TITLE: The Digital Reference Electronic Warehouse (DREW) Project: Creating the Infrastructure for Digital Reference Research through a Multi-Disciplinary Knowledge Base

AUTHOR(s): Scott Nicholson, R. David Lankes

PUBLICATION TYPE: Journal
DATE: 2007
FINAL CITATION: "The Digital Reference Electronic Warehouse (DREW) Project: Creating the Infrastructure for Digital Reference Research through a Multi-Disciplinary Knowledge Base" Nicholson, S., Lankes, R. David (2007) 46(3). Reference and User Services Quarterly.
KEYWORDS: Digital Reference, Data Mining

Preprint – Accepted for publication in Reference and User Services Quarterly

The Digital Reference Electronic Warehouse (DREW) Project: Creating the Infrastructure for Digital Reference Research through a Multi-Disciplinary Knowledge Base

Scott Nicholson (scott@scottnicholson.com) and R. David Lankes (rdlankes@iis.syr.edu), Information Institute of Syracuse, Syracuse University School of Information Studies, 245 Hinds Hall, Syracuse, NY 13244

Abstract

One of the valuable offerings of librarians in the digital age is the human intermediation of information needs. In physical libraries, these reference questions are answered and few artifacts remain from the transaction; therefore, the knowledge created through the work of the librarian leaves with the patron. Due to the medium of communication, digital reference transactions capture the knowledge of information professionals. There are hundreds of digital reference services generating knowledge every day; however, the lack of a schema for archiving reference transactions from multiple services makes it difficult to create a fielded, searchable knowledge base. This schema will also allow researchers to develop tools that practitioners can employ; this will create a collaborative environment for digital reference evaluation. The goal of this work is to outline the steps needed to develop this schema, present the results of a survey of digital reference services, explore some of the pitfalls in the process, and envision the future uses of this Digital Reference Electronic Warehouse (DREW).

Introduction

The future, and some might even say the present, for the library professional is the digital library. Instead of waiting for the user to come to their information containers in a physical collection, librarians select high-quality materials for users to access through the Internet. It is relatively easy to put a collection of static files online; however, the library is more than just a collection of documents. A crucial part of a library is the human intermediary – the librarian. This intermediary connects the users to the information needed, and can assist with advice about using the information retrieval systems and working with information.

However, many users turn to the Web search tools for their information retrieval needs. While these tools provide the user with Web pages that match a word on the topic, the quality of the results are questionable. Most Web search tools are for-profit companies and bombard users with advertising. In addition, search engine optimizers work to place commercial sites at the top of lists; this has resulted in many searches leading to page after page of commercial results. This commercial information is appropriate for some information-seeking needs, and this is an area where the Web search tools excel. However, it can be frustrating to find non-commercial information, and this is an opportunity for libraries.

There clearly is a need for intermediation with the location of material online. Users have turned to question-based search tools such as AskJeeves with the hopes of finding such assistance; however, these tools perform no better than a general search tool. There is another type of Web search tool that can take a user's question and match it to a set of results that are likely to be on topic with little advertising and no direct charge – a digital reference service. In fact, those teaching about Web search tools should always take the opportunity to present a digital reference service as a Web search tool with built-in intelligence.

Many libraries have started services where they allow users to submit questions via e-mail or Web forms. Librarians will then research the question and provide an answer and related documents to the user. Some libraries offer this service using a live-chat model, where the user is interacting with a librarian with little time elapsing between question and response. These services are usually free, although the user base may be limited to users that are affiliated with the library offering the services. Google has entered this domain with their "Google Answers" service where a user offers a question and sets the payment for a pre-approved Google Answerer to answer to the question.

Some digital reference services, commonly known as AskA services, connect the user directly to an expert in the field instead of to a librarian. Services such as Ask Dr. Math (http://mathforum.org/dr.math) and AskNSDL (http://nsdl.org/asknsdl) allow users to ask questions of experts in the topic. This is a different model of the reference process, but the information contained in these transactions is valuable. Lankes [1] presented a model that contrasted these two types of services in his research agenda for digital reference.

There are hundreds of these services around the world providing answers and resources in response to user needs. If collected into a knowledge base, it would be incredibly useful for researchers in exploring this process. Information seeking research has been an active line of exploration for decades, and there are many theories developed from small samples that could be explored with this larger dataset. In addition, by examining the common works referred to in different types of questions, automatically generated directories of high-quality material could be created and shared. The goal of the DREW project is to create a large database of reference transactions for researchers to better understand the process and create tools for measurement and evaluation that managers of reference services can employ.

Relationship of DREW to Similar Projects

There are several different types of digital multidisciplinary knowledge bases currently available. Precursors to today's knowledge bases are bibliographic databases such as ArticleFirst and database aggregators like DIALOG. As these tools have grown to include access to full-text resources, they become true multidisciplinary knowledge bases. The difficulty in using these databases comes through the methods of retrieval. Searchers have to match the words used by the author when searching free-text fields such as the title, abstract, and text of the document. Conversely, searchers could attempt to match words selected by indexers such as subject headings. Users can get frustrated with these tools, as they tend to match either too few or too many articles [2].

Another type of multidisciplinary knowledge base available is the World Wide Web. Web search tools provide a portal to this knowledge base. Most current Web search tools allow the user to search large portions of the textual data available on a conveniently-accessed subset of the Web. These search tools cannot access large portions of the World Wide Web know as the Invisible Web [3]; in fact, one study claims that the well-known search tools index only about .03% of the Web [4] .

In addition, as these search tools index the words used on the page, the user has to search on the words used by the authors of the page. Due to the commercial nature of these tools, many Web authors use Search Engine Optimization (SEO) techniques to push their pages to the top of listings [5]. If these two issues are combined – search tools only index a small portion of the Web and some companies are changing their pages to aggressively hold the top positions in the rankings of search tools – then it is expected that the typical user who only explores the first page of rankings will become frustrated with repetition of results.

One solution to these problems is human intermediation. Some search tools have integrated human intermediation through directory-based search tools; Yahoo, for example, started as a directory-based search tool. These tools allow a user to discover a small subset of resources that were selected using some type of quality criteria through a hierarchical organization structure. Over time, search tool companies have removed or reduced emphasis on these directory tools, promoting the full-text search tools in their stead.

There are some updated directory-based Web search tools that harness the power of human intermediation. The Open Directory (http://dmoz.org) and About.com (http://about.com) use experts to select Web sites on a topic and provide users with a directory-based access method. For scholarly research, Infomine (http://infomine.ucr.edu) is a high-quality directory out of the U.S., and BUBL (http://bubl.ac.uk/link) is focused on academic Web-based information from the U.K and Europe. The difficulty with these tools is similar to the problem with the bibliographic databases; searchers have to match either the terms selected by the authors of the pages or terms selected by the creator of the directory.

