

Chapter 2. Literature Review

2.1. Introduction

Chapter 1 began with a discussion of Shera's (1964) discussion of automation as providing the "opportunity... to analyze the reference process and re-define reference service" (p. 203). With the advent of digital reference services in the past two decades, reference service has indeed been redefined. The reference process, both at the desk and digital, has also been the subject of analysis. One of the goals of this study was to further analyze one process in digital reference service: the triage process.

This chapter provides an overview of the bodies of literature that inform this study and the analysis of the triage process. This study is grounded in the following three bodies of literature: digital reference, question answering, and classification. A review of the literature on digital reference defines this study's area of investigation, and demonstrates the need for this study. This chapter defines digital reference and triage, and explores the need for automation in both. A review of the literature on question answering serves as the foundation for this study's approach to questions. Theories of questioning behavior that have been developed and research on question answering systems that have been conducted to date are discussed. The literature on classification is reviewed in relation to questions as entities to be classified, and four question taxonomies are identified in the literature on desk and digital reference and question answering. First, however, the concept of "question" is discussed and defined.

2.2. What Is a Question?

This study is of digital reference services, and how questions received by digital reference services are handled during the triage process. This begs the question, however, of how to identify a question in the first place. This section discusses approaches to understanding the nature and function of questions. A definition of "question" will be

advanced, so that questions received by digital reference services may be identified for the purposes of this study.

2.2.1. Questions from the Perspective of Erotetic Logic

Erotetic logic, first proposed by Prior and Prior (1955), is a formal logical approach to the analysis of “all sentences that call for a reply” (Graesser and Black, 1985, p. 3). Erotetic logic decomposes questions into two parts: subject and request. The word “subject” is used here in an unconventional sense, to mean not the grammatical subject of the question, but the possible states of the world that are presupposed by the question. The request, therefore, identifies how many of these states are desired in the answer. For example, *Whether-questions* presuppose a finite set of alternatives, as in the question “Is John going home?,” for which the set of possible states of the world is that John either is or is not going home. Thus, while the grammatical subject of this question is John, the subject according to erotetic logic is John’s home-going. The request, therefore, identifies that the desired answer is one that specifies which of John’s alternative home-going states is true: that John either is or is not going home (Belnap and Steel, 1976, pp. 19-22). *Which-questions*, on the other hand, presuppose an infinite set of alternatives, as in the question “Which person is going home?” (Belnap and Steel, 1976, pp. 22-23).

This approach to questions as decomposable into a subject and a request is similar to Jahoda and Braunagel’s (1980) approach to reference questions as decomposable into a “given” and a “wanted.” According to Jahoda and Braunagel, the given is “the subject of the information need,” and the wanted is “the type of information needed about the subject” (p. 8). For example, in the reference question “I am looking for a book on herbal folklore,” the given is *herbal folklore*, and the wanted is *a book*.

Prior and Prior’s (1955) version of erotetic logic treats questions as either true or false, according to whether or not the answer to a question is a true statement. This approach is problematic, however, in that some questions do not have either true or false answers. The classic example of this is the question, “Have you stopped beating your wife?,”

which (at least when asked to a non-wife beating individual), cannot be answered either truly or falsely. This problem arises because this question presupposes something that is not true: that the individual questioned does or has beat his wife. Belnap and Steel (1976) solve this problem by suggesting a question be treated as true or false according to whether or not its presuppositions are true or false (p. 116). Belnap and Steel (1976) also suggest that there are two levels of presupposition: primary and secondary (p. 112). Primary presuppositions are those statements of fact that are implied by the question itself, and secondary presuppositions are those statements of fact that are implied by the primary presuppositions. For example, the primary presupposition of the question “Have you stopped beating your wife?” is that the individual questioned does or has beaten his wife. The secondary presupposition of that question is that the individual questioned is married.

This study takes from erotetic logic the idea that questions may be decomposed into a subject and a request.

2.2.2. Questions from the Perspective of Speech Act Theory

Speech act theory is not a theory so much as it is a model of language-in-use, in that it addresses the issue of how language is used to accomplish tasks (Austin, 1999). Speech act theory relies heavily on erotetic logic in that it relies on the ability to decompose speech acts into a subject and a request. To this end, Austin outlines three layers of a speech act:

1. Locutionary acts are “act[s] of saying something” (Austin, 1999, p. 94). A locutionary act is simply the act of uttering a series of sounds that have meaning. It is assumed in speech act theory that these sounds have meaning in the language of the speaker, and that the speaker and the hearer share this language. For example, in English, saying “mɪlplɪx”¹ or “fɑːblɒndʒet”² would

¹ Actually, “mɪlplɪx” is not an entirely meaningless locutionary act. Mr. Mɪlplɪx was a villain in Superman comics, albeit not one of the better known ones.

not be considered to be a locutionary act, as they are meaningless sounds. Saying “question” in English, on the other hand, is a locutionary act, as is the utterance of “farblondjet” by a Yiddish speaker, as these sounds have meaning in those languages.

2. Illocutionary acts are acts that are performed “*in* saying something as opposed to... [acts] *of* saying something” (Austin, 1999, p. 99). While a locutionary act merely must make sense, an illocutionary act must have force; that is, it must convey to the hearer the speaker’s meaning. Illocutionary force is therefore equated with semantic meaning, as opposed to the lexical meaning equated with locutionary acts. Searle (1968) claims that the same locutionary act (e.g.: “I am going to do it”) can “have the force of a promise, a prediction, a threat, a warning,” etc. (p. 406). The “question-ness,” as it were, of a question is due to its illocutionary force.
3. Perlocutionary acts are acts of “saying something [that] will... produce certain consequential effects upon the feelings, thoughts, or actions of the audience” (Austin, 1999, p. 101). An illocutionary act has a specific sort of force to it that gives it semantic meaning, and it is this illocutionary force (combined perhaps with other elements of the communication that neither Austin nor Searle elaborates on) that causes an effect in the audience. For example, the locutionary act “I am going to do it,” when intended with the illocutionary force of a promise, brings about the effect in the audience of expectation that the speaker will do “it,” while the same locutionary act intended with the force of a threat brings about the effect of generating apprehension in the audience.

In short, a locutionary act has meaning, an illocutionary act has meaning and force, and a perlocutionary act has meaning, force, and consequences. These three types of acts are not different *types* of speech acts, but rather are three *aspects* of any given speech act. Taken as a single entity, the speech act is treated in speech act theory as “the basic unit of

² Yiddish for “wandering,” or “completely off track.”

communication” (Searle, 1995, p. 21); that is, the speech act is the vehicle by which tasks are accomplished using language.

This study takes from speech act theory the idea that a question has force: that in asking a question, the questioner expects a certain action to be performed by the person questioned. The questioner expects that the action that will be performed will be to respond to the question, and it is irrelevant whether the person questioned is face-to-face with the questioner, or remote, and the question mediated by a computer interface.

2.2.3. Questions from the Perspective of Discourse Analysis

Discourse analysis relies heavily on speech act theory in that it is based on the premise that the speech act is the basic unit of communication (Schiffrin, 1994, p. 90). Speech act theory, however, presents a monologic model of communication: it treats speech acts as individual utterances without considering their connection to other utterances. Clark and Wilkes-Gibbs (1986) refer to this model of communication as the “literary model”: that is, communication in which “speakers refer as if they were writing to distant readers” (p. 3). According to this literary or monologic model, the force of a speech act lies entirely in the speaker; the effect of a speech act on the audience is assumed to be whatever was the speaker’s intended effect, and no provision is made for feedback from the audience to the speaker.

Discourse analysis, on the other hand, takes feedback explicitly into account. Clark and Wilkes-Gibbs (1986), in opposition to the “literary model,” propose what they refer to as the “conversational model.” The difference between the conversational model and the literary model is illustrative of the difference between discourse analysis and speech act theory: that the listener is not “mute or invisible during an utterance,” but rather is an active participant in a conversation, and indeed “speakers may alter what they say midcourse based on what addressees say and do” (p. 3). Indeed, Clark and Wilkes-Gibbs state that the listener is a collaborator in the performance of any speech act, and that for a speech act to have any force or effect at all, it must be “mutually accepted” by both the

speaker and the listener – or, more specifically, any speech act must be accepted by the listener, and the listener’s acceptance must then be accepted by the speaker.

Clark and Schaefer (1987) build on this idea of mutually accepted speech acts, and propose what they refer to as a *contribution*. A contribution is a combination of a speech act – that is, the utterance of some meaningful content – and the acceptance of that content. This acceptance occurs when “the speaker and addressees mutually believe that the addressees have understood what the speaker meant” (p. 20). Clearly, there must be some feedback mechanism at work in this model of conversation, since once the addressee accepts the speaker’s speech act, the speaker must be able to observe that acceptance and accept it for himself. In a conversation, the addressee must, further, be able to observe this acceptance by the speaker, so that mutual acceptance exists and is known and acknowledged. When that is accomplished, the original speech act achieves the status of “common ground” between the speaker and the addressee, for the purposes of the conversation.

Roberts and Bavelas (1996) propose a three-step model of the process of communication based on contributions to conversation and feedback between the speakers, who they refer to as “interlocutors.” This model is presented in Figure 2-1. This model is based on the premise that meaning is negotiated between the speakers, and there is a minimum of three steps that must be accomplished in order for any communication to be successful:

1. The Utterance is, quite simply, any single speech act, including non-verbal communication acts. So far, this is no different than the view of communication acts as viewed by speech act theory.
2. The Reaction is the addressee’s response to the initial speech act (again, it may be verbal or non-verbal). This reaction reflects the addressee’s interpretation of the initial speech act, which may be either correct (in that it corresponds to the speaker’s interpretation of his or her own speech act), or incorrect. Roberts and Bavelas state that the Reaction “may consist of a

request for clarification, a formulation or reformulation, or other explicit comment,” or it may simply be an “appropriate continuation” to the conversation, thus implicitly signaling that the Utterance was understood and that the conversation can move on, assuming the Utterance as common ground.

3. The Confirmation is the speaker’s reaction to the addressee’s reaction. Again, this may be verbal or non-verbal, and may be one or more of the options stated above. In the Confirmation, the speaker validates that the addressee’s interpretation of the initial speech act was either correct or incorrect, and, with the Confirmation statement, may correct the addressee’s interpretation at the same time. In this way, mutual acceptance of the initial speech act is accomplished.

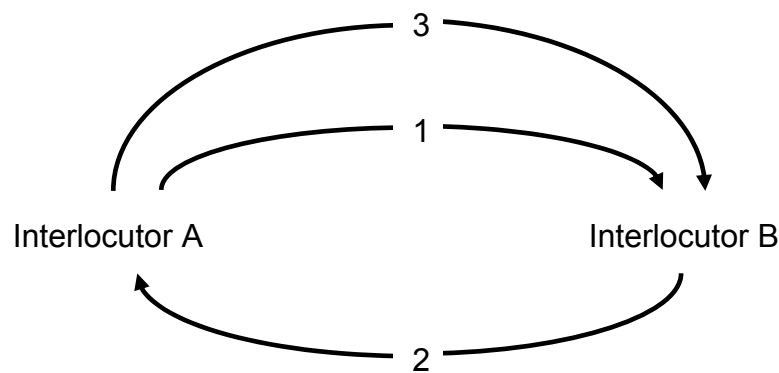


Figure 2-1: The Process of Semantic Collaboration (Roberts and Bavelas, 1996, p. 143)

This study takes from discourse analysis the idea that a question does not stand alone, but is rather one step in a larger conversation. Indeed, from the perspective of discourse analysis, a question cannot be understood only from the initial speech act; a question can only be understood by participating in, or observing, a process of communicative collaboration. Through this collaboration, a speech act may be interpreted and accepted

as a question. This interpretation and acceptance of a speech act as a question allows for two additional points that this study takes from discourse analysis. These two points are actually two sides of the same coin:

1. Questions are not necessarily phrased as questions. That is, a question can be phrased as a statement, but be interpreted and accepted as a question. For example, the statement (taken from the set of questions from phase two of this study) “I am looking for the name of the historical figure who said, ‘The difference between a politician and a statesman is, a politician looks to the next election while a statesman looks to the next generation,’” may reasonably be interpreted as a question asking: “To whom is the following quote attributed: ...?”
2. Statements phrased as questions are not necessarily questions. That is, a question may be a rhetorical question, for example, an interjection such as “What?” to indicate surprise, or a gripe phrased as a question such as “Why me?” (Graesser and Black, 1985, p. 27).

