

# Deja Vu: A Knowledge-Rich Interface for Retrieval in Digital Libraries

**Andrew S. Gordon**

The Institute for the Learning Sciences  
Northwestern University  
1890 Maple Avenue  
Evanston, IL 60201 USA  
+1 847 491 3500  
gordon@ils.nwu.edu

**Eric A. Domeshek**

Alphatech, Inc.  
50 Mall Road  
Burlington, MA 01803 USA  
+1 617 273 3388  
domeshek@alphatech.com

## **ABSTRACT**

Providing access to digital libraries will require interfaces that effectively mediate between the retrieval needs of library users and the materials that the library has to offer. This paper describes *Deja Vu*, a new interface for retrieval in digital libraries. Rather than relying on traditional query-based techniques, *Deja Vu* allows users to browse through the subject terms used to catalog library materials to find ones that meet their particular retrieval needs. The browsing process is facilitated by a new knowledge structure introduced in this paper called *Expectation Packages*. *Expectation Packages* group together subject terms based on the commonsense knowledge of library users to provide a richly interconnected browsing space. An example application of *Deja Vu* is described, which incorporates the Library of Congress Thesaurus for Graphic Materials to provide access to online image collections.

## **Keywords**

Retrieval Interfaces, Digital libraries, Thesaurus Browsing

## **INTRODUCTION**

The exciting new possibilities presented by the growth of online media collections and digital libraries bring with them a new set of challenges to overcome. While there are certainly technical problems involved in the storage and delivery of digital media that we have yet to solve, some of the most challenging problems are less obvious. Online collections fundamentally change the way that users search through a library's offerings. Users that can access libraries only through computers do not have the opportunity to browse through shelves or stacks, or talk to a reference specialist who could help mediate between their needs and what the library has to offer. Online catalogs and computer search tools, which were previously only a part of a library's rich set of access methods, become the only retrieval means offered to collection users.

This new reliance on online catalogs and computer search tools continues to push research in the area of Information Retrieval (IR). Although the multimedia nature of modern online collections prohibits the use of some IR techniques (those designed solely for text-based collections), many researchers have been exploring approaches that operate on the indexes used to catalog media materials rather than the full text of documents. One tactic has been to design systems that utilize the controlled indexing vocabularies, or thesauri, that library archivists have developed to aid in the cataloging of library materials by subject matter. Thesauri have become numerous and well utilized in the last decade; commonly used thesauri include the National Library of Medicine's Medical Subject Headings, the Library of Congress's Thesaurus for Graphic Materials, and the Department of Education's Educational Resources Information Center Thesaurus.

Early attempts at utilizing thesauri in IR systems has focused on automated methods for calculating and minimizing the distance between users' text-based input queries and archive materials. Kim and Kim [4] have developed techniques for using the hierarchical structure of thesauri to compute distance metrics between queries and archive materials. Their work advances the techniques employed by Rada and Bicknell [6] by assigning weights to terms and relationships in the thesaurus which are used to produce distance values that more closely approximate human assessments. Chen et al. [1] describe an IR system that utilizes both manually constructed thesauri and algorithmically generated thesauri in one integrated system. By using a blackboard-based design and a spreading activation algorithm, the system elaborate users' queries to match collection materials.

Recently, automated methods for thesaurus navigation have been criticized. Jones et al. [3] offer studies which cast doubt that automated thesaurus navigation techniques could mirror human search behavior, and argue instead for retrieval interfaces that give users tight control over the process of thesaurus navigation and plenty of appropriate terms to choose from. Johnson and Cochrane [2] and Schatz et al. [8] follow this advice in their retrieval systems, which employ thesauri as browsing spaces for user-directed

exploration of the terms used to catalog archive materials. The system described by Schatz et al. uses statistically calculated lists of related terms to supplement the browsing space provided by the manually constructed thesauri. While enriching the browsing space with relevant related terms greatly improves the ability of users to traverse the thesaurus network, the statistical techniques employed by Schatz et al. operate on text-based collections and produce links to terms that are not part of the thesauri, prohibiting their use in collections of non-textual material.