The setting for the current paper is in digital reference, which is human intermediation provided in direct response to user's query. Most of the time, the answer to a digital reference questions contains text as well as links to Web pages, journal articles, and other high-quality information. Therefore, the answer will connect the same types of resources discussed in the previous few paragraphs. The transaction will also have some metadata, such as subject headings, attached to it by either the user or by a staff member during the digital reference process.

In addition, the resources selected by an expert during the digital reference process will be of high quality. By gathering answers from many different resources, directories of these quality materials can be automatically generated. By appending commonly used query terms into the directory, the directory can be more easily searchable. Therefore, the knowledge base created through the archiving of digital reference transactions will be more easily searchable, contain references to high-quality resources, and provide indirect access to the human intermediation process of librarians and experts from a multitude of backgrounds.

Other Digital Reference Archives

Most reference services maintain some type of archive. That archive may be accessible only to the administrators, it may be a useful archive for those answering questions, or it may be available to users of the system. There are a few existing publicly accessible projects of archiving digital reference queries. A number of projects, such as Ask-A-Scientist (http://www.madsci.org/) and Google Answers (http://answers.google.com/answers), allow anyone to search their internal archive of question/answer pairs. While this is useful, it lacks the richness available if the transactions are collected by multiple services.

One of the largest shared archives of reference transactions is QuestionPoint's KnowledgeBase [6]. The purpose of the QuestionPoint KnowledgeBase is to provide reference librarians and their patrons with a repository for hard-to-find answers, answers to frequently asked questions, pathfinders and bibliographies on specific subjects, and the intellectual content resulting from aiding scholars in their research. Use of QuestionPoint's Knowledge Base is limited to those institutions participating in the QuestionPoint service, which allows for collaborative reference work.

This is a notable project because it is a large-scale shared reference depository with over 7,300 edited transactions as of July 2004; in addition, this knowledge base is growing as there are more than 11,000 transactions submitted and awaiting review (P. Rumbaugh, personal communication, July 6, 2004). Transactions are selected in two ways: any question submitted to the global network of reference librarians for an answer is considered, and individual libraries have the ability to select any local transaction and submit it to QuestionPoint for consideration. Once identified, the transactions are cleaned, removing all personal information about both the user and the librarian. The text of the question and answer are cleaned for clarity, free-text keywords assigned, and classification headings assigned from the top two levels of the Library of Congress Classification scheme. After ensuring that that there are not similar transactions on the topic area, the transaction is placed in the knowledge base. At this time, a "review" date can be set to trigger a manual review of the information in the transaction to ensure it is up to date.

One goal of the DREW project is to maintain a relationship with other major reference archives such as QuestionPoint. Examining these similar projects allows us to determine the needs of DREW and learn from the exploration of others. Due to the time and resources invested by OCLC and the Library of Congress in the development of the

QuestionPoint KnowledgeBase, their process and policies can serve as a model to libraries creating a cleaned archive to aid patrons and librarians. DREW, being a project to provide data for researchers about the process, requires a different type of warehouse. The transactions will not be edited for content, although personally identifiable information will be removed. Transactions on the same topic are desired, as that will allow the discovery of trends and changes over time. One of the areas of exploration, to be discussed later, is automation of several of the cleaning processes such as assignment of subject headings.

Therefore, DREW will complement these archives and knowledge bases focused on aiding librarians and their users directly. In order to do this, one goal of DREW is to create a schema that is compatible with different existing knowledge base projects. The challenge of this project is overcoming the complexity of many different services and user types. The landscape of digital reference is one of many types of services, librarians, and users interacting with a similar base of resources. There will be patterns across services, although teasing them out of the complex data is a challenge. The authors turn to complexity theory as the theoretical support for the success of this project.

Complexity Theory and DREW

To date, knowledge base work in digital reference has been primarily a deductive process. That is, either a service makes every transaction searchable, or an extensive transformation process of question selection, editing and incorporation into a pre-determined subject hierarchy. These deductive, and largely manual, processes have obvious scale problems. Further, these processes tend to be input only systems, in that they must be manually weeded of outdated information. Other issues in the deductive construction of knowledge bases are:

- Context Dependencies: Information in knowledge bases is very context dependent. It is quite possible that the only application of the information in a digital reference transcript is to that given interchange between librarian and patron.
- Metadata Creation: Time, labor and money are involved in creating metadata for transcripts and digital reference interchanges so that they may be later discovered and retrieved by end users. While some of this effort may be part of the reference process itself (for example classifying a question for distribution in QuestionPoint), it may still require effort to confirm and refine this classification data for inclusion in a knowledge base.
- Chunking: It is well known that users will ask several questions in both real-time and asynchronous transactions. How those questions and answers are “broken apart” is often dependent on human intervention and a great deal of interpretation.
- Fact Shifting and Temporal Dependencies: Answers to reference questions are often time dependent. From the name of the U.S. President to the height of Mount Everest, answers to even simple questions change. These changes, while concrete, are often hard to track over time. This does not even take into account “grey” areas where an answer or fact to apply to a question is a matter of choice among equally good options.

This is a simple and incomplete list. Issues of quality have not even been mentioned. These facts alone have stymied knowledge base builders; and this in an environment where true scale has not even come into play. How will any team of humans be expected to maintain a collection of questions and answers in an environment of million possible records? This is arguably a more difficult problem of maintaining a collection of any other type of documents for the simple fact that a knowledge base is not conceptualized as a set of documents with provenance and date, but as a collection of the more nebulous "knowledge."

While the use of full-text approaches such as vector-based information retrieval may mitigate some of these problems, they do not solve core difficulties of fact shifting, nor do they take into account the dynamic nature of the information presented. While the knowledge base grows the relationship between information may change as well. This situation is complicated when archives from different services are combined.

The authors argue that attempting to devise, scale. and equip a deductive approach to knowledge bases is ultimately unworkable. The authors further argue it is time to try a radically different, inductive approach. Simply put: let the knowledge base, or more specifically, the agents representing digital reference output, organize themselves.

Complex Adaptive Systems

The inductive approach proposed in this prospective is grounded in Complexity Theory and, more specifically, the concept of Complex Adaptive Systems as conceptualized by Holland. The authors will not explain the whole of complexity theory or delve any further than an operational explanation of complex adaptive systems in this document. For a deeper understanding of complexity theory see Waldrop [7]; for complex adaptive systems see Holland [8]; and for the application of complexity to digital reference see Lankes [9].

