Indexing training and workflow on large digitization projects

Kimberly A Schroeder

The GM Media Archives (GMMA) is undertaking one of the largest digitization projects in one of the busiest archives in the world. (See The Indexer 21 (1) April 1998, 11-14 for a description of the GMMA indexing system.) Clients that use the more than 3,000,000 images at GMMA include worldwide media, documentary videographers, advertising agencies and authors. In coordinating the reformatting workflow of up to 20,000 images concurrently, the GMMA has experimented with various approaches.

Project overview

The GMMA digital workflow has unique constraints. This includes material up to 100 years old or created yesterday, it may be created anywhere in the world. It may be completely identified or not identified at all. Most significant to workflow, it may be pulled from any step in the process at any time, to respond to client needs.

Below is an overview of GMMA workflow. This understates the complexity faced day to day, but should suffice to give an overview.

- **Processing** is defined as preparing the images for digitization. This includes preservation and searching for duplicate images.
- **Product identification**: Product experts determine the exact terminology of the vehicles in each image, as in a 1955 Chevrolet Bel Air Hardtop.
- **Batching** is what the processors do when all the identification is gathered on the prepared images. This includes filling out a logsheet that travels with the images. Batches are approximately 100 images, for ease of record keeping.
- **Scanning** reformats the original image into a digital form. This may be accomplished by directly scanning the original or reshooting and then scanning it depending on the original photographic format.
- **Indexing** includes attaching all the metadata to attain optimal retrievability.
- **Refile** includes breaking apart the batches so that the original media can be stored in an archival vault with the digital content accessed through the database.

Indexing training assumptions

In coordinating workflow for a large project such as this, the following assumptions were made for indexing training.

1) Write it and they will come
2) Change it and they will understand
3) Include it in the handbook and they will find it

We wrote an extensive manual to communicate Indexing Policy. This 80 page handbook was tabbed, graphically illustrated and covered each substep of Indexing, as well as a high level overview of related steps, such as Processing and Scanning. This was everything you need to know about Indexing at GMMA. Or was it?

What indexers really need to know

The detailed manual that was developed contained only one third of what the indexers needed to know.

In practice, there are three main aspects to indexing training.

1) Professional knowledge
2) Subject knowledge
3) Client needs knowledge
First, the **professional knowledge** must exist as a backbone to their decision making. It is preferred that indexers take the indexing course offered in their library science program. The majority of both our full-time and student interns are from Wayne State University and have taken indexing in the last four years with Elin Jacob or Louise Spiteri. Both of these instructors have done an excellent job in preparing indexers. All of the intellectual analysis and theory background has been instilled in each student. Rarely has further indexing theory had to be reinforced with the indexers.

Next is the **subject knowledge**. There are two types in the GMMA.

The first is the general departmental information on workflow, procedure changes, who to ask what, etc. This is contained in the indexing manual. The second is product related. In this case, with the General Motors collection, indexers cannot become **product experts** overnight. The identification process must be carefully examined in regard to communicating detailed product information to the indexers. The following is the text of a sample of the Product Expert Sheet that is included in the batch with the imagery.

<table>
<thead>
<tr>
<th>Image #</th>
<th>Model Year:</th>
<th>Merchandising Name:</th>
<th>Division:</th>
<th>Model Series:</th>
<th>Trim Level:</th>
<th>Body Style:</th>
<th>Product Expert Note:</th>
</tr>
</thead>
</table>

This may look very straightforward, but automotive terminology is very complex and is often difficult for both product experts and indexers to decipher. For example, when Cadillac markets a vehicle as a “coupe” to reflect its “sporty” nature, and the product expert looks at the vehicle itself and identifies its actual body style as a “sedan”, there is a conflict. Indexers have to create a schema to reflect both the marketing name and the actual body style.

Lastly, and perhaps the greatest source of intangibility is how to prioritize the **client needs**. All indexers realize that to the company, the company’s product is the most important thing in that image, but in what order of importance are all the other objects in that image? In this project priorities are further complicated by the fact we are indexing images from a variety of cultural and historical perspectives. Many issues would look different in different countries and with different political views. Issues such as War or Women’s Rights are very difficult to index. The burden falls on the indexer to take the emotion out of emotional issues and still convey the facts and composition of the visual.

As templates developed to better educate our indexers on prioritizing objects in images, there was still a feeling among the staff that something was missing. What was not realized was that the templates captured and defined the objects, but there was still no definitive education on the intangibles (as in emotion) associated with imagery. When we heard Nicolette Bromberg’s speech in October 1996 at the Midwest Archives Conference our indexing education became complete. In this speech Bromberg discussed the challenges of visual material description and the problems with conveying aesthetics through a “cataloging language”. Her speech aptly described the missing element in GMMA’s indexing training. This brought a term for what the training was lacking, “visual literacy”.

This encouraged an expansion of training to spend more time teaching the “Implication Layer” to indexers. Terms such as “Wow Shot” have been added to convey images with obvious visual strengths. Though this term sounds less than academic, it is the only practical way to point the researcher to an image that is higher in value than others with the same separate objects or content composing an image. Our clients’ needs for retrievability reach farther than just the objects, and include the overall value of the image, which includes composition, action, emotion, impact and interpretation.

