Automated Access to the NASA-JSC Image Archives

GARY A. SELOFF

ABSTRACT

THE FILM ARCHIVES at NASA's Johnson Space Center in Houston, Texas, houses the still-frame and motion picture documentation of the U.S. Manned Space Flight Program. While few people would have difficulty recognizing the milestone images stored here, the majority of the collection serves a variety of very specific interests. Access to these images has long suffered from limited intellectual control, but currently an effort is underway to remedy this situation. This article will provide background information on the recent state of the Film Archives and will discuss the goals and methodology of the effort to establish intellectual control over the storage and retrieval of images in this collection.

INTRODUCTION

Prior to, and particularly following, the explosion of the space shuttle, Challenger, in January 1986, it had become apparent throughout the National Aeronautics and Space Administration (NASA) that its visual archives were growing out of control. The instant and urgent requirement for all relevant comparative material to the Challenger launch sequence and the failed booster rocket could not be adequately met by the manual retrieval systems highly dependent on the corporate memories of a few dedicated individuals. Image collections, once manageable, had become large and unwieldy.

Gary A. Seloff, TGS Technology, Inc., JL-3/TGS, Building 8, Room 244, NASA, JSC, Houston, TX 77058
LIBRARY TRENDS, Vol. 38, No. 4, Spring 1990, pp. 682-96

The call for a new, automated approach to image management—a space-age solution—was one of the many remedies identified in the analysis of this country's worst space-age disaster.

Since 1986, solutions to the agency-wide problems of image access have been approached mostly piecemeal on a center by center basis. Several facilities quickly purchased off-the-shelf technology, gambling that it could be developed into long-term solutions. At the Lyndon B. Johnson Space Center (JSC) in Houston, Texas—the primary location of the collection of manned space flight imagery—the administration opted to study various strategies in the hopes that a universal solution would emerge.

The automation goals for the JSC film repository are: (1) to provide a standardized method for computer-assisted description and retrieval of visual materials, incorporating the wealth of corporate knowledge of the collection; (2) to create a system less labor-intensive than the manual precursor; and (3) to provide center-wide availability to electronic text and image records, accessible to knowledgeable staff members and novice end-users alike. This article describes the JSC film repository, the development of its automation strategy, and the implementation to date of its automated procedures for image access.

Description of the Collection

The JSC Film Repository houses a collection of more than 1 million negatives and transparencies, as well as around 10,000 motion picture and audio reels, documenting all aspects of the manned space flight program in this country since 1958. It is administered by TGS Technology, Inc., for NASA-JSC's Photography and Television Technology Division (PTTD). Materials archived in the repository are received from a range of sources including actual manned space flight missions, various NASA-JSC divisions, NASA contractors, other NASA centers, and other government agencies. While few people would have difficulty identifying the milestone images stored here, the majority of the images are of very limited or narrow interest. In addition to the widely disseminated space flight views, there is a broad range of subjects such as engineering studies, lunar rock samples, earth observation views, employee award ceremonies and activities, JSC facilities, distinguished visitors, astronaut portraits and astronaut training, manufacturing close-out photography, etc. New images are received at a rate of 25,000 to 65,000 per year depending on the number and nature of missions flown. Film formats include negatives and transparencies, color and black and white, in sizes ranging from 16mm to 8 × 10 inch viewgraphs to 9 inch wide rolls.

While the repository is a circulating collection—items circulate primarily to three on-site photographic labs operated by TGS Technology—it is not a browsing collection. Items are physically organized in large, rotating power files, strictly in numerical sequence by accession number, with no subject grouping. Images are cataloged according to two basic systems. On-board flight images (most of which are earth observations) are described by photographic interpreters using a very terse vocabulary and a relatively sparse format denoting latitude, longitude, altitude, date and time, and a brief reference to geographic locations or features. The resulting scene lists are indexed and published in single volumes by mission.