This paper describes *Deja Vu*, a new interface for retrieval in online digital libraries containing digital media that has been cataloged using a thesaurus. *Deja Vu* allows users to browse through thesaurus terms to find those that meet their retrieval needs and that are being used to catalog available materials. The thesaurus browsing space is enhanced by a new knowledge structure introduced in this paper, the Expectation Package, which groups related subject terms in a manner that captures the common-sense knowledge of a library's users. The use of Expectation Packages offers a knowledge-based approach to supporting the browsing process in the information retrieval task.

#### **DEJA VU INTERFACE OVERVIEW**

In *Deja Vu*, users interact with an archive by browsing through the subject terms used to catalog the library's materials. After locating an area of the thesaurus network that is related to their area of interest, users traverse the links between terms to locate those that both meet their particular needs and are being used to catalog available materials. The standard relationships provided by thesauri, which include Broader, Narrower, and Related Term links, are supplemented with links to Expectation Packages (described in the next section) which group together sets of relevant related terms. With tight integration between the thesaurus and the library records, *Deja Vu* provides the user feedback necessary to support a method of retrieval that more closely approximates the process of browsing through a collection rather than constructing a database query.

Figure 1 shows the main screen of the *Deja Vu* interface as presented to a user searching for materials in a digital library. The screen is divided vertically into two main sections. The top section organizes and displays the thesaurus subject terms, and allows users to browse through these terms to locate the ones in which they are interested. At any given time, there will be one subject term, the *focus term*, which is displayed near the upper left corner of this section (*airplanes* in Figure 1). All of the other lists displayed in the top section contain subject terms that are related to the focus term in some way. The left side of the top section contains lists for all of standard thesaurus relationships, as well as any notes attached to the focus term by thesaurus developers. The right side of this section displays the Expectation Packages associated with focus term, which are described in the next section. The bottom section provides users with a way of directly accessing

digital materials indexed by terms presented in the top section.

To understanding how to use the *Deja Vu* interface, it will be useful to examine each of the interface components with regard to the functions that they support. As an example, consider how the *Deja Vu* interface would be useful in helping a user of an online image collection find images for use in a slide show that they were creating on the topic of advances in the aviation industry.

*Specifying an area of interest:* The first task that users of *Deja Vu* face is to let the system know what general topic area they are interested in. To do this, users must identify a subject term in the thesaurus that is somehow related to their retrieval needs, and select it as the focus term. For this purpose, users can set the focus term to any term in the thesaurus in one of two ways. First, users can select a specific term from an alphabetical list accessible as a drop-down menu attached to the interface item that displays the current focus term. Second, users can use the keyboard to search for terms containing specific text. Entering a string of characters causes the system to list all of the thesaurus terms that contain the string, any of which can be selected as the new focus term. For our example user who is looking for images on the topic of advances in the aviation industry, typing the characters "airplane" would generate a list which would include the term *airplanes*, which has been selected by the user as the focus term in Figure 1.

*Browsing through thesaurus terms:* Once the user has changed the focus term to one that is in their area of interest, their next task is to examine the other terms that are linked to the focus term to locate the ones that are the most appropriate for their retrieval task. If unsatisfied with the choice of subject terms presented, users can browse to a related area of interest by changing the focus term. Double-clicking on any of the terms displayed to the user will cause it to become the new focus term, and each of the associated lists of terms will be updated. Our example user who is looking for images on the topic of advances in the aviation industry is presented with a large number of associated subject terms when the term *airplanes* is in focus. However, they could choose to change the focus term to something more specific to their interests, such as the Related Term, *airline industry*, or the Expectation Package term, *air travel*, by double clicking on either of them.

