would allow users to focus on the relevant document summaries and reformulate future queries. Berenci et al. decided to use precision and the raw number of relevant documents found as their criteria for comparison. Questionnaires were also used to measure user satisfaction, utility of the system, and the usage of the views in the VIEWER system.

The evaluation results showed that although AltaVista retrieved many more documents than VIEWER, the number of relevant summaries retrieved was very similar for both systems. Thus, AltaVista retrieved many more non-relevant documents than the VIEWER system; in short, it wasn't as efficient as VIEWER. Also, the precision values when using the VIEWER system were markedly better than those obtained when using AltaVista. Thus, Berenci et al. (2000) were able to use the measures of precision and relevant document retrieval as support for adopting an information retrieval system such as VIEWER.

Print references

William Y. Arms (2000) book titled *Digital Libraries* contains two chapters devoted to information retrieval, including some discussion of evaluation issues.

The *Journal of Information Retrieval* contains articles about information retrieval evaluation such as those written by Reid (2000) and Melucci (1999).

Online references

A paper related to information retrieval evaluation presented by Wu and Sonnenwald in 1999 at the annual conference of the Pacific Neighborhood Consortium (PNC) is available online at: http://pnclink.org/annual/annual1999/1999pdf/wu-mm.pdf.

An overview of what information retrieval is and evaluation methods associated with it can viewed at: http://cslu.cse.ogi.edu/HLTsurvey/ch13node4.html.

A general resource about information retrieval is at:
http://www.acm.org/sigir/resources.html

Online instruments, tools, guidelines, etc.

Some useful guidelines and tools can be found at the Text Retrieval Conference Home Page (TREC) at http://trec.nist.gov/.

Chapter

6

Bibliometrics evaluation

> "Although many have discussed the benefits digital library services bring to users and some efforts have been made to establish an objective basis for such claims, few techniques are available at present to quantitatively evaluate the impact of a digital libraries collection and the characteristics of its user community."
>
> - Bollen & Luce, 2002

Bibliometrics is a quantitative research method widely used in library and information science (Borgman, 1990). Its most popular applications are in fields such as the sociology of science (the study of how scientists and scientific communities engage with each other) and the study of scholarly communications (the study of how publications influence the development of science in a field). For example, a scholar may use citation analysis, a bibliometrics approach, to investigate the influence of a particular researcher's work within a field of study by determining how frequently the researcher is cited and the pattern of those citations.

There are numerous areas of bibliometric research. One area is focused on the application of bibliometric laws. The three most commonly applied laws are Lotka's law, Bradford's law, and Zipf's law. Lotaka's law is a measure of how frequently an author in a given field publishes. Bradford's law is used to determine the number of core journals in a field. Zipf's law is used to predict the frequency of which words will appear within a text.

Another area of bibliometric research involves citation and co-citation analyses. In traditional citation analysis, citations in scholarly works such as research articles and journal publications are examined and links between authors, articles, journals, etc. are established. Impact on a particular domain can be determined by counting the number of times a particular author or publication is cited. Co-citation analysis in the traditional sense refers to methods used to establish subject similarity between two documents. So for example, when two documents (A and B) are referenced in a third document (C), it is judged that documents A and B are related to each other because presumably they deal with the same subject matter. The more frequently both documents (A and B) are cited by other documents, the more related they are assumed to be.

With the development of the World Wide Web (WWW), citation analysis techniques have been applied to electronic environments. An area of research called webmetrics or cybermetrics focuses on using bibliometric techniques such as citation analysis to explore the relationship between documents on the WWW. Relationships between websites can be determined, and the impact and influence of websites can be mapped based on how frequently they are linked to other websites. More recently, bibliometric techniques are also being applied in the context of digital libraries to determine the impact and rankings of documents and journals for a library's user group.

What is bibliometrics evaluation?

For many years, the sheer size of a traditional library collection was used as the primary indicator of its quality, but more recently, the quality of a library has been judged on the basis of other factors such as perceptions of customer service and impact on education and research outcomes (Hernon, 2002). Although collection size is also an issue in digital libraries, it is already obvious that the quality of digital libraries will be judged more in terms of the degree to which they meet the needs of their patrons (Choudhury, Hobbs, Lorie, & Flores, 2002). Bibliometrics is a promising approach for evaluating the impact of digital libraries.

Bibliometrics evaluation in the context of digital libraries is a relatively novel approach whereby bibliometric techniques are applied to determine the impact and rankings of documents and journals for the users of a digital library (Bollen & Luce, 2002; Bollen, Luce, Vemulapalli, & Xu, 2003). Data collected using bibliometric methods can be used to inform decisions regarding acquisitions for the collection, organization of the collection, and services provided. For example, citation and co-citation analysis statistics can be helpful in providing information about the value of objects within the digital library collection. Citation and co-citation analysis of your digital library collection can highlight the objects contained in your collection that are highly regarded outside of your digital library user community as well as within your user community. One of the major advantages of bibliometrics as an evaluation approach is that much of the data needed may already be routinely collected within a digital library.

How do you do bibliometrics?

There are two general approaches you can take to doing bibliometric evaluation related to your digital library. The first approach would be to borrow the methodologies of citation analysis from traditional library and information science research. Using citation analysis you can determine which authors or documents within your library's collection are most frequently cited. You could then compare these values with established values, such as those determined by the *Social Science Citation Index*, the *Science Citation Index*, or the *Arts and Humanities Citation Index*. Such a comparison could help to serve as a measure of the quality of your library's collection. You could even apply bibliographic methods to see how often and where your digital library itself is cited as an indicator of its importance to a larger community such as educators or researchers.

However, a more useful approach to using bibliometric methods in your digital library evaluation might be to take a more user-centered approach. Bibliometric methods could be used to determine which documents and authors are most frequently accessed, and have the most impact, on your digital library's particular user community. These methods might be applied to specific resources or to collections of resources. The case study described below provides details about this method.

Bibliometrics case study

An innovative example of the use of bibliometrics in digital libraries is an evaluation study conducted by Bollen and Luce (2002). In an attempt to quantitatively evaluate the impact of a digital library's collection and services, and how well the collections and services addressed users' needs, Bollen and Luce used transaction log data to examine document relationships. They found that, by examining users' retrieval patterns, they could generate a community-specific measure of document impact. Specifically, they were able to determine which documents in the digital library were viewed as similar by users, and which documents were most frequently retrieved.

Before describing the method used by Bollen and Luce (2002) to generate a measure of document impact on a digital library user community, some of the ideas underlying their approach must be clarified. Here are the assumptions underlying their approach:

1. When two documents are retrieved in close temporal proximity, they are said to be co-retrieved.

2. Two documents would be co-retrieved because there is some level of similarity between them.

3. The strength of the relationship (similarity) between documents can be determined by the frequency with which the documents are co-retrieved by a community of digital library users.

4. Each time a given pair of documents is co-retrieved, the weight (strength) of the relationship between them can be increased by a small amount. The weight between pairs of documents is indicative of the degree of similarity between the documents as perceived by the community of users.

5. Document network maps can be constructed from the generated document weights. These networks can be analyzed to generate measures of document impact, such as the Journal Consultation Frequency (JCF), which is a measure based on patterns of usage rather than on frequency of citation (which is of special value in a digital library because it can include documents of various languages or media types).

With an understanding of the above ideas and assumptions, generating document relationships for user transaction logs is fairly simple:

1. Define what qualifies as a co-retrieval event. (Bollen and Luce (2002) defined a co-retrieval event as "a pair of sequential retrieval requests for a pair of documents by the same user within a given period of time.")

2. Sort your transaction logs by time and IP number. (Co-retrieval events can be reconstructed from your transaction logs once you have sorted your transaction logs by time and IP number.)

3. Generate a table of co-retrieval events. (Once your data are sorted by IP number and time, you can determine which events are co-retrieval events, i.e., those transactions whose date and time stamps differ by less than a pre-specified quantity. Bollen and Luce (2002) used a value of 3600 seconds.)

4. Generate weighted document relationships. (You can do this by increasing the relationship weight between co-retrieval documents by a small amount (r), every time they appear as co-retrieval events.)

5. Calculate document impact. (Document impact can be calculated using the JCF measure. JCF is the sum of the number of connections from other documents to the specified document (X) added to the number of connections from X to other documents in the library.)

Bollen, Luce, Vemulapalli, and Xu (2003) describe another application of this approach. They make a convincing argument for its utility as the developers of digital libraries face difficult decisions about acquisitions, especially when resources are tight.

Print references

Christine Borgman's (1990) edited volume titled *Scholarly Communication and Bibliometrics* is one of the few print resources devoted to this topic.

Online references

A basic introduction to bibliometrics is available online at:
http://www.gslis.utexas.edu/~palmquis/courses/biblio.html.

Another overview of the subject is available at:
http://www.du.edu/~jtwining/LIS4326/bibliometrics.htm.

Papers specifically related to bibliometrics and digital libraries available online include:
Bollen, J., & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns*, D-Lib Magazine*, *8*(6). Available at:
(http://www.dlib.org/dlib/june02/bollen/06bollen.html).

Bollen, J., Luce, R., Vemulapalli, S. S., & Xu, W. (2002). Usage analysis for the identification of research trends in digital libraries, *D-Lib Magazine*, 9(5). Available at: (http://www.dlib.org/dlib/may03/bollen/05bollen.html).

Online instruments, tools, guidelines, etc.

A freeware bibliometrics tool is available at:
http://www.umu.se/inforsk/Bibexcel/index.html.

Information and tools related to a related concept, “Bibliomining,” can be found at:
http://www.bibliomining.com/.

