Unobtrusive Data Analysis of Digital Reference Questions and Service at the Internet Public Library: An Exploratory Study

DAVID S. CARTER AND JOSEPH JANES

ABSTRACT

THIS ARTICLE PRESENTS THE RESULTS OF AN exploratory study attempting to establish a methodology for the unobtrusive analysis of a major digital reference enterprise. Logs of over 3,000 questions asked of the Internet Public Library in early 1999 were analyzed on the basis of questions asked (subject area, means of submission, self-selected demographic information), how those questions were handled (professional determination of subject and question nature, questions sent back to users for clarification), and answered (including time to answer) or rejected. In addition, answers that received unsolicited thanks were analyzed separately. Users seem to have difficulty in assigning subject categories to their questions, and to determine whether they are factual or require sources for assistance, and these decisions were often overridden by question administrators. The median time to answer questions was just over two days, and about one in five answers received thank-you messages from users.

INTRODUCTION

The advent of digital reference creates for librarians many new opportunities. Most of these revolve around new ways of offering service—i.e., at different times, to different populations, via different media, etc. However, since reference services delivered through the Internet are

David S. Carter, School of Information, 304 West Hall, University of Michigan, Ann Arbor, MI 48109-1092
Joseph Janes, Information School, 330M Mary Gates Hall, University of Washington, Seattle, WA 98195-2930
Library Trends, Vol. 49, No. 2, Fall 2000, pp. 251-265

mediated in a chiefly textual environment, digital reference services also afford us new ways of examining the activities of reference. At the Internet Public Library (IPL), we have been providing digital reference services to our international patron group since opening on March 17, 1995, over five years ago. During that time, we have kept an exact record of every reference interaction that we have handled, over 40,000 questions to date. In this article, our goal is to explore just what sort of things can possibly be learned by examining this record.

As this is an exploratory study, we have limited our data set of interest to the questions received during the three-month period from January to March 1999. This period provides over 3,000 questions to examine. We are also purposely limiting inquiries to rather elementary data analysis—no content analysis or direct patron inquiries—as we are primarily interested in what sort of data can be drawn out of the amalgamation of questions via automatic means. In short, we want to know if anything useful can be learned about a digital reference service without investing a huge amount of resources.

As is typical with these sorts of studies, our explorations raise as many, if not more, questions than answers. In the conclusion of this article, we examine the more complex inquiries that are suggested by our elementary data analysis. We also consider ways in which the service itself might be modified to allow for more and more complex information to be gathered non-intrusively.

Our research questions were:

- What are important characteristics of questions and users (user-assigned subjects, self-identification of users)?
- How frequently do IPL administrators override user-defined subjects and nature of questions?
- How frequently do IPL question-answerers use internal features of the question-answering system?
- How long do answerers take to answer questions?
- Who sends thank-you messages back to the IPL?
- What are important characteristics of rejected questions?

Review of Related Literature

Although digital reference services have been a part of libraries for some time, most of the literature has been anecdotal in nature. The few studies that have been done have generally focused on the nature and existence of these services (e.g., Janes, Carter, & Memmott, 1999, for academic libraries; Garnsey & Powell, 2000, for public libraries) and not any sort of qualitative or quantitative approach to the results or outcomes of these services. In a sense, this study is in the tradition of the numerous studies involving the evaluation of traditional reference services (e.g.,

Hernon & McClure, 1987; Durrance, 1989) as well as transaction log analysis (Peters, Kaske, & Kurth, 1993). The unobtrusive nature of our study shares some of the inherent limitations of transaction log analysis; as Kurth (1993) states: "Transaction log data... don't reflect, except through inference, who enters the searches, why they enter them, and how satisfied they are with their results" (p. 98). However, we are unaware of any previous studies of digital reference services and, as such, are taking the first small steps into a new area of inquiry with this study.

Overview of Internet Public Library Reference

Internet Public Library reference has been covered in detail in many other places (e.g., Lagace, 1999; Lagace & McClennen, 1998). However, we feel that it would be instructive to provide first a brief overview of the process before diving into the data.