Put simply, complex adaptive systems are grounded in the creation of autonomous agents that self-organize based on relatively simple rules. This organization is emergent, in that it is not the product of some pre-determined course, but a result of the interactions of the agents themselves. The most common analogy is that of flocking birds. Systems to simulate the flocking behavior of birds are effectively replicated by creating independent agents in a virtual space with a set of very simple rules like "you must move forward: get as close as you can to those agents near you; do not hit anything." Such simulations demonstrate very effectively that such systems produce complex results with swarms of birds on a screen avoiding obstacles…even though they were never programmed to do obstacle avoidance…or swarming.

Models using these principles have also effectively been created to simulate the activities of financial markets, traffic flows and population studies. The point is, that complex adaptive systems, consisting the interactions of autonomous agents, have been effectively used to create systems impossible to create in a deductive manner where thousands of rules and lines of code would have to be used to anticipate every possible contingency. Already artificial intelligence systems have moved away from these so-called frame-based and expert system approaches toward neural nets and inductive simulations.

These systems are also dynamic, in that the agents constantly adapt to a changing environment. They constantly seek an optimal state in changing conditions. So the virtual birds will avoid obstacles in new ways as new obstacles are added. In simulations of biological systems agents will adapt to changes in weather or food supply. It is this dynamism that makes an inductive approach particularly suitable to digital reference knowledge bases.

In order to examine the contents of DREW and develop new, inductive approaches to knowledge base analysis and construction, the research team must first define the autonomous agents in the complex knowledge base environment. These agents, according to Holland [8], must have three mechanisms:

- Tags: Mechanisms that agents utilize for aggregation and flows of information
- Internal Models: A representation of the environment used by an agent to anticipate and adapt to the environment
- Building Blocks: Components of internal models combined to build, test and re-build internal models.

The "Internal Models," and "Building Blocks" will be the result of future research. Tagging, or the mechanisms used for information flow and identification, however, are central to the present study. These tags can be thought of fields or metadata elements. By identifying common elements in digital reference transactions (knowledge base agents) these agents can be compared, clustered, and examined. In order to take the first step in building a digital reference knowledge base as a complex adaptive system the researchers turned to existing standards for representing digital reference transactions.

Standards for Exchange

The National Information Standards Organization has developing a protocol for the exchange of questions between services, called NetRef [10]. While this standard is appropriate for questions while they are being answered, it is not appropriate for the long-term archiving of the exchange. One goal of the DREW project, therefore, is to create a schema for the *archiving* of digital reference transactions once the question-answering process is complete. It is important that this archival schema be compatible with the NISO standard, and perhaps can eventually become part of that standard. Theoretically, it should be easier for systems implementing the NetRef protocol to work with the DREW archival schema.

As these questions are answered, individual reference services create archives of question/answer pairs. These are the artifacts of human intermediation, and represent valuable information that previously was lost in traditional reference. Sometimes these archives are searchable by the public, and other times they are kept as referral tools for the librarians and experts to use in answering questions. This distributed knowledge base of digital reference archives contains the expertise and knowledge of many minds; however, there is currently no way to merge these separate archives into a single knowledge base. If these reference transactions from different services could be collected, cleaned, and privatized into a single data warehouse, the amount of expertise available to users and researchers would be staggering. However, the challenges involved in creating this type of warehouse are just as staggering. The goal of this work is the present the preliminary research in determining the fields that could make up an archival schema and present current and future plans of the DREW project

Determining the Fields

The first step in creating a data warehouse is to determine the fields that will be collected. As there are many different digital reference services, any schema for capturing information from these different services will result in compromises. In order to better understand what fields would be appropriate to capture, a survey was taken of digital reference services representatives.

In order to develop the fields needed for the archiving of digital reference transactions, we start by exploring what is currently captured and then work toward implementation in an iterative manner. The first stage is a survey of digital reference services with the goal of learning:

- what fields are currently collected by services,
- what fields are services not currently collecting, but are willing to collect, and
- what fields services are not willing to collect

in each of four categories – Patron, Question, Answer, and Expert.

First, field lists were created from Janes's work [11] and a small group of digital reference services and used to develop a survey instrument. This instrument was tested with a set of volunteer librarians from these services; these librarians added additional fields to the instrument. The instrument was then delivered at the 2003 Virtual Reference Desk conference and through a Web-based survey. The online survey was promoted through the DIG_REF listserv as well as through direct contact of services doing digital reference research. If an institution had different types of reference services (such as live chat and Web form-based asynchronous), it was requested that they fill out the instrument twice.

The survey gathered demographic information such as the communication methods used for question acceptance and question resolution, number of questions received per month, platform used, and consortia information. The survey continued with a series of questions about the collection status of the fields listed in Table 1. There were other open-ended questions asked about some of the fields, such as the location of subject lists, other fields collected but not listed in each category, and other comments.

Table 1: Fields in Survey of Digital Reference Services

Patron Information	**Expert/Responder Information**
Name	Name
E-Mail	E-mail

Telephone	Telephone
City	City
State	State
Country	Country
Grade/Education Level	Title
Professional Role	Institution
Member of organization (library, school, etc.):	Qualifications
Question Information	**Response Information**
Subject (From a List)	Response Text
Subject (Free text supplied by User)	Resources consulted
Text of Question	Date of response
Purpose (e.g. How do you plan to use this information?):	Time of response
Desired form of answer	
Previously consulted sources	
Requested deadline for response	
Date of question	
Time of question	
Routing information (i.e. question referrals)	

Demographics of Respondents

There were 53 responses to the survey, which represented 49 different organizations. Respondents who had different reference services (such as chat and e-mail) that kept different archives in the same organization were asked to fill out a survey for each service. There was little duplication by members of the same consortial group in the survey responses.

Of those services that could be affiliated with an institution, slightly more than half (53%) were from academic libraries. The remaining services were fairly evenly split between public (15%), special and other libraries (17%) and AskA services without a specific library affiliation (14%).

About half (47%) of the responses were from chat-based services, 38% were from Web-based asynchronous services, and the remaining 15% used e-mail or another communication platform for reference. Combining the communication type variable with the service affiliation did show some differences, as can be seen in Table 2. For example, chat was more commonly used in academic libraries, while asynchronous Web-based form was the common method in public libraries and independent services. This would prove an interesting finding to explore on a larger basis to see if it is generalizable and to attempt to shed light on the reasons behind the differences.

