Nancy Chaffin Hunter
Colorado State University Libraries

American Library Association Midwinter Conference
San Diego, California
January 8, 2011

First, a little background.

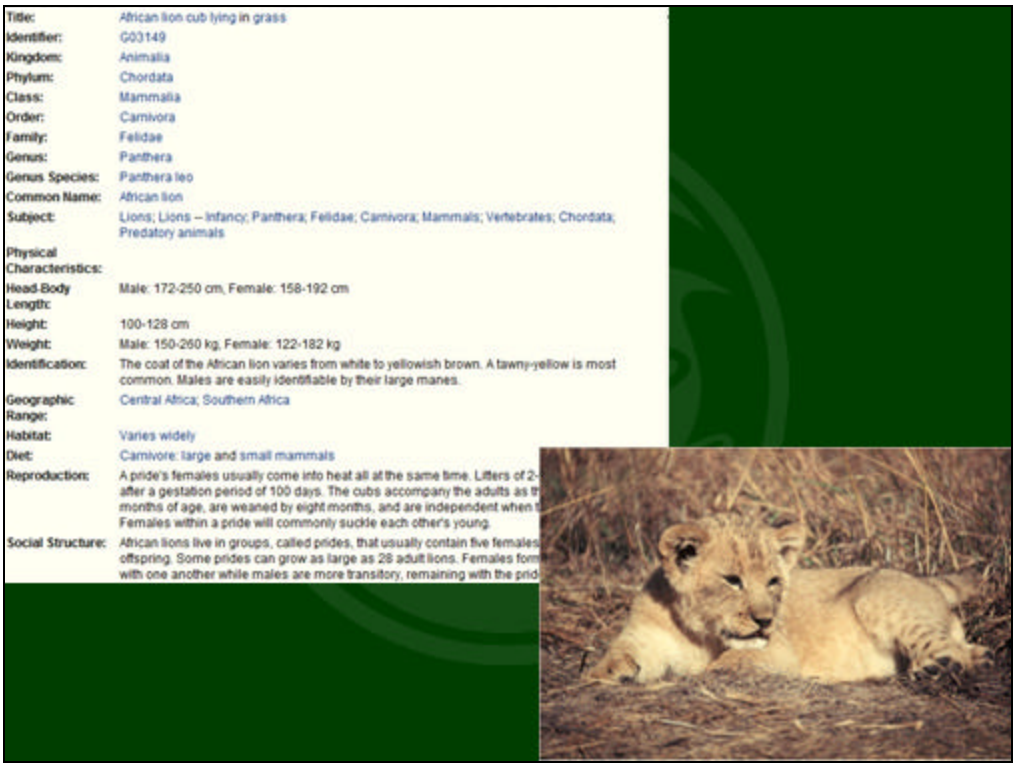
Colorado State University is the Land Grant College of Colorado. CSU was founded as the Colorado Agricultural College in 1870, six years before the Colorado Territory gained statehood. The doors opened to a freshman class of 19 students in 1879. In 1935, the school became the Colorado State College of Agriculture and Mechanical Arts, or Colorado A & M, and was renamed Colorado State University in 1957.

Currently, there are approximately 25,000 students from 49 states and 89 foreign countries. We offer degrees in virtually every area of study, including veterinary medicine, although we do not have a nursing college, a medical school, or a law school.

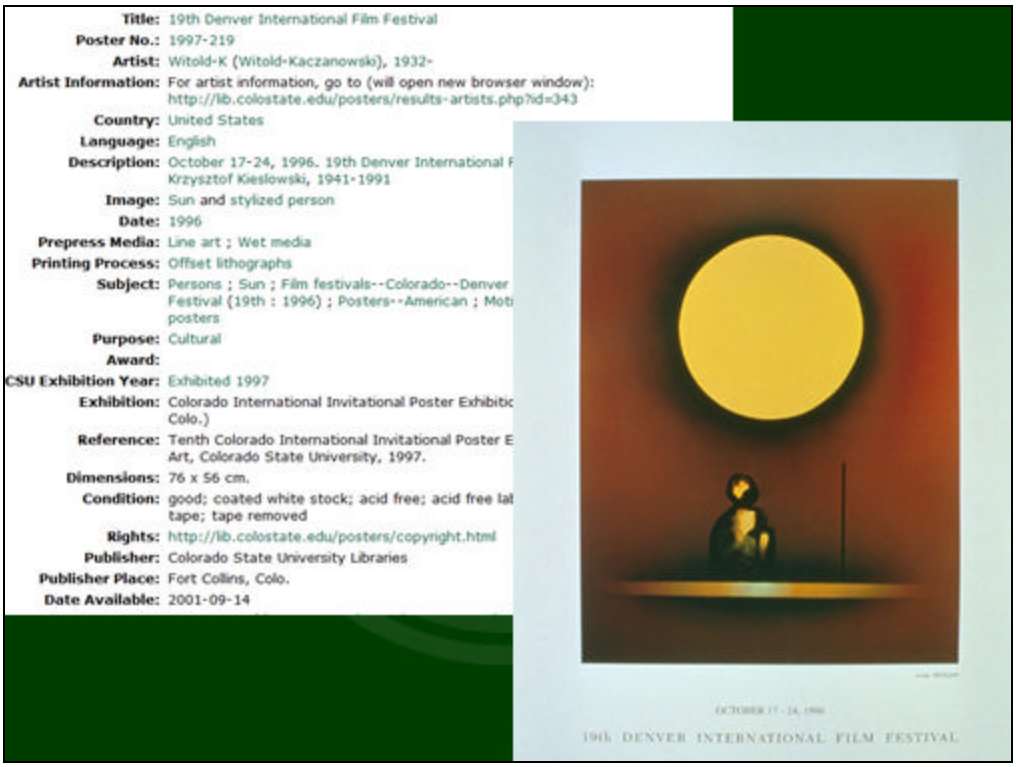
CSU libraries has been in the digitization business for over a decade. Our earliest projects were the International Poster Collection and the Warren and Genevieve Garst Photographic Collections. From the beginning, we used non Professional librarian staff to create the metadata for these projects. Students, in fact.