Users are invited to ask their questions by completing one of two forms: either a general purpose form (http://www.ipl.org/ref/QUE/RefFormQRC.html) or a youth form (http://www.ipl.org/youth/refform.html). We also take questions that have been submitted via e-mail. Users are informed that their question may be used for research purposes, as per the IPL Privacy Statement (http://www.ipl.org/about/privacy.html). All of the questions received by the IPL are entered into QRC, our Web-based centralized software used for patron interaction in general and reference administration in particular (Lagace & McClennen, 1998). Questions to QRC become items, and each item can exist in one of several categories. Questions are first relegated to an Incoming category where an IPL reference administrator (a "mucker" in IPL lingo) performs the initial tasks on the question—chiefly accepting or rejecting the question (and notifying the patron) but also assigning a subject and a subject line, verifying the e-mail address, deciding if it is a "sources" or "factual" question (see definitions below) and, if necessary, asking the patron for clarifying information. These administrators are experienced in the use of QRC, the IPL question-answering process and guidelines, and are either advanced students or volunteer professionals.

From there the question is transferred to one of two "To be Answered" categories, one each for factual and source questions. The questions are then available to be answered by the cadre of IPL reference librarians, who choose from among the available questions and CLAIM a question to indicate that they are working on it. During the process of finding an answer for the question, the librarian may post messages to herself (or, in fact, messages on others' questions as well) via a FOLLOWUP,[1] or ASK_INFO functions so as to seek further clarifying information from the patron. Finally, a question is ANSWERED by sending an e-mail response via QRC back to the patron. A patron may decide to respond back to the question, usually to ask for more information or to offer a note of thanks.

After the question has been answered, an administrator checks the answer (an important step, as IPL reference is chiefly an educational enterprise and many of the answerers are students still in the process of learning reference techniques) and then removes the item from the category. The entirety of this reference interaction is then filed away into the QRC archives.

METHODOLOGY

Although the current version of the QRC software is not built into a database, the questions, answers, and attendant interactions are stored in formatted text files. Thus it was rather straightforward to write a program in Perl to cull through the files and extract the desired data. When possible confusion arose, consultation with the reference administrators was able to clear up any points about the subtleties of the administration process. Data now in hand, a variety of exploratory analyses were performed, the results of which we will now go into in detail.

Results and Discussion

During the period used for this study, January-March 1999, 3,022 questions were submitted to the IPL. The entire corpus was analyzed using automatic processing of QRC archive files.

The first area we examined was the nature of the questions asked by Internet Public Library patrons. We looked at three areas: what means were used to ask the question (form or e-mail), the subject assigned to the question by the user, and self-identified demographic characteristics.

Table 1 shows the source of the questions received—i.e., whether the questions were submitted via the standard form, the youth form, e-mail (to any @ipl.org address), or by an unknown means (usually from another form on the IPL site—e.g., a patron might ask a reference question in a form intended to suggest a site for the IPL's Online Newspapers collection). As can be seen, the majority of the questions received, 68 percent, come from the general reference form and 26 percent arrive via e-mail. Only 4 percent of the patrons used the youth form. This 26 percent is an important number: these questions have much less structure—i.e., they do not have the field structure of questions that come in via the form and, more importantly, they do not necessarily have the information

Table 1. Source of Questions.

Source	Number	Percentage
form	2064	68.3
e-mail	788	26.1
kidform	127	4.2
unknown	43	1.4

requested on the form, which is most valuable when answering. Often, e-mail-based questions do not specify sources already consulted, motivation or reason behind the question, intended uses for the information, and so on. This has a significant impact both on policy and performance.

The two forms ask patrons to identify the subject area of their questions. Table 2 shows the distribution of these choices. Note that nearly one-third of the questioners were unable to match the subject area of their question to the list provided and thus chose "Other/Misc." (this is not the default setting on the form—the patron is forced to choose a subject area when submitting and must actively select "Other/Misc." from the bottom of the list of available choices). A comparison of the data in Table 2 to that in Table 5, the subject area chosen by the IPL reference administrators, shows a serious disconnect between the two. This has significant implications, especially in the realm of automated assistance in reference question processing—i.e., any system that relies on users to self-identify their questions will end up with a significant number of questions in the wrong places within the system, and thus the system will still require a substantial hands-on component from human beings.

Table 2. Subjects Assigned by User(chosen from form).