2.2.4. Questions from the Perspective of Information Needs and Uses

Information needs and the subsequent stages that a library patron passes through in order to fulfill that need are a well-studied area of library and information science. These topics are discussed in depth by such authors as Belkin, Oddy, and Brooks (1982a, 1982b); Dervin and Nilan (1986); and Kuhlthau (1991), and will therefore not be discussed in depth here. Briefly, however, a patron comes to an information system (one example of which is a library reference desk) with an information need that he or she may or may not be able to fully articulate, and in the process of attempting to fulfill that need progresses through several stages in which the need becomes increasingly well-defined in the patron’s mind. Taylor (1968) describes four levels of information need, the first and most basic of which is “the conscious or even unconscious need... probably inexpressible in linguistic terms” (p. 182). Belkin, Oddy, and Brooks (1982a, 1982b) develop this notion further with their ASK hypothesis, which states that “an information need arises from a recognized anomaly in the user’s state of knowledge concerning some topic or situation

and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly” (1982a, p. 62). Thus, while Taylor suggests that the library patron *may* not be able to clearly articulate his or her information need, Belkin, Oddy, and Brooks claim that this difficulty of articulation is a fundamental quality of an individual’s information need.

Dervin and Nilan (1986) cite several definitions of information needs, but all are similar in that they suggest that an individual’s information need arises out of “a conceptual incongruity... when a person recognizes something wrong in his or her state of knowledge” (Dervin and Nilan, p. 18). In other words, in the language of Sense-Making (Dervin, 1983) an individual has a gap in his or her understanding of the world. Sense-Making is founded upon the assumption that the existence of gaps is fundamental to the human condition, due to the fact that “reality is neither complete nor constant” but instead constantly changing. A gap is situational, embedded in a specific context, and may only be a gap for a specific individual. In order for the individual to gain more complete understanding of his or her situation, and to move forward in that situation, it may be necessary to bridge this gap. In order for a gap to exist, the individual must have a goal to which he or she is moving, and a lack of understanding of some element of the individual’s situation is preventing him or her from reaching that goal – thus, information need. In order to reach this goal, the individual must bridge their gap, which can only be accomplished by making use of information that is relevant to that individual in that particular situation.

Information use, from this perspective, is the use to which an individual puts a particular piece of information to assist in the bridging of his or her gap. Dervin (1983) elaborates a set of uses to which information and the “newly created sense” that an individual makes of that information may be put. Dervin refers to these information uses as Helps and Hurts, indicating that any given piece of information can be more or less appropriate to fulfill a given information need, and in fact a specific piece of information can be either beneficial or detrimental in a given situation.

This study takes from the study of information needs and uses the idea that, as Horne (1983) states, a question is “an observable behavioral act reflecting information need” (p. 5). In other words, a question is the external manifestation of a gap in an individual’s understanding of the world around him or her. The communication act of asking a question, then, is an attempt to engage another individual’s assistance in bridging that gap. This merges speech act theory’s approach to questions as having force to compel a response to a question with discourse analysis’ approach to questions as being part of a larger conversation, and adds a motivating force for the questioner to ask a question.

2.2.5. Defining the Term “Question”

Table 2-1 presents an overview of the perspectives on questions presented in the previous sections.

Table 2-1: An Overview of Perspectives on Questions

Theory	Overview of the Theory	Perspective on Questions
Erotetic logic	Formal logic of questions.	A question may be decomposed into a subject and a request.
Speech Act Theory	Decomposition of any utterance into locutionary, illocutionary, & perlocutionary acts.	A question is one step in a larger conversation.
Discourse Analysis	Communication as a process of semantic collaboration.	A question is part of a larger conversation.
Information Needs and Uses	Individuals need information to bridge gaps in their understanding of their situation.	Questions are “the observable behavioral indicators of information needs.”

To summarize the previous sections, this study takes from the perspectives presented above, the following points:

1. Questions may be decomposed into a subject and a request.
2. A question has force: in asking a question, the questioner expects that the person questioned will respond to the question.
 - a. It is irrelevant whether the person questioned is face-to-face with the questioner, or remote.
3. A question does not stand alone, but is rather one step in a larger conversation.
 - a. Questions are not necessarily phrased as questions.
 - b. Statements phrased as questions are not necessarily questions.
4. A question is “an observable behavioral act reflecting information need” (Horne, 1983, p. 5).

All of these points serve to differentiate questions from other types of speech acts, and to illuminate the function of questions as speech acts. It is clear that it is not accurate to equate the term “question,” as it is used in common parlance, with information seeking – even though information-seeking questions are treated by some as prototypical questions (Flammer, 1981; van der Meij, 1987). In order to avoid any confusion of terminology, the term “question” will be used in this study to mean information-seeking question. The vehicle by which a question is received by a digital reference service, for this study, is asynchronous electronic media – *via* an email message or a webform. The term “out-of-scope” will be used to mean any question-like speech act that either is not information-seeking, or a message that does not contain a question – in either case, does not warrant a response from a digital reference service. Once a question is received and identified by a digital reference service, it enters into the triage process, and subsequently proceeds through all of the other processes within a digital reference service, as discussed below.

Now that questions have been defined for the purpose of this study, the field digital reference will be discussed, as that is the context in which questions are studied. Before

digital reference can be discussed, however, “traditional” desk reference will be discussed, as a means to explore the process of reference.

2.3. Reference

The foundations of modern reference work were laid by Samuel Swett Green in 1876 in his seminal essay, “Personal Relations Between Librarians and Readers.” Since then, the practices involved in providing reference service have been refined, but there has never been much disagreement about the central purpose of reference service, which is to answer and provide resources to enable patrons to answer their own questions. In order to understand digital reference, it is therefore necessary to understand “traditional” desk reference.

The Reference and User Services Association (RUSA) of the American Library Association (ALA), in their Guidelines for Information Services, state that:

“Information services in libraries take a variety of forms including direct personal assistance, directories, signs, exchange of information culled from a reference source, reader’s advisory service, dissemination of information in anticipation of user needs or interests, and access to electronic information” (Reference and User Services Association, 2000, ¶ 1).

A crucial point to notice in RUSA’s definition of library information services is that it encompasses both direct and indirect reference: information services may be provided “in anticipation of user needs or interests” or as “direct personal assistance.” In other words, library information services may be proactive or reactive. Proactive services may be conceived and in place prior to any patron using them; services of this type are signage, pathfinders, and written instructions of all sorts. Reactive services, on the other hand, are provided in response to an explicit request from a patron for assistance; reference is frequently a service of this type, as in the common scenario in which a patron walks up to

the reference desk and asks a question. Most library information services have components that are direct and indirect: even a library with no pathfinders or other materials to hand out is likely to have signage indicating the different sections of the library, the bathrooms, etc. Further, the very establishment of a reference desk is proactive, even if the service provided at the desk is reactive.

Bunge and Bopp (2001) take a narrower view of reference service than RUSA does, stating that it is concerned with:

“Both personal assistance provided to individual library users (e.g., answering reference questions) and organized services provided to groups of users (e.g., bibliographic instruction)” (p. 3).

Bunge and Bopp’s view of reference service is narrower than the RUSA’s in that it is concerned specifically with direct reference. Both of these definitions, however, state that reference service is *personal*, by which they mean that a human reference librarian assists a human patron (or patrons), in an environment that places them face-to-face. Miwa (2000) refers to this form of information service as “human intermediated” reference.

A great many other definitions and operationalizations of reference service exist in the library literature: Green (1876), Taylor (1968), and Bunge (1983) are some of the better known of these. These definitions will not be exhaustively reviewed here; such a task deserves a literature review exclusively dedicated to it. What is more important, these definitions agree on the two points discussed above: reference service may be direct or indirect, and requires a human intermediary. Based on the above, therefore, reference service is defined for the purposes of this study as human-intermediated assistance provided to users in fulfillment of users’ information needs.

2.3.1. Information Needs

RUSA (2000) addresses the fact that library information services – including reference services – seek to fulfill “user needs or interests.” A thorough exploration of user information needs will not be undertaken here, as Dervin and Nilan (1986) provide an excellent overview on this subject. The current study relies on Belkin, Oddy, and Brooks (1982a) definition: “an information need arises from a recognized anomaly in the user’s state of knowledge concerning some topic or situation” (p. 62). This “anomaly” can be understood as a “gap” in an individual’s understanding of his or her situation (Dervin, 1983): a thing not known, or a misalignment of perceived reality and an individual’s mental state.

The point about user information needs relevant to this study was raised by Taylor (1968), when he pointed out that the question that an individual expresses to a system (a patron asking a question to a reference librarian or otherwise) is not necessarily an accurate expression of that individual’s actual information need. Taylor suggests that this is the case because, as an individual becomes aware of his or her own information need and expresses it to the system, the question becomes “compromised,” in the sense that it is translated from the individual’s own terms to terms suitable for interaction with the system.

Miyake and Norman (1979) propose that the very ability to ask a question depends on “the existence of appropriate knowledge structures and the level of completeness of those structures” (p. 358). Therefore, in order to ask a question, one must first possess a framework of knowledge into which to fit the question. The inaccuracy of an individual’s initial question is thus attributed to the fact that he or she is asking a question about a subject for which he or she has no framework of knowledge. This is rather a Catch-22: you can’t formulate a question about something that you don’t know enough about to formulate a question. Or, as Miyake and Norman title their paper, “To Ask a Question, One Must Know Enough to Know What is Not Known.”

2.3.2. The Reference Interview and Question Negotiation

Overcoming this Catch-22 is a crucial component of the job of the reference librarian. Green (1876) himself first suggested the practice of conducting a reference interview (though he does not refer to it using that term), as a means to elicit more information from the patron about their information need, so as to better assist the patron in fulfilling that need. It was Taylor (1968) who first suggested that the reference interview is not just a good idea, but crucial to providing reference service. Taylor suggests that the reference interview is a vehicle for a “negotiation process” during which the librarian can “filter” the patron’s original question, through conversation with the patron and utilizing increasingly more information as it is elicited from the patron, to move from the patron’s “compromised” need to approach the patron’s actual “visceral” need (pp. 182-3).

Since Taylor outlined the need for and purpose of question negotiation, a great deal of research has been conducted on the reference interview; several models of the conversation and of elicitation techniques have been developed. White (1989) suggests that there are two models of the reference interview: the Question-Oriented and the Needs-Oriented models. The Question-Oriented model attempts to understand the patron’s question, while the Needs-Oriented model attempts to understand the patron’s information need. The former model, White claims, assumes that patrons know their own information need, and question negotiation is a means to assist the patron in phrasing that need in such a way that a librarian “can work with it” (p. 634). The latter model, on the other hand, assumes that patrons do not know “what information will be useful to them” (p. 634), and question negotiation is a means to assist the patron in thinking about their information need.

One of the most important considerations for the reference librarian involved in a reference interview is what sort of questions to ask of the patron. A common dichotomy in this area is the open versus the closed question. Open questions encourage the patron to talk, in an attempt to elicit more information about the information need. Closed questions, on the other hand, present a list (usually a short list) of options, and ask the

patron to choose from among them – yes/no questions are of this type (Bopp and Smith, 2001, p. 55). Dervin and Dewdney (1986) propose a third type of question, the “neutral question,” which, as a subset of open questions, are questions that enable the librarian “to learn from the user the nature of the underlying situation, the gaps faced, and the expected uses” (pp. 508-9). Some examples of neutral questions provided by Dervin and Dewdney to learn about the underlying situation, the gaps faced, and the expected uses are as follows, respectively: “Tell me how this problem arose,” “What would you like to know about X?,” and “How are you planning to use this information?” (p. 509).

Taylor (1968) makes the bold claim that “it may be that the reference interview, the negotiation of questions is the *only* process in libraries that is not noise” (p. 194). This approach to question negotiation as indispensable is central to librarians’ attitude towards and practice of providing reference service. This attitude carried over into the development of automation for reference, in the form of knowledge-based systems for reference, and, it will be argued, is partially responsible for the failure of these systems.

2.4. Knowledge-based Systems for Reference

Ever since the 1960s, when computer-based networks began to appear widely in academic and corporate settings, there has been an interest in using them to facilitate and automate library work. Systems utilizing this nascent technology were then, and continue to be referred to as “expert systems.” Richardson and Reyes (1995), however, suggest that the term “expert system” is inappropriate, as the system does not emulate, but merely represents an expert; this difference will be discussed further below. Richardson and Reyes recommend instead the use of the term “knowledge-based system,” as the system is founded upon experts’ knowledge, both in the creation of algorithms and in the representation of the domain to which those algorithms are applied. It is for this reason that the term “knowledge-based system” is used in this study, rather than the term “expert system.”

