Beyond Information Searching and Browsing: Acquiring Knowledge from Digital Libraries

Ling Feng *Manfred A. Jeusfeld* *Jeroen Hoppenbrouwers*
Infolab, Tilburg University, PO Box 90153, 5000 LE Tilburg
The Netherlands, {ling, jeusfeld, hoppie}@kub.nl

Abstract

As one of the most complex and advanced forms of Internet information systems, digital libraries serve as an increasingly important channel to a vast array of information sources and services. However, from the standpoint of satisfying human's information needs, the current digital library systems suffer from the following two shortcomings: (i) inadequate strategic level cognition support; (ii) inadequate knowledge sharing facilities. In this paper, we introduce a two-layered digital library architecture to support different levels of human cognitive acts. The model moves beyond simple information searching and browsing across multiple repositories, to inquiry of knowledge. To address users' high-order cognitive requests, we propose an *information space*, consisting of a *knowledge subspace* and a *document subspace*. A formal description of the knowledge subspace for knowledge sharing and dissemination, as well as mechanisms for constructing the two subspaces, are particularly discussed. Such an enhanced information space extends the traditional role of digital libraries as *information provider* to *information* & *knowledge provider*. Some distinguished features in comparison with the traditional knowledge-based systems are also discussed in the paper.

1 Introduction

With the exponential growth of information in the Web, more and more people nowadays demand effective search and indexing functionalities. Digital Libraries (DLs) are a form of information technology which provides new opportunities to assemble, organize and access large volumes of information from multiple repositories, while making distributed heterogeneous resources spread across the network appear to be a single uniform federated source [29]. Under the assistance of such an information-rich system, users can move from source to source, seeking and linking information automatically or semi-automatically. From a user's perspective, DLs establish an underlying infrastructure for a bulk of digital information and information services associated with users' information acts.

Traditionally, when people retrieve information, their activities are classified into two broad categories: *searching* and *browsing*. Searching implies that the user knows exactly what to look for, while browsing should assist users navigating among correlated searchable terms to look for something new or interesting. So far, most of the major work on DL systems fall into these two categories.

However, with more digital libraries built up to handle users' information needs, we ask questions: *Are the existing information technologies powerful enough to facilitate DL users' problem solving? Or what technologies do we still need to help users better do their work? Compared to traditional physical libraries,*

are there any important features missing from the current DL systems? To answer these questions, let's first look at a scenario on the use of a DL system.

*Kooper works in a city plan & management office. He is investigating the precaution policy against flood this year. If there will be a flood in the coming summer, necessary actions (e.g., strengthing the embankment, resource allocation, etc.) must be taken now to prevent the city from sustaining losses. According to Kooper's previous experience, it seems that "*A wet winter might cause flood in summer.*" To confirm this pre-conceived hypothesis, he logins on a DL system to request information talking about the reasons of flood. The DL system returns a number of articles. He browses through the returned article list and selects three articles that look most relevant. All of the three articles mention that "*A precursor to the flooding in summer is a wet winter.*" To assure this information is also valid for the area where Kooper lives, he accesses the DL again to ask for documents reporting river flood in the city before. From the articles returned, he gets to know the latest 3 severe flood that happened in 1986, 1995, 1997, respectively, in the city. He then navigates to the meteorological repository of the DL, and accesses the weather information of the city during the winter time of these three years. He notes a tight correlation between* wet winter *and* flood summer. *Based on the information obtained from the DL, Kooper is pretty sure now about his prior flood prediction assumption for the city. He logs off the system. Having experienced a very wet winter this year, Kooper decides to draw up a city flood precaution plan immediately.*

From this scenario, we observe that a number of deficiencies exist in current DL systems. Below we outline some of the problems and highlight our work to solve them.

Problem 1 - Inadequate Strategic Level Cognition Support. The aim of DL systems is to empower users so that they can find useful information to solve their own problems. When a person goes into a library to look for something, he/she usually has certain purposes in mind. For example, he/she may want to read a specific article written by a certain author. In this situation, the target of the user is precise and clear. Sometimes, the user wishes to explore the available resources first before exploiting them. This exploration may be targeted at refining a prior understanding of a certain information context, or formulating a further concrete requirement for specific documents. Most efforts of the current DL systems aim to support these two kinds of users' behavior, namely, *searching* and *browsing*.

However, besides simply entertaining users with documents as what searching and browsing do, DL systems should also consider supporting human's strategic level cognitive work which can directly enable correct actions and problem solving. Typically, users have some prior domain-specific knowledge or pre-conceived hypotheses. They may expect the library systems to confirm/deny their existing concepts, or to check whether there are some exceptional/contradictional data source evidences against the pre-existing notions, or to provide some predictive information so that users can take effective actions. Under this circumstance, the users' information needs are not only for relevant documents, but also for intelligent answers to their questions together with a series of supporting literatures for justification and explanation, as illustrated in the above scenario. In order to distinguish this kind of information needs from the traditional searching and browsing requirements, we categorize traditional users' information searching and browsing activities into *tactical level cognition act* and the latter into *strategic level cognition act*.

As [5] says, "*The nature of an information system is to provide informational support to people as they carry out their intentional tasks.*" As one of the most complex and advanced forms of information systems, DL system designers must have a comprehensive understanding of users' information needs and their purposes

for using DLs. In the above example, if the DL system could automate user's rational exploration from the knowledge space, consisting of propositions and assertions, to the corresponding justification/explanation space, consisting of data sources in various forms like textual articles, reports, databases, etc., the information role of the DLs in helping users derive effective decisions will be greatly enforced. Unfortunately, such strategic level complex cognition support and its impact on users' work have constantly been ignored by current DL systems.

Our Work. We propose a two-layered DL function model to support both tactical level and strategic level cognitive tasks of users according to their information needs and purposes in section 2. The model moves beyond simple searching and browsing across multiple correlated repositories, to acquisition of knowledge. This knowledge subspace consisting of hypotheses, together with the associated document subspace for justifications, are particularly described in the section.

Problem 2 - Inadequate knowledge sharing facilities. Traditional libraries are a public place where a large extent of mutual learning, knowledge sharing and exchange can happen. A user may ask a librarian for search assistance. Librarians themselves may collaborate in the process of managing, organizing and disseminating information, or share experiences in using consistently-emerging new systems and tools to tackle users' search questions. Users may communicate and learn from each other by observing how others use library's resources, or by asking for help. With paper sources digitalized and physical libraries moving to virtual DLs, these valuable features of traditional libraries should be retained. We believe DLs of the future should not just be simple storage and archival systems. To be successful, DLs should become a knowledge place where knowledge acquisition, sharing and propagation take place. For example, if the DL in the above scenario could make readily available expertise and answers to strategic level cognition questions, which might require time-consuming search or consultation with experts, it can help users to better exploit the DLs and improve working effectiveness. Also, as computerized knowledge does not deteriorate with time as that human knowledge does, for long-term retention, DLs offer ideal repositories of the knowledge in the world, and make them universally accessible.