Art students were trained to do the metadata for our first digitization project, the International Poster Collection, and continue to do so under the guidance of the archivist responsible for the collection.

Biology and veterinary med students created the metadata for the Warren and Genevieve Garst Photographic Collection. The expertise of the students, in poster art and in zoology, have allowed us to create very specialized metadata for these collections – all of which used Dublin Core as a basis, but with much customization.



As you can see here, the Garst collection adds all sorts of information about the species depicted, rather than just describing the image.



In this example from the Posters Collection, the metadata is more traditionally confined to the artifact that was digitized (printing process, exhibition year, condition) but is still very elaborate compared with unqualified Dublin Core.

CSU Libraries has continued to digitize archival materials and the intellectual output of students and faculty at CSU, including The Colorado's Waters Digital Archive, CURC, Carnations, and other projects.

When each collection was begun, metadata was designed for it by the metadata librarian assigned to the project, who also created the metadata. Each project was created with usually very customized metadata. This was our standard operating procedure for several years, and only the Posters and Garst metadata was crafted by anyone other than professional MLS librarians. This customized metadata was possible due to the Digital Asset Management system we had, CONTENTdm, which allowed defining metadata fields at the collection level. We slowly recognized that this was not sustainable, but took no action, primarily because of the size of each digitized collection—generally less than 100 to 200 images/documents. Then several different but related events changed that .

The University Historical Photograph Collection was acquired.

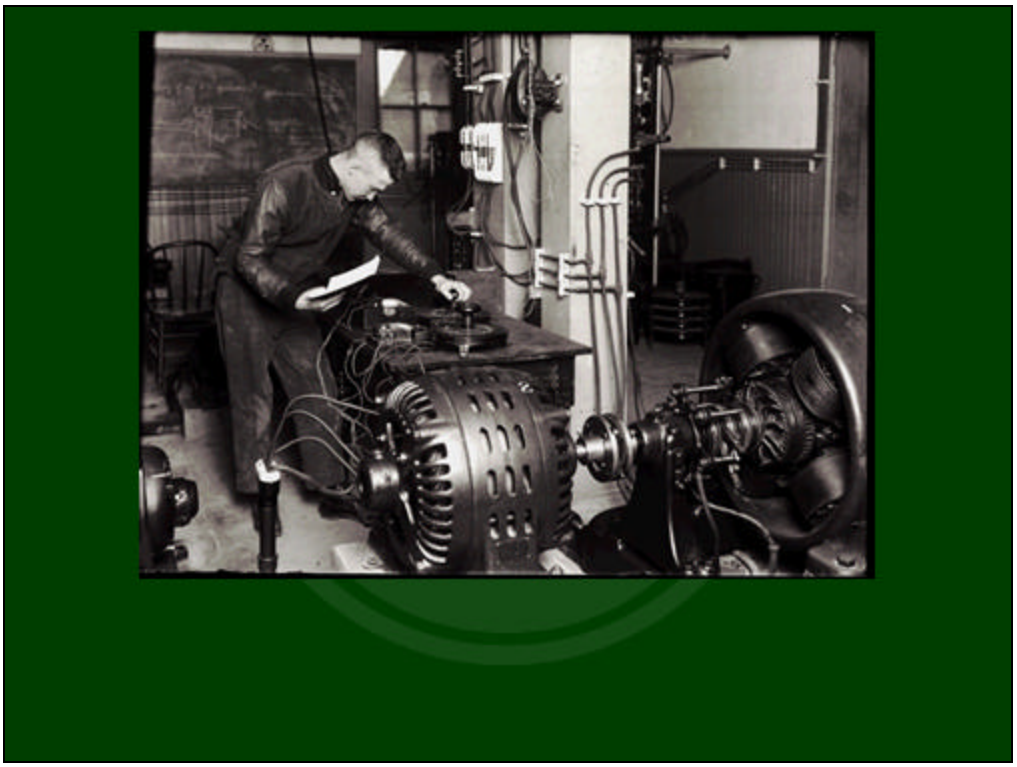
We started noticing decreases in the volume of work for our copy cataloging staff,

We acquired DigiTool for our Institutional Repository

We began the Metadata Best Practices Task Force.