2.4.1. What is a Knowledge-based System?

A simple definition of the term “knowledge-based system” is that it is “a computerized system which draws upon the knowledge of experts in a field as a foundation for its database” (Jørgensen and Jørgensen, 1991, p. 75). This definition, while simple, is sufficient for this study, as it states all of the relevant points about a knowledge-based system:

1. It is limited to a specific field – or, as it is referred to in the literature on question answering systems, a specific domain (Burger et al., 2001; Hermjakob, 2001).
2. It is computer-based, and therefore requires that all actions in this domain be decomposed into a set of algorithms, which the system employs to perform whatever actions it is designed to perform.
3. It utilizes expert knowledge in representing of the domain to which those algorithms are applied – in other words, the database upon which actions in the domain are performed.

An example will serve to illustrate these three points. One well-known domain in which knowledge-based systems are employed is chess. Computer scientists have been experimenting with chess-playing systems for decades; such systems recently achieved a great success when the IBM supercomputer Deep Junior drew Garry Kasparov in a series of six matches between 26 January and 7 February 2003 (CNN.com, 2003). Chess provides a good domain for experimentation in knowledge-based systems for two reasons:

1. It is a closed system: There is a small and explicitly specified set of rules.
2. It is a finite problem space: There are approximately 38 legal moves per turn, leading to approximately 10^{120} total possible games of 40 moves.

(Pearson, 2003).

While this is a large number of possible games, it is nevertheless finite. The important point here is that the domain of problem-solving in chess is finite: it is possible, by applying an explicit set of rules, to exhaust all possible paths of action in the domain.

Additionally, knowledge-based systems utilize expert knowledge of the domain. This knowledge is generally in the form of facts about entities in the domain in question, and rules for manipulating those entities. Some of this knowledge may be explicit, such as knowledge of the board and the six pieces, and knowledge of how the pieces can be moved on the board relative to one another. Some of this knowledge may be tacit, such as knowledge of strategy. It is, of course, the role of knowledge-based systems designers to make such tacit knowledge explicit, so as to enable the creation of algorithms based on this knowledge.

Finally, knowledge-based systems are computer-based. A human – such as Garry Kasparov – possesses knowledge, but he would not be considered a knowledge-based system. A knowledge-based system is a computer-based system – hardware or software or a combination of both – that possesses a representation of the knowledge possessed by a human expert in a specific domain. Whether or not a computer possesses knowledge is a deep philosophical question, and will not be addressed here. Therefore, for the purposes of this study, it will merely be claimed that a knowledge-based system possesses a *representation* of a human expert's knowledge.

2.4.2. The Functionality of Knowledge-based Systems for Reference

The literature on knowledge-based systems for reference goes back to the 1960s. One of the earliest published works discussing the possibility of automation in the service of reference librarianship is Shera's (1964) article, which is something of a call to arms to analyze the processes involved in reference service, so as to enable the creation of algorithms to represent these processes. The first published work discussing an actual implemented knowledge-based system for reference describes a mainframe-based system

to assist reference librarians in the selection of biographical reference books (Weil, 1968).

Since then, most knowledge-based systems for reference have had similar goals: to advise reference librarians or patrons in the selection of information sources to answer the patron's reference question. Indeed, Richardson (1995) evaluates 58 knowledge-based systems for reference created between 1968 and 1994, and the goal of all 58 is to perform this advising function. In most of these systems, this is accomplished by automatically selecting the appropriate source or format of source to answer a reference question.

Richardson (1995) argues – as would, presumably, the developers of the knowledge-based systems that he evaluates – that “the essential reference task is to classify the inquirer's question into one of the known reference formats” (p. 155), of which he describes twelve. In other words, the task of providing reference service is to match up a question with a format of information source that is likely to contain the answer to that question.

An objection to this definition of reference – and to the consequent approach to the automation of the reference transaction – is that answers are contained in specific individual information sources, not in formats of sources. Knowledge-based systems for reference that “match” questions with information sources replicate one particular type of interaction with a reference librarian: that interaction in which the reference librarian points a patron to sources and does not provide any bibliographic instruction in the use of that information source. Reference transactions such as this are of course fairly common, but they are only one type of reference transaction. A human reference librarian has the ability to assess the patron, the patron's information need, the difficulty of the question, and other factors, and determine what sort of reference transaction is appropriate; other possible outcomes of a reference transaction are that the librarian provides some bibliographic instruction in the use of the information source provided, or even an answer out of one or more sources. Existing knowledge-based systems for reference are not

capable of this range of possible outcomes; they provide an information source only – they do not provide an answer, and many systems do not even provide an explanation of why the source selected is the appropriate one to use to answer the question. In short, while human reference librarians do not always provide bibliographic instruction and answer retrieval, they at least have the ability to do so; existing knowledge-based systems for reference provide document retrieval only.

2.4.3. The Failure of Knowledge-based Systems for Reference

Attempting to match a question with the format of a source – or a specific source itself – that will contain the answer to that question is a difficult task. In an attempt to reduce that complexity, Richardson (1995) presents twelve tables of if-then rules that outline the conditions under which any of the twelve reference formats should be selected, given the nature of the question (pp. 157-164). For example, if geographical information is asked for, the appropriate source is an atlas (p. 158); if “information about words” is asked for, Richardson states that the appropriate source is a dictionary (p. 160). These simple rules are insufficient for matching questions to information sources, however: for example, the appropriate source for answering a question asking for geographical information may be an atlas, an almanac, a map, depending on what is being asked for.

Richardson’s (1995) if-then rules, however, require that the librarian possess a great deal of *a priori* knowledge of reference formats and the sort of information that those formats contain – not to mention knowledge of the specific information contained in specific information sources. In short, these if-then rules are an attempt to utilize a librarian’s knowledge of information sources, prior to any question negotiation or knowledge about either the patron or the question itself. It has been well established in library science that it is difficult, if not impossible, to answer a reference question without understanding the patron’s information need and planned use of the information (Taylor, 1968; Dervin, 1977; Dervin and Dewdney, 1986), and it has been well established in discourse analysis that it is necessary for there to be some “common ground” between conversational participants before any meaning can be exchanged (Stewart, 1978; Cissna and Anderson,

1994; Roberts and Bavelas, 1996). In short, without some knowledge of either the patron or the question, it is unlikely that a knowledge-based system can satisfactorily answer a reference question.

Whatever the theoretical shortcomings of knowledge-based systems for reference may be, the fact is that they have not succeeded in practice either. Knowledge-based systems for reference have been implemented in a variety of domains: general reference (Jørgensen and Jørgensen, 1991; Richardson, 1998), government documents (Harley and Knobloch, 1991), patents (Ardis, 1990), opera (Gerber, 1992), even recommending novels for pleasure reading (Rich, 1979, 1986). However, most of these systems were never implemented outside of their initial test-bed environments. One reason for this limitation is that knowledge-based systems for reference are designed to operate within a specific domain: the “knowledge” that such a system contains – the database upon which the system is based – is knowledge of reference sources on government documents, patents, or opera, and the rules that such a system follows are specific to matching domain-specific questions with domain-specific information sources. The approach to system development for these knowledge-based systems is therefore not extensible across domains.

In conclusion, the quest for an automated reference process is more than thirty years old, and has so far failed to produce a system that can be used across reference settings.

2.5. Digital Reference

The definition of reference service was provided above as human-intermediated assistance provided to users in fulfillment of users’ information needs. This definition deliberately does not take the medium of delivery of that service into consideration. Reference service can be and is offered *via* a number of media: in person, by telephone, by email, *via* the web, by chat or instant messaging, even by paper mail.

Early digital reference services were offered *via* email, as outgrowths of existing reference desk services in academic and special libraries (Kittle, 1985; Howard and Jankowski, 1986; Weise and Borgendale, 1986; Roysdon and Elliott, 1988; Hodges, 1989; Bristow, 1992; Still and Campbell, 1993). These digital reference services were developed both to extend the hours of availability of the reference desk, and to experiment with the new technology provided by campus-wide networks. Many physical reference desks continue to offer email-based reference services. Almost as soon as the technology became available to create a webform, digital reference services began utilizing webforms for question submission (Lankes, 1998a; Lagace, 1999). Since that time, the percentage of questions submitted to services *via* webforms has far outstripped the percentage submitted *via* email (Carter and Janes, 2000; Janes, Hill, and Rolfe, 2001). Since around 1997, a new type of reference service has begun to appear online: so-called “real-time” reference service. While the more “traditional” AskA services make use of asynchronous methods of communication, real-time reference makes use of synchronous methods of communication: chat environments (Francoeur, 2001; Kibbee, Ward, and Ma, 2002), instant messaging (Foley, 2002), and graphical co-browsing (Patterson, 2001).

In the early- to mid-1990s, reference services began to appear on the Internet that were not affiliated with a library, either physical or digital (Bushallow-Wilbur, DeVinney, and Whitcomb, 1996; Philip, 1997). Lankes (1998a) refers to services of this type of reference service as “AskA” services, “such as Ask-A-Scientist” (p. 9), since most services of this type specialize in a particular subject: for example, art (Ask Joan of Art), education (AskERIC), mathematics (Ask Dr. Math), oceanography (Ask Shamu), etc.

All of these forms of reference service, delivered by any electronic medium, are considered to be digital reference. The term “digital reference” is used in this study, rather than the almost equally popular and nearly synonymous term “virtual reference.” In her book *The Virtual Reference Librarian’s Handbook*, Lipow (2002) acknowledges that there has to date been little agreement in the use of these terms. She defines the term digital reference to indicate “a broad range of electronic reference activities that include creating and managing digital reference resources... *as well as* providing personalized

reference service *via* the Internet” (p. xix), and the term virtual reference to indicate only reference service provided *via* synchronous technologies – chat, instant messaging, voice over IP, etc. Lipow’s distinction between these terms will be used in this study, though the range of digital reference services with which this study is concerned is limited to asynchronous interactions – those conducted *via* email or webforms.

Lankes (1998a) states that reference service “can be performed either through a human expert (such as a reference librarian) or an automated interface (such as an online catalog)” (p. 25-26). What will differ between reference services offered *via* different media is the degree of interactivity between the librarian and the patron, as this interactivity is dictated in part by the limitations of the media. It is currently an open question whether or not the service offered is the same in all media (Lankes et al., 2003). There is no question, however, that responding to questions asked by patrons is one of the primary services offered by reference services, regardless of the media in which that is accomplished.

Lankes (1998a), in discussing digital reference service, states that “this question may be expressed as an e-mail request or a query to a system” (p. 25). How much of this system is human and how much automated depends on the service itself. The patron, however, may only see the interface to the system. At a library reference desk that interface (to the system of the library itself) is the reference librarian. In a digital reference service, that interface is electronically mediated: it may be a web form, a chat room or instant messaging application (Foley, 2002), a bulletin board, or a simple email address (Janes, Hill, and Rolfe, 2001). The response is generally sent to the patron *via* the same medium by which the question was submitted (Janes, Hill, and Rolfe, 2001).

The content of the response sent to a patron may be different between physical and digital reference services as well. Physical reference services and existing knowledge-based systems utilize a physical library’s collection as the corpus from which to provide documents. Nowadays, electronic materials are becoming increasingly important in library services in general, and in reference services specifically. Reference librarians are

as likely to provide bibliographic instruction in the use of online sources as they are in the use of traditional print sources. Existing knowledge-based systems, on the other hand, ironically do not provide links to electronic information sources; according to Richardson's (1995) review of 58 knowledge-based systems, discussed above, none utilize a corpus of electronic sources. Digital reference services, like physical reference services, make extensive use of electronic and online information sources. Several authors (Lankes, 1998b; Bry, 2000; Janes, Hill, and Rolfe, 2001; Tenopir and Ennis, 2001) discuss the fact that digital reference service utilize electronic sources in answering questions. The difference between the answers provided by physical and digital reference services is this: not only can citations and pointers to information sources be provided in an answer provided by a digital reference service, direct links to these sources or even copies of sources themselves (license permitting) may be provided.

The definition of reference presented above does not take the medium of delivery into consideration, and it likewise does not take the content provided into consideration. Possible variations in both of these do not change the fundamental purpose of reference. Thus, digital reference is a variation on reference: digital reference service is defined for the purposes of this study as human-intermediated assistance provided to users *via* electronic media in fulfillment of users' information needs.

2.5.1. A Model of Digital Reference

Pomerantz and others (forthcoming) present a general model of the processes involved in providing digital reference service, shown in Figure 1-1. This model consists of 5 steps:

1. Question Acquisition includes all issues related to the process of obtaining information from a user. This includes not only the user's question, but also question categorization and user identification information, *via* email, web forms, chat, or embedded applications.
2. Triage is the assignment of a question to a reference or subject expert "answerer." This step may be automated or conducted *via* human decision

support. Triage also includes the filtering out of repeat or out-of-scope questions.