Subjects	Number	Percentage
Other/Misc	869	28.8
<blank> - usually e-mail	795	26.3
Education	196	6.5
Science	186	6.2
Humanities	166	5.5
Government/Law	150	5.0
Business/Economics	121	4.0
Libraries/Librarians	105	3.5
Health/Nutrition	63	2.1
Entertainment	58	1.9
Computers	49	1.6
Internet	44	1.5
Social Services/Issues	39	1.3
Environment	35	1.2
News/Current Events	29	1.0

The general reference form gives patrons the options to identify themselves as a businessperson, a teacher, and/or a librarian. This is done so that the administrators and answerers can have a better understanding of the background of the answerer and what resources may be available to them. Table 3 shows the distribution of these choices: nearly 25 percent of the patrons using the form identify themselves as business persons, 11 percent as teachers, 7.5 percent as librarians (only 15 people chose more

than one category; nobody chose all three). These numbers should be taken with a grain of salt, as it is quite possible that a patron chooses one of these options to give themselves, and thus their question, a seeming higher level of import. We also know, from anecdotal evidence and spot-checking, that oftentimes questions from persons identifying themselves as business people are not business related and are, in fact, reflecting personal information needs.

Table 3. People who Identify Themselves as . . .

Options	Number	Percentage of Questions from Form
Business People	501	24.3
Teachers	234	11.3
Librarians	153	7.4

(only 15 people chose more than one category; nobody chose all 3)

In addition, both forms ask if the question is "for a school assignment," again so that the people answering have a better idea of how to properly respond to the question. Over half of the patrons using the reference forms identify their question as being school related (1,073 or 52 percent), indicating a high level of educational usage for the IPL reference service.

What We Did with the Question

Administration

Sources vs. Factual. Patrons can specify whether they want their question answered with a brief *factual* answer to their query, or a list of *sources* to consult to help them with their quest (or *nothing* may be indicated, especially if the question comes via e-mail).

When processing the incoming questions, the IPL administrators make this judgment. Based on the nature of the question and their own experience, a question is accepted as either *factual* or *sources*, indicating to the people answering what the most likely type of response should be given. A question may also be *rejected*—i.e., not accepted into the question pool (more discussion of this later).

Table 4 shows the distribution of *factual*, *sources*, and rejected questions, comparing the patrons' expectations with the administrators' assignments. While the patrons were very evenly split among their choices (one-third each for *sources*, *factual*, and nothing responses), the administrators were more than twice as likely to assign a question as being *sources* rather than as *factual*. It is quite likely that patrons are being overly opti-

mistic that their questions can be answered simply and directly. We also note that the rejection rate for questions is independent of whether the patrons say they want a *factual* or *sources* answer.

Table 4. Factual, Sources, Rejected Distributions.

Of the 3022 Questions,
the user said they wanted *sources* 986 times (33%)
we agreed 681 times (69%)
we reversed 70 times (7%)
we rejected 235 times (24%)
the user said they wanted *factual* answers 995 times (33%)
we agreed 357 times (36%)
we reversed 395 times (40%)
we rejected 243 times (24%)
the user said *nothing* 1041 times (34%)
we said sources 614 times (59%)
we said factual 205 times (20%)
we rejected 222 times (21%)
Overall,
1690 questions were answered with sources (56%)
632 questions were answered with factual answers (21%)
700 questions were rejected (23%)

Question Subject

IPL staff also assign subject categories to each question via subject codes that are appended to the beginning of the description line for each question. These categories are slightly different from those from which the patron can choose, but it is fairly easy to relate one set to another. (Questions that have been rejected do not receive subject codes.) The distribution of the subjects assigned by administrators is shown in Table 5. It is important to note that two of these designations, FARQ and PF, are not actually subjects but rather indicate that the question was responded to by the administrator using a standard response referring the patron to one of the IPL's Frequently Asked Reference Questions (FARQ) or Pathfinders (PF). This is also interesting since, even though patrons are encouraged to look over these resources on the IPL Web site prior to asking their questions, 13 percent of the questions are still answered in this fashion. Another important thing to note is that the number of Health and Law/Legal questions will be artificially low—as is noted in the section on Rejection below; questions on these subjects are routinely rejected for being outside the scope and purview of the IPL service.

Table 5. Subjects Chosen by Staff.