Nonflight images are initially described in the photographic work order which directs the photographer in acquiring images. After the processed negatives are sent to the film repository, catalogers follow several steps to gather more information, beginning with a visual inspection of each image, a review of related image descriptions, and contact with the requester of the image. A catalog entry is then created which includes the accession number, acquisition date, a physical description of the film type and format, key filing and cross-reference categories, and a free-form textual description of the participants and activities depicted in the scene. As recently as mid-1987, sets of cards were then manually produced and filed in an alphabetical cross-referenced card file system. (This was the first task to be automated utilizing a microcomputer and a traditional database management system.)

The Public Affairs Office (PAO), which is the liaison office between JSC and the general public, including the news media, is the largest requestor of repository images. A majority of their requests to the repository directly reference the image accession number; however, roughly one-third of the requests being made by PAO require research assistance to locate an appropriate subject. In addition, a number of requests from other sources are also received with only varying degrees of content description to indicate the desired image.

Prior to the implementation of automated systems in the summer of 1987, all requests were met through manual methods based on the 100,000 item catalog card file and the expertise of one long-time cataloger. Considering the number of requests and the size of the manual card file, the success rate in locating appropriate images was quite high; however, so was the response time as well as the cost per located image.

State of the Question

With the growing number of images and catalog cards, and with the inevitable retirement of the expert cataloger looming in the future, the manual image management system was near collapse. Serious dialogue regarding automated image retrieval systems was initiated at JSC in late July 1986 with a call for a working prototype by mid-October. That schedule proved a bit ambitious: more than three years later, the system is only now becoming a reality. In the interim, countless meetings and planning sessions, proposals, demonstrations, conferences, seminars, and small-scale projects served to educate and inform a group of image-user representatives from around the center. This group coalesced to form an ad hoc steering committee to pursue automation goals.

Expert assistance has come from the academic world at several junctures. In December 1986, an agency-wide laserdisc symposium was held in Boston with invited speakers from MIT and Simmons College, among others. A consulting team from these two schools was subsequently selected by the Photography and Television Technology Division at JSC to do a feasibility study for the implementation of an electronic visual database system. The follow-up report to the consultants' intensive two-day visit to JSC in January 1987 served to crystallize the ideas and tentative plans of the ad hoc committee. The recommendations laid out in that report clearly emphasized intellectual considerations (classification and indexing schemes) over technical considerations (hardware) as a critical factor determining flexibility for the widest range of end users. The report suggested the hiring of a consulting information expert, the development of a hierarchical thesaurus of space-related terminology, and a thorough needs assessment as practical first steps prior to actual hardware and software selection.

A poll of various potential users of an electronic image retrieval system at JSC revealed a broad interest in such a system with emphasis on image transmission to various locations and the ability to matrix multiple images on a single display screen for comparison. Projects with similar scope and retrieval criteria were currently under development at MIT (Project Athena), and The University of California at Berkeley (ImageNet). Work in progress at each center demonstrated that the technical capability to manage images in the desired manner was available, if not yet cost effective. However, there was no evidence that an innovative and effective new method for describing and retrieving images had evolved.

The inherent problem of image retrieval within many visual archives—including the NASA-JSC film repository—is both the sparsity and inconsistency of textual descriptors evoked by the visual content of the image. In the JSC film repository, where the cataloging has been done for over a decade by one competent individual, similar images separated by time can vary widely in their textual description. Partly, this can be attributed to evolving vocabulary as programs within the agency mature, but primarily it is due to the highly subjective nature of the task of describing image content. The viewpoint of the cataloger invariably changes from one week to the next and is always different from the perspective of the engineer or the scientist. To locate an item in such a system, the desired image must evoke at least a subset of the descriptors assigned by the cataloger. The wider the disparity in the points of view (for example, the engineer requests an image of a misaligned mounting bracket on a flight hardware subassembly which only exists as an image described by a cataloger as astronaut training aboard the zero-gravity aircraft), the less likely the appropriate item will be retrieved.