Chapter

7

Transaction log analysis

> The studies use a variety of research methods, including observations, surveys, interviews, experiments, and transaction log analysis. Some surveys or interviews ask questions about preference, including how users feel about the library or about specific media; others ask questions that provide information on user behavior. Observations, experiments, and logs also show what users do, but do not always reveal preferences or motivations. Each of these methods allows different types of conclusions and it is only when they are taken together that we can get a full picture of what users actually do, why they do it, what they would prefer, and what they are likely to do in the future.
>
> - Tenopir, 2003

Transaction log analysis was first developed as a means of evaluating the performance of online public access catalogs (OPAC) (Peters, 1993). As the WWW began to burgeon in the 1990s, transaction log analysis was increasingly used to study how people used and searched web sites. In recent years, as digital libraries have become increasingly widespread, user log analysis has also been employed as a digital library evaluation tool. Tenopir (2003) emphasizes the importance of supplementing the evaluation of digital libraries using transaction log analysis with other methods such as observations and experiments.

What is transaction log analysis?

Transaction log analysis is a way to track unobtrusively how users are using a digital library. As an evaluator, you may wish to analyze transaction log information as part of an overall evaluation aimed at obtaining a deeper understanding of how users are navigating through your digital library, which resources they access, and any search problems they encounter. Log analysis alone usually requires too much inference, but it provides important information that may be explained with the additional data obtained from interviews, surveys, and observations. By understanding what digital library users are doing within your digital library, you can then make informed decisions about how to better meet users' needs, perhaps by improving the quality of the underlying search algorithms or enhancing the graphical user interface.

Transaction logs are usually gathered through transaction monitoring software that is typically built into a digital library system or based on a web server that automatically tracks specific interactions. Most digital libraries maintain server logs that keep track of users' requests. These log files typically contain information such as the users' IP addresses, date and time of users' requests, search terms, and so forth. As an evaluator, analyzing user logs can provide valuable information, e.g., it can be used to create a map of what a typical user session looks like.

Here are some of the types of information that transaction log analysis can provide for evaluating digital libraries (Tenopir, 2003):

- Frequency of feature use
- Sequence of feature use
- System response times
- Hit rates
- Error rates
- User actions to recover from errors
- Number of simultaneous users
- User session lengths
- Number of transactions per session
- Location of users

How do you do transaction log analysis?

The most common type of analysis of transaction logs is the generation of usage statistics to determine which collections are accessed most often and/or which documents are retrieved most frequently. But as Bollen and Luce (2002) argue, transaction data can be used for much more than generating usage statistics. Transaction data can be analyzed to determine the structure of relationships between documents, document impact on user communities, and to reveal other characteristics of user communities. Bollen and Luce believe that user log data can be helpful in informing policies regarding acquisitions for the digital library collection as well as the organization of services provided.

Of course, before you decide to use user log data in the evaluation of your digital library, it is important to think about the kinds of decisions you hope to influence as this will directly influence the type of data you collect. For example, information about the number of users and user session lengths might be used to inform decisions about the content of the library collection or the need for advertising the collection to attract more users. Information about users' navigation choices could be used to inform decisions about page design and layout. Error rates and user actions to recover from errors may provide useful information about the skill level of typical users, and this information in turn may influence future decisions about interface design and help features of the library.

There is no step-by-step procedure to follow if you want to make use of user logs for the evaluation of your digital library. One very positive thing about user logs is that the information is usually already collected for you. The difficult part is sorting through, and making sense of the huge amounts of data collected. Here are some helpful hints for including user log information as part of your evaluation:

Know what you are looking for.

As with any evaluation or data collection technique you should try to decide up front the questions you want answered from the data and the decisions you intend to influence down the road. Particularly when dealing with a large amount of user log files, it will make it much easier for you if you know exactly what kinds of information are important to you and what kinds of information you can safely ignore.

Good software is the key.

Find a good software program to help you sort through your data. In the end it will save you a lot of time and energy if you are able to sort through your log files in an efficient way.

Look beyond the obvious.

In your analysis of transaction logs, you will certainly generate lots of statistics that demonstrate use, time of use, retrieval patterns, etc., but be sure to consider the implications of such information. For example, from your transaction logs you can generate statistics that demonstrate the typical access pattern over a weekly period. You discover that users most frequently access the digital library on Mondays but usage is very low on Sunday evenings. You can use this information to make informed decisions about the best times to conduct system maintenance or upgrades.

Transactional log analysis case study

An exemplary example of using transaction log analysis in evaluating a digital library is the evaluation done by Jones, Cunningham, McNab, and Boddie (2000) from the University of Waikato in New Zealand. Jones et al. conducted an extensive log analysis of the New Zealand Digital Library, focused on the Computer Science Technical Reports Collection. Evaluating and improving upon the user retrieval interfaces was the driving force behind conducting usability studies as well as employing transaction log analysis techniques.

Jones et al. (2000) discovered a number of interesting findings from their analyses, but what is most useful to note is that they went beyond the numbers to try to look at how the information could be used to improve users' experiences using the library. For example, their log analysis revealed that users rarely changed the default settings for query or results display options. The evaluators decided that this finding could mean one of two things: (a) the default settings were appropriate for the majority of users needs, or (b) users tended to accept the default settings regardless of what they are. Jones et al. concluded that since they could not know from only the log analysis which of these two hypotheses was true, extra care should be taken in creating the default settings and ensuring they are the most efficient settings possible, as it may be likely that users will accept the default settings as they are.

To give you more ideas of the types of analyses you can perform with your log files, here are additional examples of the analyses performed during the evaluation of the New Zealand Digital Library's Computer Science Technical Reports Collection (Jones et al., 2000).

Location and affiliations of users.

Log file analysis can provide information on the location of users, as well as their affiliation, e.g., educational institutions = .edu versus commercial interests = .com).

Boolean vs. ranked queries.

The frequency with which users opt for using Boolean or ranked search queries can be calculated. This information can help determine an appropriate default setting.

Query complexity.

The complexity of search terms used, e.g., one or two-word search terms versus five or six-word search term, can be generated from the log files to provide insight into how users approach searching in your system.

Query terms.

Analysis can be done to determine the most commonly searched terms. This information can be used for structuring term indexes in the system.

Term specificity.

Knowing whether users are using overly general or overly precise terms while searching your digital library collection can be helpful in understanding precision and recall measures.

Print references

Denise Troll Covey's (2002) handbook titled *Usage and Usability Assessment: Library Practices and Concerns*, published by the Council on Library and Information Resources, provides valuable information about conducting transaction log analysis.

Online references

Carol Tenopir's (2003) report of eight on-going studies of digital libraries, some of which involve transaction log analysis, is available at:
http://www.clir.org/pubs/reports/pub120/pub120.pdf.

Online instruments, tools, guidelines, etc.

Some useful tools for transaction log analysis are available at this website created by Virginia Tech's Digital Library Research Laboratory:
http://www.dlib.vt.edu/projects/DLLogging/

Guidelines for log file analysis are available at:
http://web.syr.edu/~jryan/infopro/statsoft.html.

Chapter

8

Survey methods

> Surveys are an effective way to gather information about respondents' previous or current behaviors, attitudes, beliefs, and feelings. They are the preferred method to gather information about sensitive topics because respondents are less likely to try to please the researcher or to feel pressured to provide socially acceptable responses than they would in a face-to-face interview. Surveys are an effective method to identify problem areas and, if repeated over time, to identify trends. Surveys cannot, however, establish cause-effect relationships, and the information they gather reveals little if anything about contextual factors affecting the respondents. Additional research is usually required to gather the information needed to determine how to solve the problems identified in a survey.
>
> \- Covey, 2002

Survey methods are the most widely used data collection technique in evaluation, so much so that some people seem to equate evaluations with surveys. Surveys enable you to collect information concerning a wide range of aspects of digital libraries, especially the attitudes people have toward them and their opinions about the various advantages and disadvantages of digital libraries.

What are survey methods?

In a general sense, survey methods encompass several major data collection strategies, including surveys of existing records, questionnaires, interviews, and focus groups. Earlier sections of this guide have presented materials related to surveys of existing records such as transaction log analysis, and a later section highlights interviews and focus groups. In the current section, the focus is on surveys that involve some sort of questionnaire.

Surveys are a way of collecting information to help you describe, compare, or explain knowledge, attitudes, and behaviors related to digital library use. Most often, within digital library evaluation, surveys are used to address issues that relate to user-centered concerns. Thus, the information derived from surveys can be used to inform decisions that will relate to issues relevant to users. Surveys are a good way to gather information about users':

- Previous or current behaviors
- Attitudes
- Beliefs
- Level of satisfaction

As with any other evaluation strategy, surveys require you to be clear about the decisions you wish to inform by collecting the information. A survey should be designed to address the questions that will help you gather information about the decisions that you or your stakeholders are facing. For example, if decisions must be made about resource cutbacks, then you would want to be sure to design your survey to include questions that will elicit as much information as possible related to the resources perceived as most valuable by your digital library patrons, and perhaps about those they consider to be superfluous as well.

How do you conduct surveys?

There are a number of issues that must be considered when one decides to carry out a survey. There are issues of sampling (Who will receive the survey? How can I motivate them to respond?); administration (How will I administer the survey? Paper? E-mail? Web-based?); design (What will my survey address? What scales do I use? What is the best wording for questions?); analysis (How will I process and synthesize the data I get?); and reporting (How can I communicate the results clearly?). All of these issues are important to consider. According to Fink (1995), the best examples of surveys have the following characteristics:

- Specific objectives
- Straightforward questions
- Sound design
- Sound choice of population or sample
- Reliable and valid instruments
- Appropriate analysis
- Accurate and timely reporting of results

Specific Objectives

Defining the objectives of your survey is important in helping you determine the questions the survey should ask and the information that you will gather. As emphasized above, the overall goal of any survey should be informing decision-making. More specifically, an objective is a statement of the intended outcome of the survey. Here are two examples of survey objectives with accompanying questions:

Objective:
Identify the technology skills and experience of typical digital library patrons.

> Sample question:
> Using a 1 (not at all) to 10 (expert), rate how familiar you are with the following technologies:
> __ personal computing __video production __digital photography __MP3 music

Objective:
Determine frequency of use of the digital library.