3. Answer Formulation includes all actions taken by the answerer to generate a response to a question, including sending the response to a reviewer or directly to the user. Factors for creating “good” answers such as age and cultural appropriateness are included in this step.
4. Tracking is the quantitative and qualitative monitoring of repeat questions for trends. Tracking allows the identification of “hot topics,” and may indicate where gaps exist in the collection(s).
5. Resource Creation involves the use of tracking data to build or expand collections to better meet users’ information needs.

This study presumes that this process model is generally applicable to all digital reference services, though different services may employ different processes at different steps. Some services may even skip steps; for example, not all services may archive previously-answered questions to create resources. Additionally, some steps may be repeated, especially if this model is seen to span more than one service; for example, a triage center at one digital reference service may receive a question and triage it to a different service, which may then triage it to an individual expert.

2.5.2. The Triage Process

The English word “triage” is derived from the French word “trier,” which means “to divide, to separate.” According to the Oxford English Dictionary (OED) (www.oed.com), the word “triage” means “the action of assorting according to quality” (quality in the sense of value, worth, or goodness). This meaning of the word dates back to the early 1700s; for example, coffee beans are triaged to separate the whole beans from the broken. The more commonly understood meaning of the word these days is the assignment of injured persons to treatment according to the urgency of their injuries, as in wartime. Again according to the OED, this meaning of the word “triage” dates back only to the beginning of the 20th Century and presumably originated during World War I. According

to the On-Line Medical Dictionary (cancerweb.ncl.ac.uk/omd), “triage” is defined as “the sorting out and classification of patients or casualties to determine priority of need and proper place of treatment.” Despite its medical bent, this definition makes two points that are crucial to the definition of triage for this study: the process of sorting or classification, and the determination of a place of treatment. Replace “casualties” with “questions,” and this definition is almost precisely the one that is meant by the use of the word “triage” in this study: the sorting out or classification of questions to determine the proper place of treatment.

In desk reference services, the librarian who happens to be at the desk when a patron approaches is generally the librarian who responds to that patron’s question. This librarian may not, however, be the most appropriate person – in terms of subject expertise, reference experience, or other service criteria – to respond to that particular question. In some cases – for example, when the question is outside the service’s scope – rather than simply turn a patron away without an answer, a reference librarian will refer a patron to another reference service or organization. In such cases, the burden is generally on the patron to contact that other service and complete the referral. Digital reference services similarly forward questions to other services, though the difference between referrals in desk reference services and triage in digital reference services is who has the responsibility for completing the referral: in digital reference, if triage is performed, it is not the patron who is sent from one service to another, but the patron’s question. Thus, the burden is on the service that received the question from the patron to perform triage. The differences between referrals and triage will be discussed in more depth in chapter 5.

Several researchers have pointed to the existence of a filtering stage in the processing of digital reference questions, which Lagace and McClennen (1998) and Lankes (1998a) refer to as “triage.” McClennen and Memmott (2001) describe several roles played by participants in the digital reference process. These roles are similar to roles in the traditional reference process, but with “some new twists imposed by the digital environment” (p. 143). One of these roles is the Answerer, which is “the customary role of the reference librarian;” that is, staffing the (virtual) reference desk and answering

questions. Another role is the Filterer, who decides which questions the service should accept, which can be answered with stock answers, and which are out of scope for the service, and which should be forwarded to other services. McClennen and Memmott describe these roles within the digital reference service in which they are employed, the Internet Public Library (IPL). Answerers for the IPL may select their own questions from the pool of all received questions, so triage is performed *between* the IPL and other services, but not *within* the IPL. Thus, the role of the filterer reflects the task of triage between services, but not within. In order to encompass both within- and between-service triage, the term “triager” will be used in this study, rather than the term “filterer.”

An additional complication to the triage process is that the email message or other communication received by a digital reference service may not contain a question at all. Silverstein and Lankes (1999) describe 16 types of out-of-scope messages received by the U.S. Department of Education’s web-based Information Service Centers. Among the types of messages that may be considered out-of-scope are viruses, advertisements, server error messages, spam, and a variety of other types. Indeed, even “thank you” messages from patrons (that do not contain a new question) may be filtered out, as they may not receive a reply from a digital reference service. As will be discussed in section 3.3.2, approximately 15% of the total emails received by the Virtual Reference Desk (VRD) are non-questions, and it seems reasonable to assume that other digital reference services percentages are similar. Thus, approximately 85% of the total emails received by digital reference services are legitimate questions.

The triage process, as presented in the discussion above, can be decomposed into three sub-processes, which can be thought of as filters:

1. Separating non-questions from legitimate questions,
2. Separating in-scope from out-of-scope questions, and
3. Determining the appropriate answerer for each question.

These decision points in the triage process and the actions that may be taken at each point are presented in Figure 2-2. A more detailed model of the triage process, based on the findings of the present study, is presented in Figure 4-8.

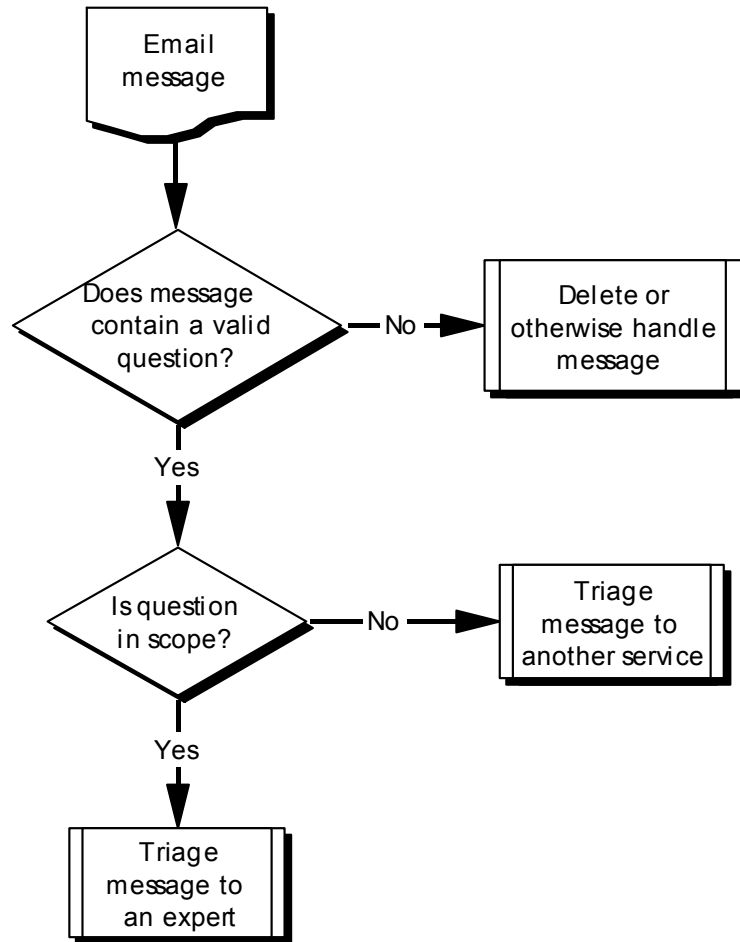


Figure 2-2: Steps in the Triage Process

2.5.2.1. Variations on the Triage Process

Two variations on the triage process are employed by different digital reference services (Lankes 1998a, Pomerantz, Nicholson, and Lankes, 2003). The first variation is concerned with the mechanism by which the decisions in the triage process get made: either by a human triager or by an automated process. In either case, criteria must exist for deciding what messages are legitimate versus non-questions, and which questions are

in- versus out-of-scope. If a human is performing this triaging, however, these criteria may be more heuristic, whereas if an automated process is performing this triaging, these criteria must be rigorously codified in software. The QuestionPoint service (formerly called the Collaborative Digital Reference Service (CDRS)) is an example of automated triage, employing an algorithm to perform routing and assigning of questions to other digital reference services. This algorithm assigns questions “on the basis of such data elements as hours of service, including time zones, subject strengths, scope of collections, types of patrons served, etc.” (Kresh, 2000, How Does CDRS Work? section, ¶ 2).

The second variation on the triage process involves how a question is “triaged” to an answerer: questions are either assigned to specific answerers by the triager (either human or automated) (Bry, 2000), or questions are stored in a “triage area” and self-selected by answerers. Lagace and McClennen (1998) state that answerers for the Internet Public Library self-select questions, and that “sludge patrol” is performed daily to locate and respond to questions that have not been selected by an answerer.

Of these two variations on the triage process, this study is concerned with the first: human versus automated triage. Triagers in digital reference services have the task of deciding how to triage incoming electronically submitted reference questions both within a service, to individual answerers, and between services. Pomerantz, Nicholson, and Lankes (2003) conducted a Delphi study to determine what factors influence the sorting of questions to answerers. One of the factors employed by all services that Pomerantz, Nicholson, and Lankes identify is the type of a question. This study seeks to achieve a greater understanding of question types, which may then be utilized as a means for improving, and ultimately developing specifications for designing algorithms to automate the triage process.

2.5.2.2. Triage in Real-Time Services

There has been no literature on real-time reference indicating that any services are performing triage on incoming real-time requests, though Francoeur (2001) states that

more sophisticated question management applications allow an administrator to “transfer” incoming requests to available librarians (p. 193). In principle, there is no reason why triage *could not* be performed on incoming real-time requests, though given the synchronous nature of the transaction, a human intermediary would have to be standing by in the recipient service to receive the triaged request, just as Patterson (2001) and Kibbee, Ward, and Ma (2002) state that librarians were on duty to receive incoming requests in the services that they report on.

The fact is, however, that at this point in time, triage is not performed in real-time reference services, beyond the rudimentary form of triage that Francoeur (2001) refers to as “transferring.” Transferring bears a closer resemblance to referrals at a reference desk than to triage in asynchronous digital reference services, though even that analogy is strained. As discussed above, referrals at reference desks occur when the librarian to whom the patron asked a question is not, for whatever reason, the most appropriate person to respond to that particular question. Referrals, like triage, may be made to individuals within the same service or to another service (Bopp, 2001). There has been no literature indicating that any services are transferring incoming real-time requests to other services; real-time transferring therefore bears a resemblance to a referral to another librarian at a reference desk.

Referrals at reference desks and transferring of incoming real-time questions may both be considered to be forms of triage, according to the general model of digital reference service shown in Figure 1-1; the significant distinction between these forms of triage is the media by which the question is triaged. This study, however, was concerned with only those services that receive questions and perform triage *via* asynchronous media.

2.5.3. The Opening Question in Digital Reference

Miyake and Norman’s (1979) premise that in order to ask a question, one must first possess a framework of knowledge into which to fit the question was discussed above. Belkin, Oddy, and Brooks (1982a) apply this premise to information retrieval (IR)

systems. Belkin, Oddy, and Brooks refer to this gap in an individual's framework of knowledge as an "anomalous state of knowledge" (ASK), and suggest that "IR systems should be designed with the non-specifiability of information need as a major parameter" (p. 66). Eliciting information about an individual's information need that he or she may not even know or be able to express has been a standing challenge for IR system developers since Belkin, Oddy, and Brooks proposed their ASK hypothesis.

Eichman (1978) and Dewdney and Mitchell (1996) both claim that the function of the initial question asked by a patron of a reference librarian "is not primarily to ask the question; rather, it is a form of phatic communication" (Dewdney and Mitchell, p. 8). Phatic communication is communication that serves the function of maintaining social relationships (Laver, 1975). Thus, not only does a patron's initial question not accurately reflect his or her information need, that question may not even be intended to serve the function of a question – it may be intended merely as small talk, to establish a relationship between the patron and the librarian.

Unfortunately, in digital reference, often the initial question is all that the librarian has. Digital reference services have found that asynchronous media do not lend themselves well to question negotiation: Carter and Janes (2000) report that if an expert from the IPL replies to a user's question with a request for clarification, 29.5% of users do not ever reply with that clarification – and judging by the author's conversations with digital reference answerers at a number of other services, Carter and Janes' finding is a remarkably small percentage. (Synchronous media for digital reference may change this for digital reference services, but further study is required to determine whether this will prove to be so.) The question has even been raised whether the reference interview has a place at all in digital reference (Wilson, 2000). Ross and Nilsen (2000) report that reference librarians in desk reference services conduct a reference interview "only about half the time" (p. 150), so the lack of a reference interview is certainly not unique to digital reference. The problem of the "disappearing questioner" (Janes and Silverstein, 2003) on the other hand, is likely to be more common in asynchronous digital than in desk reference.