*Selecting terms to retrieve materials:* Subject terms that are being used to index materials in the library's holdings display an asterisk next to the text of the terms. These terms can be selected by the user in order to retrieve the materials that use the term as an index. When the user finds a term that is both selectable and meets their retrieval needs, the term can be selected by clicking on the term where it is displayed in the top section, and pressing the *Add Selected Term* button. This action adds the selected term to the list at the left side of the lower section. On the right side of this section, the titles of all of the library items

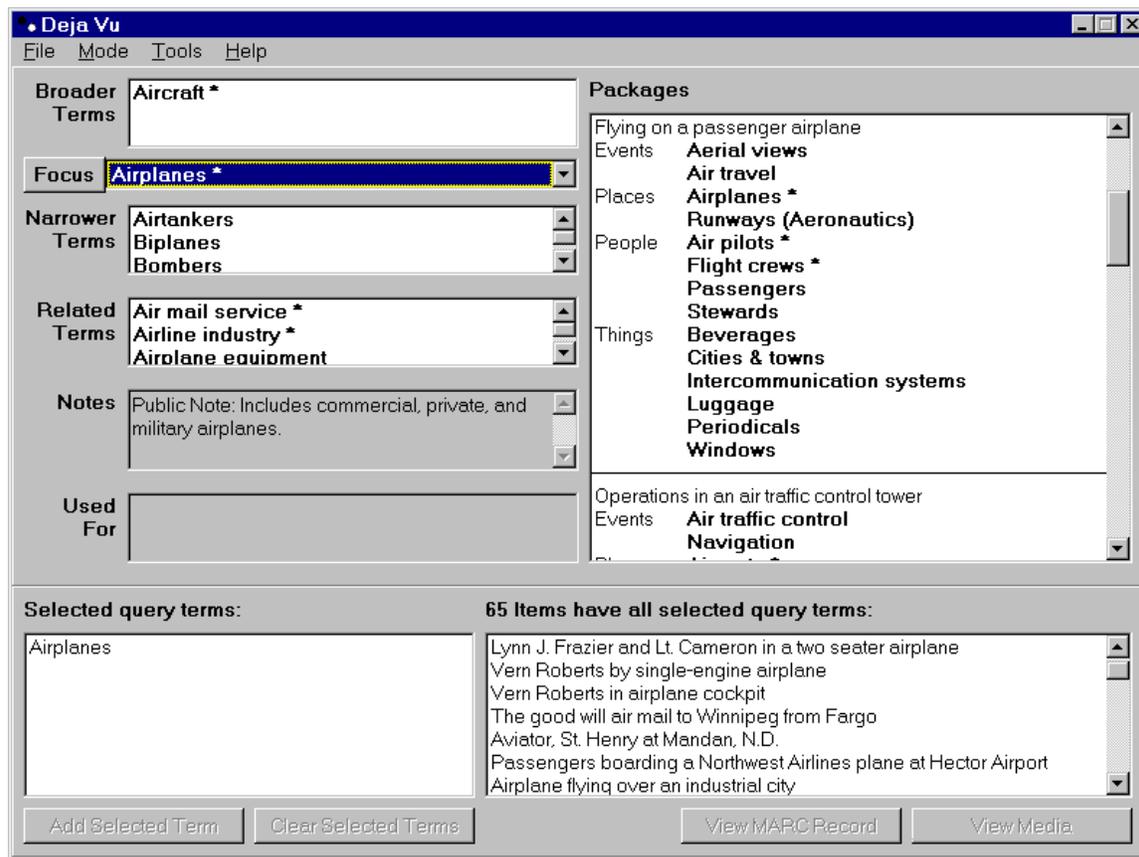


Figure 1. The DeJa Vu Retrieval Interface

that use the selected term as an index are immediately displayed. Figure 1 shows what the interface would look like after our example user selected the subject term *airplanes*.

*Narrowing the set of retrieved materials:* Often a single index term will result in the retrieval of too many library items. DeJa Vu allows users to continue to browse for subject terms which can be used to narrow the set of retrieved items to a more reasonable number. To make browsing for narrowing terms easy, DeJa Vu automatically updates the display of asterisks when a term is added or removed from the set of selected terms. Asterisks are only displayed next to subject terms that are being used as indexes for some subset of the materials that have been returned by the current set of selected terms. Since only terms displayed with asterisks can be selected, the user is prevented from over-specifying their request by selecting an unavailable conjunction of terms. Notice that in Figure 1, the 65 items retrieved by the term *airplane* could be reduced to a smaller number by selecting one of the other terms marked with an asterisk (*aircraft*, *air mail service*, *airline industry*, *air pilots*, or *flight crews*). Our example user could choose one of these terms, or change the focus term to another term such as *runways (Aeronautics)* in order to find selectable terms related to this topic that they could use to narrow the retrieved set.