Our Work. In section 3, we provide a formal description of the DL's knowledge subspace. Two methods for the construction of DL's knowledge and document subspaces are proposed in section 4. One of them is based on the manual input from experienced human users. Another approach addresses (semi-)automatic knowledge acquisition by correlating and analyzing data sources from multiple repositories in DLs. The distinguishes between such an enhanced DL system with knowledge elements and traditional knowledge-based information systems are discussed in section 6.

2 A Two-Layered DL Cognitive Function Model

Supporting users' information searching and browsing has been the focus of DL research for a long time. However, as social humans, their information expectations for a DL are more than pure searching and browsing. In this section, we extend the traditional role of DLs from *information provider* to *information* & *knowledge provider*. Figure 1 illustrates the function model of DLs in response to various information requests, which are categorized into *tactical level cognition act* and *strategic level cognition act* in the model.

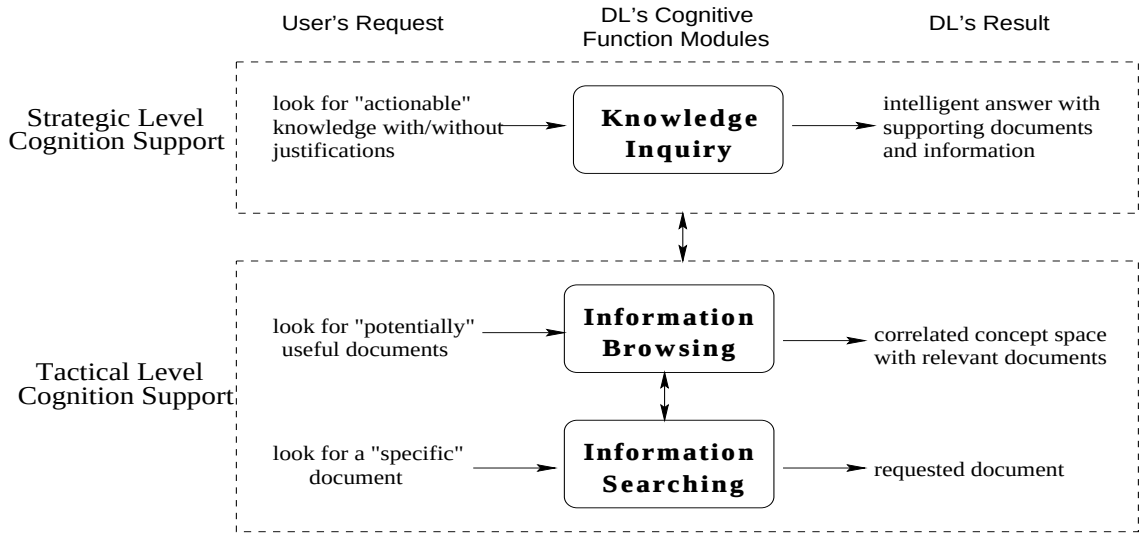

Figure 1: A two-layered DL model

2.1 Tactical Level Cognition Support

We view traditional DL searching and browsing as tactical level cognitive acts.

Searching. The target of searching is towards certain specific documents. One searching example is "*Look for the article written by John Brown in the proceedings of VLDB'88*." As the user's request can be precisely stated beforehand, identifying the target repository where the requested document is located is relatively easy. Primarily, the ability to search indexes of repositories can support this kind of searching activities.

Browsing. Different from searching whose objective is well-defined, browsing aims to provide users with a conceptual map, so that users can navigate among correlated items to hopefully find some potentially useful documents, or to formulate a more precise retrieval request further. For instance, *a user reads an article talking about a water reservoir construction plan in a certain region. He/she wants to know the possible influence on ecological balance. By following semantic links for the water reservoir plan in the DL, he/she navigates to the related "ecological protection" theme, under which a set of searchable terms with relevant documents are listed for selection.* To facilitate browsing, DLs must integrate diverse repositories to provide users with a uniform searching and retrieval interface to a coherent collection of materials. The capability that enables navigation among a network of inter-related concepts, plus the searching capability on each individual repository, constitute the fundamental support to browsing activities. Thereby, we can view a user's browsing activity as navigation plus searching, i.e., *browsing = searching + navigation*.

As the techniques of searching and browsing have been extensively studied and published in the literature, we will not discuss these any further here.

2.2 Strategic Level Cognition Support

In contrast to tactical level cognition support which intends to provide users with requested documents, strategic level cognition support not only provides documents but can also intelligently answer high-order cognitive questions, and meanwhile provide justifications and evidences. Taking the scenario in section 1 as an example, the purpose of Kooper on the use of the DL is to confirm his prior knowledge about "*Wet winter causes flood*

in summer". Instead of retrieving documents using dispersed keywords like "*wet winter*", "*flood*", "*cause*", etc, the user would prefer to pose a direct question Q_1 as follows, and expect a confirmed/denied answer from the DL system rather than a list of articles lacking explanatory semantics and waiting for his further checking.

Q_1: "*Does wet winter cause flood in summer*?"

Other high-order cognitive request examples are like:

Q_2: "*Give me articles which talk about the **cause** of flood*."

Q_3: "*Give me articles which talk about the **influences** of wet winter*."

In response to different questions, it is desirable for DL systems to provide knowledge-level answers and related justifications for holding the answers. For example, the justifications for Q_1 will consist of a series of articles talking about "*wet winter causes flood in summer*", as well as evidence articles which talk about, respectively, "*wet winter*" in certain years and "*summer flood*" in the next years in certain particular regions.

The provision of strategic level cognition support adds values to DLs beyond simply providing document access. It reinforces the exploration and utilization of information in DLs, and advocates a more close and powerful interaction between users and DL systems. With this high-order cognitive assistance, DL users people will be able to find things to solve their real information problems themselves. From the viewpoint of DLs, to realize such a strategic level cognition function, substantial information analysis needs to be done. This inevitably involves the navigation and correlation of information items across multiple repositories in DLs, and production of intelligent knowledge in answering users' strategic level cognitive questions. Thereby, compared to the tactical level cognition support by information searching and browsing, the strategic level cognition support involves the efforts on information searching, navigation and analysis, i.e., *knowledge inquiry = searching + navigation + analysis*.