On the brighter side, Bushallow-Wilber, DeVinney, and Whitcomb (1996) and Tenopir and Ennis (2001) report findings that reference librarians answering questions submitted by email find these questions to be more accurately phrased and more sophisticated than initial questions asked at reference desks. These findings fly in the face of the ASK hypothesis. Why this discrepancy occurs is an interesting question, and bears further study. Nevertheless, while the initial question received by a digital reference service may not accurately reflect the asker's information need, it is usually all that the service has to work with. There is therefore a need for techniques to wring as much information as possible from the initial question received by a digital reference service. This study proposes that classifying questions according to taxonomies along various dimensions of analysis are one such technique.

2.5.4. Digital Reference vs. Knowledge-based Systems for Reference

Knowledge-based systems for reference attempt to match a question with a type of source or a specific source that is likely to contain an answer to the patron's question. In short, knowledge-based systems for reference attempt to automate the entire reference transaction, from the receipt of the question to providing an information source.

At the other extreme is Lankes' (1998a) survey of six digital reference services. Five of these six services relied entirely on human expertise to answer questions; none of these five services employed automation for question answering or triage. Essentially, these six digital reference services provide precisely the same service as a more traditional desk reference service: answering of a question by a human expert. The only difference is the media by which that service is delivered.

A middle ground between knowledge-based systems for reference and digital reference provided entirely by humans is the sixth of the services studied by Lankes (1998a), the Mad Scientist Network (MADSci). Bry (2000) describes the MADSci's use of automation in the triage process. When the user submits a question, a CGI script searches

the MADSci archive of previously-answered questions. Bry states that “approximately 63 percent of questions are matched with archived files” – however, “only 25 percent of users deem their questions answered by this process (15 percent of all submitted questions)” (p. 118). Voorhees (2001) states that the worst-performing question answering system in TREC (discussed further below), under conditions of strict evaluation, answered 32.3% of questions correctly (p. 162). Bry does not describe the algorithm used by the CGI script to search the archives, and MADSci and TREC provide answers from different corpora of documents, so it is difficult to make comparisons. Additionally, questions received by MADSci are questions “from the field,” as it were, while TREC questions, while taken from web search engine logs, are cleaned up considerably (Voorhees, 2001, p. 161). It is possible that attempting to match reference questions to answers may be as complex a task as attempting to match questions with formats of sources, and may similarly require world knowledge of the domain about which the question asks. No research exists that has determined under what conditions question answering systems are unable to provide an answer and the question must be sent to a human reference expert to be answered, though the author is currently involved in a project that will attempt to determine precisely that (Croft, Lankes, and Koll, 2002).

These two extremes – knowledge-based systems for automating the retrieval of answer sources, and digital reference systems that employ no automation at all – leave a great deal of room, into which this study fits. This study takes the position that the Mad Scientist Network is on the right track by attempting to automate the triage process. The triage process is the first step performed by a digital reference service after the receipt of a question, and therefore arguably the most important step in the digital reference process, since the outcome of processes “downstream” may depend on the outcome of the triage process. This study also takes the position that the results of analyses of questions would be a useful input to any triage process, whether manual or automated. The criteria for analysis of questions employed in this study will be discussed below.

2.6. Question Answering Systems

QuestionPoint and the Mad Scientist Network (MADSci) were discussed above. These two digital reference services are the only services that, as of this writing, employ automation in any steps in the process of providing digital reference. QuestionPoint employs an automated triage process, assigning questions “on the basis of such data elements as hours of service, including time zones, subject strengths, scope of collections, types of patrons served, etc.” (Kresh, 2000, How Does CDRS Work? section, ¶ 2), and MadSci employs a CGI script to search the MADSci archives for previously-answered questions that may answer the question submitted by the current patron. This is the extent of the literature review of the utilization of automation in digital reference that can be conducted at this point in time. Digital reference services simply have not yet adopted a great deal of automation. There are two types of automated systems that have been developed, however, that may have utility for digital reference: question answering, and automatic classification systems. This section discusses the former, and the latter will be discussed below, in the section on Classification.

The idea of automated question answering systems is nothing new. As discussed above, knowledge-based systems for reference have been discussed and developed since the 1960s. Such systems do not actually *answer* questions, however, but rather match a question with an information source that will contain an answer. A different approach exists.

2.6.1. QUALM

Perhaps the first attempt to automate the process of formulating answers to questions was made by Lehnert (1978). Lehnert’s QUALM was a story understanding system that attempted to replicate the process by which humans understand and answer questions. These questions concerned short stories of a few simple sentences, made up mostly of facts (for example: “John went to a restaurant. The hostess seated John.” etc.) (p. 19). The QUALM system divided this process of understanding and answering questions into four

sub-processes, the first two of which are concerned with understanding the question, the last two with generating an answer:

1. Conceptual parse: Interpretation of the question at a “literal or naïve” level of understanding (p. 44). In this step the question is interpreted *prima facie*, without taking any context or world knowledge into account. For example, QUALM would interpret the question “Can you tell me what time it is?” as a question requiring a yes/no answer, rather than a request for a statement of the current time.
2. Memory internalization: The decomposition of the question to identify all of the facts presented in it. For example, a question about John would “place in the actor slot the word JOHN” (p. 46). This pointer could point to a pre-defined “record” containing additional information about John, if one existed.
3. Conceptual categorization: The decomposition of the question into “two descriptive components: a question concept and a conceptual question category” (p. 47). The question concept is the subject of the question. Lehnert describes 13 conceptual categories; for example, a question asking whether a fact is true or false, or a question asking why an agent performed some action.
4. Inferential analysis: Analysis of the context within the story that the question is asking about, to determine if the question makes implicit assumptions. For example, the question “What haven’t I packed?” is inferred to mean “What haven’t I packed that I should have packed?” (p. 48).

It is step 3, Conceptual categorization, which is of particular interest for this study. The thirteen conceptual categories described by Lehnert are, in fact, the basis for the taxonomy of the functions of the expected answers to questions, discussed below.

2.6.2. The TREC Question Answering Track

The development of question answering (QA) systems got a boost when the Text REtrieval Conference (TREC) added a QA track in 1999. These QA systems took an

entirely different approach to responding to questions than had been taken by information retrieval (IR) systems until that time. Voorhees (1999) states that:

“The goal in the QA task is to retrieve small snippets of text that contain the actual answer to a question rather than the document lists traditionally returned by text retrieval systems. The assumption is that users would usually prefer to be given the answer rather than find the answer themselves in a document” (p. 77).

While this assumption is debatable, the fact is that QA systems provide passage retrieval, while traditional IR systems provide document retrieval. Additionally, the nature of the QA task in TREC is quite different than the story understanding performed by QUALM. While the task of a QA system for story understanding is to answer questions about a set of facts presented as a story, the task of the TREC QA systems is to retrieve portions of documents from a corpus that contain answers. This sets QA systems apart from “traditional” IR systems, which attempt to retrieve from a corpus of documents one or more documents that (hopefully) contain an answer to the question put to the system. In this way, traditional IR systems are much like knowledge-based systems for reference, in that they both attempt to match a question to an information source, while QA systems are like desk and digital reference services, in that they attempt to provide actual answers, not just sources.

Because every QA system in TREC is developed at a different institution and independent of each other, every one is implemented differently. There are, however, important similarities among the different systems. The most important of these similarities, for the purposes of this study, are the steps that systems employ in processing and answering questions. In fact, these steps are similar to Lehnert’s four processes, discussed above. A summary of the processes employed by many of the TREC QA systems is as follows:

1. Question processing: This step is composed of two sub-processes:
 - a. Logical representation of a question (logical is used in opposition to natural language in this context). This logical representation may be a query in a form such it could be used as the input to a search engine or a database (Diekema et al., 2000), or in a format unique to the QA system (Moldovan et al., 1999).
 - b. Recognition of the question focus, to determine the subject or forms of the expected answer. This step is much like Lehnert's (1978) Conceptual categorization step.
2. Document processing: Documents are retrieved from the corpus of all documents as in a traditional IR system, in most systems by keyword matching. Returned documents are ranked according to system-specific algorithms.
3. Paragraph or segment finding: Documents are then broken up into paragraphs or segments of some size. The aim of this stage is to retrieve the segments of the retrieved documents that best match the question. Indeed, this step is conceptually similar to the Document processing step, only on a smaller scale: the "document" in this case is the paragraph or segment, and the "corpus" is the set of all paragraphs or segments in the set of documents retrieved in the Document processing step.
4. Answer finding: Paragraphs or segments are then broken up into smaller segments. The keywords in the logical representation of the question and/or the question focus are searched for within a "window" of 50 or 250 bytes. These smaller segments are ranked according to system-specific algorithms. The top-ranked segment is considered to be the answer. Again, this step is conceptually similar to the previous two steps, only on a smaller scale again: the "document" in this case is the segment in the "window", and the "corpus" is the set of all segment windows in the set of retrieved paragraphs or segments.

(Diekema et al., 2000; Hovy et al., 2000; Chen et al., 2001; Hovy, Hermjakob, and Lin, 2001).

Different systems call these steps by different names, and some systems break these steps down into sub-processes, but these four steps provide a general overview of the process employed by all QA systems that compete in TREC.

Recognition of the question focus is the step that is of particular interest for this study. In reviewing all of the QA systems in TREC, three ways of determining question focus emerge: the subject of the question, the taxonomy of *wh*- words, and the taxonomy of the functions of the expected answers to questions. Some systems utilize a mixture of classes from more than one of these taxonomies – as, for example, does the Webclopedia system (Hovy et al., 2000; Hovy, Hermjakob, and Lin, 2001). The Webclopedia system classifies questions first by the *wh*- word, and then subdivides these types according to a representation of “the user’s intention” (Hovy et al., 2000, p. 659) – a system-specific version of the taxonomy of functions of expected answers. The point is, all current QA systems employ one or more of these taxonomies as the basis for identifying the focus of a question. These three taxonomies are discussed in depth below.

2.6.3. The Possible Future of Question Answering Systems

In the TREC QA track, systems were allowed to generate answer strings of 50 or 250 bytes in length – that is, 50 or 250 characters (Voorhees, 1999, 2000); this was reduced to only 50 characters at the 10th TREC conference (Voorhees, 2001). In general, this short a length for an answer does not leave much room for providing a context for the answer – “supporting documentation” to provide justification for *why* the answer generated by the system is correct. Indeed, Liddy (2002) states that:

“discussion at the TREC 2001 Workshop on QA intimated that the QA track in TREC 2002 will accept as correct only fragments which contain the minimal answer to the question. Any explanatory text, even within the 50 byte limit, will cause the answer to be marked as incorrect” (p. 1).

There is some disagreement in the QA system development community concerning the wisdom of this direction being taken by the TREC QA track to limit answers to the bare minimum. Liddy (2002) goes on to state that such minimal answers are useful “in only a subset of the contexts in which QA systems are truly needed” (p. 2). Burger and others (2001) make the same claim, but add that “for all the other cases, instead of returning 50 or 250 bytes of text, [QA systems] should return the answer in a format natural to the questioner” (pp. 17-18). Hirschman and Gaizauskas (2001) propose six criteria for the evaluation of answers, one of which is justification: “the answer should be supplied with sufficient context to allow a reader to determine why this was chosen as an answer to the question” (p. 20).

If QA systems were to be used in digital reference, then the justification of an answer would be an important feature of such systems. Of course not all reference librarians provide justification for their answers to all reference questions, but to do so is generally encouraged and is taught to library students as the ideal for providing answers to reference questions (Bopp, 2001). Digital reference is only one possible application of QA systems, but it is one that encompasses several aspects that Burger and others (2001) propose are necessary directions for the future of QA system development: the question answering process occurs within a context, formulating answers may involve retrieving and combining data from multiple documents, and the domain of questions is open – that is, questions may be on any subject. Digital reference and QA systems could, in the future, arrive at a highly fruitful symbiosis: digital reference services could provide a useful testbed for future QA systems implementing increasingly sophisticated functionality, and QA systems could automate some answering of reference questions, thus enabling a digital reference service to scale up to handle an increasingly large number of questions. This study is not concerned with the utilization of QA systems in digital reference question answering, however; what is of concern is the question taxonomies utilized by QA systems.

There are two ways that QA systems could evolve in the future: with and without context provided with the answer. It may be that both types of QA system will be developed,

each to be used as appropriate. It may indeed be true, as Burger and others (2001) state, that standalone facts are sufficient to answer some question types – though it seems likely that questions for which an answer may be provided with no context are in the minority. Another consideration for the evolution of QA systems is the requirements of the users: studies of user information needs and uses are called for to determine which type of QA system is desired by different communities of users.