*Selecting materials to view:* In order to view one of the items returned by the selected terms, the user simply clicks on one of the titles presented to them in the lower right and presses the *View Media* button. This action can have a number of different effects depending on the digital media type of the item and how DeJa Vu is connected to the library's digital records. DeJa Vu is set up to launch a helper application to display digital media items that exists as accessible files. A number of device drivers can be used as well to control various multimedia devices such as laserdisc players in order to access materials not stored as computer files. For our example user, selecting on one of the titles (as seen in Figure 1) such as *Airplane flying over an industrial city*, and pressing the *View Media* button would cause DeJa Vu to load the digital image file for this record and display it using a graphics viewer application. Alternatively, the catalog record for the retrieved item can be viewed by pressing the *View MARC Record* button. After this, the user can continue viewing other returned items, or retrieve different items by adding or removing subject terms from the selected list.

#### EXPECTATION PACKAGES: GROUPING THESAURUS TERMS USING AN ORGANIZATIONAL THEORY

The utility of a browsing interface depends entirely upon the quality of its links. In DeJa Vu, where users traverse a

browsing space of thesaurus terms, it is necessary to provide users with a rich set of interconnections between terms to support the browsing process. There are three primary browsing functions that these interconnections serve. First, the browsing space must provide a set of links that a retrieval user can traverse in order to locate a term when they do not know its particular textual label. This allows the user to locate the best term for their retrieval needs once they have entered a portion of the browsing space that is close to their area of interest. Second, the browsing space should offer users reasonable possibilities that they hadn't considered. When presented with terms that are closely related to the ones they were initially going to use, users may refine or change their retrieval goals to take advantage of an opportunity facilitated by a rich browsing space. Third, the browsing space must offer users alternative terms when the ones they hoped to use do not retrieve adequate materials. A good browsing space will link the desired term to all of the terms in the thesaurus which may be used to index other materials which are relevant.

These functions each require that the browsing space be well organized and richly interconnected. However, the degree of interconnections found in existing thesauri often leave much to be desired. A recent release of the Library of Congress's Thesaurus for Graphic Materials (LCTGM) consisted of 4,421 reflexive taxonomic links (Broader and Narrower term pairs) and 5,888 reflexive associative links (Related term pairs) [5]. Added together, there are a total of 20,618 unidirectional links divided amongst 5,760 authorized terms, or only 3.58 links per term on average. If existing thesauri like the LCTGM are to be used as a basis for browsing interfaces to digital libraries, the degree of interconnections between terms must be improved.

Designing appropriate and numerous links between thesaurus terms is a labor-intensive knowledge engineering task, even for a moderately-sized thesaurus like the LCTGM. The best that developers can hope to do is maximize the payoff of every bit of knowledge engineering that goes into interconnecting thesaurus terms. One way to greatly reduce the amount of work necessary to provide a rich browsing space is to change the way that developers link terms. Currently, thesaurus developers generate one-to-one connections between thesaurus terms (specified as lists of Broader, Narrower, or Related Terms). An alternative approach is to generate connections by clustering thesaurus terms into fully connected groups. That is, designers can identify groups of terms where each member of the group should be linked with every other member. For every fully-connected cluster that is assembled of size  $n$ , the number of new unidirectional links created is equal to  $n(n - 1)$ . In contrast, identifying  $n$  number of one-to-one relationships generates only  $2n$  new unidirectional links (assuming thesaurus links are reflexive).