> Sample question:
> How often do you use the ____ digital library?
> __This is my first visit. __Rarely (less than once a month)
> __Monthly (1 to 3 times a month) __ Weekly (1 to 3 times a week) __Daily

Straightforward Question and Responses

Of course, for your respondents to be able to provide you with the information you are seeking, you should ask questions in as clear, precise, and straightforward a manner as possible. For the most part, questions should be focused (dealing with only one thought or issue at a time). Failing to use correct grammar and syntax will decrease the survey's credibility and dampen participation. Questions can take one of two forms: open-ended or closed. Open-ended questions require respondents to generate their own answers using their own words. Questions in which the respondent is required to select responses from a pre-determined set of answers are closed questions. Open questions are useful when you do not know the types of responses to expect to a question, and for gaining a respondent's unique perspective about an issue in his or her own words. The difficulty with open-ended questions is in the analysis of the responses, which requires training in qualitative research. Closed questions are often more practical because their results lend themselves to statistical analysis. However, closed questions are more difficult to write because the response choices must be known in advance. Here are examples of open and closed questions:

Open-ended question:
What do you value most about the digital library? ________________________________

Closed question:
How many times during the past week did you use the digital library?
__never __1 time __2-3 times __4-5 times __6 or more times

Sound design

Choosing the type of survey design you will employ depends upon the decisions you need to make and the aligned objectives of your research. You can choose to do a comparative design or a descriptive design. A comparative design involves surveying two or more groups distinguished by variables of some importance to your evaluation. For example, you might wish to compare the attitudes of professors versus students toward a digital library. Descriptive designs are employed when you are looking to gather information from whole groups, e.g., a survey of the population of your patrons concerning needed extensions to the digital library.

Sampling

Ideally, you might want to administer a survey to the entire population of your users, but this is rarely feasible or even necessary. One of the most important things is to ensure you have a representative sample. A representative sample shares all the important characteristics (such as age, gender, skill level, etc.) of the larger population that you are interested in studying. To select a representative sample there are a number of sampling techniques that can be used, but before you decide on a sampling technique you should establish the eligibility criteria for your study. The eligibility criteria are those characteristics that you deem respondents need to have in order to complete the survey. For example, if you are interested in how middle school teachers are using your digital library in their classroom, you might set the eligibility criteria to include:

- teachers of grades 6-8 (because you are only interested in middle school teachers)

- teachers who have used the digital library for at least 6 months (because you are only interested in experienced users of the digital library)

Establishing eligibility criteria helps you determine who among the general population is eligible to be included in your sample population. Once you have established those individuals eligible to participate, you need to choose a suitable sampling technique to find participants for your survey. There are advantages and disadvantages to using different methods for selecting a sample. Typical sample selection methods include: random, cluster, convenience, and snowball.

Random sampling: The basis of random sampling is that every individual has an equal chance of being selected. One way to generate a random sample is to apply a table of random numbers to a list of prospective participants. Or you can use a random number generator available on the Web at: http://www.random.org/. An advantage of using a random sample is that the results are relatively unbiased because of the equal probability for participants to be selected.

Cluster sampling: A cluster is a naturally occurring unit such as a school, a county, city, state, etc. With cluster sampling you can randomly select from among the clusters, and then survey all the members of the cluster or a random subset of them. Note here that the resulting sample may not be representative of the other clusters, as well as not representative of aspects not covered by the cluster.

Convenience sampling: This method relies on using an already available group of individuals. For example, surveying the professors and students in a university with which you are associated may be considered a convenience sample.

Snowball sampling: Sometimes it is difficult to find participants who meet your criteria for inclusion in your evaluation. However, most often you are able to find at least one or two participants who meet your criteria. Subsequently, these participants will likely be able to identify one or two other individuals who will also meet your criteria, and thus your sample snowballs.

Reliable and Valid Instruments

Reliable survey instruments allow you to obtain the same information each time that you use it (assuming no intervening circumstances). A reliable survey instrument is said to be relatively free of "measurement error," which is important in ensuring that results represent individuals' "true" attitudes, opinions, etc. You can increase the reliability of your survey instrument by doing some of the following: (a) ensuring the reading level of your survey is appropriate for your population, (b) ensuring your questions are clearly written, and the directions are easily understood, and (c) ensuring you administer the survey in appropriate ways and in appropriate environments (to ensure environmental factors have a minimal impact on participant responses).

Valid survey instruments must be reliable first, but in addition they should measure what they are intended to measure. For example, if a survey's aim is to find out about the pedagogical beliefs of teachers who use your digital library, the results from your survey should be judged by content experts to measure teacher pedagogical beliefs as well as be consistent with other measures of pedagogical beliefs. Validity is often discussed along four dimensions: content, face, criterion, and construct.

Content validity is the degree to which a measure appropriately assesses what it was designed to measure (as briefly discussed above). Content validity is often established by basing survey construction upon models or conceptual frameworks found in the literature. For example, if you were interested in surveying digital library users' information-seeking behaviors, you could base your survey on the literature about how individuals typically seek out information in digital libraries.

Face validity deals with how well a measure appears to address what it was intended to address. Does it ask the important questions needed, and does it use the appropriate language to do so? Unlike content validity, face validity is not grounded in the literature. Judging face validity is subjective, but the process is enhanced when experts are used.

Criterion validity is focused on one of two things: predicting future performance (known as predictive validity such as is found in test like the Graduate Record Exam (GRE)), or comparing responses to those from more well-established surveys (known as concurrent validity).

Construct validity is the degree to which a survey is able to distinguish between participants who do and do not have certain characteristics. Construct validity can typically be established in two ways: 1) establishing that your survey has convergent validity (correlates with measures of similar characteristics) and has discriminant validity (does not correlate with measures that are not similar), and 2) establishing that your survey can discriminate between groups of individuals on important variables.

Appropriate analysis

Depending on the type of survey you have constructed, analysis of survey data can employ statistical or qualitative techniques. Surveys that are primarily composed of closed questions lend themselves to statistical analysis. Typical goals of statistical analyses are to produce:

- Descriptions (e.g., the backgrounds of respondents; responses to questions)
- Relationships (e.g., connections between variables)
- Comparisons (e.g., between subsets of your sample population)
- Predictions (e.g., of how individual variables such as gender or age relate to reported behaviors)

There are numerous qualitative data analysis techniques that can be used to analyze the data from open-ended questions, but a common approach is to categorize common responses. You can go through all participant responses to categorize them based on similar main ideas or issues. By doing this, you will gain a better sense of which responses were most frequently given by respondents and thus have some indication of the more important issues for your survey sample.

Accurate and timely reporting of results

You can conduct the best survey in the world, but if you fail to report the results accurately and in a timely matter, it is all for naught. It is easy to mislead people, intentionally or unintentionally, with graphs and tables that present data in a skewed manner. After 50 years, Darrell Huff's (1954) book *How to Lie with Statistics* remains a popular critique of how statistics and graphs are often used to misinform. It is doubtful that evaluators of digital libraries would intentionally mislead their stakeholders, but care must be taken to present the results completely, warts and all. It is equally important to report survey results in a timely manner so that the decisions that the evaluation is intended to inform haven't already been made. Rather than waiting to compile results into a long printed report that few decision makers will read, it is often more effective to report findings in brief bulletins or executive summaries.

The above characteristics of good survey research are important for all types of survey research including written surveys, interviews, and focus groups. Next, let's take a more in-depth look at how written surveys can be used in digital library evaluation research.

Questionnaires

There are a number of important issues to keep in mind when designing a written questionnaire, but one of the most important is the construction of the questions. Here are several tips, based partly on guidelines suggested by Fink (1995), to help you design a questionnaire that will be effective in obtaining the information needed to influence decision making concerning your digital library:

Keep questions short and specific. Try to avoid asking two things in one question such as "When and how often do you use the digital library?" Be as specific as possible when asking questions to ensure that respondents will give you the type of information you need. For example, if you are interested in the times of day that individuals are using your digital library asking "When do you use the digital library?" might not be the best question to ask. Asking more specific questions such as "What times of day do you most often use the digital library?" will prompt respondents to give specific times rather than other ambiguous answers. Asking vague questions only results in difficult-to-interpret or unhelpful answers. Also try to avoid using run-on sentences.

Use vague qualifiers with caution. Vague qualifiers are adverbs like "usually" that can mean different things to different people. Using such words makes it difficult to interpret participants' responses because you cannot be sure how they interpreted the qualifier (e.g., "often").

Use jargon, abstract terms, and acronyms with caution. Unless you are able to carefully assess that your survey participants understand jargon (e.g., mirroring), abstract terms (e.g., economy of execution), or acronyms (RMB for right mouse button), you should avoid them in your questionnaire. Alternatively, you can provide definitions of terms with which you suspect your participants may be unfamiliar. You will also need to define any abstract terms that may have multiple meanings to clarify the particular meaning that you want participants to apply in completing your survey.

Organize questions from easiest to more difficult. Typically survey designers order the questions from easiest to most difficult and complex. Partly this is to ease participants into the survey and not scare them off with the first question! Although participants can answer printed questionnaires in any order they chose, this "rule" is typically still followed by most survey designers. Web-based surveys can be designed to gradually reveal questions, perhaps from simple to complex.

Organize questions in a logical order. Related questions should be grouped together, and the questionnaire should be organized in such a way that questions are asked in an order that makes logical sense. Typically asking general questions before more specific ones works well.

Have a rationale for where you place demographic questions. Demographic questions are most commonly placed at either the beginning or the end of a questionnaire. Survey researchers differ on where they believe demographic questions should be placed. Those who maintain they should be placed at the beginning of the questionnaire believe so for two reasons: (a) demographic questions are easy for respondents to answer (thus it falling line with the reasoning of starting with easier questions), and (b) respondents who turn in incomplete questionnaires tend to leave the last part, rather than the beginning unfinished. Advocates for placing demographic questions at the end of the questionnaire disagree with this logic and instead insist that: (a) since most questionnaires are accompanied by a cover letter describing the topic of the survey and encouraging participation, starting with demographic questions negates the purpose of the cover letter, (b) many people find demographic questions boring and thus may not be motivated to complete the survey if they are the first questions asked, and (c) beginning with easy questions that engage the participant in the topic of study will increase the response rate and reduce missing data. Wherever you choose to place your demographic questions, be sure that you have thought about the potential implications of placing them where you do.

