

# An On-line Image Data Base System: Managing Image Collections

Malchus B. Baker, Jr.<sup>1</sup>, Daniel P. Huebner<sup>2</sup>, and Peter F. Ffolliott<sup>3</sup>

**Abstract.**—Many researchers and land management personnel want photographic records of the phases of their studies or projects. Depending on the personnel and the type of project, a study can result in a few or hundreds of photographic images. A data base system allows users to query using various parameters, such as key words, dates, and project locations, and to view images matching the query. This application helps select and locate images for presentations, reports, and publications. The image data base is also available on the World Wide Web.

---

## Introduction

---

It is often desirable or necessary to obtain photographic records of the phases of a study. A picture can be worth a thousand words, and these records provide documentation of both minor and major changes that are not accurately recalled from memory. Depending on the personnel and the type of project, a study can result in a few or hundreds of images. At best, the slides, negatives, or prints accumulate and are stored in a folder with brief notation and are later deposited in a filing cabinet.

This paper presents an archive system that combines image information and a medium resolution electronic version of an image in a searchable data base. Users can search the image data base using parameters, such as key words or subjects, dates, and locations, and then view a medium resolution version of images that meet the search criteria. This procedure provides a relatively easy method of locating a photo or slide for a presentation, report, or manuscript. The system provides a quick means to retrieve records and provides others with the ability to search and access images through the World Wide Web (Web).

---

## Preliminary Considerations

---

The initial objective in creating this data base was to develop a relatively easy method of documenting, storing,

---

<sup>1</sup> *Research Hydrologist, Rocky Mountain Research Station, U.S. Department of Agriculture, Forest Service, Flagstaff, AZ*

<sup>2</sup> *Biological Technician, Rocky Mountain Research Station, U.S. Department of Agriculture, Forest Service, Flagstaff, AZ*

<sup>3</sup> *Professor, School of Renewable Natural Resources, University of Arizona, Tucson, AZ*

and locating images for presentation, reports, and manuscripts. We wanted a process that could quickly locate a set of images from a collection at any future date. Therefore, we needed a system that could easily be searched using keywords to identify characteristics desired in an image and that could locate the image once identified.

In making images available for retrieval, copyright issues are an important consideration. However, the images we planned to include were in the public domain, thus eliminating the copyright issue. In collections not in the public domain, copyright would need to be determined early.

It was anticipated that the image archive system would be used by personnel at our lab who may have collections of different emphasis. An important consideration was to develop a system that a user could search all available collections with one query. This virtual pooling of collections increases the likelihood that users can find a suitable image, and it increases collaboration with colleagues. The desire to perform these cross-collection searches imposed some limitations on system design. To facilitate these searches it was decided to use the same data base structure for all collections. Using the same data base structure for diverse collections required that the structure be generalized. Figure 1 shows the basic structure for this data base.

The initial process of documenting the image was the most time consuming, and the ease of entering these data likely determines whether users will adopt the system. A major consideration was how to assign contents to the primary search field "subject." If one assigns subjects in a haphazard manor, searches are less effective and more difficult to formulate. A controlled vocabulary is often used to address this issue. An example of a controlled vocabulary is the Library of Congress Thesaurus for Graphic Materials I: Subject Terms (TGM I). While the uniformity of application with a controlled subject vocabulary would be advantageous to searchers, it is also time consuming to assign these terms and the scheme may not be precise enough for specialized collections. Requiring catalogers to apply TGM I for subject listings would likely discourage use by local custodians of important collections. To help local custodians create and apply a subject list that was concise and applicable to their collection a drop down subject list was developed on the data entry form. When catalogers begin an image collection, the subject list is empty. As the cataloger inputs subjects, they are added to the drop down list. The cataloger can

then review the current list of subjects and either apply the appropriate one or add a new one. This system helps to minimize use of synonymous subject terms.

Another consideration was the format for storing the electronic images. Because we were going to share these images on the Web, our choices were effectively limited to GIF and JPG file formats. The JPG file format was selected because of the high level of file compression available. Since thousands of images were included in the collections, using small, compressed files was essential. We decided to store images at about 600 x 400 pixels, with a file

size of about 150K bytes per image, which provides a balance between usability and storage space. This size image can be effectively used on a Web page or in an on-screen presentation. If a higher quality image is required, users would go back to the original image. With data base structure and image file format decided, we developed the data input system.