Providing answers with no context flies in the face of the ideal for reference work, both desk and digital. As discussed above, the reference librarian should not just provide a standalone answer, but also some justification for that answer. The reference librarian is not intended to be the final arbiter of information, but rather an interface to a collection of information sources. Of course some digital reference services do not provide answers, but rather provide lists of citations to documents that discuss particular subject areas – in other words, such services provide the context in which the answer may be found, and require the patron to find the answer him- or herself. Therefore, if the methods and technology from QA systems are to be brought to bear on the automation of reference, it must be QA with the provision of context.

Table 2-2 presents the different types of systems (both human and automated) that exist in the arenas of reference, IR, and QA, and the approach that these systems take to answering questions.

Table 2-2: Types of Systems and Their Approach to Question Answering

Type of System	Approach to Question Answering
IR system	Document retrieval
Knowledge-based system for reference	
Desk reference service	Document or answer retrieval, depending on the specific question and the policy of the service
Digital reference service	

QA system	Answer retrieval
-----------	------------------

2.7. Classification

This section discusses the literature in the field of classification theory from which this study draws a framework for classifying questions.

2.7.1. Classification and Categorization

A distinction is drawn in this study between classification and categorization. Jacob (1991) makes this distinction as follows:

“The process of classification involves the systematic assignment of entities to groups or classes according to an established set of principles. ... Categorization, on the other hand, refers to the process of dividing the world of experience into groups – or categories – whose members bear some perceived relation of similarity to each other” (p. 74).

In other words, classification is performed according to a deductive, pre-existing scheme, while categorization is performed according to an inductive, perhaps idiosyncratic, scheme.

This distinction is an important one to make because the terms “classification” and “categorization” are often used synonymously, as are the terms “class” and “category.” This study classifies questions according to pre-existing taxonomies. This study also, however, inductively discovers categories of attributes of questions that affect the triage process. Thus, both classification and categorization are utilized in this study, and it is therefore important that these two processes be identified as related, but separate.

2.7.2. Classification of Intellectual Entities

The entities classified by most classification schemes are physical entities: print materials and other media, artifacts, and the like. For example, the domain of the Library of Congress Classification is the material in the Library of Congress collection (Chan, 1990), and the domain of the *Art & Architecture Thesaurus* (AAT) is all of the “material culture of the Western world from antiquity to the present” (Petersen, 1994, pp. 29-30). Against this background, questions are unusual entities for classification, as they do not possess the quality of necessary physical existence in the sense that print materials or other objects have.

The term “intellectual entities” is used in the classification literature to mean the physical materials held by a library’s collection. Books and other materials are referred to as intellectual entities because they contain intellectual content. As such, representing intellectual entities is the *raison d’etre* of library classification. The term “intellectual entity” is used in this study in a slightly non-standard way, to mean not books and other library materials, but questions submitted to digital reference services. Representing electronic materials is a problem that has concerned library classification for several decades (Vizine-Goetz, 2002). However, some consensus seems to be developing on this issue: intellectual entities should be classified in the same way, regardless of the container in which they are manifested (Dublin Core Metadata Initiative, 2002). The use of the term “intellectual entity” in this study to refer to questions submitted to digital reference services is therefore consistent with its prior use: a question contains intellectual content, whatever the medium by which it is submitted to a reference service.

2.7.3. Mutual Exclusivity

Graesser, Person, and Huber (1992) claim that questions can be assigned to one or more categories in their taxonomy (p. 171), discussed in section 2.8.3. This flies in the face of the strict requirements of hierarchical classification schemes, according to which a given entity can belong to one and only one class (Kwaśnik, 1999, p. 26). Other forms of

classification scheme may or may not have this requirement: the part-whole relationship employed by some classificatory trees would seem not to require mutual exclusivity (for example, a valve may be part of both an automobile's tires and gasoline compressor). A taxonomy is a hierarchy, however, and therefore subject to the strict mutual exclusivity requirement. In this study, therefore, questions will be assigned to one and only one class in each taxonomy. Classification of questions might benefit from schemes in which questions could belong to more than one class, and future research is called for into such classification schemes for questions.

2.7.4. Automatic Classification and Automatic Indexing

There are two distinct bodies of research and development in the automation of classification tasks. These tasks are referred to in these bodies of literature as automatic classification and automatic indexing, though these terms are not used at all consistently in this literature. Automatic classification is the task of grouping documents based on similarities or measures of association, and automatic indexing is the task of assigning predetermined terms to documents. According to the distinction between classification and categorization discussed above, a more appropriate name for automatic classification would be automatic categorization, and a more appropriate name for automatic indexing would be automatic classification. But for the sake of clarity the received terms will be used here. Both of these bodies of research and development will be discussed in this section.

Van Rijsbergen (1979) provides an excellent overview of the technical details of automatic classification, and so those details will not be discussed here. The important point is that the premise behind automatic classification is what van Rijsbergen refers to as the cluster hypothesis: that "closely associated documents tend to be relevant to the same requests" (p. 30). The purpose of automatic classification is to create clusters of documents, either *a priori* or in response to a specific query, in which all documents are likely to be relevant to an information retrieval query. Different systems for automatic classification use different methods for creating these clusters, including graph theory

(Augustson and Minker, 1970) and vector analysis (Salton, Wong, and Yang, 1975). Some recent work that takes automatic classification in an interesting new direction is presented by Bergmark and colleagues (Bergmark, 2002; Bergmark, D., Lagoze, C., & Sbityakov, A., 2002) in their work on automatic collection generation in digital library environments.

While automatic classification appears to be a useful direction for automation in improving search engine and digital library functionality, such systems do not perform the sort of classification that was performed manually in this study. That sort of classification – the assignment of classes to entities bears a closer resemblance to the task performed by automatic indexing systems.

Some of the earliest work on automatic indexing is presented by Maron (1961), Borko and Bernick (1963, 1964), and Borko (1964), describing experiments in which documents were automatically assigned predetermined terms, and those assignments then compared with human indexing assignments. Subsequent work on automatic indexing follows this model of judging the accuracy of the system *via* human judgment, either 1. comparing the automatically assigned terms with human indexing assignments or 2. judging the relevance of the documents indexed with specific terms to the subjects described by those terms. Recent work in automatic indexing has experimented with assigning index terms from classification schemes in a variety of subject areas: medicine (Larkey and Croft, 1996), case law (Thompson, 2001), even the semantic web (Allan, 1996).

There are two features of automatic indexing that are important to this study: 1) the indexing terms are predetermined, often from a preexisting classification scheme, and 2) the accuracy of the system is judged by comparison to human performance of the same task or relative to human relevance judgments. Experiments in automatic indexing in subject areas in which the preexisting classification scheme was inadequate, experimenters created their own set of indexing terms (Maron, 1961; Borko and Bernick, 1963, 1964). For this study, no preexisting classification scheme of factors that affect the process of triage existed, so the researcher identified a set of fifteen such factors *via* a

Delphi study of digital reference triagers (Pomerantz, Nicholson, and Lankes, 2003) (see section 4.2.2.4 for the list of these fifteen factors). This set of fifteen factors was expanded to thirty-eight during the course of Phase 1 of this study. This classification scheme of thirty-eight classes served as the scheme according to which factors that influence the triage process were classified during Phase 1. This study proposes a system to automate part or all of the triage process, which would be analogous to automatic indexing systems: 1) the classification scheme used would be the thirty-eight factors that influence the triage process, and 2) the accuracy of such a system could only be judged according to human judgment of whether a question has been triaged correctly.

2.7.5. Faceted Classification Schemes

The idea of faceted classification schemes was originated by Ranganathan (1965). While this approach to classification has gained a great deal of popularity in the classification literature since it was proposed, at the time of its introduction it was revolutionary.

Faceted classification flies in the face of traditional classification schemes, which, as Ranganathan states, “provide... either for one free facet only, or for one frozen and one free facet” (p. 11). In other words, traditional classification schemes allow entities to be classified either, 1) according to only one aspect of the entity, as interpreted by the classifier, or 2) according to two aspects of the entity, only one of which is open to the interpretation of the classifier, and the other aspect is predetermined by the structure of the classification scheme itself.

Ranganathan’s (1965) idea of faceted classification was to allow entities to be classified according to several aspects of its content. Ranganathan proposed five “Fundamental Categories,” or dimensions, along which entities could be classified: Personality, Matter, Energy, Space, and Time. Ranganathan proposes a “Postulate of Sequence,” stating that the basic facet – which he himself admits is a “nebulous” concept (p. 273), but can be taken to mean the most important dimension along which an entity is classified – “should be put first; and the other facets should be arranged thereafter in the sequence of the decreasing concreteness of the Fundamental Categories of which they are respectively

taken to be manifestations” (p. 67). Ranganathan uses the diesel engine as an example: Personality, according to Ranganathan, is the basic facet, as “it is a piece of iron made functional and endowed with a personality of its own” (p. 212). After that, in decreasing order of importance, a diesel engine is an object of Matter, then a manifestation of Energy, then manifestations of Space and Time.

Other classificationists have built on and expanded Ranganathan’s ideas on faceted classification schemes. A great deal of this work was done by the Classification Research Group, and reported on by Vickery (1960, 1966). There are two expansions of faceted classification that are relevant to this study: 1) faceted classification schemes are “locally produced” (Vickery, 1966, p. 16), and so may be specifically designed to fulfill particular classificatory requirements, and, as a corollary, 2) the facets may be any categories that are logical for the entities being classified – Personality, Matter, Energy, Space, and Time are not the only possible categories.

The *Art & Architecture Thesaurus* (AAT) is an example of a faceted classification scheme, designed to meet the particular requirements of its domain. Indeed, it is not confined to five facets; AAT has seven facets: Activities, Agents, Associated Concepts, Materials, Objects, Physical Attributes, and Styles and Periods. Each facet is then broken down according to its own logic: for example, the Agents facet allows for the nominal designation of “people, groups of people, and organizations” (Petersen, 1994, p. 32), while the Styles and Periods facet is arranged chronologically. Any given object classified according to AAT is classified along each of the seven facets, enabling a rich description of that object to be created.

This approach to classification enables this study to propose the possibility of a faceted classification scheme for questions. Such a faceted scheme may be “locally produced” in the sense that it may be specifically designed to fulfill the particular classificatory requirements of digital reference services. The facets of such a scheme may be existing question taxonomies discussed in the subsequent section, or they may be other question

taxonomies altogether. The possibility of other question taxonomies than those discussed in the subsequent section will be discussed in chapter 5.

2.7.6. Evaluation of Classification Schemes

There is no universally accepted methodology for evaluating classification schemes. Indeed, the evaluation of classifications schemes is not a large branch of the literature on classification, and every work that performs this evaluation does so according to slightly different criteria.

Polan's (1975) evaluation of classification schemes developed specifically for use in children's libraries relies on one criterion: utility. Polan states that the only criterion on which a classification scheme need be evaluated is the use that may be made of books on the shelf – that is, how well the organization of books on the shelf matches the manner in which the patrons think about those books and their subjects. This one-criterion evaluation is rather simplistic. In addition, it is entirely user-based: it requires that in order to evaluate a classification scheme, observations must be made of users' interactions with materials classified according to that scheme. While user-based evaluation is a highly fruitful method for the evaluation of any system, an entirely user-based approach does not allow for the existence of absolute criteria: that there may be features that a classification scheme *must* possess in order to be considered useful.

On the other side of the coin, Merrett's (1982) evaluation of map classification schemes takes six criteria into account, all absolute. These criteria are:

- Notation: Syntax of the classification vocabulary (e.g., alphanumeric, punctuated, natural language).
- Area: How regions on the map are listed (e.g., alphabetically, by continent).
- Theme: How the classes in a classification scheme are divided into sub-classes.
- Scope for area-theme synthesis: Syntax of the vocabulary for representing map regions.

- Bias: National focus of the map (e.g., American, international).
- Suitability: In what sort of a library would the classification scheme be appropriate.

While some of these six criteria (e.g., Area and Bias) are phrased such that they are most appropriate for the evaluation of map classification schemes, all of these criteria may be extended to the evaluation of any classification scheme.

Kwaśnik and Liu (2000) present the most comprehensive set of criteria for the evaluation of a classification scheme, encompassing all of the evaluation criteria discussed above.