We sketch a DL's information space, comprised of a *knowledge subspace* and a *document subspace*, in Figure 2 to address users' strategic level cognition requests.

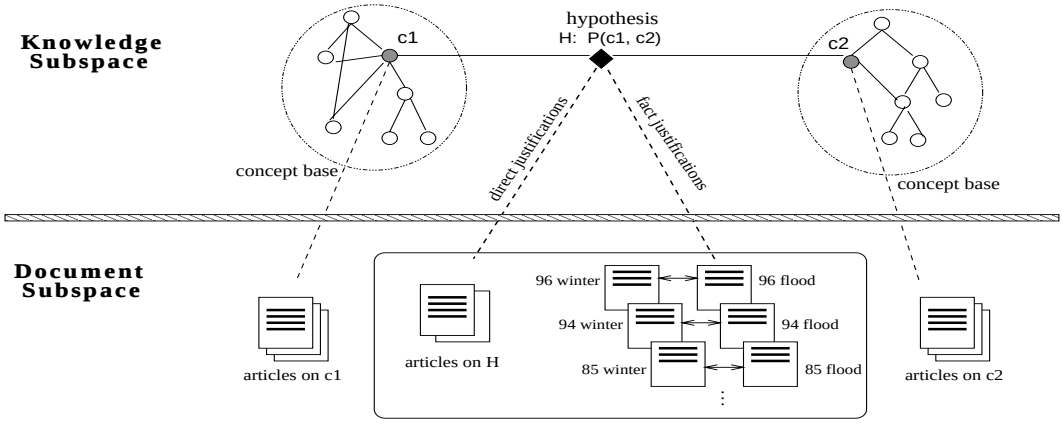

Figure 2: A DL's information space for strategic level cognition support

1) The Knowledge Subspace. The basic constituents of the *Knowledge Subspace* are knowledge, such as hypotheses, rules, beliefs, etc. [1] Each piece of hypothesis describes a certain relationship among a set of concepts. For example, the hypothesis "H: *Wet winter may cause summer flood*" explicates a causal relationship between a cause "C_1: *wet winter*" and the effect "C_2: *summer flood*" it has. A more general hypothesis in respect to H is like "H': *wet winter may cause river behavior*". Based on different concept relations (e.g.,

[1] In this initial study, we focus on hypothesis knowledge in empirical sciences.

is-a, *part-whole*, *synonym*, and *antonym*) in the concept base, we can build up inter-relationships of relevant hypotheses, formulating a hypothesis lattice around one theme. A detailed description of the knowledge subspace is given in Section 3.

2) The Document Subspace. Under each hypothesis is a justification set, giving reasons and evidences for the knowledge. These justifications, made up of articles, reports, data, etc., constitute the *Document Subspace* of the DL's information space. Here, we differentiate two kinds of justifications, namely, *direct-justifications* and *fact- justifications*. Direct-justifications give documents that support the knowledge straightforward. Taking hypothesis H for an example, the article mentioning exactly that "*wet winter is an indicator of summer flood.*" is a direct-justification. On the other hand, fact-justifications provide evidences for believing the knowledge. For instance, by illustrating abreast wet winter meteorological reports and summer flood reports in the same years, the system can hopefully confirm users of this hypothesis. It is worth notice here that the document subspace challenges traditional DLs on literature organization, classification, and management. For belief justifications, we must extend the classical **keyword** based index schema, which is mainly used for information searching and browsing purposes, to **knowledge-justification** based index schema, in order that the information in DLs can be easily retrieved by both keywords and knowledge. Details for constructing the knowledge and document subspaces are described in Section 4.

The information space consisting of both knowledge and documents as described above enables different levels of cognition solutions. For instance, referring to the above questions Q_1, Q_2 and Q_3, the DL system can provide not only intelligent answers according to hypothesis H, but also a series of links to the justification articles for further reference.

3 A Formal Description of DL's Knowledge Subspace

In this section, we define the basic constituent of the DL's knowledge subspace - *hypothesis*, starting with its two constructional elements, i.e., *concepts* and *relations* among the concepts.

3.1 Concepts and Concept Terms

We distinguish two kinds of concepts: **atomic concepts** and **composite concepts**. Atomic concepts are the building blocks of sentences (e.g., "*dog*", "*animal*", "*traffic-jam*", "*wet-winter*", "*summer-flood*", etc.), which convey the most fundamental cognitive knowledge in human society, while composite concepts are built up from atomic concepts through a concept conjunctive operator $\sqcap$. One composite concept example is "*warm-winter* $\sqcap$ *wet-winter*". A **concept term** can be either an atomic or a composite concept.

<concept term> ::= <atomic concept> | <composite concept>
<composite concept> ::= <atomic concept> $\sqcap \ldots \sqcap$
<atomic concept>

Throughout the paper, we use $C = \{c_1, c_2, \ldots, c_s\}$ to denote a set of atomic concepts in the universe of discourse, and T to denote a set of concept terms taken from C.

3.2 Concept Relations

Based on the substantial work on lexicography and ontology [19, 12, 25, 14, 15], four **primitive relationships** between atomic concepts are considered in our current study. They are *Is-A*, *Part-Whole*, *Synonym*, and *Antonym*, each of which is denoted using a binary predicate. For example, *Is-A* ("*summer-flood*", "*river-behavior*") represents the notion that "*summer-flood*" is a "*river-behavior*", and *Antonym* ("*wet-winter*", "*dry-winter*") represents a pair of antonym concepts, "*wet-winter*" and "*dry-winter*".

Property 1 *The four primitive relationships of atomic concepts have the following properties:*
. Is-A *is reflexive and transitive.*
. Part-Whole *is reflexive and transitive.*
. Synonym *is reflexive, transitive and symmetric.*
. Antonym *is symmetric.*

Based on the primitive relationships of atomic concepts, we define four concept-term-based relationships. Assume $t = x_1 \sqcap \ldots \sqcap x_n$ and $t' = y_1 \sqcap \ldots \sqcap y_m$ are two concept terms in the following definitions, where $t \in T$, $t' \in T$, $\forall i\ (1 \leq i \leq n)\ (x_i \in C)$ and $\forall j\ (1 \leq j \leq m)\ (y_j \in C)$.