Utilize closed questions whenever possible. Although you may be tempted to add open questions (e.g., What are the benefits of using our digital library?), be cautious. Open questions are easy to pose, but often time-consuming and difficult to analyze. In addition, if you have very many of them in your questionnaire, your response rate will go down because people simply do not have the time to respond.

Craft closed questions with care. Closed questions are easier to analyze, but only if they have been well-written in the first place. Perhaps the most difficult challenge is creating an exhaustive list of responses for closed questions. If you are not sure that your categories are exhaustive, consider use of an "other" answer category that allows respondents to clarify what "other" is for them. It is often difficult to create mutually exclusive answer categories, but you should strive to do so. At the same time, you must attempt to keep answer alternatives short and precise. It is also essential to inform respondents whether they can select only one answer or multiple answers.

Pilot test your questionnaire. Never distribute a questionnaire before you have pilot tested it with several different audiences. No matter how experienced you are at preparing questionnaires, there will always be things you miss. Start testing the questionnaire with colleagues, and gradually increase the testing with people who are more like the eventual targeted respondents. Expect to go through three to ten versions of your questionnaire, depending on its complexity and length.

Survey methods case study

Cherry and Duff (2002) report a follow-up survey of users of the Early Canadiana Online/Notre Memoire En Ligne (ECO) digital library. They employed a web-based questionnaire that can be viewed at: http://informationr.net/ir/7-2/p123qre.html. Their survey confirmed the results of an earlier survey that found users highly valued ECO. Interestingly, it also reported that respondents to the second survey requested the same improvements as those on the first survey, perhaps suggesting that the survey results have not informed decision makers at ECO sufficiently. The authors recommend surveying digital library users over time.

Print references

The best investment you can make in this area is to buy the recently updated *Survey Kit* edited by Arlene Fink (2002).

Online references

An online guide to survey design is available at:
http://www.surveysystem.com/sdesign.htm.

Covey's (2002) *Usage and Usability Assessment: Library Practices and Concerns* provides valuable information about designing surveys. It is available at:
http://www.clir.org/pubs/reports/pub105/contents.html

Online instruments, tools, guidelines, etc.

Some useful guidelines for survey design can be found at:
http://www.pearsonncs.com/research-notes/2.

More information about the technical aspects of survey design can be found at:
http://www.ubmail.ubalt.edu/~harsham/stat-data/opre330Surveys.htm.

Hints for writing effective questions are at:
http://edresearch.org/pare/getvn.asp?v=5&n=3.

Recent examples of web-based surveys can be found at:
http://www.library.usyd.edu.au/borrowing/docdel/questionnaire1.html,

http://healthcybermap.semanticweb.org/questionnaire.asp,

http://www.bcpl.net/~dcurtis/digital/quest.html, and

http://resources.theology.ox.ac.uk/library/content/feedback.html.

Chapter

9

Interviews & Focus Groups

> Data gathered from focus groups are used to inform decision making, strategic planning, and resource allocation. Focus groups have the added benefit of providing good quotations that are effective in public relations publications and presentations or proposals to librarians, faculty, university administrators, and funders. Several DLF (Digital Library Federation) respondents observed that a few well-articulated comments from users in conjunction with quantitative data from surveys or transaction log analysis can help make a persuasive case for changing library practice, receiving additional funding, or developing new services or tools.
>
> - Covey, 2002

Interviewing is a frequently used data collection method in evaluations of all kinds. In the context of digital library evaluation, interviews can serve to meet a variety of goals, such as measuring user satisfaction levels, getting user feedback, seeking user input, etc. Interviews can be conducted on a one-on-one basis, or in group settings. In the latter settings, interviews are often called focus groups. Interviews and focus groups can be used as part of a number of other evaluation methods described elsewhere in this Guide such as with usability testing and service evaluation.

What are interviews and focus groups?

Interviews and focus groups essentially boil down to asking people questions to which they respond verbally as opposed to in writing. Patton (1990) identifies three types of interviews typically used in evaluation or research. These are: the informal conversational interview, the standard open-ended interview, and the interview guide approach. Each of these types of interviews serves its own individual purposes well.

Informal conversational: This type of interview is the most flexible and open-ended. Informal conversational interviews depend on the natural flow of interaction between two people and allows the evaluator to pursue questioning in any direction, not having to rely on a script. These informal interviews will rarely be appropriate within a substantive evaluation, but they may be useful when you are in informal settings and meet patrons of your digital library.

Standard open-ended: This type of interview relies on a standard set of questions (called a protocol) that has been created ahead of time to elicit in-depth responses from participants. This type of interview is typically used when there is a team of evaluators, and you want to limit the variation between interview experiences. Using a standard open-ended interview allows for easier comparison between participants' responses, since each respondent is asked the same questions, typically in the same order.

Interview guide: The semi-structured, guided interview is a combination of the informal conversational and standard open-ended interview. There is an interview guide or protocol that serves as a checklist of topics that should be covered during the interview. However, there is no set order in which the questions need to be asked, and some questions may be skipped and others may be added. It is semi-structured, in the sense that there is a set of topics that needs to be covered, but the evaluator has the flexibility to explore certain questions in greater depth as he or she deems appropriate.

Interviews conducted in a group setting are often called focus groups. Focus groups can be used much like the individual interview, to elicit in-depth responses to topics of importance to the digital library evaluation. Generally focus groups consist of 6 to 8 participants and a moderator, and last anywhere from 30 minutes to 90 minutes. Typically the interaction is audio-taped, or sometimes video-taped, and later transcribed. A focus group protocol is very similar to the semi-structured, guided interview protocol, in that the moderator typically has a list of questions and topics to discuss with participants, but the order and depth in which the topics are discussed is flexible.

How do you do interviews and focus groups?

Interviews provide participants in an evaluation more opportunities to speak in their own voice instead of merely responding to the categories of questions that others have defined for them, as they might with a questionnaire. Getting started with interviews and focus groups inevitably involves preparing a protocol of questions. It is important to plan your interview protocol carefully. Refinement of an interview protocol will often involve several trial interviews and subsequent revisions of the questions. Interview protocols generally have two types of questions. The first are the major questions you wish to address. Under each of these, there are usually several secondary questions that can be used to prompt participants when they are not adequately responsive to the primary question. For example, suppose you asked a user, "What are the primary reasons you use our digital library?" If the user is unable to articulate specific reasons, you might ask secondary questions such as "Do you use the digital library for education? For research? Or for something else?"

Recording interview and focus group data presents challenges. Most interviews and focus groups are recorded. Audio or video recorders may be used, but you should be aware that these devices may intimidate interviewees to such a degree that their responses are limited. Some evaluators prefer to take brief notes during an interview, and subsequently record a more detailed transcription of the interview immediately after its conclusion. If you adopt the notes strategy, it is advisable to write everything you can recall concerning one interview before starting to conduct another. Otherwise, you are apt to confuse the responses of one person with those of someone else.

Here are ten steps to follow in carrying out interviews and focus groups:

1. Organize a team of colleagues to assist in developing the interview/focus group protocol.

2. Determine the purposes of the interview/focus group (e.g., collecting digital library user input into a redesign process).

3. Identify a representative sample with whom to conduct the interviews or focus groups.

4. Generate a list of draft questions, initially focusing on "brainstorming" as many good questions as possible, and later selecting the best ones.

5. Construct a draft interview or focus group protocol with the questions in the order that seem to make the most sense.

6. Test the protocol with a small sub-sample of your representative sample, and look for misunderstandings and dead-ends. Expect to make changes.

7. Revise the protocol and retest if necessary.

8. Carry out the interviews or focus groups with the rest of your sample.

9. Process and analyze the data using qualitative data analysis methods.

10. Report and use the results to influence decisions in a timely manner.

Analyzing the qualitative data involves going through the transcripts of what the respondents said (and/or your notes) to look for themes, patterns, or categories. Suppose you are conducting interviews with scientists to determine their willingness to trust the data provided by digital libraries of real time data. Some themes that might emerge are concerns about validity, reliable access, and transmission errors, as well as appreciation for the value of having such data for their particular scientific community. It is beyond the scope of this guide to go into much detail about qualitative data analysis, but Miles and Huberman (1994) provide extensive guidance in this area.

Interviews and focus groups case study

Peterson and York (2003) report their user evaluation of the Montana Natural Resource Information System (NRIS), a digital library of natural resource information used by diverse user groups including federal, state, and local government employees, academicians and scientists, and private citizens. To obtain a representative sample of the various user groups, the evaluators employed snowball sampling whereby people interviewed early in the process nominated others who should be interviewed.

For this digital library evaluation, a total of fifty interviews were conducted throughout the state of Montana. Although transaction log analysis had already indicated a wealth of hits for NRIS (nearly 2,000 sessions per day), the interviews allowed the evaluators to provide decision makers with valuable information about how NRIS was actually used. Among the more surprising results were that many users preferred to access raw data from NRIS rather than the data processed with the NRIS-supplied applications. This evaluation study illustrates the importance of triangulating your evaluation results with multiple methods of data collection. Relying upon transaction log analysis alone in this case might have yielded misleading interpretations of the use of this digital library.

Print references

There are numerous books available about interviews as a research and evaluation method. One of the volumes in *The Survey Kit* edited by Arlene Fink (2002) is titled *How to Conduct In-Person Interviews for Surveys*. Krueger and Casey (2000) wrote a useful text called *Focus Groups: A Practical Guide for Applied Research*. Michael Quinn Patton's (1997) book, *Qualitative Research and Evaluation Methods* is an excellent resource concerning how to analyze data from interviews and focus groups, especially when supplemented with Miles and Huberman (1994) volume, *Qualitative Data Analysis: An Expanded Sourcebook*.

Online references

A valuable online reference about interviewing can be found at:
http://ag.arizona.edu/fcr/fs/cyfar/Intervu5.htm.