Projects under the auspices of the Getty Center (*Art and Architecture Thesaurus*) and the Library of Congress (*MARC for Visual Materials*) evidence a recognized need for standards in image description, but no current system was available for uniformly classifying space-related images or for providing an appropriate retrieval interface for the variety of potential users at JSC. Again, support was enlisted from the academic world with the award of a grant to fund an unsolicited proposal from the Graduate School of Library and Information Science at The University of Texas at Austin. This proposal to design and develop an automated image management system for the JSC archives was based on prior research undertaken through Project Icon at the University of Texas, directed by Mark Rorvig. (Project Icon team members involved with the NASA Visual Thesaurus project, working under the direction of Mark E. Rorvig, include Chris Ladoulis, David McClelland, Jesus Moncada, Richard Reed, Jeff Skaistis, Charles Hudson Turner, and Charles Young.)

Like the projects at MIT and Berkeley, Project Icon is directed toward the automation of image management for academic applications, but its primary focus is on automation of image description and retrieval processes. Part of the experimental work undertaken in the Project Icon lab sought to demonstrate the substitutability of images themselves for the textual descriptions of images in an image retrieval system (Rorvig, 1987). This line of experimentation led to the idea of a "visual thesaurus" as an aid to catalogers and researchers of visual materials, much as a traditional thesaurus is used in textual retrieval systems. While the pragmatists on the ad hoc committee at NASA felt somewhat uncomfortable with some of the theoretical elements of this approach, it presented a relatively low-cost alternative to unproven multimillion dollar turnkey systems from commercial vendors.

Implementation

Meetings between the Project Icon team and the JSC image retrieval committee established the priorities and goals for the automated image retrieval project for the first two years of the grant. Highest on the list was the development of a hierarchical thesaurus tailored to the users of JSC imagery. Once developed, the thesaurus would be incorporated into an easy-to-use, intuitive, system-independent interface. Then, selection would begin for the assembly of hardware and software system components.

For much of the first year, thesaurus development was an all consuming task. Considerable effort was expended to develop or convert algorithms for the automation of the thesaurus term selection processes, based on the work of Gerard Salton (1968). The working vocabulary for this task was drawn from 5,000 randomly selected and machine-scanned cards from the JSC film repository catalog card file (representing twenty years worth of cataloging activity), 5,000 records from a recently implemented computer-based version of the film repository catalog card file, and 4,000 word-processor documents (image descriptions) from the JSC Public Affairs Office.

Some half million terms were processed to derive occurrence counts for each unique term to eliminate noise words (e.g., of, and, the) and to discount terms occurring too frequently or infrequently to be significant. From this analysis, each document was identified by the significant terms it contained, and a term-document matrix was generated. By cross checking each term by each document in the matrix, a coefficient was assigned to all term-pair combinations based on the frequency of co-occurrence of the terms within the documents (McClelland, 1988). Those pairs with coefficients above a set threshold were selected as related terms for the thesaurus. The list of terms and term-pairs was then reviewed manually by a team of four collection administrators and catalogers to delete irrelevant terms and related pairs, and to assign hierarchical relationships. The resulting substantially reduced set of terms and related pairs was then enriched by massaging this list against the much larger *NASA Thesaurus* (1985) of scientific and technical terms to derive additional appropriate related terms from that extensive source.

While work on the thesaurus proceeded, a parallel effort was established to create the user interface incorporating the concept of a visual thesaurus. Acknowledging the progress of the construction of the term thesaurus, the Project Icon team still expressed reservations that any controlled vocabulary could be effectively used in the JSC environment due to the diverse base of images and users. This concern was compounded by limited cataloging resources and by the general idiosyncratic nature of responses of catalogers and users to images in the collection.

Conclusions drawn from experiments testing the substitutability of images for textual descriptions of images in an image retrieval system suggest that human judgments obtained by exposure to images were more robust and more quickly obtained than those derived by exposure to textual descriptions (Rorvig, 1987). This theoretical base supported the proposal to construct a visual thesaurus utilizing equivalent images from the JSC film repository to duplicate the relationships established in the linguistic thesaurus. The two components, if presented together in a single user-interface, would provide the cataloger/searcher with a flexible alternative to the traditional controlled vocabulary utilized for indexing text documents.