A data input station was established to build the data base and scan the images. This consisted of a computer running Windows 95, a Hewlett Packard (HP) PhotoSmart scanner and a Microsoft (MS) Access application developed for this purpose. MS Access was selected as a data base because of prior familiarity. The HP PhotoSmart scanner was selected because of its low cost and ability to easily scan 35mm slides and prints up to 4" x 5".

A custom form for data input was developed to make the process as easy and efficient as possible. The form (figure 2) provides a comfortable format for the user to input data into the fields. Drop down pick lists are used where appropriate for the user's convenience and to minimize typographic errors. The subject field drop down list helps to enforce a quasi-controlled vocabulary, as previously described.

| Field Name   | Data Type | Description   |
|--------------|-----------|---|
| collection   | Text      | name of collection  |
| collectionid | Text      | collection "id" number (information)  |
| filename     | Text      | name of the image file  |
| filename     | Text      | name of the original image file   |
| media        | Text      | is black and white print, color negative, color slide                                       |
| media number | Text      | photo number on image or mount (slide number, USFS number, etc)                             |
| subject      | Text      | collection assigned subject(s)  |
| comments     | Memo      | information that does not fit in another field (permissions, slide mount, print info, etc.) |
| people       | Text      | names of people in the photograph   |
| photographer | Text      | name of photographer  |
| year         | Number    | year photo was taken (4 digit)  |
| month        | Text      | 1-12 if year was not entered  |
| month        | Number    | month photo was taken (3 digit)   |
| day          | Number    | day of month photo was taken (2 digit)  |
| location     | Text      | location of the subject   |
| state        | Text      | state or province photo was taken in  |
| country      | Text      | country photo taken in  |
| copyright    | Text      | copyright info  |
| copyright    | Text      | person who knows about or has access to collection  |
| quality      | Number    | image quality 1 to 5 (used to order images best first in query results)                     |

Field Properties

General | Lookup

Field Size: 255

Indexed:

Required:

Default Value:

Validation Rule:

Validation Text:

Required:

Allow Zero Length:

Indexed:

The field description is optional. It helps you describe the field and is also displayed in the table bar when you select the field as a field. (Press F2 for help on properties.)

Figure 1. Basic data base structure.

Open Image Input Form

Collection Name:  add subject

Collection Cost #:

Slide/Image #:

Media Type:

Photographer:

Subject:

Comments:

Study Site:

State:  Months / Day / Year

Country:   /  / 1965

People:

Copyright:

Filename:  quality:

Result:  to  of 1000

Figure 2. Data input form.

---

## Using the System

---

With system components in place, a strategy for entering data was designed. The first step is to scan the images and to capture any information that was available on the slide frame or on image notes and enter this information into the data base. This cataloger function could be done by someone with technical competence. The second step is to review the records and apply appropriate subject terms or other information that were not captured in the first phase. This function must be done by a subject matter expert, with knowledge of the collection. Usually, this is the custodian. A two phase system like this minimizes the time commitment by the subject matter expert.

As an example, we will use a collection of slides that a hydrologist has accumulated over a 30-year period named watershed management (wm). The first slide in the watershed management collection is identified as wm000001, allowing for a total of 999,999 images in the collection. The cataloger enters this identifier and the .jpg extension in the "filename" box in lower left corner of the data entry form (figure 2.) This number wm1 is written on the slide. The image is then scanned and saved with the same filename (wm000001.jpg). The cataloger then captures any other available information about the image, and stores it in the appropriate fields on the form. A box for Comments is provided for entering information that does not comfortably reside in any other field. The cataloger then repeats the process for additional images.

Once the cataloger has scanned each image and added the available information about the image to the data base, the second step begins. The subject matter expert reviews each record (now consisting of an electronic image and associated information) and adds appropriate subject categories and any other information he or she recollects about the image.