Table 2: Type of library versus communication method of reference service

	Chat	**Web form**	**Email / Other**
Academic	54%	30%	17%
Public	29%	71%	0%
Special/Other	50%	50%	0%
Independent	34%	50%	17%

Another question was the average number of transactions per month. Upon examination of this field, it was noted that the answers ranged from 10 to 30,000 (for Tutor.com's Online Classroom). This range of answers is represented in the data in Table 3. In each case, the standard deviation is greater than the mean, which means the data are badly skewed. The median was calculated to give a less biased idea of the central point of the data. The median number of Web-form based questions was 80 per month, and the median number of chat questions was 120 per month. The non-normal nature of this data makes a trustworthy generalization difficult to produce.

Table 3: Mean, standard deviation, and median of reference questions answered each month

	Mean	**Standard Deviation**	**Median**
Chat	1906	6410	120
Web form	164	192	80
E-mail	30	31	18

Another demographic collected was the platform used by the reference service. The results after cleaning the data are in Table 4. The entries for E-mail, Web form + E-mail, and In-house tool may refer to the same type of service – some type of system using existing e-mail and Web servers. If these are combined, then there are three clear popular choices – Question Point, Tutor.com, and some type of in-house use of existing resources.

Table 4: Percentage of respondents using each reference tool

Platform / Software	**Percentage of Respondents**
(E-mail, Web form, or In-House tool combined)	*27%*
Question Point	23%
Tutor.com	21%
E-mail	13%
24/7	8%
Web form + E-Mail	8%
In-house tool	6%
Altarama RefTracker	4%
QABuilder 2.0	4%
Docutek VRL Plus	2%
eAssist NetAgent	2%

ExpertCity's Desktopstreaming	2%
LivePerson (HumanClick)	2%
Open Ask A Question	2%
PHP Live Support	2%

Exploration of Communication Forms

Much of the upcoming analysis is split on the distinctions of communication form used, as the types of fields collected in chat may be different than the fields collected via a Web form and via e-mail. The eventual goal is to create one schema that will serve all of these communication platforms.

A series of questions on the survey sought information about the communication practices of different service types. For example, all surveyed e-mail and Web form-based services e-mailed a copy of the answer or transaction to the patron; however, only 72% of the chat-based services regularly sent a copy of the transaction to the user.

A similar set of questions explored through which format questions are eventually resolved. These results, in Table 2, show that there is not much crossover between formats. Chat reference is resolved in chat about 80% of the time, and Web form questions are resolved via Web forms or e-mail most of the time. The high percent of other forms of answers that started as chat reference is probably because the synchronous connection has already been made, and it is then convenient to complete the transaction via the phone.

Table 5: Formats of Final Resolution of Reference Transactions

Incoming Question Format	E-mail answer	Web form answer	Chat answer	Other form (telephone, visit)
E-mail	98%	0%	1%	1%
Web form	23%	74%	0%	3%
Chat	7%	3%	80%	10%

Fields Collected by Services

In order to understand what information is being collected by services, the analysis is presented in two parts. First, the fields currently collected by services are presented. Following that, the discussion turns to the data that informs the rest of this schema: what fields are services either currently collecting or willing to collect?

Table 6 lists the fields, sorted by category and overall usage, of what services currently collected during the reference process. Looking at the overall results, the most common set of fields currently collected about a reference transaction are: patron e-mail and name; question text, date, and time; and the response text, date, and time. This aggregate set of fields disguises patterns that appear when the results are broken out by communication method used.

Since the two common communication methods are Web form and chat, they will be examined individually.   Chat services tend to be more freeform, and therefore may not explicitly collect many fields. Some services as the user to set up an account before the chat session; this will result in more information about the patron, but not more information about the specific information need behind a reference transaction. Even though chat services tended to collect less information than average, many still collect the patron name and e-mail; question text, date, time, and referral/routing information; and the response text, date, and time. One field of note here is the above-average collection of referral/routing information. Many chat services reported capturing fields like IP address, which was the most common information put into the "Other" open ended survey questions. In addition, as seen earlier, chat sessions end in a different communication channel 20% of the time; they therefore have a stronger need to capture this type of transferal information.

The group of Web form reference services captured more information on average than other types of services; this is not surprising as the process of asking a question via a Web-based form is more structured than asking the same question via e-mail or chat. The most common fields currently collected via Web form-based asynchronous reference are: patron e-mail, name, country, and state; question text, date, and time; response text, date, time, and responses collected. Since the information is collected in small fielded pieces, it is then easier to keep in those pieces in a data warehouse. It is because of this that DREW will start by aggregating Web form-based services, and then move to more free-form services as the warehouse develops.

One interesting pattern is the lack of information collected about the person answering the question during the process. There are two types of individuals who answer questions – those who are trained to do research and answer a question from existing resources (such as librarians) and those who are able to answer questions in a specific topic area because they are trained experts in that area. Librarians are trained to provide citation information, and document the authoritativeness of an answer through the support of external works. Experts, on the other hand, provide the authority for their answer based upon their credentials. If services do not keep information about the person who answered the question, then the authority behind an expert-answered question disappears. Because of this, it is important to encourage experts who are answering questions to supply references to works that would contain the answer to the question, even when they know the answer without looking anything up. As these experts may not have been trained as librarians, the administrator of the system needs to ensure that training is available in the basics of created a response that will have supported authority with no identity of the answerer.

Table 6: Percentage of services currently collecting specified fields

	Overall	**Web form**	**Chat**	**E-mail/Other**
Patron Information				
E-mail	77%	90%	68%	67%
Name	72%	80%	68%	50%
Country	36%	65%	20%	0%

State	34%	55%	24%	0%
Member of Organization	34%	35%	32%	17%
City	32%	55%	20%	17%
Educational level	30%	40%	28%	0%
Phone number	23%	25%	16%	17%
Professional Role	23%	30%	16%	0%
Question Information				
Text of question	93%	100%	88%	83%
Date	91%	95%	92%	67%
Time	85%	85%	92%	50%
Routing/Referral information	45%	30%	60%	17%
Subject (free-text)	43%	35%	44%	83%
Deadline for answer	17%	30%	4%	17%
Desired form of Answer	11%	10%	8%	17%
Purpose	9%	20%	4%	0%
Previously consulted resources	9%	10%	8%	0%
Subject (from a list)	8%	10%	8%	0%
Responder Information				
Name	53%	50%	60%	33%
E-mail	45%	35%	52%	50%
Institution	45%	45%	52%	0%
State	34%	40%	32%	0%
Country	32%	40%	28%	0%
City	28%	35%	28%	0%
Title	25%	30%	24%	0%
Telephone	17%	20%	16%	0%
Qualifications	17%	20%	16%	17%
Response information				
Date	93%	90%	96%	83%
Text of response	89%	95%	88%	67%
Time	87%	80%	96%	67%
Resources consulted	51%	65%	40%	33%

Fields that Services are Willing to Collect

Another way of looking at the data is to explore which fields services either collect now or are willing to collect in the future. The data was recalculated using this new model, and the results are in Table 7. This is important in aiding the development of the DREW schema. While services may not be currently collecting information, they may be more willing to collect the information if they perceive that the data will be useful in improving their service and the understanding of the field.