Prototyping of the Visual Thesaurus was carried out on an Apple Macintosh II microcomputer with four megabytes of random access memory, utilizing Hypercard® software—an authoring environment well suited to interactive video applications. Two hundred popular images, supplied by the Public Affairs Office, were digitally scanned and linked to associative thesaurus terms. Selection of a descriptive term from the thesaurus retrieved its associated image, as well as broader, narrower, and related terms along with their associated images. The same selections could be triggered by choosing any image on the display screen, or images could be browsed in sequence. In each case, selected terms were added to a temporary buffer which became the search query to be applied against the full textbase of cataloging records or became the key descriptors of a new image being cataloged. Reaction to the prototype by the ad hoc committee was mixed, but eventually the go-ahead was granted to continue this two-pronged approach to system design.

Work continued on the textual thesaurus, while development of the visual interface was switched to a relational database management system called 4th Dimension®, from Acius, Inc. Hypercard had proven an excellent prototyping environment, but the number and complexity of the relationships needed for the actual interface called for a more efficient method of data handling. 4th Dimension offered the same image handling capabilities as Hypercard, combined with a powerful programming language and a flexible data retrieval engine. Several months and almost a hundred pages of codes later, an integrated 4th Dimension application was delivered to the JSC film repository in spring 1989 for beta-testing (see Figure 1).

In this database format, each thesaurus term is a database record with relational links to each broader, narrower, and related term, and to any corresponding digitized images, which are also stored as records. (Since the interface itself is in database form, new thesaurus terms, images, and relationships can be easily added or modified through a maintenance module, shown in Figure 2.) Similar in operation to the prototype, when a user enters a term the system responds with a list of related entries (see Figure 3). Wherever available, the associated images are also displayed. As the user proceeds, one term or image conjures another, thus enriching the descriptive process, and with each selected category, a textual entry is added to a temporary storage buffer. When the process is complete, the buffer is appended to a new cataloging record as a string of keywords, or applied against the database of existing catalog records as a search query, depending on the mode of operation. In fact, until this term/image selection process is completed, there is no distinction between the activity of cataloging versus searching the database.

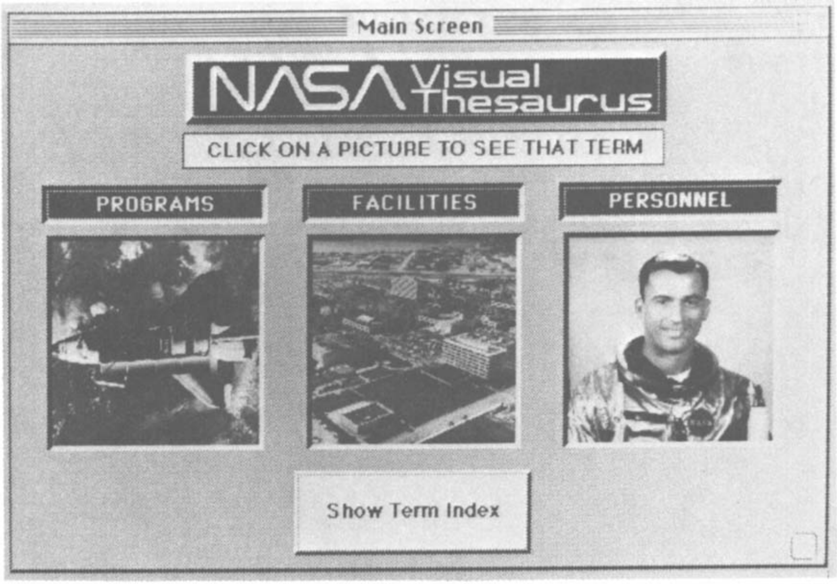
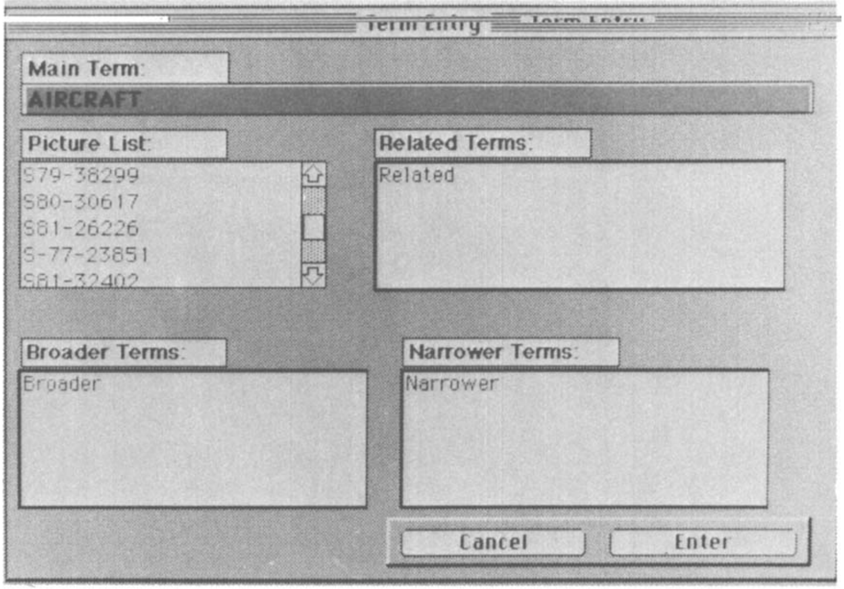