### Indexing workflow

In the last three years three distinct workflows have been tried in indexing. The first was the most obvious, Complete Step Indexing. The images were handed off to the Indexers and they had complete responsibility to check them in, analyze them, correct errors, load them into the system and index them. The reasoning behind this was to create an economy of knowledge about each batch of images. This way each batch had an expert that understood every area and would be able to solve problems. It was thought that there would be a time-saving in the same individual following a batch through all substeps of indexing, thus decreasing an orientation period that each new person handling a batch would undergo. The problem was that Indexers felt overwhelmed, there were so many steps that the pressure could be intimidating. No one felt competent at each substep.

The second workflow was an Assembly Line style. The Loading function was broken off and professionals were hired with digital scanning backgrounds to perform this task. Then indexers were trained to become specific field experts. This was a divergence from the previous generalist approach. Since an average record has 23 fields, that alone could seem overwhelming to indexers. Some fields were toggles, some where free form and some were controlled vocabulary. The fields with similar content were grouped together. Three categories were established, Physical Description, Thesaurus Information, and Memo Information.

These were defined as:

- **Physical Description**: Image number, Format, Film Type, etc.
- **Thesaurus Information**: Objects Defined (e.g. 1955 Chevrolet Bel Air Hardtop)
- **Memo Data**: Internal Record Keeping (e.g., Product Expert Note: This vehicle has incorrect headlights. These headlights were used only on concept vehicles, never in production.)

Initially, there was reluctance to handle the images more than once to index them, but after our staff did a test, we learned that our accuracy increased. It did not show a difference in time spent indexing per image. It did teach, however, that specific training and periodic workshops on each field in the Indexing process were important for refining skills.
The third is a method that involves a Check-In Workflow. This has created Check-In experts. Again, this responsibility was pulled out of the indexing step and has evolved into a step on its own. This step has broadened from its initial definition of examining whether all the images are there and marked correctly on the Log Sheet. Now, the Check-in person performs this task, as well as making additional notes on things to call to the attention of the indexers. They also put the appropriate templates in with the batch. This has resulted in the most amazing gains in productivity. The number of minutes to index each image has decreased by a more than 30%. The Indexer also has someone to ask questions of: the Check-In Person who has seen that batch and analyzed it. This has taken a huge burden off of Indexers working with a very diverse collection.

To give you an idea of the breadth of knowledge needed, GMMA right now is processing collections in Testing, Design, Powertrain, Opel (German), GM Japan (1920s), Auto Shows, Futurama, 1999 Press Kits, 1998 Product Catalogs, General GM Product (all eras) and the History of several domestic divisions. Not only would it be impossible for the indexers to know about all of these specific collections, most are only a few hundred images which is less than a weeks work. Training twelve indexers on all the details of each collection is not viable in such transitory subjects. Training the Check-In people to recognize the basics of each collection and know as a standard what is needed for the Indexer to perform well is much more plausible. If all the information is not in that batch, it will be returned to the product experts for more information.

**Error Rate**

The second advantage to this Check-In Workflow method is the ability to examine GMMA's error rate. Until recently, the error rate had not been examined in detailed statistics. Though there is no excuse for this, all organizations get wrapped up in day to day survival and can fail to look at the big picture. Anyone who has managed a large database project knows that each error that you see is like being hit by an anvil! When one glaring error was found, worry spread as to whether we were being accurate enough. GMMA inputs as many as 2400 records a week and multiplying that by 23 average fields per image, means that there were over 55,000 data fields per week being added. The error rate began to be put into perspective. I asked our Quality Check Indexers to track error rates. The following are the highline results:

<table>
<thead>
<tr>
<th>Sample</th>
<th>10,000 records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product expert</td>
<td>100</td>
</tr>
<tr>
<td>Indexing supervisor review (No. errors)</td>
<td>1</td>
</tr>
<tr>
<td>Percentage error per record</td>
<td>23</td>
</tr>
<tr>
<td>Average number fields per record</td>
<td>230,000</td>
</tr>
<tr>
<td>Total number of possible fields in sample</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

Currently in process is a further definition of types of errors and their weighting by critical value. Thus determining which errors are more critical to avoid than others. For instance, if the Image Number is wrong, the retrievability of the image is dangerously low. Numerical errors do not allow you to approximate the location and browse. A Subject error would allow serendipity (e.g. if a Pontiac Bonneville is indexed as a Pontiac LeMans, a user could still browsePontiacs to find the image). An error in a limited field, as the Image ID without the hierarchy behind it, however, simply fails to provide a browse capability.

**What We Know Now**

What we have learned is that your instinct is not always correct in determining processes of this magnitude. We had to conduct several experiments and speak each week to indexers to get their opinions, as to how to improve their knowledge, their comfort level and therefore their capabilities. I will not profess that we are there yet. I will say that at each weekly meeting, I learn something positive and something negative. At least we are open to learning. I know without the incredible support of my supervisory staff and the inspirational and tigerlike attitudes of our indexers, this enormous task would not have been possible.

I understand now, after nearly five years into this project why so few large reformatting projects are done, or done well. It is full of a plethora of details that can weigh down the system, if they are not correctly prioritized. Most indexers would have been too discouraged, most companies would have compromised on precision, but our perseverance has allowed us to create the most accurate and detailed database for a company that is the most diverse in the world. This new corporate asset has resulted in a self-supporting department that has reached out and preserved a part of global history. That is the power of an Indexer!

Kimberly A. Schroeder's company Archive Impact is a consulting firm specializing in indexing, training and workflow issues on imaging projects.

Special thanks to Pam Gannon, Mies Martin, Tom Freiman and Margie Jankowski Friant; thanks also to the team at the GM Media Archives.