Definition 1 ***Specific Relation SPEC*** (t, t')**.**
Concept term t is more ***specific*** *than concept term* t'*, iff* $\forall y_j \in \{y_1, \ldots, y_m\}\ \exists x_i \in \{x_1, \ldots, x_n\}$*, either* $(x_i = y_j)$ *or Synonym*$(x_i,\ y_j)$ *or Is-A*$(x_i,\ y_j)$.

Example 1 *SPEC (*"wet-winter $\sqcap$ warm-winter"*,* "wet-winter"*); SPEC (*"wet-winter $\sqcap$ warm-winter"*,* "warm-winter"*); SPEC (*"warm-winter $\sqcap$ *rare"*,* "abnormal-winter $\sqcap$ *seldom"*), since* Is-A *(*"warm-winter"*,* "abnormal-winter"*) and Synonym ("rare", "seldom").*

Definition 2 ***Equivalent Relation EQ*** (t, t')**.**
Concept term t is ***equivalent*** *to concept term* t'*, iff the following two conditions hold:*
1) $\forall y_j \in \{y_1, \ldots, y_m\}\ \exists x_i \in \{x_1, \ldots, x_n\}$*, either* $(x_i = y_j)$ *or Synonym*$(x_i,\ y_j)$*;*
2) $\forall x_i \in \{x_1, \ldots, x_n\}\ \exists y_j \in \{y_1, \ldots, y_m\}$*, either* $(y_j = x_i)$ *or Synonym*$(y_j,\ x_i)$.

Example 2 *EQ (*"warm-winter $\sqcap$ *rare"*,* "warm-winter $\sqcap$ *seldom"), since Synonym ("rare", "seldom").*

Definition 3 ***Intersect Relation INSE*** (t, t')**.**
Two concept terms t and t' ***intersect****, iff* $\exists y_j \in \{y_1, \ldots, y_m\}\ \exists x_i \in \{x_1, \ldots, x_n\}$*, either* $(x_i = y_j)$ *or Synonym*$(x_i,\ y_j)$.

Example 3 *INSE (*"warm-winter $\sqcap$ *disease"*,* "warm-winter $\sqcap$ wet-winter"*), with* "warm-winter" *as the common concept.*

Definition 4 ***Opposite Relation OPSI*** (t, t')**.**
Concept term t is ***opposite*** *to concept term* t'*, iff the following two conditions hold:*
1) $\forall y_j \in \{y_1, \ldots, y_m\}\ \exists x_i \in \{x_1, \ldots, x_n\}$*, either* $(x_i = y_j)$ *or Synonym*$(x_i,\ y_j)$ *or Antonym*$(x_i,\ y_j)$*;*
2) $\exists y_j \in \{y_1, \ldots, y_m\}\ \exists x_i \in \{x_1, \ldots, x_n\}$*, Antonym*$(x_i,\ y_j)$.

Two concept terms are opposite if their concepts contained are either the same, synonym, or antonym. The second requirement in the definition indicates that there exists at least one antonym concept pair between the two concept terms.

Example 4 *OPSI (*"dry-winter $\sqcap$ flu", "wet-winter $\sqcap$ flu"*), since Antonym (*"dry-winter", "wet-winter"*) and the rest concept "flu" exists in both concept terms.*

Property 2 *The four context-term-based relationships defined above have the following properties:*
. SPEC is reflexive and transitive.
. EQ is reflexive, transitive and symmetric.
. INSE is reflexive and symmetric.
. OPSI is symmetric.

3.3 Hypotheses

A hypothesis communicates a human's cognitive idea or thinking about things in existence, such as the causal connection of situations, the sequential occurrence of events, etc. Here, we describe each piece of hypothesis through a predicate with concept terms as its arguments. At the moment, we focus our study on binary predicates associated with two concept terms - a left-side concept term and a right-side one. For example, the hypothesis "*Wet winter may cause summer flood*" can be expressed as *Cause* ("*wet-winter*", "*summer-flood*"). "*Air pollution may cause acid rain*" is another hypothesis example which can be described as *Cause* ("*air-pollution*", "*acid-rain*"). The DL's knowledge subspace is made up of a number of this kind of hypotheses.

Definition 5 *A **hypothesis** H is a binary predicate $H = P(t_l,\ t_r)$, where P is the predicate name, and $t_l, t_r \in T$ are the left- and right-side concept terms of the predicate, respectively.*

By means of the concept-term-based relations, we can formulate the inter-relationships among hypotheses as follows.

Definition 6 *A hypothesis $H = P(t_l,\ t_r)$ is more **specific** than a hypothesis $H' = P(t'_l,\ t'_r)$, written as $H \preceq_h H'$, iff one of the following conditions holds:*
1) *EQ* $(t_l,\ t'_l)$ *and SPEC* $(t_r,\ t'_r)$*;*
2) *EQ* $(t_r,\ t'_r)$ *and SPEC* $(t_l,\ t'_l)$*;*
3) *SPEC* $(t_l,\ t'_l)$ *and SPEC* $(t_r,\ t'_r)$.
*Conversely, H' is called **more general** than H, written as $H' \succeq_h H$.*

Axiom 1 *Let $H = P(t_l,\ t_r)$ and $H' = P(t'_l,\ t'_r)$ be two hypotheses, where $H \preceq_h H'$. H is true implies that H' is true. In other words, $P(t_l,\ t_r) \rightarrow P(t'_l,\ t'_r)$ and $\neg P(t'_l,\ t'_r) \rightarrow \neg P(t_l,\ t_r)$.*

Note that using the definition of speciality/generality between hypotheses, we can be sure that if a hypothesis is consistent with a set of documents, any generalization of it will also be consistent with this document sets. Conversely, if a document does not justify a hypothesis, it cannot justify any specialization of that hypothesis either.

Example 5 *Hypothesis* $H_1 = Cause$ (“wet-winter $\sqcap$ warm-winter”, “summer-flood”) *is more specific than hypothesis* $H_2 = Cause$ (“wet-winter”, “summer-flood”), *which is also more specific than hypothesis* $H_3 = Cause$ (“wet-winter”, “river-behavior”). *That is,* $(H_1 \preceq_h H_2)$ *and* $(H_2 \preceq_h H_3)$, *since SPEC* $(H_1.t_l,\ H_2.t_l)$ *and EQ* $(H_1.t_r,\ H_2.t_r)$, *SPEC* $(H_2.t_r,\ H_3.t_r)$ *and EQ* $(H_2.t_l,\ H_3.t_l)$.

Definition 7 *A hypothesis* $H = P(t_l,\ t_r)$ *is* ***equivalent*** *to a hypothesis* $H' = P(t'_l,\ t'_r)$, *written as* $H \equiv_h H'$, *iff EQ* $(t_l,\ t'_l)$ *and EQ* $(t_r,\ t'_r)$.