Subject Code	Number	Percentage	Subjects Chosen by Staff
FARQ	228	9.8	(Frequently Asked Reference Question)
SCI	225	9.7	Science
HIS	201	8.7	History
LIT	184	7.9	Literature
BIO	147	6.3	Biography
HUM	147	6.3	Humanities
LIB	129	5.6	Libraries
MSC	96	4.1	Miscellaneous
GEO	79	3.4	Geography
PF	78	3.4	(answered with IPL Pathfinder)
BUS	77	3.3	Business
POTUS	74	3.2	Presidents of the United States
ENT	60	2.6	Entertainment
SOC	57	2.5	Social Science
EDU	53	2.3	Education
GOV	49	2.1	Government
GEN	47	2.0	General Reference
INT	39	1.7	Internet
COM	38	1.6	Computers
HEA	27	1.2	Health
MUS	23	1.0	Music
LAW	19	.8	Law
DIY	16	.7	Do-It-Yourself
POL	14	.6	Politics
MIL	13	.6	Military.
PSY	13	.6	Psychology
REL	10	.4	Religion

Answering

When answering questions, IPL students, volunteers, and staff have several options. They may CLAIM a question, indicating that they are working on the question; UNCLAIM, indicating that they aren't anymore; mark a question as NEED_HELP, requesting assistance from others; or ASK_INFO, indicating that they have asked the patron for further clarifying information. The QRC system also allows anyone to post internal messages (known as FOLLOW_UPs), either as temporary notes to oneself during the process of searching for an answer or as assistance to others in answering.

Of the 2,322 questions answered (700 were rejected), 669 (28.8 percent) were answered before being posted to a "To Be Answered" category; these were answered directly by an administrator (nearly half via a FARQ or PF message) and thus will not have CLAIMs, NEED_HELPs, ASK_INFOs, and so on.

Thus, 1,653 "regular" questions were answered. Tables 6-10 show an analysis of those questions. It can be seen by these data that the majority of the questions are answered in what would be considered a "standard" fashion—i.e., CLAIMed once, never UNCLAIMed, with no FOLLOW_UPs from either the answerer or others. However, nearly 15 percent of the questions are worked on by more than one person (i.e., CLAIMed more than once), 35 percent of the questions have FOLLOW_UPs by the answerer, and 25 percent have FOLLOW_UPs by someone other than the answerer. The average number of self-FOLLOW_UPs is 0.63, and the average number of FOLLOW_UPs by others is 0.44. Only a small fraction of the questions were ever marked NEED_HELP or ASK_INFO—by corollary, IPL question answerers offer help far more often than it is requested.

Table 6. CLAIMed Questions (Being Worked On).

Number of Times CLAIMed	Number	Percentage
0	30	1.8
1	1401	84.8
2	165	10.0
3	48	2.9
4	7	0.4
5	1	0.1
6	1	0.1

Table 7. UNCLAIMed Questions (have Stopped Working on the Questions).

Number of Times UNCLAIMed	Number	Percentage
0	1449	87.7
1	160	9.7
2	36	2.2
3	7	0.4
5	1	0.1

Table 8. FOLLOW_UPs by Eventual Answerer.

Number of Times Followed Up by Eventual Answerer	Number	Percentage
0	1088	65.8
1	311	18.8
2	135	8.2
3	61	3.7
4	26	1.6
5	18	1.1
6	11	0.7
7	2	0.1
10	1	0.1

Table 9. FOLLOW_UPs by Others.

Number of Times Followed Up by Others	Number	Percentage
0	1232	74.5
1	260	15.7
2	93	5.6
3	33	2.0
4	15	0.9
5	9	0.5
6	7	0.4
7	3	0.2
8	1	0.1

Table 10. Questions Marked . . .

Options	Number	Percentage
NEED_HELP	53	3.2
ASK_INFO	66	4.0
both of these	7	0.4

Time to Answer

One important measurement of a digital reference service is the time it takes to respond to the patron with an answer (patrons are promised their answer within one week of posting). To evaluate this, we examined the time to answer the question as measured in days, as recorded automatically from the time the question was received at the IPL to the time an answer was posted back to the patron. Figure 1 shows the distribution of these answer times, while Table 11 gives the average time to answer as well as the quartiles. (These results do not include questions that were answered directly by administrators, only those posted to a "To Be Answered" category.)

Table 11. Time to Answer (Measured in Days BetweenTime Question Posted to IPL and Time Answer Posted Back to User).