Clustering thesaurus terms to create richly interconnected browsing spaces requires an organization theory to assist knowledge engineers in determining which terms should be grouped together. We propose that different organizational theories may be appropriate for different thesauri, but that each should be based on the commonsense expectations that library users have about the relationships between thesaurus terms. These clusters, which we refer to as Expectation Packages, should group together terms based on our best understanding of how their corresponding concepts are represented and organized in the minds of library users. Accordingly, creating Expectation Packages is a process similar to the knowledge representation task in the field of Artificial Intelligence, where theories of memory organization guide the representation of knowledge used in intelligent reasoning systems. An example of how a particular organizational theory is applied to the development of a full set of Expectation Packages for a thesaurus is described in the next section.

In *Deja Vu*, an Expectation Package is displayed to a user when any of its member terms are selected as the current focus term. Figure 1 shows an example of an Expectation Package, entitled *Flying on a passenger airplane*, which appears because one of its member terms, *Airplanes*, is the current focus term. In *Deja Vu*, each Expectation Package is represented and displayed as a simple structure consisting of a distinguishing textual title and a list of thesaurus terms. The list of thesaurus terms is divided into a set of labeled sub-lists, which categorize the terms by the roles that they play in the Expectation Package. In the example Expectation Package displayed in Figure 1, these sub-lists are labeled *events*, *places*, *people*, and *things*. The sub-lists are displayed in a fixed order for each Expectation Package to provide a consistent layout. These sub-lists are constant for each of the Expectation Packages developed for a given thesaurus, and are determined by the particular organizational theory used to guide Expectation Package development.

The *Deja Vu* system incorporates a set of tools to be used by knowledge engineers for the purpose of creating a set of Expectation Packages for a particular thesaurus. The basic process for creating Expectation Packages requires that knowledge engineers analyze each of the thesaurus terms to determine if they belong to an Expectation Package which has not yet been created. Each new Expectation Package is specified in full when the first member term is identified, which promotes a development cycle where the majority of Expectation Packages for the entire thesaurus are created after the first fraction of terms have been analyzed. That is, in the beginning of the process, each new thesaurus term may spark the creation of several new Expectation Packages, but later in the process, the remaining terms will have already been assigned to appropriate Expectation Packages created earlier.

## EXPECTATION PACKAGES FOR THE LCTGM

As an example of the use of Deja Vu with a standard thesaurus, we have outfitted Deja Vu with the subject terms from the LCTGM and created a full set of Expectation Packages to enhance the browsing space that it provides. The resulting Deja Vu system offers an effective retrieval interface for any visual media archive that uses the LCTGM for cataloging purposes.

Developed by the Library of Congress Prints and Photographs Division, the LCTGM contains over five thousand authorized subject terms for indexing graphic materials such as photographs, drawings, posters, cartoons, and other still image media formats. It contains terms that refer to visual elements that can be found in graphic materials, including people, place, things, and activities, as well as the topics or abstract ideas which these materials address. The vocabulary of the LCTGM was developed on an as-needed basis while indexing materials in the Prints and Photographs Divisions' collections and continues to grow with contributions from the many institutions that use the LCTGM for their own image collections.

To guide the development of Expectation Packages for the LCTGM, an appropriate organizational theory was necessary. Given the topic area of the LCTGM, we needed a theory of memory organization that captured the interrelationships between everyday activities, things, people, and places in the world. We selected Schank and Abelson's theory of scripts [7] which postulates that our expectations about everyday activities, things, people, and places are organized by representations of culturally shared knowledge about common and ritualized activities. Initially developed as a means of controlling inference in natural language processing programs, the theory of scripts is well suited to the task of guiding Expectation Package development.

The original representations of scripts were very detailed, as they were to be utilized directly by natural language processing programs. Each script described the series of events that would take place in common activities such as going to a restaurant for a meal, attending a wedding at a church, or flying on a commercial airplane. Turning these rich representations into Expectation Packages requires that knowledge engineers identify thesaurus terms which are semantically identical to the concepts contained in these scripts. The thesaurus terms that correspond to concepts in a particular script are then categorized into sub-lists which identify the role that each term plays in the Expectation Package. For the Expectation Packages developed for the LCTGM, the five roles were used to categorize thesaurus terms. These were *Places*, *People*, *Things*, *Events*, and *Misc*, each of which correspond to components of the original representations of scripts.