Definition 8 *Two hypotheses* $H = P(t_l,\ t_r)$ *and* $H' = P(t'_l,\ t'_r)$ *are* ***supplementary***, *written as* $H \simeq_h H'$, *iff either of the following two conditions holds:*
1) *EQ* $(t_r,\ t'_r)$ *and INSE* $(t_l,\ t'_l)$.
2) *EQ* $(t_l,\ t'_l)$ *and INSE* $(t_r,\ t'_r)$.

The first condition of the definition states that the two hypotheses have the same left-side, but the right-side of the predicate intersect, while the second condition is in reverse.

Example 6 $H_1 = Cause$(“wet-winter $\sqcap$ warm-winter”, “summer-flood”) *and* $H'_1 = Cause$(“wet-winter”, “summer-flood”) *can be viewed as a pair of supplementary hypotheses* $(H_1 \simeq_h H'_1)$, *since EQ* $(H_1.t_r,\ H'_1.t_r)$ *and INSE* $(H_1.t_l,\ H'_1.t_l)$.

So is the pair of $H_2 = Cause$(“wet-winter”, “summer-flood”) *and* $H'_2 = Cause$(“wet-winter”, “summer-flood $\sqcap$ hot-summer”) $(H_2 \simeq_h H'_2)$, *since EQ* $(H_2.t_l,\ H'_2.t_l)$ *and INSE* $(H_2.t_r,\ H'_2.t_r)$.

Definition 9 *A hypothesis* $H = P(t_l,\ t_r)$ *is* ***opposite*** *to a hypothesis* $H' = P'(t'_l,\ t'_r)$, *written as* $H \propto_h H'$, *iff either of the following two conditions holds:*
1) *EQ* $(t_r,\ t'_r)$ *and OPSI* $(t_l,\ t'_l)$.
2) *EQ* $(t_l,\ t'_l)$ *and OPSI* $(t_r,\ t'_r)$.

Example 7 $H_1 = Cause$(“wet-winter”, “summer-flood $\sqcap$ hot-summer”) *and* $H_2 = Cause$(“wet-winter”, “summer-flood $\sqcap$ cool-summer”) *can be viewed as a pair of opposite hypotheses* $(H_1 \propto_h H_2)$, *since EQ* $(H_1.t_l,\ H_2.t_l)$ *and OPSI* $(H_1.t_r,\ H_2.t_r)$.

Property 3 *The relationships defined on hypotheses have the following properties:*
. Specific $\preceq_h$ *is reflexive and transitive.*
. Equivalent $\equiv_h$ *is reflexive, transitive and symmetric.*
. Supplementary $\simeq_h$ *is symmetric.*
. Opposite $\propto_h$ *is symmetric.*

Hypotheses and the defined relationships among them constitute DL's knowledge subspace, on which knowledge inquiry, navigation and induction can be performed to support users' tactical requests. If a new user inquiry has the form of a hypothesis, the above relationships like *equivalent, specific/general, supplementary* can be explored to find matching hypotheses in the knowledge subspace. The hypotheses together with the backing documents (see Figure 2) are returned to the user as a part of the answer to his/her inquiry.

4 The Construction of DL's Knowledge & Document Subspaces

The construction of such a knowledge subspace and its associated justification (document) subspace can be done in two ways: 1) *human-centered knowledge acquisition.* Experienced humans input hypothesis knowledge manually, based on which justifying articles are collected by performing searching and browsing on DL systems; 2) *machine-centered knowledge acquisition.* Under the assistance of users, DL systems automatically deduce hypothesis knowledge by correlating and analyzing data sources. We discuss these two methods in detail in the following subsections.

4.1 Human-Centered Knowledge Acquisition

The central problem in all attempts to leverage a digital library by adding *knowledge* to mainly *symbolic data* (either in the form of raw data or document collections) is to understand the data. Currently, machines hardly extract proper knowledge from data alone. Therefore, any practical application of the techniques described in this paper involves human knowledge acquisition efforts.

The sheer amount of data available through current digital library systems, including the Web, prevents "charting of the knowledge space," as traditionally done by librarians and other knowledge workers. Their efforts are still invaluable, as all "portals" on the Internet prove: few true search engines without human knowledge charts survive to date. However, humans alone on the knowledge provider side cannot properly chart all documents and data that becomes, or already is, available to (potential) knowledge consumers.

What is needed is a human input from the *consumer* side. As even detailed logs of search engines prove, these logs (taken at the tactical level of information seeking) only reveal attempts of users to find what they want using coarse, dumb tools [30]. It is like attempting to reconstruct the history of the great pyramids in Egypt by looking only at the cuts the tools of the stone carvers made on the sand stone blocks. What is totally missing from search engine query logs is the actual *strategic intention* of the user behind the keyboard.

Acquiring this strategic intention is difficult. Few users will volunteer a semi-formal description of their intentions when there is no single form of short-term reward for their efforts. Users will never take the time to learn a formal language to express their strategic target in, and any user interface trying to guide them will likely become either too complex to understand or too coarse to be of use.

In order to overcome at least some of these problems and to be able to run some experiments on strategic user intention acquisition, we intend to concentrate on a very limited subset of knowledge acquisition (hypothesis support/denial). Such a limited goal can be pursued with limited tools, making it suitable for a small group of users with a well-defined type of knowledge need. On top of the limited, but understandable and simple strategic intention acquisition tool, we intend to implement some kind of immediate reward system that makes it attractive for users to specify their strategic intentions to the system before embarking on their traditional, tactical search effort. Currently we think of a system that extends the user's quota for certain activities, typically inter-library loans or other kinds of priced assets that are required for the work, but are in short supply due to the costs involved. Users who are active in supplying proper strategic targets are rewarded by increasing quota for related activities, facilitating their work. Existing systems which use this "return on investment" policy have showed that such a policy has a self-regulating property [2], and indeed may lead to a

[2] http://www.slashdot.org

coarse but useful filtering of data items (documents) useful to support a strategic goal.

Our future work will concentrate on finding a proper test case for such a system, and to gather data on the success of a "return on investment" policy, the collected strategic knowledge targets in respect to the retrieved documents, and the re-usability of these targets and the document result sets for other users with comparable information needs.