Kwaśnik and Liu's set of evaluation criteria is as follows:

- Scope of the classification – The breadth of the domain covered.
- Exhaustivity – Completeness of coverage of the domain. In other words: how big is the Miscellaneous class?
- Expressiveness – The accuracy of the reflection of the domain covered. In other words: how “lifelike” is the classification in its representation of the scope?
- Granularity – The level of specificity of the classes. In other words: are the scopes of the classes at an appropriate and consistent level of specificity?
- Hospitality – Ease of accommodation of new classes. This criterion is related to Granularity in that it refers to the hospitality of existing classes to subclasses. This criterion is also related to Exhaustivity in that it also refers to the hospitality of the scheme to new top-level classes.
- Structure – The relationships among the entities in classes, both implicit and explicit.
- Vocabulary – The depth and consistency of the terms used to refer to the classes.
- Coherence – The logic and unity of the Vocabulary and Structure.
- Consistency – The predictability of the Vocabulary and Structure.
- Usability – Ease of use, for the classifier. In other words: how user-friendly is the scheme for the task of classifying entities in the domain?

- **Browsability** – Ease of use, for the user. In other words: how user-friendly is the scheme for exploration of the domain?

This set of criteria for the evaluation of classification schemes is particularly useful, as it is specific enough to account for most of the relevant aspects of a classification scheme, but broad enough that it looks at a classification scheme as a whole, not at the specifics of how any individual entity is classified. This is the sort of evaluation that was required for Phase 2 of this study: an evaluation that allows each question taxonomy to be judged on its merits of structure and utility, without overmuch concern for the actual process of the classification of any particular entity.

There are, however, two criteria that may be added to Kwaśnik and Liu’s (2000) set. These are as follows:

- **Partitioning** – The ability of the classification scheme to distinguish between entities along significant and useful dimensions. In other words: What are the criteria that differentiate one class from another?
- **Prototype characteristics** – The ability of the classification scheme to create classes based on useful criteria of association. In other words: What are the defining characteristics of the prototypical entity in a class?

Kwaśnik and Liu’s (2000) set of criteria, plus these additional two, break down into four groups based on the focus of the criterion. These groups of criteria are as follows:

Table 2-3: Thirteen Criteria for the Evaluation of Classification Schemes

Criteria Concerned with	Criteria
The domain of the entire classification scheme	Scope
	Exhaustivity
	Expressiveness
The domain of classes	Granularity
	Hospitality

Relationships between classes and entities in classes	Structure
	Partitioning
	Prototype characteristics
The terminology used in the classification scheme	Vocabulary
	Coherence
	Consistency
The utility of the classification scheme	Usability
	Browsability

The evaluation of the question taxonomies in this study will be performed using this set of thirteen criteria.

2.8. Survey of Existing Question Taxonomies

Research question 2 for this study was: How does question type correlate with the action taken on a question in the triage process? In order to answer this question, question types must be identifiable. In short, this study required classification schemes for questions, according to which the questions received by digital reference services could be classified.

An extensive review was conducted of the bodies of literature from several fields that deal with questions. This review was conducted to identify question taxonomies that exist in these bodies of literature. In the course of this review, four question taxonomies were identified in these bodies of literature. These four question taxonomies are as follows:

1. Subjects of questions
2. Wh- words
3. The functions of expected answers to questions
4. The forms of expected answers to questions

The term “taxonomy” is used throughout this study to refer to these four classification schemes for questions. Indeed, the term taxonomy is used more or less synonymously with the term “classification scheme” throughout this study. The term taxonomy is used in this study because it is the term used in the following two bodies of literature:

- 1) the literature on the cognitive processes of questioning and question-answering behavior (Lehnert, 1978; Graesser, Lang, Horgan, 1988; Graesser, Person, and Huber, 1992), and
- 2) the literature on QA systems research and development (Moldovan et al., 1999; Hovy et al., 2000; Hovy, Hermjakob, and Lin, 2001).

The term taxonomy is therefore used throughout this study in order to position this study relative to research in questioning behavior and development of QA systems, as an extension of that work into the arena of digital reference.

Kwaśnik (1999) suggests that a taxonomy must be organized as a hierarchy. The example that Kwaśnik uses is Linnaean taxonomy, the traditional zoological classification of animals. Of the taxonomies of questions utilized in this study, however, only two of three are hierarchies at all: the taxonomy of forms of expected answers is flat. Furthermore, the taxonomies of wh- words and functions of expected answers are only minimally hierarchical, with few classes containing sub-classes. This could be viewed as an argument against the use of the term taxonomy, in favor of the more general term classification scheme. The term taxonomy is chosen, however, because these taxonomies show signs of developing into more fully hierarchical forms. All three taxonomies were flat as they appeared in the literature, but two of three developed hierarchical characteristics as they were developed and clarified (this is discussed in detail in chapters 3 and 4). No claim is made here that, during the course of this study, these three taxonomies have been as fully developed as it is possible for them to be. The nature of an hierarchical classification scheme is it may grow and expand as more and more comes to be known about the domain it represents; this is one reason why the Linnaean taxonomy has proven to be robust enough to endure nearly 200 years of use. As future studies

reveal more about the nature and types of questions, it seems likely that these taxonomies will develop more fully hierarchical characteristics.

2.8.1. Subjects of Questions

The term “subject” is used throughout this study instead of the term “topic,” though in common usage these two words are frequently treated as synonymous. This was done to remain consistent with the precedent set in the literature on classification theory. Beghtol (1986), in a review of approaches to the study of the “aboutness” of documents, states that what a document is *about* is its “intrinsic subject” (p. 85). Similarly, in studies of inter-indexer consistency (Leonard, 1977; Hutchins, 1978) the term “subject” (as in “subject heading”) is consistently used to describe the terms assigned to an entity by an indexer. The term “subject” is used in this study in keeping with the use of that term in the field of classification.

There are many classification schemes that organize entities according to their subjects. In the field of classification, these frequently take the form of thesauri, both subject-specific (such as the *Art & Architecture Thesaurus* (Petersen, 1994) and the ASIS Thesaurus of Information Science (Milstead, 1999)) and general (such as the Library of Congress Subject Headings (Library of Congress, 1992)). The function of a thesaurus is to represent an intellectual entity by means of a set of subjects, from the most general to the most specific.

Organization by subject has been a common means for classifying documents since Melvil Dewey first conceived of his subject scheme in 1873. Once the leap was made in libraries to thinking about other types of artifacts as intellectual entities (such as art and architecture), then it was a smaller step to thinking about questions (non-print and indeed, immaterial entities) as intellectual entities that could be classified. And indeed, organization by subject was the first approach taken to the classification of questions. Perhaps the earliest example of a classification scheme for questions dates back three-quarters of a century. Conner (1927) used the ten main classes of the Dewey Decimal

Classification (DDC)³ to classify questions recorded by the reference department of the Carnegie Library of Pittsburgh. Conner analyzes the percentages of questions received by the reference department in her library in order to make recommendations as to the content of courses in library school, so that future librarians will be prepared to answer questions on the most commonly asked subjects. What is perhaps most interesting about Conner's classification is that she applies the same scheme (DDC) that is used to classify materials in the library to also classify reference questions. The assumption made by Conner's approach to classifying questions is that any question received by the reference desk can be answered by material in the library's collection. This assumption proceeds further: since there is a one-to-one correspondence between the question and the source from which an answer is derived, the same scheme that is used to classify materials from which answers may be derived is appropriate for questions.

Taylor (1968) describes five "filters through which a question passes" (p. 183) in the mind of the reference librarian, that enable the librarian to interpret the question, understand the patron's information need, and proceed to formulate an answer. The first of these steps is the determination of the subject of the question. Although Taylor does not discuss taxonomies of subjects of questions, such a taxonomy is implied by this first filter. As innovative as Taylor was in the reference community, however, still he was heir to the tradition of thinking about reference questions first and foremost in terms of their subject.

To be fair, reference services to this day tend to think about reference questions exclusively in terms of their subject. Many reference evaluation tools require that the subjects of questions be recorded (King, 1982; Murfin and Gugelchuk, 1987), the intention being to identify subjects on which it is difficult for reference librarians to answer questions, or to which reference librarians frequently cannot give accurate answers (Rothstein, 1964; Crews, 1988). Many digital reference services have resources available on their websites (often frequently asked questions edited for general

³ DDC 12 was published in 1927. Conner does not state what edition of DDC she used, but it is reasonable to assume that it was edition 11 or 12.

consumption), which are organized in subject hierarchies. The VRD service, from the archives of which the questions in the data set utilized in Phase 2 of this study are drawn, organizes resources according to the Gateway to Educational Materials (GEM) Subject element.

There are many subject classification schemes in existence; the Library of Congress Subject Headings, ERIC thesaurus, ASIS Thesaurus of Information Science, and GEM are only a few that have been designed to perform specific tasks in specific environments. There are so many subject classification schemes, in fact, that, as discussed in chapter 4, many different digital reference services are already using different ones. Part of Phase 3 of this study is to classify questions according to the taxonomies identified in this section. Given the number of choices of subject classification schemes available, however, it is a matter of preference which one is used, and different digital reference services use different schemes. Given the availability of several subject classification schemes, the choice of the particular subject classification scheme was left to the particular reference service and its unique requirements and classification practices. This study therefore classified questions according to only three of the four taxonomies identified in the literature: the taxonomies of wh- words, functions of expected answers, and forms of expected answers.

2.8.2. Wh- Words

“The Five Ws” is a simple and common classification of questions in English. It is learned in school at an early age, at least in the United States, as the standard way to construct a question in English. The five Ws are also a tradition of journalistic writing style, going hand-in-hand with the “inverted pyramid”: in the traditional news story, the most important elements must be included in the opening paragraph. These most important elements are Who, What, When, Where, Why, and How. The fact that the Five Ws are actually five Ws and an H is merely a reflection of the perversity of spelling in English.

Perhaps because it is such a common and intuitive way of thinking about questions (at least for English speakers), it is popular in the literature on questions and question answering. Indeed, Robinson and Rackstraw (1972a, 1972b) wrote two entire volumes investigating wh- words, the forms of questions based on these words, and the forms of answers to these questions. Robinson and Rackstraw define wh- words as “the total set of lexically marked interrogative words” (1972a, p. 2). By “lexical set” Robinson and Rackstraw mean a distinct group of words that can be used in “similar linguistic environments” (p. 39). Thus, for Robinson and Rackstraw, the wh- words are a set of words that can all be used to form an interrogative sentence – in other words, a question. As a corollary, any sentence containing a wh- word is a question.

Despite a certain elegance in the simplicity of the five Ws, Robinson and Rackstraw (1972a) add one class to this set. Robinson and Rackstraw draw a distinction between open and closed questions and, by extension, between questions phrased using the words What and Which. According to Robinson and Rackstraw, the word “what” indicates a question about an infinite or undefined set, while the word “which” indicates that the question asks for the identification of a particular entity out of a finite set of entities.

The taxonomy of wh- questions, according to Robinson and Rackstraw (1972a), is as follows:

- Who
- Which
- What
- When
- Where
- Why
- How

Several QA systems in TREC use wh- words as the primary criterion for the logical representation of questions and the recognition of question focus (Moldovan et al., 1999; Hovy, Hermjakob, and Lin, 2001; Kwok et al., 2001). Some of these systems subdivide

wh- types – for example, into such classes as WHY-FAMOUS-PERSON and ABBREVIATION-EXPANSION (Hovy, Hermjakob, and Lin, 2001, p. 167). The purpose of this subdivision of wh- types is to enable QA systems to recognize “the semantic type of the desired answer” (Hovy, Hermjakob, and Lin, 2001, p. 167). Such QA systems treat this subdivision as a hierarchical one: the first level of classes are wh- words, and the second level are semantic types. This study makes use of these same two classification schemes, but treats them not as two levels of one hierarchy, but as two facets of a multi-faceted scheme.

2.8.3. The Functions of Expected Answers

Due to the complexity of some questions, Hermjakob (2001) points out that “it is not sufficient to analyze the wh-group of a sentence.” Hermjakob suggests that to determine how best to answer a question, “it is important to classify questions with respect to their answer types” (p. 18).

There is a fine balance to be struck here, on the following issue: this study investigates the types of *questions* received by digital reference services, not the types of *answers* provided to those questions. Therefore, a taxonomy of answer types was avoided in this study. On the other hand, an information-seeking question is an attempt by the questioner to elicit a response from the person questioned; therefore an information-seeking question requires an answer.

Thus we arrive at the taxonomy originally developed by Lehnert (1978). This taxonomy is not, strictly speaking, a taxonomy of questions, but rather a taxonomy of the possible functions of an answer in fulfilling a questioner’s information need. On the other side of the coin, Dervin (1983) describes six measures of the fulfillments of information needs. These six measures are facets along which information needs may be described, and in fact may be viewed as taxonomies of information needs.