An online introduction to interviewing can be found at:
http://jan.ucc.nau.edu/~mid/edr725/class/interviewing/.

Another online tutorial focused on interviewing is available at:
http://www.roguecom.com/interview/.

A valuable online reference about focus groups can be found at:
http://www.soc.surrey.ac.uk/sru/SRU19.html.

Online instruments, tools, guidelines, etc.

Guidelines for conducting structured interviews are available at: http://edresearch.org/pare/getvn.asp?v=5&n=12.

Additional interview tools can be found at Bill Trochim's Research Methods Knowledge Base at: http://trochim.human.cornell.edu/kb/intrview.htm.

The "Basics of Conducting Focus Groups" website provides guidelines and tools: http://www.mapnp.org/library/evaluatn/focusgrp.htm.

Chapter 10

Observations

> When looking at the same scene or object, different people will see different things. What people "see" is highly dependent upon their interests, biases, and backgrounds. Our culture shapes what we see, our early childhood socialization forms how we look at the world, and our value systems tell us how to interpret what passes before our eyes. How, then, can one trust observational data?
>
> - Patton, 2002

Obeservations are controversial evaluation methods because many people view them as too subjective, as noted in the quote above from the evaluation guru, Michael Quinn Patton. It is true that everyday observations are notoriously untrustworthy as shown in the divergence of testimony given by different "eyewitnesses" to the same event in judicial trials. However, rigorous, disciplined observations are essential to the research conducted in social science fields such as anthropology and sociology as well as in nature sciences such as zoology and entomology. Observations also have a long history as an evaluation method. Although at first glance, observations may seem to have limited applicability within the context of digital libraries, they can have important utility as described below.

What are observations?

Observation is a data collection method used to gather detailed information about a situation or event. Observation data is used to describe the setting, activities, participants, and the meaning of the observations from the observer's perspective (Patton, 2002). Observation data should be factual, accurate, and detailed, but not so detailed as to include irrelevant or trivial information that makes the description difficult to understand. The best observational data allows the reader to fully understand the situation described (Patton, 2002). Observational methods are ideal for gathering data related to user-centered issues, such as the usability of your digital library. Observational methods are also ideal for providing information about the impact and uses of your digital library in real-life settings. This chapter focuses on the latter goal of observations.

How do you do observations?

The value of using trained and skilled observers in your digital library evaluation will be evidenced in the quality of data they gather. For purposes of reliability, most evaluations will involve multiple observers conducting multiple observations. Doing observations takes considerable skill and training, but trained observers can report the same event with accuracy and reliability. Patton (2002) states that to become a skillful observer, training is required in the following areas:

- Paying attention: seeing what there is to see, and hearing what there is to hear

- Writing descriptively

- Discipline in recording field notes

- Separating detail from trivia to achieve notes that are detailed and not overwhelmed by the trivial

- Using rigorous methods to triangulate and validate observations

- Reporting the strengths and limitations of one's own perspective

If you want to use observational methods in the evaluation of your digital library, you probably desire to learn more about how your digital library is implemented in a real-life setting like a classroom or the spaces where students study or collaborate such as coffee shops. For example, suppose you are interested in how your digital library is used in an undergraduate introductory biology class. If so, there are a number of dimensions on which observational approaches vary that can be used to shape the design of the observational approach to be used in the digital library evaluation. Patton (2002) outlines six different dimensions to consider:

1. Role of the observer: What role will the observer (evaluator) play in the setting in which the observation is to take place. The observer can act as a full participant in the evaluation context, or be an outside onlooker, unobtrusively making observations. For example, will the evaluator participate in all activities in the biology course like a fully participating student or be a removed observer with no role in the class, or some combination in-between?

2. Insider versus outsider perspective: Will the evaluator approach the observation with the goal of capturing the insider perspective – recording both what it is like to be part of the class as a student, and also what is happening to everyone else in the class? Or will an outsider perspective be taking by the evaluator, focusing on capturing the separate events and their relation to each other, from a distance?

3. Who does the observing: A single evaluator, a team of evaluators, or a combination of evaluators and participants (in the example of the biology class, this would include students and/or the instructor) could gather observational data.

4. Disclosure of the observer's role to others: The extent to which participants in the setting are aware of the purpose and role of the evaluator can vary from full and open disclosure to no disclosure at all. In the biology course example, open disclosure could involve introducing the evaluator at the beginning of class to students, informing them that this person will be sitting in on some classes taking notes for a clearly stated purpose. Partial disclosure would involve identifying the observer, but not clarifying the purpose. No disclosure is rarely used, but it may occur in situations where the evaluator is trying to gain an insider perspective and there is concern that if other participants were aware of the true identity of the evaluator, they would behave differently. Obviously, there are a number of ethical issues that need to be considered before choosing to adopt the no disclosure approach.

5. Duration of observations and fieldwork: The extent and number of observations will depend on the questions and focus of your evaluation. This can range anywhere from short, single observations to long-term or multiple observations. The duration of your observations will also depend on the amount of resources you have. For a comprehensive view of how your digital library is used in the undergraduate biology class, it would be ideal to have an evaluator observe classes throughout the course, however you may not have enough time or resources for this to be possible. Some sort of schedule for sample observations would have to be worked out in advance.

6. Focus of observations: The focus of your observations is important to determine early on, although it is always subject to change. You can choose to adopt a broad focus, including almost all aspects of the setting. Or, you can choose to focus on a very narrow and specific event or behavior. Again, this depends on the specific questions of interest in your evaluation. Using the biology course example, you may choose to focus your observations on teacher use of the digital library's resources during classes, making note of things such as: (a) comments made during use, (b) any technical issues experienced during use, and (c) tasks for which digital library resources are used. It is often helpful to make an observation guide or some sort of form that lists the types of behaviors or events you are particularly interested in observing.

Here are some other factors that should be considering when conducting observations:

Setting – Where are you? A library? A classroom? A lab? An auditorium? Starbucks? A home? What does it feel like to be there? Is the space amenable to the use of a digital library? Are people comfortable in this space?

Objectives – Why are people here? Who is in charge (if anyone)? Are activities self-directed? Who maintains goal directed behavior? To what degree are people on task? What is the assumed or stated goal of using a digital library in this context?

Implementation – Are things going as planned? Does the digital library technology work? Who handles problems? Are people confident in their use of the digital library?

Interactions – How do people interact among themselves? Is the atmosphere formal? Informal? Friendly? Unfriendly? How are you viewed? Do people share information about what they are finding with the digital library?

Nonverbal Behavior – What does body language tell you about this digital library? Are people interested? Going through the motions? Intimidated? Confused? Exhilarated?

Unobtrusive Factors – What areas of the digital library are heavily used? What areas are less used? Can people hear audio components? Are they able to print easily?

Unexpected Things – What did you see you didn't expect? What surprised you? What delighted you? What worried you?

Before going somewhere to conduct field observations, it is useful to write down your expectations. Try to describe any biases or prejudices that might affect your observations. Having these expectations on record before observing provides you with a better basis for interpreting what you actually saw when you observed. Thinking back on the biology course example, you might write:

> I am going to observe undergraduate students using a digital library in a large section biology course at a huge state university. I expect the course to be innovative because all students are expected to bring their laptops to the classroom. There is wireless Internet access across this entire campus. The instructors in this course have a reputation for innovative applications of technology in their teaching. In fact, these instructors developed one of the major collections for this digital library. All the students may not be "whiz kids" with respect to technology, but most have higher than average computer literacy. Their enthusiasm about biology may vary in that for many of them this is a required course. It will be interesting to see how the digital library collections are integrated into the course activities. I am a little nervous about being in an undergraduate classroom again. After all, it has been a long time. I wonder if I can relate to the students of this generation. Fortunately, there are nearly 300 students in the large lecture hall, and I should be able to blend in.

There are several guidelines for recording your observations:

- Collect field notes. This is not optional! Do not trust everything to your memory. You'll forget things or worse invent things you did not really see.

- However, don't try to write down everything. No one can do this! Record your notes in some sort of outline or shorthand style that fits your experience and skills. Flesh out your notes with more details as soon as you can.

- Make your notes as descriptive as possible. Do not try to interpret what you are seeing at the same time you are describing it. Try to separate description from interpretation.

- Add interpretations and your own feelings later.

Separating observations from interpretations is not easy. Here is an example of poor observation notes followed by a better example:

Poor notes:

- The lecture hall is a warm and friendly place. The lecturer clearly loves her students. There are lots of cool things the instructor does with the digital library resources.

Better notes:

- Large, well designed lecture hall, well-equipped with technology, including a large projection screen and wireless capacity to send and receive data among faculty and students

- A little noisy at times, but neither the lecturer nor the students seem bothered

- Most students appear very attentive – very few appear to be off task – only a few are checking email and a few are instant-messaging each other despite instructor's request not to do so during class

- Teaching assistants very active – helping students with technical and content problems – moving around the lecture hall during whole period

- during the 55 minute lecture, the instructor directs students to observe video simulations downloaded from digital library on three occasions – students record reflections on simulations and their lecture notes using the course management system tools provided

As soon as possible after you have completed your observations, you will want to go over your notes and write down everything else you can recall concerning what you have observed. It takes practice to develop good descriptive skills, and it will inevitably take some time and training to develop these skills.

Observations case study

A good example of the application of observations in a digital library context is the evaluation of the Perseus Digital Library (Marchionini, 2000). The Perseus Digital Library (PDL) (http://www.perseus.tufts.edu) is dedicated to providing digital resources for the Humanities. It has been under continuous development since 1987, and represents one of the rare digital libraries that have been intensively evaluated. Data collection methods used in the evaluation of the PDL included observations, interviews, document analysis, and learning analysis. Marchionini describes five types of observational methods used in the PDL evaluation: (a) baseline, (b) structured, (c) participant, (d) think-aloud, and (e) automatic screen journaling.

- *Baseline observations* were semi-structured and consisted of the observers sitting in classrooms or labs taking notes of the ongoing activities. The purpose of conducting baseline observations was to help the evaluators become situated within the setting and build relationships with those individuals to be observed.