Time	Number of Days
average	2.96 (s.d. 2.70)
Q1 (25th percentile)	0.77
median	2.05
Q3 (75th percentile)	4.89
skew	1.13

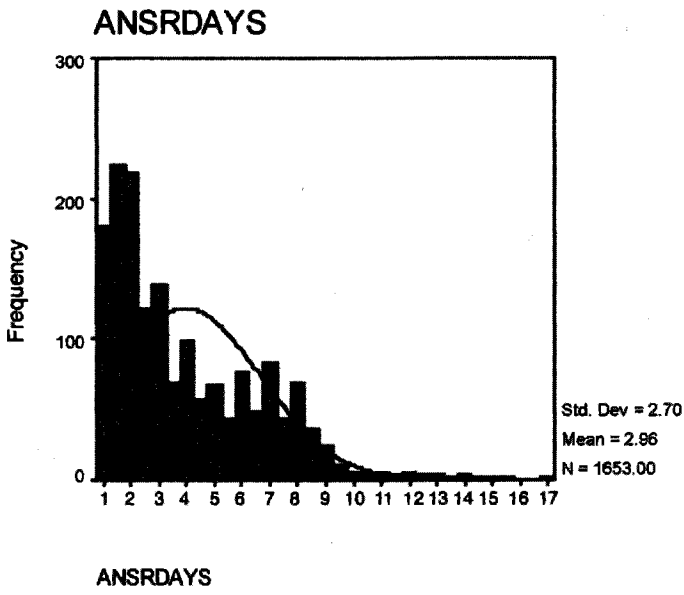

Figure 1. Distribution of Times-to-Answer.

As can be seen from the data, it takes on average nearly three days for a question to be answered, with nearly half of the questions answered within two days and more than a quarter answered within one day.

The questions answered directly by administrators averaged only 0.44 days to answer; when including these in the analysis, the overall average time to answer for all questions is 2.26 days (median 1.07 days).

We also compared time to answer with other characteristics of the question. The average time to answer was 2.10 days for a factual question and 2.31 days for a sources question—no significant difference. The fastest questions to answer were factual questions received from e-mail (average 1.69 days, n = 166); the slowest to answer were sources questions from

e-mail (average 2.38 days, n = 465). This discrepancy can partly be accounted for by the fact that a factual question is more likely to be answerable directly by an administrator by employing a standard FARQ response.

Thanks

Of the 2,322 questions answered, 458 (19.7 percent) received unsolicited thanks from users. The thank rate was 24.4 percent for factual questions, and 18.0 percent for sources questions (the difference is significant at the .001 level, C = .071).

Table 12 shows the thank rate by the subject area of the question. In general, humanities subjects rank near the top, physical sciences near the middle, and social sciences near the bottom. Unsurprisingly, users whose questions were answered with a standard FARQ or PF response were far less likely to express gratitude.

Table 12. Thank Rate by Subject Area of Question.

Subject Area	Percentage
LIB	26.2
MUS	25.9
EDU	23.8
LIT	22.3
HUM	22.2
BUS	19.0
COM	18.8
MSC	18.5
SCI	18.3
POTUS	18.3
HIS	18.2
GEO	18.1
INT	17.4
GOV	17.0
GEN	16.4
SOC	16.2
ENT	15.8
BIO	15.4
LAW	13.3
HEA	12.5
PF	0.4
FARQ	0.3

Table 13 shows the thank rate by questioner type. There does not appear to be much of a difference in the thank rate for those who do or do not choose to identify themselves as part of one of these groups, nor among the three groups.

Table 14 shows the thank rate by the question source. While there is no significant difference between questions submitted via e-mail and from

Table 13. Patron Thank Rate by Questioner Type.

Questioner Type	Percentage
librarians	25.4
business people	25.3
teachers	21.1
school asst	15.0

Table 14. Thank Rate by Question Source.

Question Source	Percentage
e-mail	20.7
form	20.3
kidform	7.6

the regular form, questions submitted using the youth form receive thanks at a far lower rate. The thank rate for questions identified as being for a school assignment was also significantly lower, 15.0 percent, leading one to a conclusion that kids send thanks along far less often than adults.

The thank rate for questions answered in less than the median time (2.05 days) was 18.4 percent; far less than the median time, 22.6 percent—a significant difference at the .01 level. This at first seems counter-intuitive in that the longer it took to answer a question, the more likely the patron was to send back a note of thanks. However, further examination of the data suggests a different factor at work. The thank rate for questions answered before posting by an administrator was 10.8 percent (72/669). The thank rate for questions (not answered before posting) with one or more FOLLOW_UPs was 28.9 percent; for those with no FOLLOW_UPs, the rate was 16.9 percent. As these factors can be taken as a measure of question difficulty (i.e., the harder a question, the longer it takes to answer, the more notes made to oneself, the more assistance offered, and so on) it can be inferred that the harder a question is, the more appreciative the patron is for the answer provided.