The image data base is now ready to use. We share our data base over our local area network (LAN), so that our in-house colleagues can search and access it. Because many cooperators are unable to access our LAN, we added a web server with software to establish a common gateway interface to provide the link between Web pages and data bases.

### Searching the Data Base

Our image data base is accessible at <http://www.rms.nau.edu/imagedb/>. To locate an image, users can apply a simple tool that locates their search term in any field in the data base. A search for **beaver** would match records where the subject contained beaver, or

where the location contained beaver (as in Beaver Creek), etc. A more focused search tool, which is also available, allows users to search using a combination of fields. Figure 3 shows how users would use this tool to find images of riparian areas taken before 1976. Figure 4 shows the results of that search. To speed transfer of these images across the Internet, we use these smaller, thumbnail pictures on this preliminary results page. These thumbnails are about 160 x 100 pixels, with a file size of about 10K bytes. When searchers see an image of interest, they can click on the thumbnail to view the larger, scanned image (figure 5).

### Managing the Physical Collection

Once the image is documented and digitized, it must be stored to facilitate future location. Hanging folders can store 20 35 mm slides, and a standard filing cabinet drawer can easily hold 6,000 slides. Our watershed management collection is stored in 2 filing cabinet drawers. It is essential that the collection custodian can physically locate an image when given the unique identification number (e. g., wm000001).

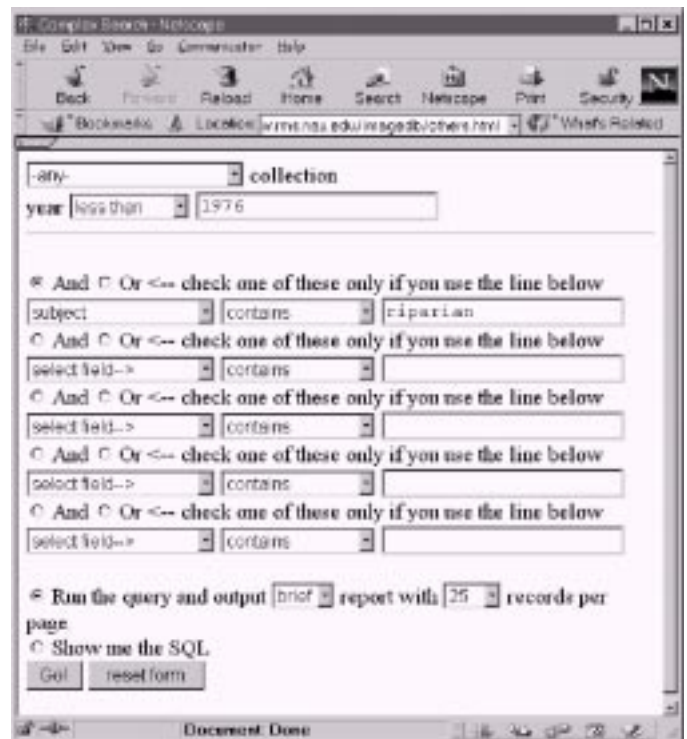


Figure 3. The more focused search tool.