- *Structured observations* involved systematically observing behavior and recording notes for a selected sample of students in a classroom or lab setting. The observations followed an established protocol, for example, the evaluator selected five students and alternated observations between the five students every three minutes. In the context the PDL evaluation, details such as whether students were taking notes or looking at the instructor, were recorded. Specifics of what details to record in any observation will depend on the setting and specified evaluation goals.

- *Participant observations* involved sessions in which the evaluator was allowed to interact (ask or answer questions) with the students being observed. All sessions were audio-taped, and evaluators had a semi-structured protocol to guide their interactions.

- *Think-aloud observations* as used in the PDL evaluation were much like the think-aloud protocols described in Chapter 4 – Usability Evaluation of this guide. In Marchionini's evaluation of the PDL, students were audio-taped and asked to think aloud while working on various tasks.

- *Automatic screen journaling* described in the PDL evaluation was akin to transaction log analysis as described in Chapter 7 of this guide. Data collected from the automatic screen journaling was used to determine interaction patterns such as the number of requests for different resources and temporal patterns of access.

The use of these five observational data collection methods, combined with the other data collection methods such as document analysis, interviews, and learning analysis allowed Marchionini and his team of evaluators of Perseus to gather quality information to guide decision about refining and extending this notable digital library. A number of important recommendations for the PDL as well as for digital libraries in general have been made on the basis of the findings and interpretations stemming from this comprehensive evaluation.

Print references

There are numerous books published about observational methods in fields such as anthropology (Spradley, 1997) and sociology (Schutt, 2003). A book specifically focused on classroom observations is *An Introduction to Classroom Observation* (Wragg, 1999). Observational methods are also described in Michael Quinn Patton's (2002) textbook titled *Qualitative Research and Evaluation Methods*.

Online references

An online introduction to participant evaluation can be found at:
http://jan.ucc.nau.edu/~mid/edr725/class/observation/.

An innovative approach to using video as a research and evaluation tool is described at:
http://www.pointsofviewing.com/.

Online instruments, tools, guidelines, etc.

Some useful guidelines for conducting observations are available from the University of Wisconsin Extension Service at:
http://cecommerce.uwex.edu/pdfs/G3658_5.PDF.

Chapter

11

Experiments

> The logic of this [experimental evaluation] design is foolproof. Ideally, there is no element of fallibility. Whatever differences are observed between the experimental and control groups, once the above conditions are satisfied, must be attributed to the program being evaluated.
>
> - Suchman, 1967, pp. 95-96

Experimental methods, usually associated with the "hard sciences" such as physics and chemistry, can also be used as a method within the context of evaluating digital libraries. The experimental (or more often, quasi-experimental) model is a widely accepted and frequently employed evaluation approach within the fields of computer science and information science. In addition, it has had, and continues to have, many proponents within education and other social sciences (Campbell & Stanley, 1966; Rossi, Lipsey, & Freeman, 2003; Tate, 1990).

What are experiments?

In the ideal world, experiments require randomized assignment of evaluation participants, often called "subjects" in the design of such a study, to different treatments (e.g., two different digital library interfaces, one that uses only icons and another that combines icons with text labels). In the real world, subjects in an experimental group are more often assigned to some sort of treatment (e.g., access to a digital library) while subjects in a control group receive no treatment. The following figure illustrates the design of the latter form of evaluation.

	Time			
		1 (Pre)		2 (Post)
Experimental Group	R	0	X	0
Control Group	R	0		0

R = Random Assignment 0 = Observation X = Treatment

When a methodological authority such as Suchman (1967) (quoted on the previous page) and his adherents speak in terms of an evaluation approach being foolproof or infallible, people listen. Hence, it is not surprising that the experimental approach to evaluation remains deeply entrenched in the minds and actions of many social scientists and evaluators today as well as many evaluation clients in the context of digital libraries. For decades, experimental methods have been held up as the "gold standard" for evaluation by some experts for whom every other approach is viewed as inferior (Campbell & Stanley, 1966). This model remains the method of choice for educational research and evaluation in certain circles today (Shavelson, Towne, & the Committee on Scientific Principles for Education, 2002).

However, the continuing advocacy of experimental methods by many evaluators (e.g., Fitz-Gibbon & Morris, 1987) stands in contrast to the critique of these methods by contemporary evaluation theorists. For example, Guba and Lincoln (1989) claim that evaluation should be concerned with understanding the nature of human phenomena (such as digital libraries) from multiple perspectives, emphasizing the roles of culture, gender, context, and other factors in the construction of "reality." With regard to evaluation methodology, many contemporary evaluation experts are more likely to recommend anthropological or ethnographic methods rather than experimental ones.

Nonetheless, it is important to understand experimental methods of evaluation. Many clients view experimental methods as the only way of providing credible evidence of the effectiveness or impact of educational innovation such as digital libraries. In addition, in a digital library development context, small scale experiments can be useful for providing evidence of the relative effectiveness of some digital library design features over others (Maeda, 2002).

How do you do experiments?

If you apply experiments in a digital library evaluation, you (or your clients) may desire to be able to make some sort of causal statements about the library or some of its features. If so, this usually involves the specification of some sort of hypothesis. For example, you might hypothesize that undergraduate students with access to digital libraries will include more references in their term papers than students who only have access to traditional libraries. It would be feasible, but necessarily advisable, to design an experiment whereby college students would be randomly assigned to different courses, some of which promote the use of digital libraries and others that limit students to the use of traditional libraries. Both groups of students could be given an identical term paper writing assignment, and after all the papers are collected, the numbers of references could be counted, and the support for the hypothesis (or lack thereof) could be calculated. Statistical analysis would be applied to determine whether any differences found were statistically significant (i.e., did not occur by chance).

There are obvious weaknesses in this example of an experimental (or quasi-experimental) approach to evaluation. First, the control of treatment variables, as required by experimental methodologies, is impractical in most contexts where digital libraries are implemented. Although, students might be admonished to only use digital libraries in some courses and traditional libraries in others, there is no guarantee that there would not be considerable variance in library usage within the two treatment groups. Second, the emphasis on what appears to be a clear cut quantitative outcome measure, number of references, is flawed by the failure to establish the importance or relevance of this outcome indicator. Suppose that it was found that the students using digital libraries had more references than the students in the other courses. Such a result would say nothing about the quality of references. It could be that the students using traditional libraries had fewer references, but had better ones in terms of quality and relevance to the topic of the term paper. Third, the experimental approach can only support or fail to support pre-stated hypotheses; it cannot discover unexpected effects of a digital library or other innovation. Perhaps access to digital libraries increased the number of references used in the term papers, but also increased plagiarism within the papers. Fourth, randomized experiments can be unethical in some situations. Restricting access to one type of library or other might be viewed as limiting the learning potential of the participating students.

Perhaps the most serious problem with experimental methods is that their application often requires a stripping away of contextual variables. The use of digital libraries (or any other innovation) is greatly influenced by the context in which it occurs (Guba & Lincoln, 1981). The requirements of experimental evaluation designs demand that contextual aspects be controlled by random assignment of subjects to treatments, but it is these contextual factors that may be most important. In actuality, the vast majority of evaluations conducted with this model are "quasi-experimental," a compromise that introduces many difficulties with respect to the analysis and interpretation of findings. As a result, evaluators operating within the experimental model frequently fall back upon designs that can be most easily managed, focus on variables that are easiest to measure, apply statistical methods without meeting the assumptions underlying their use, and draw conclusions that have little or no practical application (Schwab, 1970).

Experimental methods case study

In the online *Journal of Digital Information*, Salamapsis and Diamantaras (2002) describe an experimental evaluation of two different search system architectures for digital libraries. The authors were seeking to determine the relative effectiveness of the open hypermedia system (OHS) for retrieving information in comparison to web browser-based searching (WWW).

Twenty-four subjects were randomly assigned to either the OHS treatment or the WWW treatment. Each subject was tested individually by being given the same information query problem. They each had thirty minutes to find as many relevant documents as possible from a predefined digital library. Recall (the proportion of relevant documents that are retrieved from the collection of all relevant documents) and precision (the proportion of documents retrieved that are relevant to the information being sought) were calculated for the search results of each participant.

Although it was found that the subjects using the OHS treatment were more effective in terms of both recall and precision, the results were not statistically significant. Salamapsis and Diamantaras (2002) presented several arguments for the importance of their findings, but in the end, they concluded that "because the results cannot be validated statistically, the views and statements reported in [their] paper should be regarded as indicative and tentative." The "no significant differences" problem has been evident in decades of research and evaluation in educational contexts (Clark, 2001). Even if this evaluation had revealed statistically significant differences in the OHA and WWW search tools, there would be no guarantee that the results found in such a controlled experiment would generalize to the rough and tumble world of real world digital library usage.

Print references

A basic evaluation textbook that describes experimental approaches is *Evaluation: A Systematic Approach* (Rossi, Lipsey, & Freeman, 2003), now in its seventh edition. Fitz-Gibbon and Morris' (1987) book, *How To Design a Program Evaluation*, is part of a ten-volume *Program Evaluation Kit* published by Sage Publications that encompasses experimental methods as well as alternative models (http://www.sagepub.com/).

Online references

Scientific Research in Education, a volume published online by the National Academy Press in 2002, provides guidance for evaluators who choose to use experimental methods similar to those employed in medical trials:
http://www.nap.edu/books/0309082919/html/R1.html.

Online instruments, tools, guidelines, etc.

This website from the World Bank clarifies the difference between true experimental and quasi-experimental methods:
http://www.worldbank.org/poverty/impact/methods/designs.htm.

Tools for establishing the rationale for experimental evaluation approaches can be found at:
http://www.children.smartlibrary.org/NewInterface/segment.cfm?segment=2446.