4.2 Machine-Centered Knowledge Acquisition

With more and more digital information in a wide variety of disciplines accumulated in DLs today, the automatic and/or user-assisted semi-automatic extraction of inherent knowledge from such a large volume of data becomes indispensable. In this subsection, we outline a framework for machine-centered knowledge discovery across multiple repositories in DLs. Basically, it proceeds in 6 steps: 1) *set up knowledge discovery targets,* 2) *identify relevant resources,* 3) *filter out interesting concepts from identified resources,* 4) *correlate concepts according to contextual information,* 5) *extract knowledge and justifications from correlated concepts,* and 6) *evaluate the discovered knowledge and justifications.*

Phase-1: Set Up Knowledge Discovery Tasks

Before extracting knowledge and associated justifications, we need to know what kind of knowledge and justifications are expected from users. Does the user want to know the inherent *associations* of some of concepts in certain areas? Or does the user want to obtain some knowledge on *classifying* or *discriminating* objects? Or is the user interested in the *sequential evolution* of certain objects in order to make predictions? Here, a friendly user interface is important to the natural and accurate specification of such knowledge acquisition targets.

Phase-2: Identify Relevant Resources

Confronted with a huge collection of data sources scattered in heaps of repositories, we must identify repositories which contain the most likely relevant resources in respect to a given knowledge request. Otherwise, the following knowledge discovery process will be lengthy, aimless and inefficient. To do this, we must categorize the information content in each repository. By querying the concepts, which are elicited from the user's request and background, against these meta-data catalogs, we may identify relevant data sources. For example, if a user is interested in the correlations between "*summer flood*" and "*meteorological factors*", the meteorological repository in the observatory headquarters, and the repository in the river management office, will be identified as mostly relevant.

Phase-3: Filter Out Interesting Concepts from Identified Resources

The identified resources from Phase-2 could be in various formats. They can be unstructured articles, multimedia documents, formatted reports, database records, semi-structured files, or hypertexts, etc. To make knowledge discovery across multiple heterogeneous repositories possible, we need to transform the original data sources into a uniform format. A *record structure* consisting of a set of keywords can be exploited to describe each documental entity. These keywords explicate the major concepts conveyed by the corresponding documents. For example, from a textual article which mentions "*high rainfall amounts in November and December in 1996*", we can filter out "*wet winter*" concept "*in 1996*". Note that the transformation from heterogeneous resources into keyword-based records solicits a wide range of techniques, including natural language processing, information analysis, categorization and summarization, textual and multimedia data

mining, etc.

Phase-4: Correlate Concepts According to Contextual Information

Data records filtered out of diverse resources must be logically linked together so that their inherent associations can be detected. This could be done based on their common contextual information. For example, to find possible relationships between *season* and *river behavior* as a consequence, we can link *yearly weather record* obtained from one resource, with the corresponding *river behavior record* in the *same year*, which is obtained from another resource. Here, "*year*" conveys a kind of contextual information with which the existence of concepts "*weather*" and "*river behavior*" have concrete meaning. Table 1 shows a list of logical connection examples between these two kinds of records.

linked record	weather record		river behavior record	
ID	*Context Info.*	*Concept*	*Context Info.*	*Concept*
1	1984 :	wet winter, warm	1984 :	serious flood summer
2	1985 :	dry winter, warm	1985 :	no flood summer
3	1986 :	wet winter, cold	1986 :	no flood summer
4	1994 :	wet winter, cold	1994 :	flood summer
5	1996 :	except wet winter, cold	1996 :	flood summer

Table 1: A logical connection example between two resource records

Phase-5: Extract Knowledge and Justifications from Correlated Concepts

After preparing linked data records, we are now in a position to find their inherent knowledge and justifications. In the current study, we focus on the discovery of correlationships of concepts. The association rule technique developed in the data mining area can be applied to this knowledge extraction phase.

The problem of mining association rules from transactional data was first introduced in [1]. The application is sales data of supermarkets. It aims to discover the associations among items purchased by customers such that the presence of some items in a transaction will imply the presence of other items in the same transaction. The following is a mathematical model to address the problem of mining association rules. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of literals, called items. Let D be a set of transaction records, where each transaction record A consists of a set of items such that $A \subseteq I$. Let X be a set of items. A transaction A is said to contain X if and only if $X \subseteq A$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I, Y \subset I$ and $X \cap Y = \phi$. The rule $X \Rightarrow Y$ holds in the transaction set D with *confidence* c if c% of transactions in D that contain X also contain Y. The rule has *support* s if s% of transactions in D contain $X \cup Y$.

Applying the concept of association rules in transactional data to our linked records illustrated in Table 1, we can find the correlationship of concept terms like "*wet-winter* $\Rightarrow$ *summer-flood*". Since among the total 5 records in this small database, 3 of them contain $\{$*wet-winter, summer-flood*$\}$, thus $support = 3/5 = 60\%$. Within the 4 records stating *wet winter*, 3 of them also state *summer flood*. Therefore, we have $confidence = 3/4 = 75\%$. These supporting records, together with the *support* and *confidence* measurement values, can serve as the justifications for the extracted knowledge.

Phase-6: Evaluate the Discovered Knowledge and Justifications

Obtaining a set of knowledge and justifications is not the end. Besides using statistical indicators such as support, confidence to measure the strengths of discovered knowledge, we may also incorporate subjective

measurements to evaluate and rank their significancy in respect to users' cognitions and their real-life problems.

5 Discussions

We extend the traditional role of DL systems from information provider to information & knowledge provider by incorporating both knowledge and documents into the DL's information space. The proposed framework raises a number of issues, calling for multi-disciplinary cooperation in the information science field. In this section, we provide a brief review of relevant work from different areas, including *ontologies and lexical relations, knowledge representation, information retrieval and multi-source integration in digital libraries,* and then discuss some distinguished features of our work by comparison with the traditional knowledge-based information systems.

5.1 Related work

Ontologies and Lexical Relations. The word *ontology* in philosophy refers to a systematic description of a minimal set of concepts that a language needs to express all its other concepts. In linguistic and lexicographic contexts, an ontology is a specification of the concepts and their relationships that exist for a community of people or automated agents. Its purpose is to enable knowledge sharing and reuse. To reflect cognitive relations among the concepts as expressed by or embodied in words in natural languages, the theory of *lexical relations* was developed [19], where information is represented as a collection of nodes connected by labeled arcs that express links or relationships between the nodes. These links may represent semantic and syntactic relationships among concepts, and can be used to reflect not only major aspects of word meaning, but also their morphological relationships [25]. The most widely recognized relationships include *is-a, part-whole, synonym, antonym*, which can be denoted using either a labeled arc or a proposition node. Further, a methodology for identifying, evaluating and describing different relationships, along with their logical properties in modeling human reasoning, was presented [22]. The method begins with a study of dictionary and folk definitions to obtain the linguistic formulae used to express the link in question. Then it discovers how this link appears in anthropology, linguistics, philosophy and psychology, and the way it interacts with other links and the way it is used in conceptual information processing. A hierarchy of lexical-semantic relations is thus constructed which contains both basic and non-basic semantic relations as well as a large number of lexical relations. Many (but not all) of the lexical relations so far identified can be extracted automatically from dictionary definitions [25]. CYC developed in [21] aims to build a common sense knowledge base consisting of roughly 100,000 general concepts spanning all aspects of human reality. WordNet is another remarkable and widely used lexicon, which groups the words together by means of synonym sets. It aims to provide an aid to search in a lexicon in a conceptual rather than an alphabetical way [24].