First, CSU Libraries Archives and Special Collections acquired a collection of around half a million images from the department housing the campus photographers in 2006. These images were in all formats, from magic lantern slides and glass plate negatives to modern 35 millimeter color negatives and digital image files. In 2007 we began planning for digitizing the oldest images, because of their age, preservation concerns, and historical significance. This was around 14,000 images. This was the largest digitization project, by a significant factor, CSUL had taken on. It was also decided that these were of significant historical value and that both the finding aid, and the metadata for the digitized images, would provide item-level descriptions.

AT the same time, copy cataloging staff in the Libraries were seeing their workloads diminish. Shelf-ready purchase plans, acquiring more and more electronic resources where we were supplied with MARC record sets by the vendors, and decreasing monograph acquisition budgets all contributed to the decline of the workload. But these staff were well trained in creating MARC metadata to varying degrees according to their state classification levels. Retraining them to create non-MARC metadata would be the equivalent of *not* having all their “skill” eggs in one basket.



Also in 2007, we acquired the Ex Libris Digital Asset Management System DigiTool for our Institutional Repository. This system was different from CONTENTdm, in that although custom metadata can be accommodated, the basic premise is that most metadata would be consistent. This was going to require a new approach to metadata.

To develop this new approach to metadata, the Metadata Best Practices Task Force was created and ultimately developed the CSU Core Data Dictionary, which defined required and recommended metadata fields based on Dublin Core. Subsequent project-specific data dictionaries were developed that went beyond the CSU Core to some degree or another. The first such project-specific data dictionary was for the University Historical Photograph Collection or UHPC. A specific new element to this project was planning for metadata to be built by copy-cataloging staff.

Knots

The inventory of the finding aid was the basis for the metadata. The inventory was organized first by format, then by subject matter. The word document was manipulated and converted into a delimited text file, which was then converted to an Excel file. From the inventory, we derived:

The physical location (box #, folder #, etc.),

the title of the image &

the date if known.

This was done several times with different sections of the inventory as it was an ‘in process’ document. Using the newly written UHPC data Dictionary, I added columns in the spreadsheet for each metadata element. Prior to training staff, I had also taken the subject categories and subcategories listed on the inventory and assigned somewhat broad Library of Congress subject headings for the inventory categories, rather than assigning each image it’s own LCSH headings.

Once everything was in place:

The spreadsheet was readied for metadata

LCSH terms had been added to the inventory

Directions for entering the metadata was written,

Then staff were selected for training (with their approval, of course).

And training began.

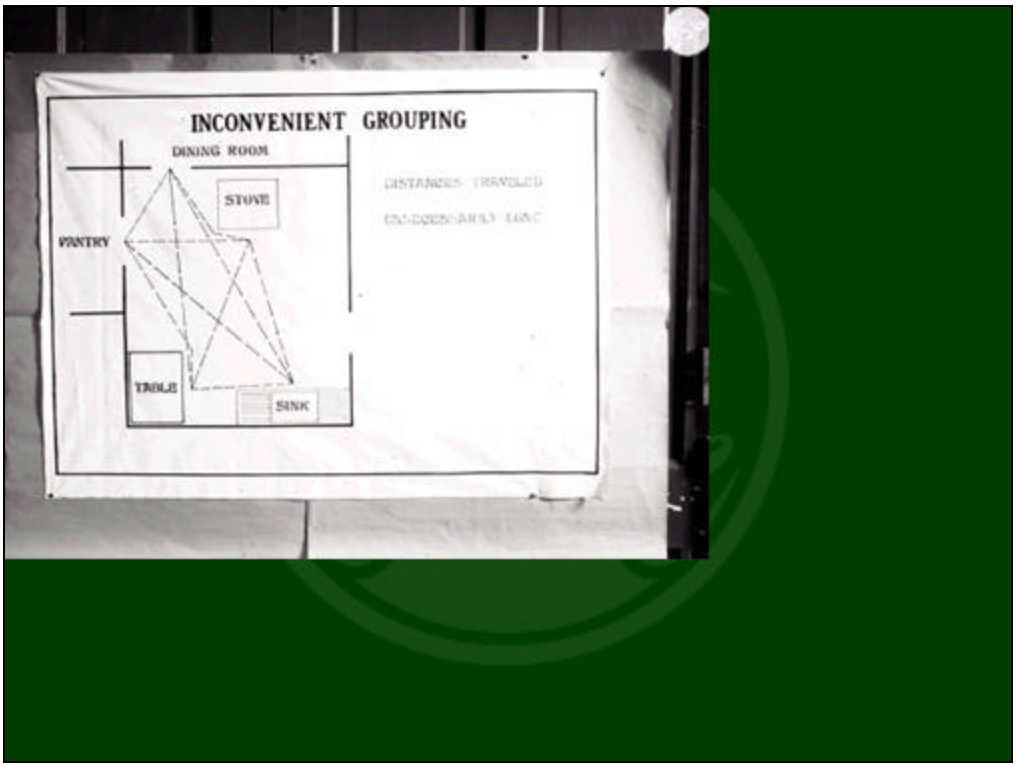
Two library technicians I, one library technician II, and one library technician III were chosen to be trained to work on the metadata. The training went fairly well; the staff understood where they were to find the information to enter into each field; there was very little ‘cataloger’s judgment’ involved. The one field they DID need to make up on their own was the description field, where I asked them to provide a brief description of what the image depicted. This was an area we had to constantly monitor to make sure descriptions were brief but meaningful. One of the technicians was a long-time CSU employee, had worked in other areas on campus, had lived her entire life (she is close to retirement) in Fort Collins, and recognized buildings, streets, structures, etc. in the images, so that she could write for this image:

A winter scene of campus viewed from the bell tower towards the southwest which includes the heating plant, the Statistics Building, and the row of trees down the center of the Oval. A team of horses pulls a sled of people around the Oval.