Chapter

12

Evaluation Reporting

> .....evaluation results were seldom compelling to the interests and ideologies of stakeholders, stakeholders usually regarded scientific input as minor in decision making, and problem solving is far from a rational endeavor.
>
> \- Shadish, Cook, Leviton, 1991

Evaluations may be planned and implemented with great care and expertise, but unless they are reported in an accurate and timely manner, they will have been fruitless exercises. As noted in the quote above, evaluation results are just one source of influence competing for the attention of stakeholders, and not always the most compelling. Evaluations are not ends in themselves, but a means to better decision making. Unless digital library decision makers (funding agency officers, advisory panels, policy committees, administrators, and so forth) receive credible information provided by an evaluation at the times when critical decisions must be made, the evaluation might as well have never been done in the first place.

What are the characteristics of good reporting?

In presenting the findings of an evaluation, remember that most stakeholders want more than "just the facts." They expect you to explain how you have collected the data and how you arrived at the interpretations and recommendations in your report. Reporting an evaluation is as much about telling the "story" of the evaluation in a convincing manner as it is about rendering sophisticated tables, charts, and statistical analyses. Frankly, people seldom remember figures and graphs, but they do recall stories. Moreover, they are much more likely to share stories, and thus, in turn, influence other stakeholders.

Traditional stories have plot components, and so do evaluation reports. Your evaluation report should include a rich description of the context for the evaluation. It should explain the unique nature of the digital library being evaluated. Include hot links to the library if the report is digital or screen captures that illustrate its features if the report is a print document. Strive to give the reader a feel for the digital library. At a minimum, an evaluation report should answer the following stakeholder questions:

- What is the background of this digital library? Who created it? How is it funded? Who does it serve? What are its unique affordances? What are its future prospects?

- Why is the purpose of the evaluation? What decisions are the results intended to inform? What questions were addressed?

- What methods were used? What is the alignment among decisions, questions, and methods? How were evaluation participants recruited?

- What worked as planned? What was changed during the implementation of the evaluation? What limitations exist that must be taken into account when reviewing the results?

- What were the results? How do the results align with the questions and decisions? How do different groups of stakeholders interpret the results?

- What recommendations can be made based upon the results? What are the anticipated outcomes of making different decisions? What trade-offs, if any, are evident?

How do you prepare evaluation reports?

The reality is that most evaluations are still reported as written documents, although they are often shared electronically as Adobe Portable Document Format (pdf) files or in other digital formats. A final written report should contain all the elements that will make it useful to the decision makers and other stakeholders. Here is an outline of a typical evaluation report:

1. Title Page
2. Table of Contents
3. Executive Summary
4. Overview and Background (what was evaluated and purpose of the evaluation)
5. Decisions (intended to be influenced by the evaluation) and Questions (that were addressed)
6. Methodology (the evaluation design and any instruments that were used)
7. Results (organized by methods, e.g. interviews, questionnaires, observations, or by questions)
8. Discussion and Recommendations
9. References
10. Appendices

Most reports start with an executive summary summarizing the findings and presenting the recommendations along with a brief rationale for each recommendation. Pay special attention to crafting a compelling executive summary because this is the only part of your report that many, if not most, decision makers will read. As illustrated in the following hypothetical example, the structure of the executive summary should emphasize the major recommendations stemming from the evaluation and provides minimal explanation wherever required.

Evaluation of the Southern Botanicals Digital Library

Introduction

The Southern Botanicals Digital Libraries (SBDL) is a digital repository of educational resources focused on enhancing K-12 science education through the study of endangered plants. In response to a mandate from the SBDL funding agency, the American Botanical Foundation, to evaluate its efficacy, an external evaluation was conducted by evaluators from Old South University using three primary methods: (1) transaction log analysis, (2) interviews with the teachers and students from selected rural, suburban, and urban school districts, and (3) focus groups conducted with curriculum directors at the annual meeting of the American Science Teachers Society (ASTS).

Overview of Results

The SBDL users download more than 9,000 resources per month. Transaction log analysis indicated that 30 percent of users come from the K-12 community, 25 percent from higher education institutions, 20 percent from other botanical sites, and the rest from the general public or undefined. Just over 80 percent of the K-12 users come from schools in the eleven southeastern states represented in the SBDL collection. Interviews revealed enthusiastic adoption of the SBDL resources in rural and suburban school districts, but minimal usage in large inner-city urban districts. Focus groups at the ASTS meeting indicated that more than half of the science curriculum developers were unaware of the SBDL. Focus group participants praised the diversity and media components of the SBDL, but complained that there were insufficient capabilities for searching for resources that met specific national, state, and district science education goals and objectives.

Recommendations

- It is recommended that the SBDL strive to incorporate more resources that permit the integration of resources in urban settings. Educators from urban districts are unable to participate in several of the well-received initiatives of the SBDL such as the planting of bog gardens in coordination with the Southeastern Wildlife Organization. Alternative urban gardening projects should be defined and appropriate educational resources should be adopted or created.

- It is recommended that all educational resources be searchable by national science standards as well as by the standards of the eleven states represented in the SBDL collection. A mechanism whereby school districts might automate the process of linking their science education objectives with the SBDL collection should also be explored.

- It is recommended that the developers of SBDL seek to become a collection included in the National Science Digital Library (NSDL). This will increase awareness of the SBDL.

In addition, most evaluation reports include appendices that provide greater detail about various aspects of the evaluation. Appendices often include copies of the instruments used in the evaluation and even transcripts of original source data.

Given the nature of digital libraries, consider utilizing alternative reporting formats such as Web pages and video to present your results in the most compelling way. A well-designed Web report would include links to the library itself and to specific features of the library that have been evaluated. An online report can be easily linked to online discussion forums to allow all stakeholders to participate in on-going discussions of the evaluation results. Such discussions can be especially powerful in helping the results of an evaluation to be transformed into action.

Video evaluation reports may require additional resources, but a professional quality video report can have an enormous impact on decision makers. Videos can also be used to kick-off focus group discussions of evaluation reports involving critical groups of stakeholders. Although they are not focused on digital libraries per se, the video reports of educational technology integration initiatives produced by the George Lucas Foundation (http://www.glef.org) provide excellent models for video evaluation reports of digital library projects.

Sample Reports

There are numerous examples of evaluation reports on the Web that can serve as models for your reports.

A report of an evaluation report focused on a digital library resource can be found at:
http://it.coe.uga.edu/~treeves/RSUSeval/.

An early report of the evaluation of the American Memory Project digital library is available at: http://memory.loc.gov/ammem/usereval.html.

Columbia University's evaluation report concerning the Online Books project is at: http://www.columbia.edu/cu/libraries/digital/texts/about.html. This report is available in three different formats: Microsoft Word, HTML, and Adobe pdf.

Several evaluation reports of the Perseus digital library are available on the Web at:
http://www.perseus.tufts.edu/FIPSE/.

A variety of evaluation reports from the California Digital Library project are available at: http://www.cdlib.org/inside/assess/evaluation_activities.html.

An example of an external evaluation report is the Project JSTOR evaluation at:
http://www.mnprivatecolleges.com/jstor/images/jstor_finalreport_pdsce.pdf.

An excellent way of spreading the word about an evaluation of a digital library project is to publish a report in a journal such as Budhu and Coleman (2002) in: http://www.dlib.org/dlib/november02/coleman/11coleman.html.

Print references

Morris, Fitz-Gibbon, and Freeman's (1987) book, *How To Communicate Evaluation Findings*, is part of a ten-volume *Program Evaluation Kit* published by Sage Publications (http://www.sagepub.com/).

Most evaluation textbooks will include a chapter or section about evaluation reporting and occasionally a sample report in an appendix. Two of the best evaluation texts have been authored by Michael Quinn Patton (1997, 2002): *Utilization-Focused Evaluation: The New Century Text* and *Qualitative Research and Evaluation Methods* (both in their third editions). Adding these volumes to your evaluation resources collection would be a good investment.

Online references

Chapter Five from the *User-Friendly Handbook for Project Evaluation* developed by the National Science Foundation contains examples of project evaluation reports at: http://www.ehr.nsf.gov/rec/programs/evaluation/handbook/chap5.pdf.

A 2002 edition of the National Science Foundation's *User-Friendly Handbook for Project Evaluation* is available at: http://www.nsf.gov/pubs/2002/nsf02057/start.htm.

A *User-Friendly Handbook for Mixed Method Evaluations* from the National Science Foundation is at: http://www.ehr.nsf.gov/EHR/REC/pubs/NSF97-153/start.htm.

Online instruments, tools, guidelines, etc.

Some very useful guidelines can be found in Gary Marchionini's (2000) paper titled "Evaluating Digital Libraries: A Longitudinal and Multifaceted View" available online at: http://ils.unc.edu/~march/perseus/lib-trends-final.pdf.

References

Arms, W. (2000). *Digital libraries*. Cambridge, MA: MIT Press. Available at: http://www.cs.cornell.edu/wya/DigLib/index.html.

Bias, R. G. (1994). The pluralistic usability walkthrough: Coordinated empathies. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 63-76). New York: John Wiley & Sons.

Bollen, J., & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns*, D-Lib Magazine*, *8*(6). Available at: (http://www.dlib.org/dlib/june02/bollen/06bollen.html).

Bollen, J., Luce, R., Vemulapalli, S. S., & Xu, W. (2003). Usage analysis for the identification of research trends in digital libraries, *D-Lib Magazine*, 9(5). Available at: (http://www.dlib.org/dlib/may03/bollen/05bollen.html).

Borgman, C. L. (Ed.) (1990). *Scholarly communication and bibliometrics*. Newbury Park, CA: Sage Publications.

Budhu, M., & Coleman, A. (2002) The design and evaluation of interactivities in a digital library. *D-Lib Magazine*, *8*(11). Available at: http://www.dlib.org/dlib/november02/coleman/11coleman.html.

Bush, V. (1945) As we may think. *Atlantic Monthly*, *176*, 101-108. Available at: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm.

Campbell, N. (2001). *Usability assessment of library-related web sites: Methods and case studies* (Guide # 7). Chicago: American Library Association.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Casti, J. L. (1989). *Paradigms lost: Images of man in the mirror of science*. New York: William Morrow.

Chen, H. (1990). *Theory-driven evaluations*. Newbury park, CA: Sage.