Knowledge Representation. Knowledge is an important element of any AI application. A number of knowledge representation schemas are designed in the AI field so that the knowledge can be applied in the reasoning process to solve problems. These techniques can be roughly categorized as either *declarative methods*, in which most of the knowledge is represented as a static collection of facts accompanied by a small set of general procedures for manipulating them; or *procedural methods*, in which the bulk of the knowledge is represented as procedures for using it. Typical declarative methods include predicate logic, semantic nets,

frames and scripts. 1) Predicate logic involves using standard forms of logical symbolism to represent real-world facts as statements, written as well-formed formulae, which are made up of predicates, constant terms, variables, logical connectives, and quantifiers. 2) Semantic nets take the complex structure of the world into consideration, where information is represented as a set of nodes connected to each other by a set of labeled arcs, representing relationships among the nodes. 3) Frames and 4) scripts serve as general-purpose knowledge structures that represent some common features of things or sequence of events in a particular context. Some procedural knowledge representation methods include procedures and production rules. In our study, we explore the use of predicate logic and semantic network mechanisms to represent hypotheses and their inter-relationships in DL's knowledge subspace.

Information Retrieval and Multi-Source Integration in DLs. To support efficient information *searching* activity, many efforts have been made in developing retrieval models, building document and index spaces, extending and refining queries for DLs [13, 9]. In [11], index terms are automatically extracted from documents and a vector-space paradigm is exploited to measure the matching degrees between queries and documents. Indexes and metadata can also be manually created from which semantic relationships are captured [10]. Furthermore, the information space consisting of a large collection of documents can be semantically partitioned into different clusters, so that queries can be evaluated against relevant clusters [32]. According to topic areas, a distributed semantic framework is proposed to contextualize the entire collection of documents for efficient large-scale searching [27]. To improve query recall and precision, several query expansion and refinement techniques based on relational lexicons/thesauri or relevance feedback have been explored [31]. A recent work incorporates knowledge about the document structures into information retrieval, and the presented query language allows the assignment of structural roles to individual query terms [33]. Applying AI to library science, many library-oriented expert systems have been developed in the literature [20]. Most of these systems essentially aid in carrying out the support operations of libraries, such as descriptive cataloging, collection development, disaster planning and response, reference services, database searching, and document delivery, etc. [20]. On the other hand, observing one DL usually contains lots of distributed and heterogeneous repositories which may be autonomously managed by different organizations, in order to facilitate users' *browsing* activities across diverse sources easily, many efforts have been engaged in handling various structural and semantics variations and providing users with a coherent view of a massive amount of information [29, 6]. The concept extraction, mapping and switching techniques, developed in [4, 6], enable users in a certain area to easily search the specialized terminology of another area. A dynamic mediator infrastructure [23] allows mediators to be composed from a set of modules, each implementing a particular mediation function, such as protocol translation, query translation, or result merging [26]. [28, 17] present an extensible digital object and repository architecture FEDORA, which can support the aggregation of mixed distributed data into complex objects, and associate multiple content disseminations with these objects. [18, 26] employ the distributed object technology to cope with interoperability among heterogeneous resources. The experiences in designing and implementing digital libraries including the archival repository architecture, user interface, and cross-access mechanism, etc. are extensively described in [16, 8, 7, 3].

5.2 Some Distinguished Features

Compared to the traditional expert-like knowledge-based information systems, the DL systems enhanced with knowledge elements have the following distinguished features.

Functions. Knowledge-based systems simulate human behavior by making deductions using the rules of logical inference. Most businesses have processes which are based on rules and company policies. Knowledge-based systems are designed to apply these rules to make judgement in processing business routines and come up with a conclusion to a certain pre-defined problem [2]. For example, a production rule used in knowledge-based systems always has the format: IF x THEN y, whose IF part states a premise and THEN part refers to the conclusions or consequences. On the contrary, the mission of a DL system equipped with a knowledge subspace is to make expertise knowledge widely available to the public. We can view such a system as an **information** & **knowledge dictionary**, since a huge body of knowledge of various kinds in the world, together with their justification documents, is preserved, classified, and maintained inside its knowledge subspace. Turning on a DL system, users can acquire not only the requested documents, but also the knowledge in response to their high-order cognitive questions.

Scopes. A knowledge-based system intends to solve problems in a specific area/domain (e.g., company delivery charge, heart disease diagnosis, etc.) The rules stored in its knowledge base are thus only limited to a particular field of interest. Comparatively, the scope of the knowledge embraced within the DL's knowledge subspace is very extensive, covering a wide spread of scientific and enginering disciplines. Users from different backgrounds can turn to the library for expert-like helps in carrying out their work.

Unitization. With the continuing developments in storage and communication technologies, a tremendous amount of structured, semi-structured, and unstructured information assets is collected and maintained within a DL. While we extend the DL's information space to incorporate knowledge, such a huge body of documents constitutes knowledge justifications for users' further reference. In comparison, this is not the case for traditional knowledge-based systems, which provide only a limited amount of rules and facts in a particular field of expertise.

6 Conclusion

Motivated by the problems - (i) inadequate strategical level cognition support; (ii) inadequate knowledge sharing facilities - with the present-day digital library systems, we first introduce a two-layered digital library function model to support different levels of human cognitive acts. The tactical level cognition support aims to provide users with requested relevant documents, as searching and browsing do, while strategic level cognition support can provide not only documents but also intelligent answers to users' high-order cognitive questions. Second, to address users' high-order cognitive requests, we propose an information space comprised of a *knowledge subspace* and a *document subspace*. Finally, two mechanisms for constructing the two subspaces are discussed to enable knowledge sharing and propagation among DL users. The major contributions of the paper are twofold. First, the presented DL information space extends the traditional role of DLs from *information provider* to *information* & *knowledge provider*. Second, the traditional simple keyword-based index schema is expanded to strategic knowledge-based level, consisting of inter-related hypotheses that are backed by documents. One can say that the essence of the documents is reconstructed in the DLs' knowledge level.