Figure 1. The Main Screen of the NASA Visual Thesaurus



Main Term:
AIRCRAFT

Picture List:
S79-38299
S80-30617
S81-26226
S-77-23851
S81-32402

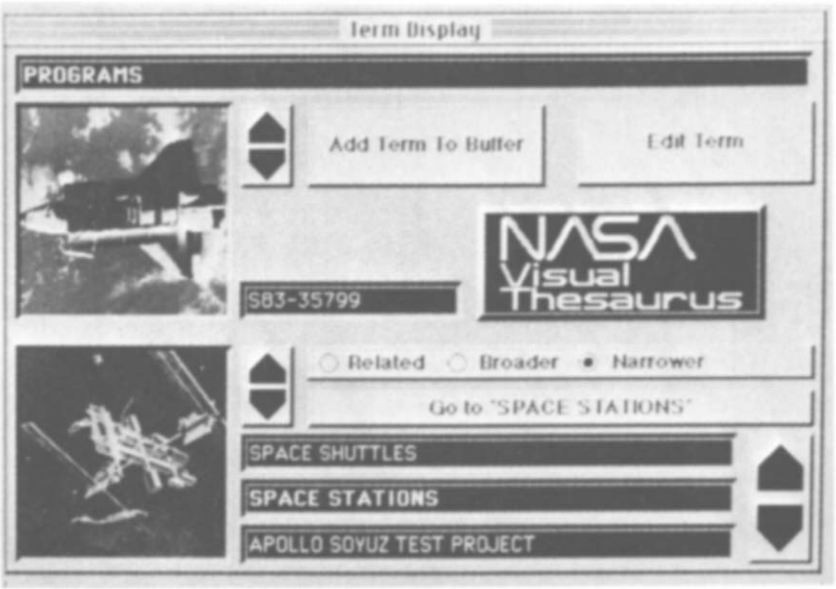
Related Terms:
Related

Broader Terms:
Broader

Narrower Terms:
Narrower

Figure 2. Adding Image Numbers in the Maintenance Module

One of the design goals in creating the visual interface was to maintain independence between it and the data retrieval engine used to search the catalog records, so that the same interface could address existing databases operating on various platforms. For this reason, in part, the interface development was continued in the Apple



PROGRAMS

S83-35799

○ Related ○ Broader ● Narrower

SPACE SHUTTLES

SPACE STATIONS

APOLLO SOYUZ TEST PROJECT

Figure 3. Thesaurus Terms with Related Images

Macintosh environment while the cataloging records were maintained in an MS-DOS environment on IBM-AT compatible machines. In searching, when sufficient search terms have been selected and added to the buffer, the query is passed to the MS-DOS machine (via a null-modem cable and a communications package) to be processed by the data retrieval engine. The results of the search are then passed back and displayed in a window in the visual thesaurus interface, with the entire sequence transparent to the user.