Rejection

Of the 3,022 questions received during the examination period, 23 percent (700) of the questions were rejected—i.e., not accepted to be answered. Table 15 shows the distribution of the reasons for rejection by the administrators. More than half of the questions were rejected because the service was over quota—i.e., the service received more questions that day than could reasonably be answered by the service. Another 18 percent were rejected because the patron wanted an answer faster than the service could provide. Still another 7 percent were rejected because the patron supplied an invalid e-mail address (and the administrator could not ascertain what the correct address was).

TABLE 15. Reason for Rejection.

Reason	Number	Percentage
quota	374	53.4
date-passed	125	17.9
bounce	51	7.3
no reply	33	4.7
law question	34	4.9
medical question	23	3.3
scope	14	2.0
closed	14	2.0
not your library	8	1.1
rerout	20	2.9

Of the 112 questions that were marked as ASK_INFO, 33 did not reply, a dropout rate of 29.5 percent. This is an indication of the difficulty of establishing any sort of dialog between the patron and the librarian in an e-mail-only environment.

Table 16 shows the rejection rate based on the source of the question. Questions submitted via the youth form were the most likely to be rejected, questions received via e-mail the least, with those from the regular form in the middle. When shown these data, the IPL reference administrators were quite surprised, as their anecdotal evidence suggested that the exact opposite was true.

Table 16. Source of Rejected Questions.

Source	Number	Reject Rate Percentage
form	496	24.0
e-mail	156	19.8
kid form	35	27.6

Table 17 shows the rejection rate by self-identification of the patron. Of note here is that perhaps it does not pay to identify oneself as a business person.

Table 17. Reject Rate by Type of Questioner.

Questioner Type	Percentage
Business	28.3
Teacher	23.1
Librarian	22.9
School	24.7

CONCLUSION

As can be seen in the earlier analysis, there is a potential wealth of information that can be culled from the data surrounding a digital reference question. However, one is obviously limited to the data collected. This may seem to be an obvious point but, when designing a reference question intake form, librarians should consider not only what they will need to answer the question, but also what sort of automatic data analysis they may wish to do in the future.

An interesting phenomenon that shows in the study is the existence of a tiered reference service: a number of questions are rejected, common inquiries are responded to via standard answers (FARQs and Pathfinders), quick questions are handled by the administrators, and "regular" questions are handled by the reference librarians. These tiers were not designed into the system, but rather have evolved from experience and are evident in the analysis.

While the data analysis is, in many aspects, interesting, in its own right it can also serve as a powerful tool for further exploration. Armed with such knowledge, we can now dive into other avenues of exploration—such as content analysis of the questions, a patron satisfaction survey, librarian attitudes, and so on—with a much better background than can be accomplished in evaluating "traditional" reference services.

Another fairly obvious extension of this analysis would be a longitudinal approach: looking at a similar period of time from 1998 and 2000 could give a picture as to how things at the service have changed (or not).

Furthermore, comparisons of data between and among other libraries, as well as other "AskA" services (e.g., "Ask A Space Scientist") and commercial question and answer services would also be instructive.

NOTE

[1] Words in all caps here are designations of question status within the QRC system.

REFERENCES

Durrance, J. C. (1989). Reference success: Does the 55 percent rule tell the whole story? *Library Journal, 114*(7), 31-36.

Garnsey, B. A., & Powell, R. R. (2000). Electronic mail reference services in the public library. *Reference & User Services Quarterly, 39*(3), 245-254.

Hernon, P., & McClure, C. R. (1987). *Unobtrusive testing and library reference services.* Norwood, NJ: Ablex Publishing.

Janes, J.; Carter, D.; & Memmott, P. (1999). Digital reference services in academic libraries. *Reference & User Services Quarterly, 39*(2), 145-150.

Kurth, M. (1993). The limits and limitations of transaction log analysis. *Library Hi Tech, 11*(2), 98-104.

Lagace, N. (1999). Establishing online reference services. In *The Internet Public Library handbook* (pp. 153-183). New York: Neal-Schuman.

Lagace, N., & McClennen, M. (1998). Questions and quirks: Managing an Internet-based distributed reference service at the University of Michigan School of Information. *Computers in Libraries, 18*(2), 24-27.

Peters, T. A.; Kaske, N. K.; & Kurth, M. (1993). Transaction log analysis. *Library Hi Tech Bibliography, 8,* 151-183.