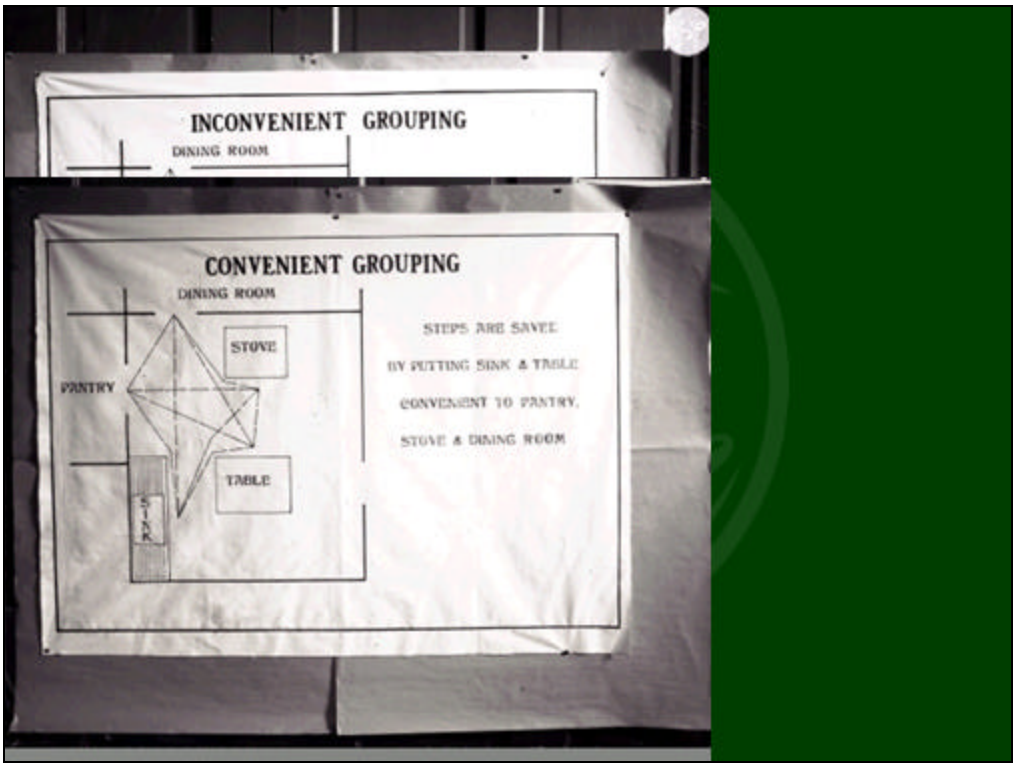
The one element I had not considered in planning was the level of skills the staff had with Excel software, viewing images, using server documents and so forth. There was a learning curve for 3 of the 4 staff (some had a steeper curve than the others); fortunately, the LT II was also in charge of collecting and presenting departmental statistics, and had excellent Excel skills. She served as the one the others would go to if they were having spreadsheet problems. Initially we used the Shared Document feature of Excel, but after one disastrous incident where the rows got mangled, I created separate spreadsheets for each staff person and did not make them shareable.

All data was keyed in manually during the first year or so of the project. Staff had clear instructions where to find the information to enter into each field. If they had questions, they left notes in a column labeled Notes for Nancy with a brief description of their problem. This was cleaner than using e-mail and although I tried to be available when they were working on the metadata, that proved problematic, so we developed this asynchronous system.

Over the course of the project, the LT I's and II created most of the metadata, and for a short while so did the Library Tech III. Initially I did all the QC on the metadata as it was completed, but with 4 people working a total of 24 hours a week on metadata creation, I found I could not keep up. Only a couple of months into the project, I trained the LT III on how to QC the metadata, and passed that work on to her and only worked on the knotty problems (and there were many!) we encountered.



Over the course of the project we developed several processes to make the metadata entry easier, more accurate, and much quicker.