Table 7: Percentage of services currently collecting or willing to collect specified fields

Field	**Overall**	**Web form**	**Chat**	**E-mail / Other**
Patron Information				
E-mail	83%	90%	80%	67%
Name	79%	85%	76%	67%
State	70%	75%	64%	67%
Member of Organization	70%	65%	72%	67%
City	68%	75%	60%	83%
Country	66%	80%	52%	67%
Phone number	59%	65%	52%	50%
Educational level	59%	55%	64%	50%
Professional Role	49%	45%	48%	50%
Question Information				
Text of question	100%	100%	100%	100%
Date	100%	100%	100%	100%
Time	94%	90%	100%	83%
Routing/Referral information	83%	75%	92%	67%
Subject (free-text)	76%	60%	80%	100%
Deadline for answer	72%	75%	64%	83%
Previously consulted resources	70%	75%	64%	67%
Desired form of Answer	59%	60%	52%	67%
Subject (from a list)	51%	45%	52%	50%
Purpose	51%	55%	48%	50%
Responder Information				
Name	79%	75%	88%	50%
Institution	79%	70%	92%	50%
E-mail	70%	60%	80%	50%
State	64%	55%	72%	50%
Country	62%	55%	68%	50%
Title	62%	55%	72%	50%
City	59%	50%	68%	50%
Qualifications	53%	45%	60%	50%
Telephone	51%	45%	60%	33%
Response information				
Text of response	98%	100%	100%	83%
Date	98%	95%	100%	100%
Time	94%	90%	100%	83%
Resources consulted	77%	80%	76%	67%

Looking at the Overall column, one can see that services are willing to collect much more information than they currently collect. One obstacle is the fact that patrons are less likely to ask a question if they have to fill out more fields. The patron and expert information need be collected only once, then matched to each question through a logon process. The question and response information would need to be gathered every time.

In order to develop the proposed DREW schema, we will now explore each area of the survey and discuss the usefulness of the fields to research needs. There are two types of research needs that are important: the needs of administrators in understanding their own digital reference system, and the needs of researchers in looking at the larger-scale picture.

Transaction Information

One of the challenges of DREW is that it will hold different forms of intermediation. The goal is to collect questions from all types of digital reference services – Chat, E-mail, Form-based, etc. Therefore, at the center of the DREW record will be the information from the transaction. For a chat transaction, the body of the chat will be included. In an e-mail transaction where there was little restriction on the information in the e-mail, the e-mail text will be included. If a Webform was used to collect fielded information, then the Question and Response will be divided and included. There will also be a field to identify the type of transactional data in the record.

Using this structure will make it difficult for some researchers to explore relationships between questions and responses. A priority for researchers is to develop algorithms that will divide the large textual chat and e-mail transcripts into separate questions and answers.

Patron Information

Even though services are willing to collect considerable patron information, little of this is actually needed in understanding the question-answering process. In fact, it is important to mask personally identifiable information about the patrons. Therefore, most of the patron information will not be part of the DREW schema. There are a few useful fields about the patron that more than half of the services would be willing to collect. Information about the location of the patron, such as *Country* is important, especially as different countries have different laws about intellectual property. QuestionPoint has faced many of these problems, and it is expected that as DREW grows, international intellectual property issues may arise (P. Rumbaugh, personal communication, July 6, 2004). One of the common fields that was a write-in was *Zip Code*; this field combines city and state information and can be used to map DREW to a demographic database, but does not intrude upon the personally identifiable information about the patron.

Another area of interest is the patron's organizational membership or educational level. As different services cater to different age and educational levels, it would be useful to have some basic knowledge about the patron. An important distinction for DREW is the intended age level attached to a question, which might be different than the level of the patron asking a question. For example, questions asked by another for a child would need to be identified as a child-level question. For this field, services will have to map their own data collected about their questions to an *Educational Level* field, which would have the broad choices of:

- Child (elementary school, primary school),
- Pre-Teen (middle school, junior high),
- Teen (high school),
- College (undergraduate),
- Adult, or
- Unknown.

Individual services will have to use their best judgment in mapping their own fields to these choices.

One of the products of DREW will be customized reports for their own service. In order to aid in this process, there will also be a *Custom Patron Type* field, which will allow a service to enter a different classification with local meaning for their service.

Question Information

It is more important to collect information about the question than information about the patron, as seen with the Educational Level field above. Fields such as *Date*, *Time* and *Previously Consulted Sources* are all potentially useful. Some type of *Free-Text Subject* and *Category* information is also useful, and one of the areas of research is to attempt to automatically map this to a common list. Services are willing to share *Referral Information*; the key information for DREW is if the question was:

- Internal (answered in the same service where it was asked),
- External - Sent (sent out to a different service to be answered), or
- External - Received (a question received from a different service).

In addition, there will be a *Referral Service* field, where the original service can indicate the name of the service involved in the referral. This data will be useful in understanding patterns of referral between services.

Just as before, in order to aid services in their own reporting, there will be a *Custom Question Type* where services may add an internally useful categorical variable.

Responder Information

As before, services are the least willing to collect information about the person answering the question. QuestionPoint actively removes this information (P. Rumbaugh, personal communication, July 6, 2004), and maintaining no information about the expert will protect the privacy of the individual. Given the issue discussed earlier, one field about the expert would be useful: *Responder Role,* with the choices of

- Subject Expert (someone answering the question because they are an expert in a topic area),
- Librarian / Researcher (someone answering the question because they know how to find information in resources),
- Unknown / Other.

Another field involved with the responder identify is the *Service Name* field. This will be useful in conjunction with the referral information above, as well as in creating individual reports for participating services. This field will undergo authority control as participants are added to DREW; eventually, this same authority file will be used for the *Referral Service* field.

Again, there will be a single *Custom Responder Type* that can be used by an individual service for categorical data to aid in reporting.