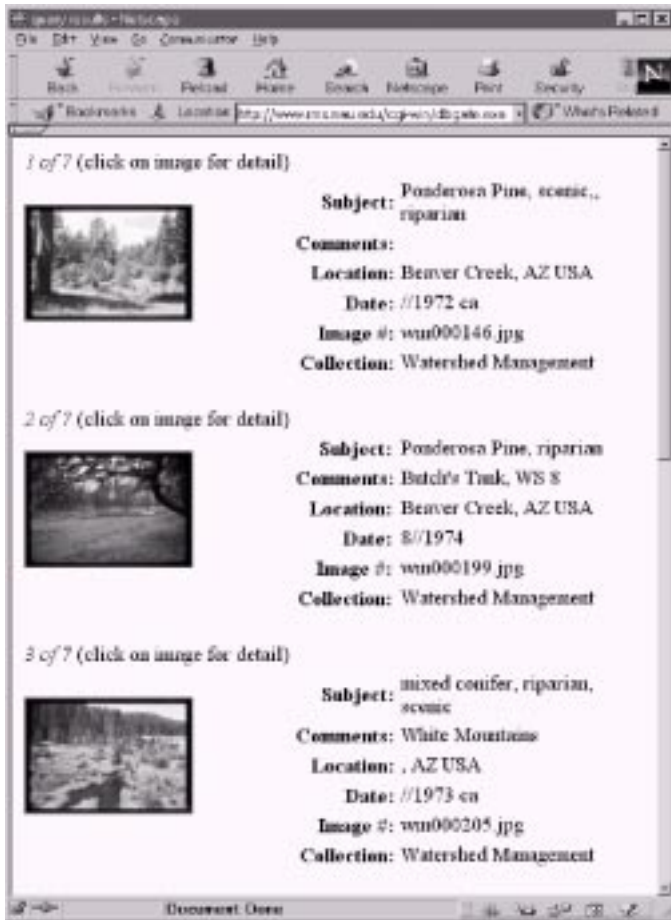


Figure 4. Results from the query in figure 3.

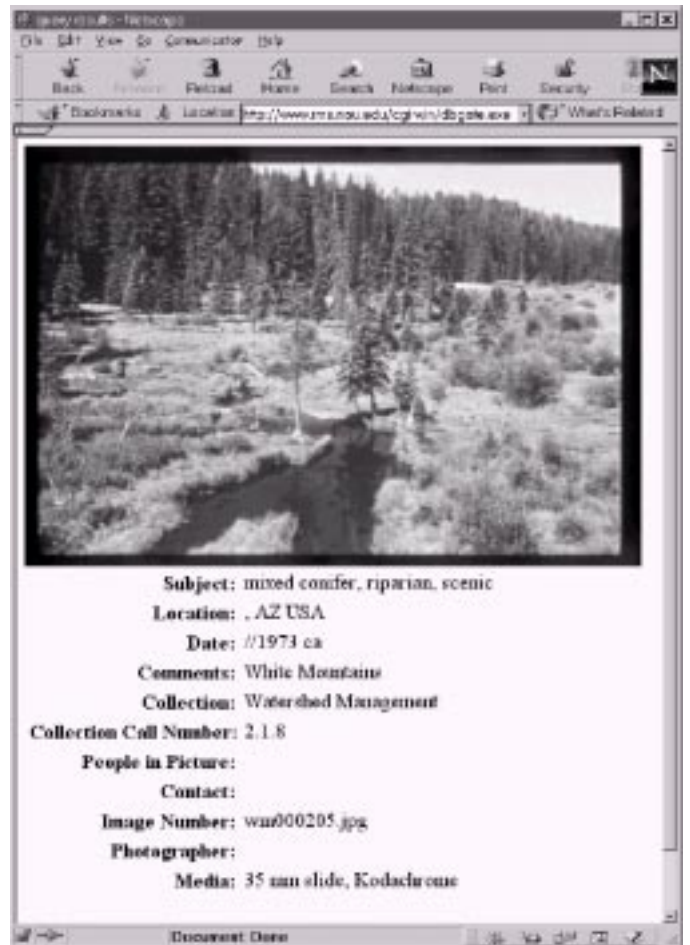


Figure 5. The larger scanned image that results from clicking on a thumbnail in figure 4.

---

## Management Implications

---

Availability of a searchable image data base is a valuable asset to researcher, educators, and interested publics, particularly when it is available on a Web site. Researchers and educators are frequently asked to give technical and informational presentations to various interest groups. These presentations are more interesting when accompanied by slides. However, as the number of available images increases, the more time consuming it becomes to retrieve, use, and refile these images for future application. Although we are not advocating solicitation of images from individuals outside your work unit, our process allows for the exchange of images between colleagues.

With the advances in computer technology, it is expeditious and responsible to spend a little time documenting and archiving graphical data so the original expense of collecting is not lost and to ensure that these data are more readily available.

---

## Acknowledgments

---

The authors wish to thank Linda M. Ffolliott, Information Systems Specialist, College of Agriculture, University of Arizona, Tucson, Arizona, and David W. Huffman, Research Specialist, School of Forestry, Northern Arizona University, Flagstaff, Arizona, for their comprehensive technical reviews of this paper.