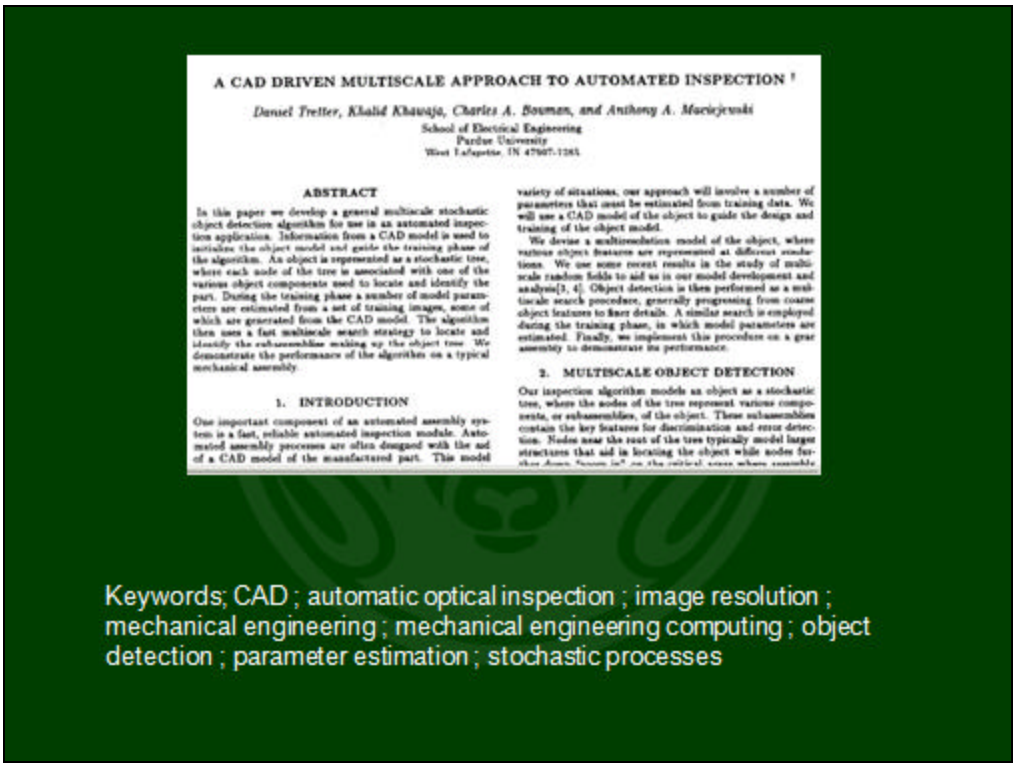
An investment of my time to use the power of Excel to copy, ‘grow’, sort, etc. allowed me to create spreadsheets with more metadata filled-in and subsequently allowed staff to more than triple their output and cut down on errors tremendously.

This ‘experiment’ with using copy cataloging staff to create metadata was so successful that we began looking at other projects to see what could be done by classified staff.

But the problem identified was the software skills. Metadata can be batched into our IR from comma separated value files (i.e. Spreadsheets), MARC-XML (not MARC), and METS-XML. (There are probably other methods, but these are the 3 we have developed to date). Additionally, metadata can be entered using the Meditor staff tool in our IR. It was unreasonable to think that staff who had minimal Excel skills (at the beginning) would have XML skills in place. So to create or use these types of files and systems, staff need training in the software. We are certainly not shy about asking staff to learn new software, systems, what have you. In the 20 years I've been a librarian there has been a constant influx of upgraded, revised, brand new software--constantly. So much so that I believe the motto of tech services departments everywhere is "Change-R-Us". But we do want to capitalize on skills already in place and making sure that works before engaging wholesale upgrades.

One staff member has turned out to be our guinea pig for these other efforts. The success we have had with her work has encouraged us to start training other staff. She has good computer skills, excellent at learning new things, sees workflows intuitively, and enjoys challenges.

This Library Tech III is highly accomplished with our ILS, which is Innovative Interfaces. We have several current projects where we are digitizing CSU publications, as well as print theses. The analog manifestation of these works had already been cataloged. I developed the procedure to repurpose these MARC records; that is, output them from our catalog; make modifications necessary to make the metadata fit the requirements of our IR using Terry Reese's excellent MarcEdit software, and still using that software, covert the records from MARC to MARC-XML. She had previously been trained on MarcEdit after we started a MARC record service for our e-journals, as we did certain global changes to the records in MarcEdit before loading updated records. With minimal training and good written procedures, she has repurposed records for several series of CSU publications, including the Atmospheric Science Bluebooks; such as the one on the screen.



A CAD DRIVEN MULTISCALE APPROACH TO AUTOMATED INSPECTION †

Daniel Tretter, Khalid Khawaja, Charles A. Bouman, and Anthony A. Maciejewski
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907-1285

ABSTRACT

In this paper we develop a general multiscale stochastic object detection algorithm for use in an automated inspection application. Information from a CAD model is used to initialize the object model and guide the training phase of the algorithm. An object is represented as a stochastic tree, where each node of the tree is associated with one of the various object components used to locate and identify the part. During the training phase a number of model parameters are estimated from a set of training images, some of which are generated from the CAD model. The algorithm then uses a fast multiscale search strategy to locate and identify the subassemblies making up the object tree. We demonstrate the performance of the algorithm on a typical mechanical assembly.

1. INTRODUCTION

One important component of an automated assembly system is a fast, reliable automated inspection module. Automated assembly processes are often designed with the aid of a CAD model of the manufactured part. This model variety of situations, our approach will involve a number of parameters that must be estimated from training data. We will use a CAD model of the object to guide the design and training of the object model.

