Good afternoon. It’s a pleasure to talk to you today and to have the opportunity to tell you about the first time I was responsible for something many of you are familiar with, a mass metadata migration. And the focus of this story is how in a very short time frame we significantly enriched our data in several critical ways as it was moved from one system to another.
Specifically, after years of delay, the MoMA Archives decided to migrate its image catalogue from a hopelessly antiquated system into The Museum System, the same place it takes care of all the Museum’s artworks. We call it MAID, and today it contains the metadata for all the high resolution digital images in the Archives. Of 44,000 records at the time of migration, 33,000 were installation photographs, one of MoMA Archives key historical assets. The other 10,000 or so records were for other photographs, archival documents, and miscellaneous items from across the Archives that were digitized for publications or other use by request of staff and researchers. While the migration had been years in the making, my role in the work lasted only about two months.
For the sake of time I’m going to pass over the various motivations for this move though you can see some here.

The basic migration workflow was that our IT department would extract the data from the legacy software and hand it to me, I would do what needed to be done to get it ready for the new system, and MoMA’s TMS team would import it into The Museum System.

The old database was a black box to me and to our IT department. For a while IT wasn’t sure they could export any data out of the system in a coherent way. But to everyone’s relief, with some work, they were able to produce a single spreadsheet with one image record per line.

Batch importing records into TMS isn’t that easy. But a while ago TMS’s parent company Gallery Systems had created a custom script for MoMA that essentially allowed a one click upload from an excel spreadsheet.
So when our IT department handed over a 44,000 line spreadsheet the first and easiest task was to re-shape the data to fit into the import spreadsheet. This mainly involved renaming some of these spreadsheet columns and merging a few different fields into one.

The more important and complex question was how quickly and to what extant could we enrich these records, adding metadata that would be both valuable to our users and helpful to us in integrating these resources into the larger data environment. What fields were necessary to add, what were possible within this one-click uploader script, and what had to be considered off the table?
The uploader allowed for one unique identifier or “object number” but it also allowed for one “alternate number”. So we took advantage of this by retiring our current object number to the “alternate number field” and inventing a new slightly more rational system of object numbers in its place.

Next, TMS relies on a classification field to very basically categorize objects. We never had this capability before, it’s very, and the uploader allowed us to add it. MoMA uses a fairly limited list of classification terms and adding or changing any causes a lot of work as it affects numerous reports and functions downstream. So we couldn’t go wild in choosing new classification terms. Ultimately we negotiated six new terms as well as using a few already present in the system.
The next major metadata enhancement involved dates. It was late breaking news to us that the data importer allowed for standardized date/time fields in addition to a free text display date field. When I heard this I think I jumped out of my chair. We had never been able to search or sort on dates in our old image system, only by keyword. But this meant we had to convert forty thousand display dates into normalized start and end dates. Let me tell you, this isn’t as hard as it looks but requires somewhat advanced Excel skills, or in my case a lot of googling “how do I do that?” Briefly put, the key to this kind of activity is to recognize the patterns the specific data falls into, use that pattern to divide your dataset into smaller manageable groups, and tackle each group one by one. For best throughput I tend to start with the easiest largest groups first and save the most unique and tedious issues for last. For example, out of 44,000 records; there are 5000+ records where the value is blank. 75% of the rest feature two date phrases separated by the word through. We can convert them using one transformation. And before you know it, every record has a normalized date. Except, y’know, for everywhere a month was misspelled or there was some other abnormal permutation.
Most importantly, the TMS uploader allowed us to add two name entries to each record. Records in the old system were very flat and there was only one uncontrolled field for “artist” that was barely used and another for photographer. By adding and populating these additional name fields to a line in our spreadsheet, the upload script would connect that object record to existing authority records in TMS. This was a more laborious process than our other actions, but it was also of more critical importance for connecting our archival records to all other objects in the system through related names.

I’m a big proponent of Microsoft Access as a tool and it enables this type of complex data entry to be simplified and streamlined. I imported the Excel spreadsheet of MAID records into Access. Then I had the TMS admin send me a spreadsheet of all constituents name records in the system and imported that also into Access. In Access I could build a simple custom interface where each MAID record could be more simply displayed. The new added name fields are on the right.
The entire list of 53,000 existing constituent names populated these drop down menus.
Each name was assigned a relator term from a controlled list.
Most of this work was done by an intern so instructions were simple. Pick two names from the description, always include the photographer if there is one, and select them from the drop down menu. Describe them with the appropriate relator term. If no individual names are mentioned, move to the next record. If the names aren’t already in the drop down menu, move to the next record. The emphasis was on speed and results while avoiding getting bogged down in problematic edge cases. Ultimately through this process and some mass updating of installation photo data, we added nearly 50,000 name authority connections to these records. 95% of all the records in the migration have at least one name link. This work took about four weeks. Once all this work was done it was uploaded to the new system by our TMS team.
What was the result? Firstly, I can’t overemphasize what a big step it was for us to have our image records in a modern system.

Secondly, we can for the first time search by format or within a date range. More importantly the materials are fully integrated with artworks and other collection objects across the museum. These are search results for any object with Calder as a related name. In the highlighted record he’s only the subject of the photograph. But you can see in the search results how works from the Archives appear next to sculptures, drawings and items from other parts of the Museum.

There is of course much work that can still be done. More names to add, legacy errors and omissions to correct, and finding new ways we can exploit the new system we’re in.
I haven’t today talked at all about front ends. But let me mention that months after our data migration, MoMA’s TMS team were able to customize an off the shelf TMS web access module to create a new user interface that, for so that for the first time all of our digital images are freely accessible on the web.
Thank you.