The data retrieval engine selected for this project is almost as unusual as the visual thesaurus interface. Personal Librarian®, from Personal Library Software, is a commercial adaptation of the Syracuse Information Retrieval Experiment (SIRE) developed in the mid-seventies at Syracuse University. Avoiding traditional Boolean search logic, Personal Librarian utilizes a variety of statistically based search algorithms to select documents—ranked in order of relevance—in response to natural language queries. The more terms included in the Personal Librarian query, the better the response, both in recall and precision. An actual search against the JSC film repository 14,000 record cataloging database—a patron requiring all images of "Shuttle Emergency Landing Sites"—contrasts the results of this approach to the standard practice.

A Boolean keyword search for "shuttle" or "emergency" or "landing" or "site" against a traditional database returned over 4,600 records in physical storage order with hits mostly on the commonly used term "shuttle." A modified search for "emergency" or "landing" or "site" brought the number of hits down to 100, again in no meaningful order. A Boolean AND search for "emergency" and "landing" and "site" returned eight records, all right on target but not a comprehensive list.

The same search for "shuttle emergency landing sites" against the Personal Librarian database also returned over 4,600 hits but displayed them in order of statistical relevance. The first eight records were the same as the Boolean AND search, followed closely by "abort landing sites," "trans-Atlantic landing sites," and "ELS," for a total of about twenty relevant records. To deal with more difficult searches, Personal Librarian provides a variety of search functions including an "expand" capability to suggest additional related search terms.

Promise and Pitfalls of a Visual Thesaurus

After several months of thorough beta-testing and modification, a finished version of the visual thesaurus was delivered to the JSC film repository in July 1989. Shortly thereafter, all cataloging and searching activities were switched to the new system. The results to date have been mostly positive: catalog records have become denser with a broader and more consistent descriptive vocabulary for key term assignment while still retaining the flexibility and comprehensiveness of the free-form textual descriptions. (Both key terms and free-form descriptions are searched by the Personal

Librarian retrieval engine.) But while the thesaurus serves as a prompter for the cataloger, suggesting related filing categories, it does not yet compensate for lack of cataloging expertise.

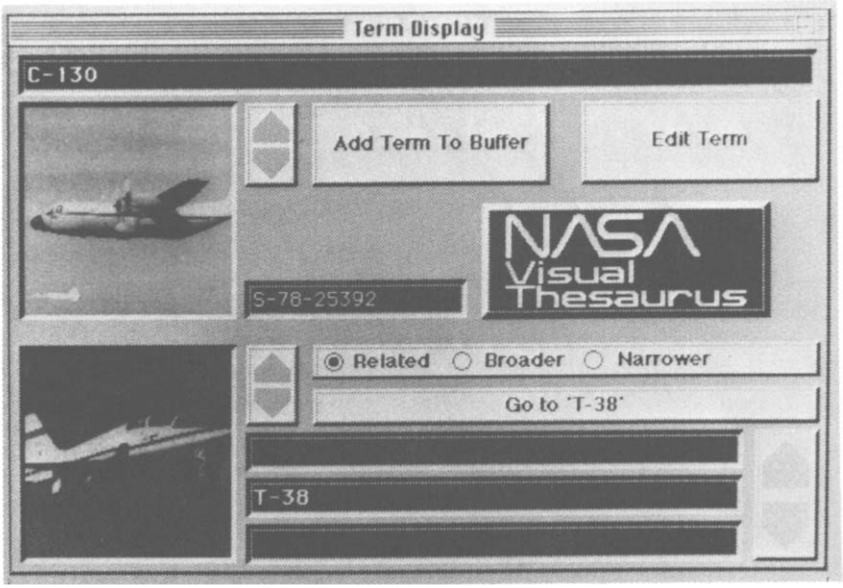
The visual component of the thesaurus is still driven by only about 200 images. Upon completion of the term thesaurus, a routine was written to automatically select the best representative image from the film repository holdings corresponding to each thesaurus term. The approach was intended to isolate statistically the catalog record for which a given term was the most significant descriptor and to pull the corresponding image. Unfortunately, examination of the statistically chosen images revealed, in a majority of cases, ancillary subjects to the target term. For example, the image selected through this algorithm for a term describing a particular test apparatus depicted the item to be tested rather than the apparatus itself. The procedure for scanning and linking new images to terms is in place and is quite simple, but a dedicated effort is needed to manually select the several thousand images required to make this a robust system.