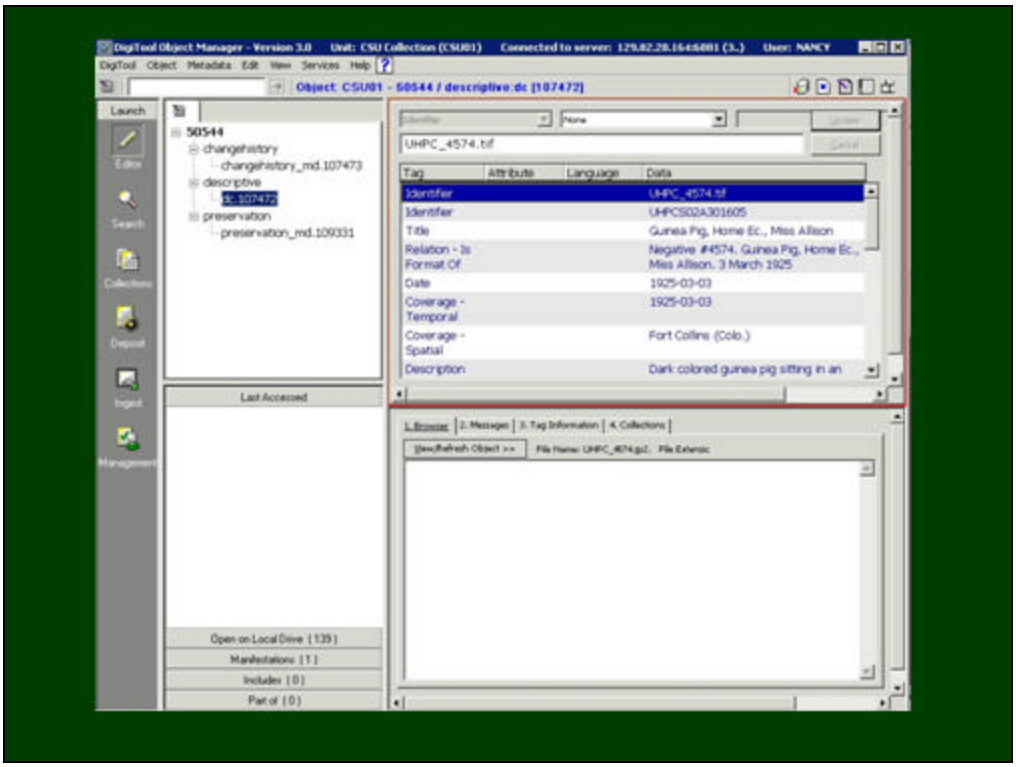
We devise a multiresolution model of the object, where various object features are represented at different resolutions. We use some recent results in the study of multiscale random fields to aid us in our model development and analysis[3, 4]. Object detection is then performed as a multiscale search procedure, generally progressing from coarse object features to finer details. A similar search is employed during the training phase, in which model parameters are estimated. Finally, we implement this procedure on a gear assembly to demonstrate its performance.

2. MULTISCALE OBJECT DETECTION

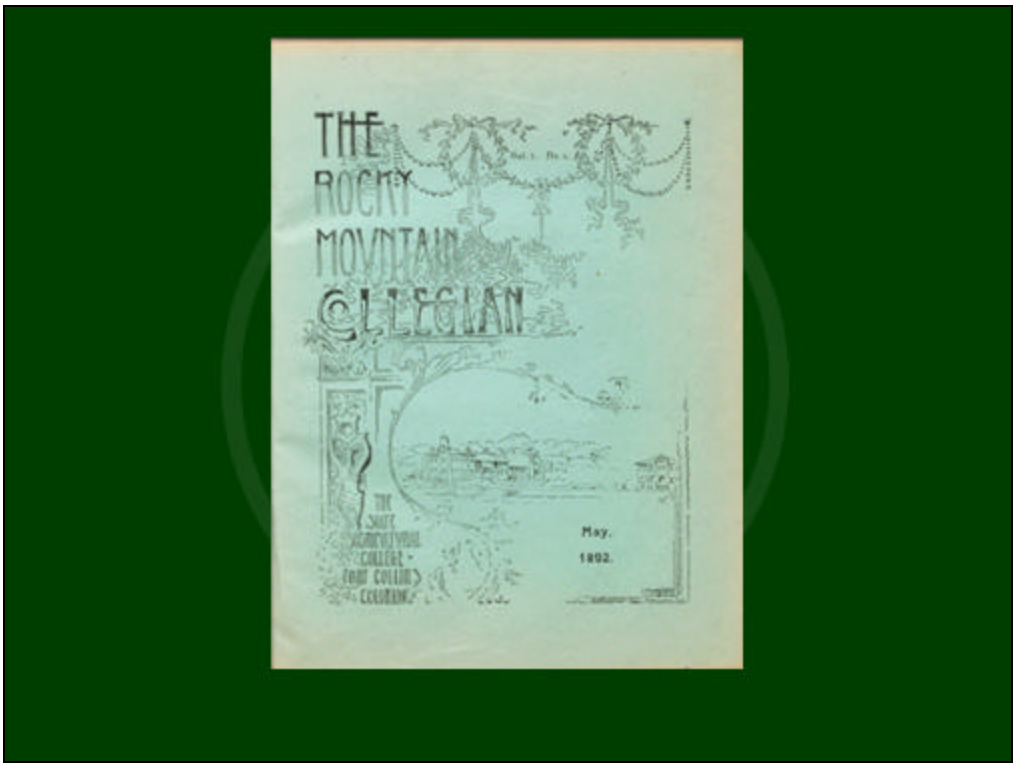
Our inspection algorithm models an object as a stochastic tree, where the nodes of the tree represent various components, or subassemblies, of the object. These subassemblies contain the key features for discrimination and error detection. Nodes near the root of the tree typically model larger structures that aid in locating the object while nodes further down "zoom in" on the critical areas where assembly