Figure 2 shows a collection of sample Expectation Packages of LCTGM terms in the Deja Vu system. What unifies these examples is that they all include the LCTGM

term *coins* as a member. Accordingly, each of these would appear in their entirety in the list of packages displayed to the user when the subject term *coins* was the current focus term.

A total of 770 Expectation Packages were created for the LCTGM. The development work was done by one computer science graduate student working half time over four months (approximately two person-months). On average, each Expectation Package contained 12 terms. The set of Expectation Packages generated just over 100,000 unique unidirectional links, adding an average of over 17 additional links for each authorized term in the thesaurus. While no attempt was made to verify the psychological validity of any of the particular Expectation Packages that were created, an effort was made to exclude Expectation Packages that were not thought to be shared by members of American or Western culture.

## INSTALLATIONS AND EVALUATIONS OF DEJA VU

For the purpose of evaluative user testing, Deja Vu has been installed at two sites containing large collections of digital images. The first installation is at the North Dakota Institute for Regional Studies (NDIRS) at North Dakota State University where Deja Vu is being used as an interface to over 11,000 images captured on a video laserdisc. The second installation is at the Library of Congress Prints and Photographs Division in Washington, DC, where Deja Vu is being evaluated as an interface to a collection of over 25,000 digital images accessible online as part of the National Digital Library initiative.

### Installation at NDIRS:

NDIRS is a moderately sized historical archive associated with the library of North Dakota State University. It is dedicated to the preservation of the history and culture that is unique to the northern plains region of the United States. The institute's photographic archive is used by the university and the general public primarily for the purposes of historical research, commercial advertising, and book publishing. In 1988, the institute completed the production of a video laserdisc containing more than 15,000 of the 50,000 historical photographs in the institute's collection, each captured as an individual video frame and accessible by serial frame numbers. After completing the video laserdisc, the institute began the process of indexing each of the images on the video laserdisc using the LCTGM as the main source of subject terms. When Deja Vu was first installed at the institute in March of 1997, more than 11,000 of the images had been indexed.

After the institute provided the MARC records for their video laserdisc collection, Deja Vu was installed on a computer at the institute where it could be used both by the institute's staff and by the general public. To integrate the Deja Vu interface with the images available on the video laserdisc, custom software was developed to allow computer control of the institute's laserdisc player for instant access to the collection.

<p><b>Begging for money on a city street</b></p> <p>Events <i>Pleading (Begging)</i></p> <p>Places <i>Business districts</i> <i>Commercial streets</i></p> <p>People <i>Amputees</i> <i>Beggars</i> <i>Blind persons</i> <i>Homeless persons</i> <i>Mentally ill persons</i></p> <p>Things <i>Coins</i> <i>Tin cups</i></p> <p>Misc <i>Charity</i></p>	<p><b>Going gambling in a casino</b></p> <p>Events <i>Card games</i> <i>Gambling</i> <i>Victories</i> <i>Wagers</i></p> <p>Places <i>Casinos</i> <i>Flatboats</i></p> <p>People <i>Losers</i></p> <p>Things <i>Coins</i> <i>Crossed fingers</i> <i>Electric signs</i> <i>Playing cards</i> <i>Slot machines</i></p> <p>Misc <i>Organized crime</i> <i>Wealth</i></p>
<p><b>Getting a soft drink or snack out of a vending machine</b></p> <p>Places <i>Automobile service stations</i> <i>Cafeterias</i></p> <p>Things <i>Aluminum</i> <i>Candy</i> <i>Cans</i> <i>Carbonated beverages</i> <i>Chewing gum</i> <i>Vending machines</i></p> <p>Misc <i>Beverage industry</i></p>	<p><b>Making a wish at a wishing well</b></p> <p>Events <i>Wishing</i></p> <p>Places <i>Wishing wells</i></p> <p>Things <i>Coins</i> <i>Pails</i> <i>Ropes</i></p> <p>Misc <i>Magic</i></p>
<p><b>Going to the bank to make a transaction</b></p> <p>Events <i>Banking</i> <i>Document signings</i> <i>Saving &amp; investment</i></p> <p>Places <i>Banks</i></p> <p>People <i>Bankers</i> <i>Guards</i></p> <p>Things <i>Cash registers</i> <i>Coin counting machines</i> <i>Coins</i> <i>Identification photographs</i> <i>Money</i> <i>Security systems</i> <i>Wages</i></p> <p>Misc <i>Paydays</i> <i>Wealth</i></p>	<p><b>Taking your dirty clothes to the laundromat</b></p> <p>Events <i>Cleaning</i> <i>Laundry</i></p> <p>Places <i>Apartments</i> <i>Laundries (Rooms &amp; spaces)</i></p> <p>Things <i>Baskets</i> <i>Clothing &amp; dress</i> <i>Coin operated machines</i> <i>Coins</i> <i>Household soap</i> <i>Washing machines</i></p> <p>Misc <i>Time</i></p>