Response Information

All four fields listed on the survey are useful for research and many services are willing to collect them; therefore *Response Date, Response Time,* and *Response Resources* are all part of this proposed schema.

If available, the field *Response Type* will be added with the following choices (based on the NETREF standard):

- Answer (where the response answers the question),
- Clarification (where the response is a request for clarification),
- Out of Scope (where the question was not answered), or
- Other (Thank yous and other types of transactions).

Finally, there will be a *Custom Response Type* available for services to use for categorical data.

Observations

While those doing chat reference currently collect the least amount of information, they were the most willing to collect additional information for this research. Conversely, the Web form services were less willing to collect additional fields. There are several hypotheses as to the reason behind this finding. Administrators who filled out this survey for a chat service may be frustrated by the lack of data currently collected about a chat reference service, and thus are willing to collect more information if there is an opportunity. Conversely, those running Web form services may have noted that collecting more fields results in fewer questions. In addition, as a Web form-based service requires much more planning to develop fields that are collected, administrators of these services may be less willing to make changes. Further research is needed to explore these hypotheses.

Summary of Fields

Based upon this research, the types of information going into the DREW archival schema for digital reference transactions includes:

- Service Name
- Question Educational Level
- Patron Zip Code
- Patron Country
- Question Free-Text Subject
- Question Category
- Question Date

- Question Time (standardized)
- Previously Consulted Sources
- Question Referral
- Referral Service
- Responder Role
- Response Date
- Response Time (standardized)
- Response Resources
- Response Type
- Transaction Text
    - Question Text and Response Text/Response Type, or
    - Transaction Text (in the case of E-mail), or
    - Chat Transcript
- Transaction Type
- Custom Patron Type
- Custom Question Type
- Custom Response Type

These elements will form the “tagging system” mentioned earlier in the discussion of complexity theory.

Current Challenges

The survey provides a starting point for exploration, by providing the fields that will define the service. There are three current research challenges for this project: NISO standards and threading, subject list authority, and privacy.

NISO Standards and Threading

One goal of this process is to create a schema for archiving that is compatible with the networked reference services protocol NISO AZ, a.k.a. NetRef [10]. In its current configuration, this standard is designed to assist with the operational needs of passing questions from one service to another.

As with most data warehousing applications, the data kept for archiving is usually in a different form than the data used in the operation of the system. In addition, the timing of the application of the standards is important; NetRef is applied when the question is passed to another service, while the DREW schema is applied to the transaction after it is completed.

It is important, therefore, that the archival schema be compatible with the operational standard. This is critical in improving participation with the DREW program; making a data warehouse structure compatible with the NISO standard will make it easy for services using the NISO standard to supply transactions for the warehouse.

One significant issue in the transition from the operational standard to archival form is the de-threading and cleaning of a reference transaction to extract the important components of the transaction needed for data warehousing. The structure of the data warehouse will be based upon the key elements of the transaction – question and initial response. If the thread continues, that will be separated into a second record that links back to the initial thread. There are several possibilities of what this second transaction could be – a new or follow-up question from the patron or a request for more information from the expert. As the data warehouse grows, one line of exploration will be to attempt to automatically classify transactions; this will prove useful in creating cleaner search mechanisms and automating reference processes.

The NetRef threading issues are a harbinger of the problems to come in attempting to incorporate chat reference into this type of knowledge structure. In chat reference, there is not usually a clearly defined question and answer; rather, these two parts of the transaction may be presented throughout the interchanges. One intriguing line of research is to use natural language processing to automatically extract “the question” and “the answer” from a chat transcript. A more realistic solution would be to take the chat platform and build in markup tools so that a librarian answering a question could quickly mark key phrases and components of an interchange for later cleaning and archiving. In addition, if the important parts are marked and the full text is available, it then becomes much easier to train systems using machine learning techniques to successfully pick out the key parts of the conversation.

Subject List Interoperability

One of the current challenges in crossing the boundaries between digital reference services, as well as other knowledge management systems, is that of subject assignments. Most services assign a subject term to a question at some point in the process: the user may assign a subject when the question is asked, the administrator may select a subject explicitly through a field or implicitly through expert assignment, or the expert may assign a topic during the answering process. Many times, these subjects come from a list that is unique to that service.

Different approaches to this problem of creating subject lists for multi-source knowledge bases were reviewed by Zeng and Chan [12] in their review of interoperability between knowledge organization systems. In order to select a method for subject list interoperability, the key factors of this particular setting must be enumerated. The individual digital reference services will either be a general service or a subject-specific service. The subject-specific services will have a specific subject list, and it is important to maintain that specificity so that subject-specific services on similar subjects can take advantage of that detail. However, the question classification term list used by a subject-specific service could be rolled up to a higher-level term that would be appropriate for the general service.

Therefore, it is important to maintain the original selection and subject list established by the reference service to aid that service in management and reporting and to help that service work with similar services. From a knowledge-base perspective, however, it is important to map these varying subject lists to a common list to aid in interoperability.

Returning to the various approaches presented by Zeng and Chan (2004), there are several possibilities. The first is called a satellite thesaurus, which starts with a superstructure thesaurus that would be appropriate for a general reference service. Then, where specialized thesauri are available, they are attached to a node of the general superstructure. This allows the maintenance of the individual specialized subject lists while maintaining some relationship between them.

Another approach is direct mapping, where terms from different vocabularies are mapped to each other. This is then built into the system, and whenever a search is performed on one term, it is mapped to the other terms. This does require more time to plan, but would make it easier for similar services with different subject lists to come together into one knowledge base. The danger with this effort comes with the general services, as it would prove challenging to map all general reference service thesauri to each other.

A third approach is switching, where all individual subject lists are mapped to an intermediary subject list. This is similar to direct mapping, except that everything is mapped to one list instead of trying to map all lists to each other. This is currently the approach used by several large multi-disciplinary knowledge base projects, such as HILT (High-Level Thesaurus Project) [13] and National Library of Medicine's Metathesaurus[14].

The HILT project is an intriguing one for the DREW project. Over the last few years, researchers funded by the Joint Information Systems Committee in the United Kingdom have been creating a thesaurus to link resources from different information systems. They have based their work on the Dewey system, and this thesaurus is available at http://hiltpilot.cdlr.strath.ac.uk/pilot/top.php. If the DREW project uses this thesaurus as the base for the switching approach, where other services map to this general thesaurus, it will serve several purposes. First, the thesaurus will be the result of research and testing on multiple systems, so is stable and accepted. Second, it will raise the possibility of interoperability between DREW and other information services using the HILT subject list. Therefore, we are investigating the feasibility of using the HILT thesaurus as the DREW master subject list.