Still, several observations can be made concerning use of the visual thesaurus. Most obvious is the visual component of the interface itself: it is far more interesting and inviting, with its constantly changing display, than a purely textual interface. New users, however, can be somewhat distracted by the images, gravitating toward illustrated categories instead of more appropriate descriptors that currently lack related image representations. This fixation on the image seems to diminish with experience on the system, and reliance shifts more heavily to the textual thesaurus entries.

Initial reactions suggest that, as the interface is more fully developed with images, an appropriate balance will be struck between image and textual stimulus, mitigated, perhaps, by the familiarity of the user with aerospace terminology. For example, a direct search for the terms, "aircraft: T-38" and "Aircraft: C-130" might provide more facility to veteran NASA personnel, whereas visual cues (such as in Figure 4) prompted by the key term "Aircraft," might better serve the neophyte trying to locate or identify "astronaut flight training views" (the T-38 jet trainer, not the C-130 cargo transport).

Search queries constructed with the visual thesaurus work adequately and transparently with the Personal Librarian database, although some of the flexibility available from the Personal Librarian command line interface is lost in the visual thesaurus environment. To date, the search function of the visual thesaurus has suffered because most of the catalog records in the database were entered without the benefit of the controlled key term vocabulary which the visual thesaurus taps. For this reason, experienced collection researchers have preferred searching directly from the Personal

Librarian command line. As the majority of catalog records shifts to the controlled key term vocabulary, the effectiveness of the visual thesaurus interface should improve.



C-130

S-78-25392

● Related ○ Broader ○ Narrower

T-38

Figure 4. Browsing the Visual Thesaurus

These caveats aside, the real promise of this system is that even in its infancy, it has delivered quantifiable improvements in the cataloging process and has imbued a sense of order and control where chaos threatened to reign.

FUTURE PLANS

In spite of the work carried out at MIT and other centers, image management hardware and technology has remained far ahead of the intellectual control processes required to harness its power. Thus the development and implementation of the visual thesaurus cataloging and retrieval interface was the major hurdle in the plan to automate many aspects of image acquisition and retrieval at Johnson Space Center. The next major tasks in this project are the development of an electronic image capture workstation and the development of an electronic image viewing station. Once configured, the viewing workstation will be replicated at various locations around the center and linked together via high-speed communications networks. All new images accessioned into the collection would then be captured and stored electronically, and referenced to the corresponding textual catalog entry, so that images as well as text could be sent to any image viewing station in response to a search. By establishing a uniform system configuration, other NASA centers could also join this network and share both image and data resources.

The benefit of this plan is increased utilization of this valuable and underutilized collection. Cost for access to the collection will be reduced due to more efficient cataloging and searching capabilities. Cost to view or preview selected images will also be reduced: currently, hard copy 8 inch X 10 inch color photographic prints are made to satisfy many requests for which electronic "quick-look" images would suffice. Combined with other current projects which automate work order tracking and circulation of film between the repository and the photographic labs, this effort should provide a generally more responsive system, anticipatory rather than reactionary, to users' needs and effective for years to come. Like many other programs carried out by the National Aeronautics and Space Administration, it may also provide spin-off benefits for a wide spectrum of institutional image collections.

REFERENCES

McClelland, D. (1988). *The Information Retrieval Thesaurus: A brief overview of its role and its construction.* Unpublished paper presented at NASA-Johnson Space Center, March 31.

National Aeronautics and Space Administration. (1985). *NASA thesaurus.* Houston, TX: NASA.

Rorvig, M. E. (1987). The substitutability of images for textual descriptions of archival materials in an MS-DOS Environment. In K. Lehman & H. Strohl-Goebel (Eds.), *The application of microcomputers in information documentation, and libraries* (pp. 407-15). Amsterdam: North Holland Press.

Salton, G. (1968). *Automatic information organization and retrieval.* New York: McGraw-Hill.