**Figure 2.** Example Expectation Packages for the LCTGM.

### **Installation at the Library of Congress:**

The Library of Congress Prints and Photographs Division is the United States' premier archive for historical research involving graphic materials, with an estimated 13.5 million prints and photographs in its holdings. The Prints and Photographs Division has a sizable staff, which includes reference specialists who assist library users in locating materials, catalogers who process the materials in the library's collection, as well as a staff of collection curators. The collections are utilized primarily by historical researchers, publishers, and advertisers, who gain access to the collections via a variety of different electronic and traditional library search tools.

In order to reduce damage to prints and photographs due to handling, the library has invested heavily in digitizing its collections so that they will be available for online access in conjunction with the National Digital Library initiative. The Prints and Photographs Division provided the MARC records for one of their largest digital collections consisting of over 25,000 photographs of American life from the years 1880 to 1920 obtained from the Detroit Publishing Company. Deja Vu was then installed at the Library of Congress where it is accessible to reference librarians assisting library users, as well as other library staff members. Deja Vu utilizes pointers to online images provided in the MARC records to allow direct access to the collection over the Internet using a standard web browser interface.

### **Evaluation of Deja Vu:**

To assess the value of Deja Vu, we have compiled the written comments provided to us by the catalogers and reference specialists who have examined and utilized Deja Vu system at these two institutions. The extent to which Deja Vu has been used by this population and the users that they support at these institutions is very different. At NDIRS, where the previously available search resources were modest, Deja Vu has been integrated as a primary tool to aid reference specialists and collection users in finding photographic materials. At the Library of Congress, whose large staff was well equipped to handle user requests before the introduction of Deja Vu, the system has not been significantly utilized as a search resource. Although the extent of use has differed greatly between these two institutions, the comments of the catalogers and reference specialists at both institutions were consistent with each other.

The positive comments of our evaluators center primarily around the interface functionality of Deja Vu. Library catalogers and reference specialists typically must cope with a sharp division between library materials and their cataloging records. In Deja Vu, the tight integration of cataloging information with digital media is viewed as the primary advantage of the system over existing search tools at these institutions. The functionality that this tight integration affords, mainly that users of the system can

immediately see what is available to them in the archive while searching through the thesaurus terms, is seen as a significant improvement over current access techniques. In addition, our evaluators feel that the process of browsing through thesaurus terms improves users' understanding of the relationship between archive materials and the cataloging resources that archivists employ.

The primary criticisms of the Deja Vu system focus on its limited application to the full range of requests that reference specialists must service. Since Deja Vu was designed to facilitate topic-based searching of archive materials, it is of very limited value to users looking for materials created by a specific person, whose subject matter is best referred to by proper name, or that are distinguished by some information concerning their production. These types of retrieval needs account for a significant portion of the requests that these two institutions receive, and Deja Vu did little to improve access in these cases. Furthermore, because Deja Vu operates using a single thesaurus (in this case, the LCTGM) it does not effectively support topic-based searching for materials whose cataloging may have been supplemented with terms from other major thesauri or (in the case of NDIRS) with terms that were created locally to meet retrieve needs specific to a particular collection.