The implementation would be that individual services would work with DREW to develop an appropriate mapping to the HILT subject list. In addition, the original subject terms would be captured in the data warehouse. As the project grows, there may be the need to create secondary, more specific, metathesauri to allow the mapping between different services focusing on the same topic area.

Eventually, this mapping will take place either as part of the data cleaning process through mapping algorithms developed between DREW and each institution or it will occur with the host institution mapping their subject headings to the shared thesaurus before submitting the transactions to DREW. It is expected that the number of services participating in the warehouse will be small enough that mapping programs could be created at the start of the integration of results from a new service with the aid of that reference service. An important consideration with this warehouse-side mapping is that if a service changes their subject list it is updated in the warehouse; however, this would not prove a challenge through automated notification when DREW receives a new, unmapped, subject.

Privacy

One of the constant concerns about library data is that of patron privacy. The library has traditionally been a safe place for users to gather information. Legislation such as the USA PATRIOT act threatens the privacy of patron histories, as it gives government bodies the right to access patron records without the patron knowing they are being watched through a roving wiretap [15]. In response to this, some libraries are actively deleting and shredding records [16]. As digital reference services typically collect an e-mail address for a patron, it is possible that they also can be targets for a roving wiretap. If the archives of the service contain personally identifiable information about a patron, then the service would be required to turn over transactions if requested by the appropriate authorities.

In this case, the archival schema for DREW provides a method of protecting the personally identifiable information about a patron while still maintaining the useful information included in the transaction. In addition, the information needed to make administrative decisions is kept. Therefore, the data warehouse balances the need to protect the patron and the need to maintain a data-based history of the service's activities.

This type of data warehouse is typically used in bibliomining (data mining for libraries) to support decision-making across the library. However, there are some challenges in digital reference transactions that do not occur in other types of library transactions. Since patrons ask a free-text question or have a flowing discussion, it is possible that patrons include personal information within the text of their question. There are currently no automated solutions to strip out the personally identifiable information from a reference transaction.

This is similar to the problems of deidentificaiton of medical records [17] where personal information is removed while the useful information from the records are maintained. An active research area in natural language processing is the automated identification and replacement of this personal information in medical records. As this research agenda is advanced and solutions are created, we will adapt these medical informatics tools to reference transactions.

Safe Harbor Policy Compliance

One of the goals of DREW is to involve other countries; therefore there are international privacy guidelines to which DREW will adhere. These guidelines were originally created by the European Union, and have been adopted by the United States. This policy, known as the Safe Harbor Privacy Principles, is made up of seven areas that ensure those individuals whose data is in the data warehouse are properly protected. These areas form the basis of the DREW privacy policy:

- Notice – Each service participating in the DREW project will add to its existing privacy policy a statement about DREW, the subset of transaction information transferred to DREW, what the data is used for, who is using the data, and how they can opt-out of the project.

- Choice – Users of digital reference service, including both patrons and experts, have the ability to request that their information be removed from the data warehouse. Due to the anonymous nature of DREW, this request will be initiated at the service where the question was asked, and the service will pass along the record ID to be removed from the warehouse.
- Onward Transfer – In order to comply with this area of Safe Harbor, the digital reference services participating in the DREW warehouse must comply with the Notice and Choice clauses of this policy. This means that each service will notify their users about the DREW project and allow their users to be able to remove information from the warehouse. Any additional participants, such as library science researchers, will also have to verify that they offer a level of protection equivalent to that offered through this policy.
- Access – The users and experts involved can request to see their DREW records through the service that submitted the question to DREW. After seeing these records, they can request to have them adjusted or removed from the archive.
- Security – Access to records in the DREW warehouse will be controlled through password-protection and firewalls. Researchers working on topics related to the reference process may request data from DREW. Participating libraries will be able to receive their own transactions, as well as reports generated using the data from their transactions. As the DREW project grows, participating institutions will be made aware of the change in advance and be allowed to remove their transactions at any time.
- Data Integrity –There is no personal information kept in DREW. If mistakes were made in transmittal, the submitting service can correct the DREW records. In addition, if the information in a transaction is incorrect, DREW participants can submit annotations to be added to a transaction.
- Enforcement – The DREW advisory board will serve as an external body to ensure that DREW is complying with the Safe Harbor Policy. If needed, the DREW advisory board may contact an external group from another organization such as the American Library Association to investigate privacy concerns.

The Usefulness of DREW

This warehouse of digital reference transactions will allow a level of understanding about library services previously unavailable for researchers and educators. In addition, administrators of participating services will gain access to customized reporting and management information tools as they are developed.

Support of Current Teaching and Research

There are a number of lines of human intermediation research that would be advanced through the availability of DREW records. One of the challenges for digital reference researchers is getting access to large amounts of cleaned data; DREW will provide a robust source of transactions for these researchers. Those seeking to understand information seeking behavior or how experts use resources in answering questions would be able to rapidly improve the generalizability of their models through access to data on this scale.

Another line of research that would be benefited by this data warehouse is the measurement and evaluation of digital library services. Tools such as bibliomining, or data mining for libraries, require large amounts of cleaned data (Nicholson, 2003). DREW is an ideal place for bibliomining research, and the results will allow the development of new measurement tools for digital reference services and the discovery of novel and actionable patterns existing in the transactions. One goal of this line of research is to create a Management Information System that can be applied to the entire database for research purposes, and that participating libraries can access to learn more about their own services.

Informing Service Management and Decision-Making

One of the challenges facing individual services is the need for informed management decisions. This call is embodied in Evidence-Based Librarianship, which implores librarians to use the best available evidence when making decisions for their library. In addition, librarians are asked to justify their services on a regular basis; many are too busy running their service to step back and create the tools needed to analyze their services appropriately.

As researchers develop methods of measurement and evaluating digital reference services, these tools and models can be integrated into DREW. As these tools are created, managers of individual services can request any of the reports created for the entire warehouse to be run on just the data from their own system. This creates a significant reason for services to participate in the DREW project, as they will then have access to a strong management information system associated with DREW.

Digital reference consortia will also be benefited by this relationship, as they can get the same reports and information about their entire consortia. This type of information was previously challenging to discover, but is essential to strong decision-making. As consortia make decisions that can have long range impact and may not be able to change those decisions easily, it is important that these decisions be powered by the best evidence available.