We view this work as a first step, with a number of interesting problems and challenges remaining for future work. (1) To facilitate strategic level cognitive activities, efficient storage and management of the knowledge & document subspaces is very important and must be carefully planned. This demands effective indexing

strategies for both knowledge and justifying documents. (2) Efficient knowledge inference and navigation mechanisms must be built to support users' question-answering. (3) A flexible and easy-to-use query language is to be designed to help DL users make the best of information and knowledge assets in solving their problems. (4) Scalability is a big issue in any DL, which usually contains tens of thousands of repositories of digital information. Discovering knowledge from such a huge amount of heterogeneous resources and maintaining the discovered knowledge are a big challenge. (5) Libraries exist in a social and economic framework. Intellectual property and data security are a big concern when performing cross-repository correlation and analysis. One possible way here is to conduct multi-leveled information analysis, so that users of different authentication levels can have different views of analysis results. (6) Our eventual goal is to develop a practical DL system, which can empower human with real actionable knowledge in solving their information problems.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the 1993 ACM SIGMOD International Conference on management of data*, pages 207–216, Washington D.C., USA, May 1993.

[2] R.G. Anderson. *Information and Knowledge Based Systems: an Introduction*. Prentice Hall, 1992.

[3] K. Beard, T. Smith, and L. Hill. Meta-information models for georeferenced digital libraries. *Journal on Digital Libraries*, 1(2):153–160, 1997.

[4] N. Bennett, Q. He, C. Chang, and B.R. Schatz. Concept extraction in the interspace prototype. Technical report, Dept. of Computer Science, University of Illinois at Urbana-Champaign, 1999.

[5] P. Checkland and S. Holwell. *Information, Systems and Information Systems - making sense of the field*. John Willey & Sons, Inc, 1998.

[6] H. Chen. Semantic research for digital libraries. *D-Lib Magazine*, 5(10), 1999.

[7] B. Cooper, A. Crespo, and H. Garcia-Molina. Implementing a reliable digital object archive. Technical report, Standford University, 2000.

[8] A. Crespo and H. Garcia-Molina. Modeling archival repositories for digital libraries. Technical report, Standford University, 1999.

[9] F. Crestani, M. Lalmas, C. van Rijsgergen, and I. Campbell. 'is this document relevant? ... probably': A survey of probabilistic models in information retrieval. *ACM Computing Surveys*, 30(4), 1998.

[10] T. Dao. An indexing model for structured documents to support queries on content, structure and attributes. In *Proc. of the IEEE Forum on Research and Technology Advances in Digital Libraries*, pages 88–97, California, USA, April 1998.

[11] M. Dunlop and C. van Rijsbergen. Hypermedia and free text retrieval. *Information Processing and Management*, 29(3), 1993.

[12] M.W. Evens and R.N. Smith. A lexicon for a computer question-answering system. *American Journal of Computational Linguistics*, 83:1–93, 1979.

[13] W.B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, 1992.

[14] T. Gruber. A translation approach to portable ontologies. *Knowledge Acqusition*, 5(2):199–220, 1993.

[15] N. Guarino and C. Welty. Ontological analysis of taxonomic relationships. In *Proc. of the 19th International Conference on Conceptual Modeling*, USA, October 2000.

[16] J. Hoppenbrouwers and H. Paijmans. Invading the fortress: How to besiege reinforced information bunkers. In *Proc. of the IEEE Advanced in Digital Libraries*, pages 27–35, USA, May 2000.

[17] R. Daniel Jr and C. Lagoze. Extending the warwick framework: From metadata containers to active digital objects. *D-Lib Magazine*, 3(11), 1997.

[18] R. Kahn and R. Wilensky. A framework for distributed digital object services. Technical report, Corporation for National Research Initiatives, 1995.

[19] F. Kiefer, editor. *Studies in Syntax and Semantics*, chapter Semantics and Lexicography: Towards a New Type of Unilingual Dictionary (Y.D. Apresyan, I.A. Melcuk and A.K. Zolkovsky). Holland, 1969.

[20] F.W. Lancaster and B. Sandore. *Technology and Management in Library and Information Services*. University of Illinois Graduate School of Library and Information Science, 1997.

[21] D. Lenat. CYC: a large-scale investment in knowledge infrastructure. *Communications of ACM*, 13(11), 1995.

[22] J.A. Markowitz, J.T. Nutter, and M.W. Evens. Beyond is-a and part-whole: More semantic network links. *Journal of Computers and Mathematics with Applications*, 23(6-9):377–390, 1992.

[23] S. Melnik, H. Garcia-Molina, and A. Paepcke. A mediation infrastructure for digital library services. Technical report, Standford University, 2000.

[24] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet: an on-line lexical database. Technical report, Princeton University, 1993.

[25] J.T. Nutter, E.A. Fox, and M.W. Evens. Building a lexicon from machine-readable dictionaries for improved information retrieval. *Journal of Literary and Linguistic Computing*, 5(2):129–137, 1990.

[26] A. Paepcke, R. Brandriff, G. Janee, R. Larson, B. Ludaescher, S. Melnik, and S. Raghavan. Search middleware and the simple digital library interoperability protocol. *D-Lib Magazine*, 6(3), 2000.

[27] M. Papazoglou and J. Hoppenbrouwers. Contextualizing the information space in federated digital libraries. *SIGMOD Record*, 28(1):40–46, 1999.

[28] S. Payette, C. Blanchi, C. Lagoze, and E.A. Overly. Interoperability for digital objects and repositories: The cornell/cnri experiments. *D-Lib Magazine*, 5(5), 1999.

[29] B. Schatz and H. Chen. Digital libraries: Technological advances and social impacts. *IEEE Computer*, 32(2):45–50, 1999.

[30] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large Alta Vista query log. Technical report, 014, Systems Research Center, USA, 1998.

[31] B. Vélez, R. Weiss, M.A. Sheldon, and D.K. Gifford. Fast and effective query refinement. In *Proc. of the 20th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 6–15, Philadelphia, USA, July 1997.

[32] P. Willett. Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(5), 1988.

[33] J.E. Wolff, H. Flörke, and A.B. Cremers. Searching and browsing collections of structural information. In *Proc. of the IEEE Advances in Digital Libraries*, pages 141–150, USA, May 2000.