Our evaluators agreed that the Expectation Packages provided for the LCTGM enriched the browsing space significantly. When reference specialists offered Deja Vu as a tool to library users, they described Expectation Packages as an extension of the set of Related Term links. Our evaluators report that they (and the users they support) found the Expectation Packages to be intuitive and helpful for finding particular terms that they would not have otherwise thought of. However, our evaluators were critical of the subjectivity of the provided set of Expectation Packages, which contrasts with the rather objective nature of the standard relationships found between thesaurus terms. Also, the Expectation Packages seem to capture a modern view of the world that is discordant with the historical nature of the particular collections used in Deja Vu's evaluation. In addition, several evaluators expressed concern that the Expectation Packages would be difficult to maintain given that thesauri like the LCTGM are constantly being revised with new terms and new organization.

### **CONCLUSIONS**

The success of digital libraries will depend heavily on the quality of the retrieval systems provided to their users. The best retrieval systems will offer online users retrieval tools that adequately replace the rich set of retrieval resources offered to visitors of traditional libraries. The ability to browse through thesaurus terms used to catalog library materials gives users a way to develop and satisfy their retrieval needs by considering the materials that are available in online collections. The design of the Deja Vu system incorporates the functionality and content necessary to facilitate this method of retrieval for existing collections.

The major challenge in developing systems that allow users to browse through thesaurus terms is offering them a rich and intuitive browsing space to traverse. The links provided by existing thesauri are often too sparse to support the tasks that a browsing space must service. The use of Expectation Packages to cluster thesaurus terms provides us with the means of creating a rich and well-organized browsing space given a tractable amount of knowledge engineering effort. By applying our best theories of human memory organization and representation as organizational theories for Expectation Packages, we can work to organize thesaurus terms in a manner that is most intuitive for library users.

The Deja Vu system demonstrates how IR problems can be addressed by knowledge-rich approaches. By integrating knowledge representations directly into the interface to be utilized by users, Deja Vu provides an intelligently designed browsing space without requiring an intractable amount of representational complexity. The installation and evaluation of Deja Vu at NDIRS and the Library of Congress encourage its use at other institutions that utilize the LCTGM for cataloging purposes, and the development of Expectation Packages for other widely used thesauri.

#### **ACKNOWLEDGMENTS**

The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting. Special thanks to Arden Alexander at the Library of Congress Prints and Photographs Division and John Bye at the North Dakota Institute for Regional Studies for providing materials for this research. Also thanks to Andersen Telemedia and all of the people that helped develop the earlier versions of Deja Vu.

#### **REFERENCES**

1. Chen, H., Lynch, K. J., Basu, K., Ng, D. T. (1993) Generating, Integrating, and Activating Thesauri for Concept-based Document retrieval. *IEEE Expert*, 8(2):25-34, April 1993.
2. Johnson, E., and Cochrane, P. (1995) A Hypertextual Interface for a Searcher's Thesaurus. Editors: Shipman, F., Furuta, R. and David M. Levy, in *Digital Libraries '95*, Austin, June 11-13, 1995.
3. Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J., Walker, S. (1995) Interactive Thesaurus Navigation: Intelligent Rules OK? *Journal of the American Society for Information Science*, 46(1):52-59, 1995.
4. Kim, Y. W., and Kim, J. H. (1990) A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph. *Journal of Documentation*, 46:113-116, 1990.
5. Library of Congress Prints and Photographs Division. (1995) *Thesaurus for Graphic Materials*. Washington, DC: Library of Congress Cataloging Distribution Service.
6. Rada, R., and Bicknell, E. (1989) Ranking documents with a thesaurus. *Journal of the American Society for Information Science*, 40(5), 1989, 304-310.
7. Schank, Roger C. and Abelson, Robert P. (1977) *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
8. Schatz, B.R., Johnson, E., Cochrane, P., and Chen, H. (1996) Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval, *Digital Libraries '96*, Bethesda, MD, March 20-23, 1996.