Cherry, J. M., & Duff, W. M. (2002). Studying digital library users over time: A follow-up survey of *Early Canadiana Online*. *Information Research*, 7(2). Available at: http://InformationR.net/ir/paper123.html.

Choudhury, S., Hobbs, B., Lorie, M., & Flores, N. (2002). A framework for evaluating digital library services. *D-Lib Magazine*, *8*(7/8). Available at: (http://www.dlib.org/dlib/july02/choudhury/07choudhury.html).

Clark, R. E. (Ed.). (2001). *Learning from media: Arguments, analysis, and evidence*. Greenwich, CT: Information Age Publishing.

Cook, T. D., & Gruder, C. L. (1978). Metaevaluation research. *Evaluation Quarterly*, *2*(1), 5-51.

Covey, D. T. (2002). *Usage and usability assessment: Library practices and concerns*. Washington, DC: Council on Library and Information Resources.

Dilevko, J., & Dolan, E. (1999). *Government documents reference service in Canada: Implications for electronic access*. Available at: (http://dsp-psd.pwgsc.gc.ca/Rapports/Dilevko_Dolan/dilevko-e.html).

Durrance, J. C., & Fisher, K. E. (2003). Determining how libraries and librarians help. *Library Trends, 51(4)*, 305-334. Available at: (http://www.si.umich.edu/libhelp/DURRAFISH.pdf).

Fink, A. (2002). *The survey kit.* Thousand Oaks, CA: Sage Publications.

Fink, A. (1995). *The survey handbook.* Thousand Oaks, CA: Sage Publications.

Fitz-Gibbon, C. T., & Morris, L. L. (1987). *How to design a program evaluation*. Newbury Park, CA: Sage.

Hackos J. T., & Redish J. C. (1998). *User and task analysis for interface design.* New York: John Wiley & Sons.

Harter, S., & Hert, C. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. In M. Williams (Ed.), *Annual review of information science and technology* (pp. 3-94), Medford, NY: Information Today Inc.

Heath, F., Kyrillidou, M., Webster, D., Choudhury, S., Hobbs, B., Lorie, M., & Flores, N. (2003). Emerging tools for evaluating digital library services: Conceptual adaptations of LibQUAL+ and CAPM. *Journal of Digital Information*, *4*(2). Available at: (http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Heath/).

Hernon, P. (2002). Quality: New directions in the research. *The Journal of Academic Librarianship*, *28*(4), 224-231.

Hovater, M., Krot, M., Kiskis, D. L., Holland, I., & Altman., M. (2002). *Usability testing of the Virtual Data Center*, Usability Workshop at Second ACM+IEEE Joint Conference on Digital Libraries. Available at: http://www.uclic.ucl.ac.uk/annb/DLUsability/Hovater7.pdf.

Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. (2000). A transaction log analysis of a digital library. *International Journal on Digital Libraries*, *3*, 152-169.

Kahle, B., Prelinger, R., & Jackson, M. E. (2001). Public access to digital material. *D-Lib Magazine*, *7*(10). Available at: http://www.dlib.org/dlib/october01/kahle/10kahle.html.

Kahn, M. J., & Prail, A. (1994). Formal usability inspections. In J. Nielson and R. L. Mack (Eds.), *Usability inspection methods* (pp. 141-171). New York: John Wiley & Sons.

Keith, S., Blandford, A., Fields, B., & Theng, Y. L. (2002). An investigation into the application of Claims Analysis to evaluate usability of a digital library interface. Technical report available at:: http://www.cs.mdx.ac.uk/ridl/UET/.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Kyrillidou, M. (2002). From input and output measures to quality and outcome measures, or, from the user in the life of the library to the library in the life of the user. *Journal of Academic Librarianship*, *28*(1/2), 42-46.

Lagemann, E. C. (2000). *An elusive science: The troubling history of educational research.* Chicago: The University of Chicago Press.

Maeda, A. (2002, September). Multilingual information processing for digital libraries. In *Proceedings of the PNC Annual Conference and Joint Meetings 2002*, pp. 80-81 (abstract), Osaka, Japan. Available at: http://pnclink.org/annual/annual2002/pdf/0921/12/c211206-2.pdf.

Marchionini, G. (2000). Evaluating digital libraries: A longitudinal and multifaceted view. *Library Trends*, *49*(2), 304-333. Preprint available online at: http://ils.unc.edu/~march/perseus/lib-trends-final.pdf.

Marchionini, G., Plaisant, C., & Komlodi, A.(in press) The people in digital libraries: Multifaceted approaches to assessing needs and impact. A. Bishop, B. Buttenfield, & N. Van House (Eds.), *Digital library use: Social practice in design and evaluation.* MIT Press. Draft available at: (http://ils.unc.edu/~march/revision.pdf).

Mark, M. M., & Shotland, R. L. (Eds.). (1987). *Multiple methods in program evaluation.* San Francisco: Jossey-Bass.

Mead, J. P., & Gay, G. (1995). Concept mapping: An innovative approach to digital library design and evaluation. *ACM SIGOIS Bulletin*, *16*(2), 10-14.

Melucci, M. (1999). An evaluation of automatically constructed hypertexts for information retrieval. *Journal of Information Retrieval*, *1*(1-2), 91-114.

Miles, M. B., &Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage Publications.

Nielsen, J. (2001). *First rule of usability? Don't listen to users.* Available at: http://www.useit.com/alertbox/20010805.html.

Nielsen, J. (2000). *Designing Web usability*. Indianapolis, IN: New Riders.

Nielsen, J. (1993). *Usability engineering*. San Francisco: Morgan Kaufmann.

Norlin, E. (2000). Reference evaluation: A three-step approach-surveys, unobtrusive observations, and focus groups. *College & Research Libraries*, *61*(6), 546-553.

Norlin, E., & Winters, C. (2002) *Usability testing for library web sites: A hands-on guide*. Chicago: American Library Association.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Peters, T.A. (1993). The history and development of transaction log analysis. *Library Hi Tech*, *11*(2), 41-66.

Peterson, E., & York, V. (2003). User evaluation of the Montana Natural Resource Information System (NRIS): In-depth evaluation of digital collections using snowball sampling and interviews, *D-Lib Magazine*, *9*(7/8). Available at: http://www.dlib.org/dlib/july03/peterson/07peterson.html.

Pettigrew, K. E., & Durrance, J. C. (2001). *Public use of digital community information systems: Findings from a recent study with implications for system design.* International Conference on Digital Libraries *Proceedings of the first ACM/IEEE-CS joint conference on Digital Libraries***,** 136-143.

Reeves, T. C., & Hedberg, J. G. (2003). Interactive learning systems evaluation. Englewood Cliffs, NJ: Educational Technology Publications.

Reid, J. (2000). A task-oriented non-interactive evaluation methodology for information retrieval systems. *Journal of Information Retrieval*, *2*(1), 115-129.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Salamapsis, M., & Diamantaras, K. (2002). Experimental user-centered evaluation of an open hypermedia system and web information-seeking environments. *Journal of*

Digital Information, *2*(4). Available at: http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Salampasis/.

Saracevic, T., & Dalbello, M. (2001). A survey of digital library education. *Proceedings of the American Society for Information Science and Technology, Vol. 38* (pp. 209-223). Silver Spring, MD: American Society for Information Science and Technology. Available at: http://www.scils.rutgers.edu/~tefko/ProcASIST2001.doc.

Saracevic, T. (2000). Digital library evaluation: Toward evolution of concepts. *Library Trends*, *49*(2), 350-369.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 138-148). New York: Association for Computing Machinery.

Schutt, R. K. (2003). *Investigating the social world: The process and practice of research* (3rd ed.) Thousand Oaks, CA: Pine Forge Press.

Schwab, J. J. (1970). *The practical: A language for curriculum*. Washington, DC: National Education Association, Center for the Study of Instruction.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage Publications.

Shavelson, R. J., Towne, L., & the Committee on Scientific Principles for Education Research. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press. Available at: http://www.nap.edu/books/0309082919/html/R1.html.

Shneiderman, B. (1987). *Designing the user interface: Strategies for effective human-computer interaction*. Reading, MA: Addison-Wesley.

Shulman, L. (2000). Inventing the future. In P. Hutchings (Ed). *Opening lines: Approaches to the scholarship of teaching and learning.* Menlo Park, CA: Carnegie Publications. Available at: (http://www.carnegiefoundation.org/elibrary/docs/inventing.htm).

Spradley, J. P. (1997). *Participant observation*. New York: Holt Rinehart & Winston

Suchman, E. A. (1967). *Evaluative research*. New York: Russell Sage.

Tate, R. (1990). Experimental design. In H. J. Walberg & G. D. Haertel (Eds.), *The international encyclopedia of educational evaluation* (pp. 553-561). New York: Pergamon Press.

Tenopir, C. (2003). *Use and users of electronic library resources: An overview and analysis of recent research studies*. Washington, DC: Council on Library and Information Resources. Available at: http://www.clir.org/pubs/reports/pub120/pub120.pdf.

Wallace, D. P., & Van Fleet, C. (2000). (Eds.). *Library evaluation: A casebook and can-do guide*. Englewood, CO: Libraries Unlimited.

Wildemuth, B. M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A. & Gruss, R. (2003, June) *How fast is too fast? Evaluating fast forward surrogates for digital video*. Paper presented at the 2003 Joint Conference on Digital Libraries (JCDL '03). Available at: (http://www.jcdl.org/awards.shtml).

Wragg, E. C. (1999). *An introduction to classroom observation* (2nd ed.). London: Routledge.

Wu, M., & Sonnenwald, D.H. (1999). Reflections on information retrieval evaluation. *Proceedings of the 1999 EBTI, ECAI, SEER & PNC Joint Meeting.* Academia Sinica, Taipei, Taiwan.

Zhang, Z., Basili, V., & Shneiderman, B. (1999). Perspective-based usability inspection: An empirical validation of efficacy. *Empirical Software Engineering, 4*(1), 43-69.

Index (to be developed)