Modeling the Complex Digital Reference Landscape

One area of research stemming from the use of Complexity Theory is modeling the digital reference transactions within DREW as a Complex Adaptive System. Once the digital reference transactions have been cleaned an inductive system of clustering can be utilized to examine the self-organizing nature of digital reference knowledge bases. Each transaction will be modeled as an autonomous agent with a set of attributes (the proposed DREW element set). Some of the attributes are static (such as the text of the transaction), but some are dynamic (such as the time since the transaction was closed, or the number of times the agent is referred to by other transactions). By placing these transactions in an n-dimensional space (2 or 3 dimensions for visualizing the space for example), pair-wise comparisons between the agents can be conducted (in essence determining how similar any two agents are). Agents will move “closer” or “farther” apart based upon these comparisons. It is anticipated that these agents will inductively cluster. It is also hypothesized that these clusters will change over time as not only the dynamic attributes change (a transaction ages for example), but the agents themselves change (with new questions, or new references are added).

Creating a Infrastructure for Virtual Collaboration

One of the exciting possibilities of a DREW schema is that it empowers the infrastructure to allow for virtual collaboration between researchers and practitioners. Services will start by providing records for DREW. Researchers will then use these records to develop tools across different services. These researchers will then be encouraged to prepare their models and tools using the DREW schema so that the services participating in DREW can apply these research results to their own services. Practitioners can immediately benefit from research and will be encouraged to not only continue their involvement in DREW but also to improve their management of the digital reference service. Researchers can then test the difference these new tools and models make on reference service, and the cycle continues.

This model is currently in use in the open source community. As infrastructure and data schema are created and programmers use this to develop tools. As tools are created and release, other programmers improve on the code and the result is that the users have a much better experience. This virtual collaboration will allow digital reference to rapidly improve as a service.

Conclusion: The DREW Research Agenda

The process of creating this digital reference archive introduces a set of questions that power a research agenda. Each of these questions stem from a challenge (a.k.a. opportunity) in the process of creating, implementing, and using this warehouse of digital reference transactions. Some of these issues have been previously addressed in this paper.

- What would an archival schema for digital reference transactions look like? Will one schema work for all communication mechanism used of digital reference? What minimum subset of these fields is needed to be useful?
- What tools are needed to extract these fields from digital reference transactions? How complete of an archival record can be automatically recreated from a chat or e-mail reference transaction?
- Is there a thesaurus that would be useful in linking subject lists from different services? Can this assignment of subject headings be done inductively?
- What subset of fields will maintain the information needed for research and discovery while still protecting the privacy of patrons? What policies are needed to balance keeping a data-based history of the service with the need to protect personal information of patrons?
- How can the information space within DREW be explored through bibliomining and visualization tools? What patterns can be discovered about the process of answering questions? Can the changing space of reference transactions be demonstrated through animated visualizations?

- What is the life of a reference transaction? Are there facets that can be used to predict how long a question will be useful in a question archive? What indicators can be used to detect questions that have outdated information?
- How can digital reference be rapidly improved through the virtual collaboration of researchers and practitioners? What management tools are most effective in helping digital reference services improve? What measurable differences do these tools make?

Through reference authoring via human intermediation, libraries have the ability to produce large amounts of high-quality information. In order to understand this information and create tools that allow for the rapid creation of knowledge bases, as well as advance our conceptual understanding of the changing face of reference, researchers need a cleaned collection of transactions from a wide variety of services. The DREW project will supply researchers with this data source, as well as making it possible for participating services to quickly benefit from the results of the research.

Bibliography

1. R. David Lankes. "The Digital Reference Research Agenda." *Journal of the American Society for Information Science and Technology : JASIST* 55, no. 4 (2004): 301-12.
2. Karen M. Drabenstott. "Classification to the Rescue--Handling the Problems of Too Many and Too Few Retrievals." In *International Isko-Conference (4th :1996 :Washington, D.C.). Knowledge Organization and Change Indeks Verlag*. (West Germany, 1996), 107-18.
3. Chris Sherman, and Gary Price. *The Invisible Web : Uncovering Information Sources Search Engines Can't See*. (Medford, N.J.: CyberAge Books, 2001).
4. Michael Bergman. "The Deep Web: Surfacing Hidden Value." *Journal of Electronic Publishing* 7, no. 1 (2001): Accessed June 15, 2004, http://www.press.umich.edu/jep/07-01/bergman.html.
5. Danny Sullivan. 2002. Intro to Search Engine Optimization. Accessed June 15, 2004, http://searchenginewatch.com/webmasters/article.php/2167921.
6. OCLC. 2004. Knowledge Base. Accessed June 15, 2004, http://www.oclc.org/questionpoint/libraries/knowledge/default.htm.
7. M. M Waldrop. *Complexity: The Emerging Science at the Edge of Order and Chaos*. (New York: Touchstone, 1992).
8. J.H. Holland. *Hidden Order: How Adaptation Builds Complexity*. (New York: Addison Wesley, 1995).
9. R. David Lankes. *Building & Maintaining Internet Information Services: K-12 Digital Reference Services*. (Syracuse, NY: ERIC Clearinghouse on Information & Technology, 1998).
10. Library of Congress. 2004. Netref: Niso Committee Az: Networked Reference Services. Accessedd May 27, 2004, http://www.loc.gov/standards/netref/.
11. Joseph Janes. "Question Negotiation in an Electronic Age." In *The Digital Reference Research Agenda*, edited by R. David Lankes, Scott Nicholson and Abby

Goodrum. (Chicago, IL: Association of College and Research Libraries, 2003), 48-60.

12. Marcia Lei Zeng, and Lois Mai Chan. "Trends and Issues in Establishing Interoperability among Knowledge Organization Systems." *Journal of the American Society for Information Science and Technology : JASIST* 55, no. 5 (2004): 377-96.
13. Dennis Nicholson, Ali Shiri, and Emma McCulloch. 2004. Hilt: High-Level Thesaurus Project Phase Ii. Accessed June 10, 2004 , http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport/0HILT2FinalReport.doc..
14. National Library of Medicine. 2004. Umls Metathesaurus. Accessed Jun 10, 2004, http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.htm.
15. American Library Assocation. 2004. The USA Patriot Act & Libraries. Accessed June 17, 2004, http://www.ala.org/ala/washoff/WOissues/civilliberties/theusapatriotact/usapatriotact.htm.
16. Scott Nicholson. "Avoiding the Great Data-Wipe of Ought-Three." *American Libraries* 34, no. 9 (2003): 36.
17. Workgroup for Electronic Data Interchange. 2003. De-Identification and Limited Data Set White Paper. Accessed June 7, 2004, http://www.hipaadvisory.com/action/WEDIpapers/Deid.pdf.