Keywords; CAD ; automatic optical inspection ; image resolution ; mechanical engineering ; mechanical engineering computing ; object detection ; parameter estimation ; stochastic processes

Several years ago, she had received training on creating EADs. This, of course, meant she was trained in an XML editing software; she currently has XMetal loaded on her workstation. Our digital content librarian has trained has to create DC-XML records for faculty papers, beginning with those from our Electrical Engineering faculty. Aside: she was have some difficulty with identifying appropriate keywords for these highly technical papers (duh!). Virtually everyone of these is indexed in IEEE Xplore, so I suggested she look for the records for these papers there and use the keywords they assigned. Our engineering librarian heartily approved.



Eventually, she was also trained on Meditor, the back office metadata editing tool for our IR, so that she could enhance, correct, or in some cases add metadata to objects that had already been ingested. Our Digital Content Librarian let her know what needs to be done and she is experienced enough to proceed with little additional direction.

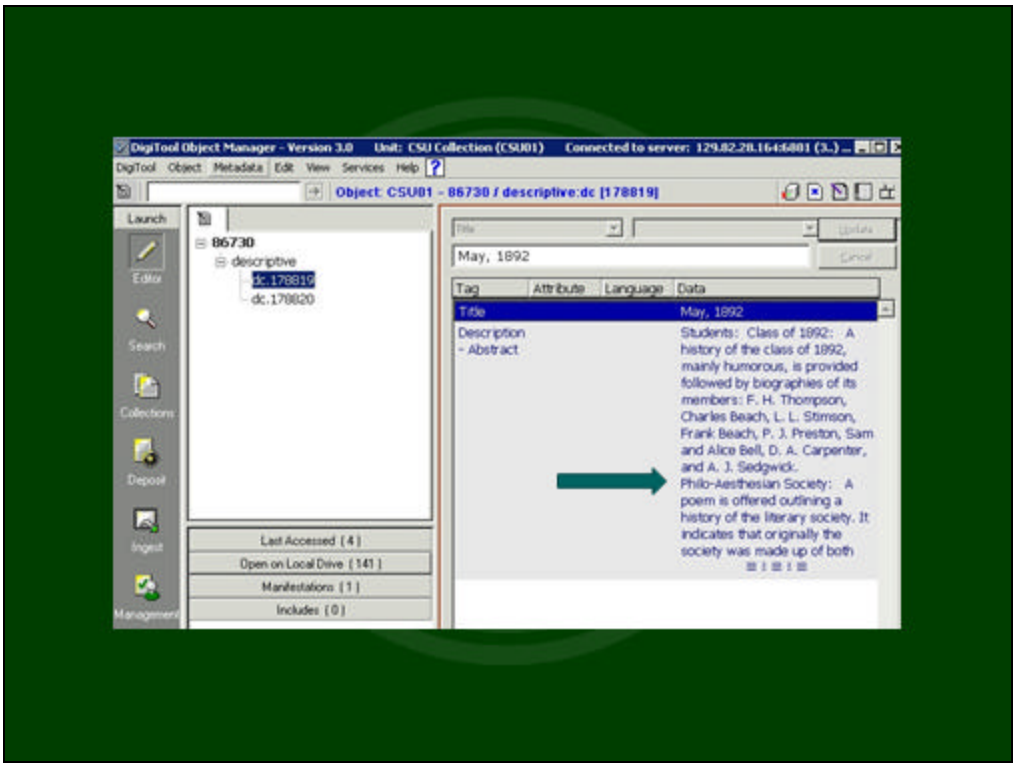


Having worked with the LT III on Meditor, we decided it was time to train another staff member on using Meditor.

We are currently digitizing our campus/student newspaper, The Rocky Mountain Collegian, beginning v. 1 no. 1 (This is not it—no. 1 was kind of 'plain Jane' so I chose v. 1 #6.)

We decided early on to create these as METS objects, using the MARC record for the serial. However, these are impossible to OCR (which would have allowed us to do full text searching in our IR) due to the fancy fonts, poor condition of the type, lots of images, etc. But we certainly wanted to improve discoverability and wanted to add description metadata to each child of the METS object.

It turns out a retired CSU History Professor had donated his papers to Archives, and he had included an incomplete draft of an index he had begun to the RM Collegian. These were modern typing (for the most part) and after testing we are able to produce a reasonably good PDF that could be OCRed, with only minor corrections needed.