The taxonomy of the functions of expected answers was originally developed by Lehnert (1978) for use in her story-understanding system QUALM, discussed above. Subsequently, her taxonomy was adopted by Graesser and colleagues for several studies analyzing questions asked by individuals in a variety of real-world settings: while reading texts, while learning a new computer system, and while watching television news (Graesser, Lang, and Horgan, 1988). Graesser, McMahan, and Johnson (1994) state that this taxonomy defines question classes “on the basis of the content of the information sought” by the question (p. 520). Indeed, “the content of the information sought” is one definition of question focus, the recognition of which is an important process employed by many TREC QA systems. Over time, Graesser and colleagues (Graesser, Person, and Huber, 1992; Graesser, McMahan, and Johnson, 1994) developed a theoretical model of question-asking behavior. This taxonomy reached its most fully developed form in Graesser, McMahan, and Johnson. In this developed form, this taxonomy is divided into classes that require short versus long answers. This taxonomy of the functions of expected answers was used to classify questions and subsequently modified in Phase 2 of this study (see Appendix D for the modified version of this taxonomy). This taxonomy, as presented by Graesser, McMahan, and Johnson, is as follows:

Table 2-4: The Taxonomy of Functions of Expected Answers

Class		Abstract specification
Short Answer	Verification	Is a fact true? Did an event occur?
	Disjunctive	Is X or Y the case? Is X, Y, or Z the case?
	Concept completion	Who? What? When? Where? What is the reference of a noun argument slot?
	Feature specification	What qualitative attributes does entity X have?
	Quantification	What is the value of a quantitative variable? How many?

Long Answer	Definition	What does X mean?
	Example	What is an example label or instance of the category?
	Comparison	How is X similar to Y? How is X different from Y?
	Interpretation	What concept or claim can be inferred from a static or active pattern of data?
	Causal antecedent	What state or event causally led to an event or state?
	Causal consequence	What are the consequences of an event or state?
	Goal orientation	What are the motives or goals behind an agent's action?
	Instrumental / procedural	What instrument or plan allows an agent to accomplish a goal?
	Enablement	What object or resource allows an agent to perform an action?
	Expectational	Why did some expected event not occur?
	Judgmental	What value does the answerer place on an idea or advice?
	Assertion	The speaker makes a statement indicating that he lacks knowledge or does not understand an idea.
	Request / directive	The speaker wants the listener to perform an action.

Several researchers have utilized the taxonomy developed by Lehnert (1978) and further developed by Graesser and colleagues to classify questions. Oddly, several of these researchers utilized this taxonomy to classify questions in medical environments. Keyes (1996) utilized this taxonomy to classify a set of queries associated with a database of

documents on cystic fibrosis. Stavri (1996) utilized this taxonomy to classify physicians' questions during the process of making preliminary diagnoses. Smith (2002) utilized this taxonomy to classify artificial questions simulating those that are asked by clinicians about patients. These three studies demonstrate the usefulness of this taxonomy for classifying questions. White (1998), however, goes a step further and utilizes this taxonomy not simply to analyze questions, but to analyze questions asked at reference desks. White's study is a landmark for both the development of this taxonomy and for the literature on library reference. Graesser, Lang, and Horgan (1988), Keyes, Stavri, and Smith demonstrated the applicability of this taxonomy to the analysis of questions generated in real-world settings; White, however, was the first to demonstrate the applicability of this taxonomy to the analysis of questions generated during the reference transaction.

2.8.4. The Forms of Expected Answers

Investigation of the question types that patrons ask at library reference desks has been a thread running through the library literature for the better part of a century (Conner, 1927; Swift, 1934). Three distinct methods have been employed for the classification of questions in reference environments:

1. The subject of the question,
2. Types of sources from which the answer may be drawn, and
3. The forms of expected answers.

2.8.4.1. The Subject of the Question

Classification according to subject was discussed in section 2.8.1, and will not be discussed further here.

2.8.4.2. Types of Sources from Which the Answer May Be Drawn

Taylor (1968) states that “when an inquirer approaches the reference desk, he has some picture in mind as to what he expects his answer to look like, *i.e.* format, data, size, etc.” (p. 187). Taylor goes on to state that it is the job of the reference librarian, through the vehicle of the reference interview, to alter the patron’s *a priori* notion of what the answer should “look like.” In many cases, the patron’s notion of what the answer should look like gives way to the librarian’s notion of what the answer should look like. It could be argued that this is as it should be, since the librarian is the expert in answering reference questions and knows what sorts of answers can be provided using the library’s collection. Others, such as Dervin (1977) and White (1989), argue that because the process of query negotiation is a process of communication, the librarian and patron should arrive at a common ground in terms of formulating an answer. In either case, however, it is ultimately the reference librarian that guides the patron to an information source or sources, and – ideally – to an answer within a source.

Consequently, much of the training that reference librarians receive – both in library school and on the job – focuses on information sources, both general types and specific sources. Entire textbooks and training manuals have been written to teach reference librarians what sources to select, when faced with questions of certain types (Slavens, 1985; Bopp and Smith, 2001). Is it any wonder that, as discussed in section 2.4, knowledge-based systems for reference attempt to match a question to an information source? These systems are designed to perform the functions of reference librarians – but reference librarians have years of training that cannot easily be turned into algorithms. Still, that does not prevent librarians from trying, and as a result, an informal taxonomy of the types of sources from which answers may be drawn has evolved in the literature. Richardson (1995) describes twelve reference formats (p. 156):

1. Abstracts
2. Atlases
3. Bibliographies
4. Biographical sources
5. Dictionaries
6. Directories
7. Encyclopedias
8. Government publications

9. Handbooks / manuals

11. Statistical sources

10. Indexes

12. Yearbooks

Attempting to match a question to an information source, however, requires a great deal of world knowledge about reference and libraries. It may be possible to convert this world knowledge into algorithms; this is, at least in part, the goal of current IR and QA systems. This study takes a step back from that goal, however, and argues that it is necessary to first analyze the question before attempting to formulate an answer. This taxonomy of the types of sources from which answers may be drawn will therefore not be utilized in this study. This taxonomy should, however, be kept in reserve, since future research and development bridging digital reference and QA systems are likely to require this taxonomy for the implementation of future reference QA systems.

2.8.4.3. The Forms of Expected Answers

As discussed above, several schemes exist for classifying entities by subject. These schemes are highly standardized, but not widely used in reference services. On the other hand, the taxonomy of the forms of expected answers to questions is widely used, but not standardized. The need for standards for measurement and evaluation of reference services has been recognized in the library profession for some time. In the mid-1970s the American Library Association (ALA)'s Library Administration and Management Association (LAMA) created standard definitions for two types of reference transactions for inclusion in their Library General Information Survey (LIBGIS) (U.S. Department of Education, 1981). These two types of reference transactions are as follows:

- “A reference transaction... involves the knowledge, use, recommendation, interpretation, or instruction in the use of one or more information sources by a member of the reference/information staff.”
- “A directional transaction... provides assistance in finding and using library services, collections and facilities.”

(White, 1981, pp. 33-34)

The LIBGIS definitions were the first standardization of types of reference transactions, and for the first time provided a classification (simple as it is) of the types of services provided at a reference desk. Also for the first time, the LIBGIS enabled reference services at different libraries, holding different collections and serving different communities of patrons, to share reference statistics (Kaske and Aluri, 1980). Additionally, researchers conducting evaluation studies of reference services utilized the LIBGIS classes (Herner, Vellucci, and Leyman, 1972; Phenix, 1983).

The classes “reference” and “directional” are, however, extremely broad. As a result, some researchers and libraries divided these classes into a variety of subclasses (White, 1981; Phenix, 1983). Rothstein (1964), predicting the classification to come, discusses grouping questions into the following types: *directional*, *ready reference*, *search* (or research), and *readers’ advisory* (p. 458). Seng (1978) discusses three question types: first, *direction*, and then two that are subclasses of the LIBGIS *reference* class, but which Seng defines in a unique way: *information* (a question that “is concerned with information resources and/or their use”), and *general* (a reference question “answered through the use of information resources”) (p. 22). Brown (1985) drops the *directional* class entirely, and divides questions into *informational* (any question that can be answered using ready reference sources such as the card catalog or telephone directory) and *reference* (any question that requires non-ready reference sources to answer it) (p. 294). Brown goes a step further and, following in Conner’s (1927) footsteps, classifies reference questions according to the ten main classes of DDC⁴. Fogarty (1995) discusses the following four types: *directional*, *instructional*, *ready reference*, and *extended reference* (p. 20).

These variations on the LIBGIS theme demonstrate that even given a standard, different services will modify and extend that standard to accommodate their specific situation and requirements. What is even more interesting, for the purposes of this study, is the amount of “convergent evolution” that has occurred surrounding this taxonomy of question types. Several researchers and libraries explicitly modified the LIBGIS classes. Equally many,

⁴ Most likely edition 19, published 1982.

however, independently developed question classification schemes that resembled the LIBGIS scheme, and either did not know of the existence of the LIBGIS scheme, or simply did not mention it (Lynch, 1978; Bunge, 1990; Dewdney and Mitchell, 1996; Stalker and Murfin, 1996; Carter and Janes, 2000). Looking across all of these variations on a theme, the following taxonomy of the forms of the expected answer to a question was developed. This taxonomy of the forms of expected answers was used to classify questions and subsequently modified in Phase 2 of this study (see Appendix D for the modified version of this taxonomy).

Table 2-5: The Taxonomy of Forms of Expected Answers

Class	The Form of the Expected Answer
Directional	Questions asking about the location of a specific information source.
Holdings	Questions about whether a specific information source or document is owned by the library.
Ready reference	Questions asking for simple, factual answers; the answer should be readily ascertainable from available information sources.
Exact reproduction	Questions asking for pictorial and textual materials, taken directly from an information source and unchanged.
Description	Questions asking for a description of something, briefer in length than the original thing (basically, an abstract).
Readers advisory	Questions asking for assistance in the choice of books or the gathering of data.
Bibliographic instruction	Questions asking for assistance in use of information source(s).
Research	Questions asking for involved answers; the answer should require some effort and wide use of information sources to

	formulate.
Citation list	Questions asking for a list of information sources on a particular subject.
Analysis	Questions asking for some form of data analysis, whatever that data might be – scientific, social, financial, etc. Questions of this type might ask for trends, pro or con arguments, cause and effect, compare and contrast, etc.
Critique	Questions asking for an evaluative discussion of a particular subject. (E.g.: a movie review, Cliffs notes-like analyses of a book, etc.)

2.8.5. Linguistic Analysis

As defined by Liddy (1998), natural language processing is “a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis” (p. 14). These levels of linguistic analysis are as follows:

7. Pragmatic: Understanding the purposeful use of language in situations, particularly those aspects of language which require world knowledge
6. Discourse: Interpreting structure and meaning conveyed by texts larger than a sentence
5. Semantic: Determining the possible meanings of a sentence, including disambiguation of words in context
4. Syntactic: Analysis of words in a sentence in order to uncover the grammatical structure of the sentence
3. Lexical: word level analysis including lexical meaning and part of speech analysis
2. Morphological: Componential analysis of words, including prefixes, suffixes and roots
1. Phonological: Interpretation of speech sounds within and across words

Several QA systems that compete in the TREC QA track utilize natural language techniques (Moldovan et al., 1999; Chen et al., 2001; Hovy, Hermjakob, and Lin, 2001). It should come as no surprise, therefore, that the four taxonomies discussed above – two of which are used by some existing QA systems (wh- words and the functions of expected answers) – correspond to the top four levels of linguistic analysis, as follows:

Table 2-6: Levels of Linguistic Analysis of Question Taxonomies

Level of Linguistic Analysis	Question Taxonomy
Pragmatic	The forms of expected answers to questions
Discourse	The functions of expected answers to questions
Semantic	Subjects of questions
Syntactic	Wh- words

Liddy (1998) states that “lower levels of language processing have been more thoroughly investigated and incorporated into IR systems” (p. 15). This is an advantage for any future QA system for reference. Such a system can incorporate existing developments utilizing the lower levels of linguistic analysis, and concentrate on implementing the upper levels, specifically utilizing the results of this study. Additionally, question taxonomies at these levels of linguistic analysis may be useful for any type of system that, like QA systems, classifies questions.

2.9. Chapter Summary

This chapter reviews the literature on digital reference, question answering, and classification. First, questions are defined so that questions received by digital reference services may be identified for the purposes of this study. Both digital reference and triage are defined and the need for automation in both is explored. A review of the literature on question answering and question answering systems is presented as the foundation for this study’s approach to questions. This leads into a review of the literature on

classification, and the idea of questions as entities to be classified. Finally, four question taxonomies are identified that are judged to be relevant to the task of automating the triage process.