We asked the LT II who had worked on the UHPC project to add Dr. Hansen's index data as description and the month/year as title to the child portion of the METS object. After minimal training (with complete procedures written by our Digital Content Librarian) she did this in Meditor after the METS objects were created. As a result, keyword search for, say, Philo-Aesthesian Society will retrieve this May, 1892 reference to the society. Another successful example of copy cataloging staff creating non-marc metadata.

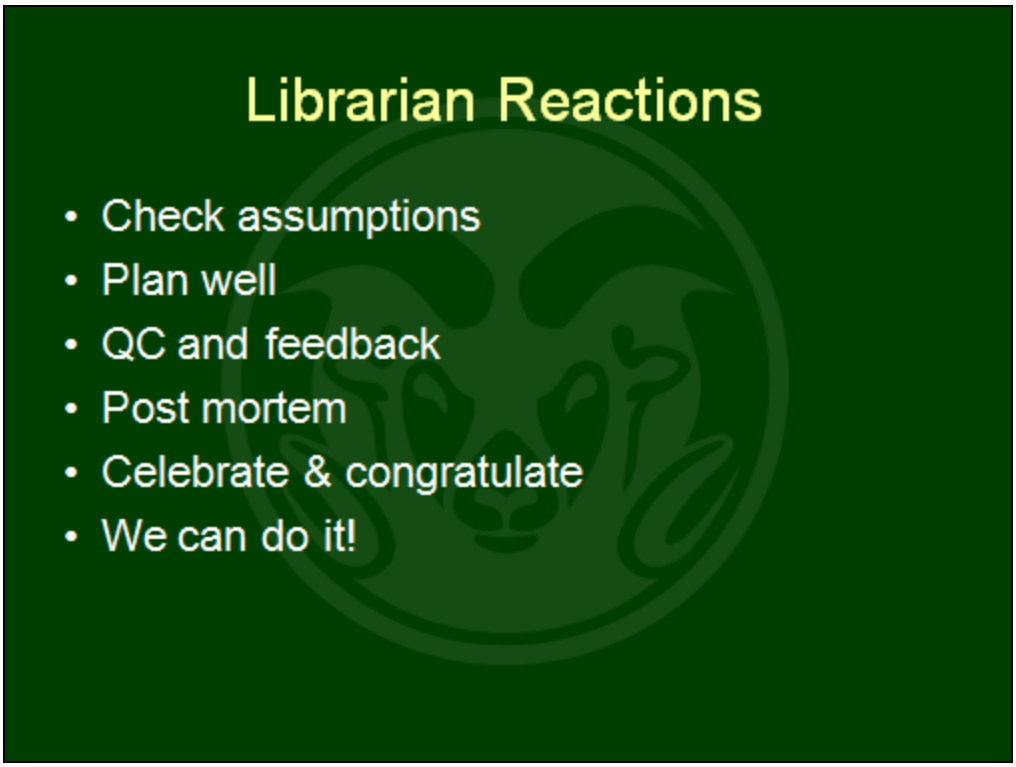
Why?

Digitization and e-resources are increasingly emphasized in our library. This allows copy cataloging staff who deal mainly with print monographs to be in the "growth industry" within the Libraries. Plus, it sends a good message to staff that you're willing to invest in training them with new skills for this Brave New World.

It's sexy! I was hired as the first metadata librarian at CSUL in 2001. I believe (but don't quote me) that Michael Gorman commented that cataloging had finally become sexy. Descriptive metadata is really descriptive cataloging in fancy new gown, but we don't have to emphasize that, do we?

In the case of the staff who participated in the UHPC, the images were fun to work with. Staff members were able to show off their expertise in the history of CSU and Fort Collins, farm animals, plants, style of dress, etc.

Additionally, this bring a different task into their everyday work. It breaks up the routine and builds their intellectual muscles. The LT II was thrilled that several months after the conclusion of the UHPC metadata I asked if she could work on the Collegian project.



MLS Librarians in tech services are becoming increasingly managers of processes (if not actual staff) rather than 'worker bees'. As such we need to act like managers.

Be cognizant of even your most basic assumptions. Regardless of how much time anyone spends at a computer 'time in grade' doesn't necessarily equate with strong computer skills. Be prepared to have to do some training to develop basic skills.

Planning, planning, planning. Make sure of your course so that copy catalogers in an unfamiliar situation don't feel as though you're making this up as you go along. Have good documentation for your procedures; test drive it on people with experience in this area to see where the holes are. And be willing to modify processes or even policies if the folks in the trenches bring up things you just haven't thought about.

Pay close attention to quality control, especially in the beginning but continually. I'm sure you do that more or less in your copy cataloging departments, so carrying that over to non-MARC metadata should be natural. And provide feedback to help develop skills.

At CSUL, most of our digitization projects are just that, projects with a conclusion. Or at least milestones. Engage your metadata staff in analyzing what went right and what needed improvement in your project at the end or at milestones.

Thank you for your time and attention. Any questions?

Thank You!

Nancy Chaffin Hunter
Metadata Librarian, Associate Professor
Nancy.Hunter@ColoState.edu