Original Article

Harvesting organically grown vocabularies: Finding user-generated controlled vocabularies within pre-existing information repositories

Chad Beer

is the Senior Digital Asset Manager at Cline, Davis and Mann. He has worked for 12 years in digital asset management and archiving in creative environments, including the Metropolitan Museum of Art Photo Studio, and Gap Inc.'s N.Y. Design Archives. He specializes in representing and educating internal end users who may be new to working with DAM. He has helped organizations move from analog to digital workflows. More recently, he has been working on maximizing (and simplifying) DAM systems. Working as a bridge between users and developers, he is focused on designing workflows and customizing systems to serve and woo end users. In August 2009, Beer established the NYC Digital Asset Managers meetup group: <http://www.meetup.com/NYCDigitalAssetManagers/>.

ABSTRACT Tasked with creating two very different vocabularies for a new DAM system, our implementation team did so at a minimum of cost and time. Instead of starting from scratch, they approached the project as interviewers, organizers and translators. Grooming previously gathered data, and leveraging the expertise of in-house professionals, became the team's points of focus. Research revealed the data, already waiting, in existing repositories ranging from a robust billing system to Excel spreadsheets. They were not called 'controlled vocabularies' by the people who created and used them, but they identified the terms and relationships that the DAM system would need. Other in-house experts were looped in to groom the existing information into the formats needed to suit the DAM system's purposes. All the while IT partners were engaged, ensuring the data would ultimately work with scripts to automatically tag files in our DAM system. Spreading the work among carefully chosen internal departments saved valuable time and costs. Leveraging user-generated tools to feed metadata into the DAM system is unifying our agency in the use of a common language.

Journal of Digital Asset Management (2010) 6, 185–189. doi:10.1057/dam.2010.19

Keywords: metadata; controlled vocabulary; automation; workflow; outreach; implementation

PLANNING A NEW DAM SYSTEM: CHALLENGES AND A SOLUTION

Like any large organization, the ad agency for which I work employs professionals from a variety of disciplines. When it came time to plan and develop a new DAM system, a team was gathered from the agency's Archiving and IT departments. The new system would need to

be useful to a variety of subject matter experts, including prepress specialists, business development teams and account executives. It would need to be robust enough to locate assets based on the varying needs of all teams. In effect, it had to speak more than one language, and our DAM implementation team would need to act as interviewers, organizers, ring leaders and translators.

Correspondence:

Chad Beer
Cline Davis & Mann LLC
220 East 42nd Street,
New York, NY 10017,
USA E-mail: cbeer@cdmworldagency.com;
Website: <http://www.meetup.com/NYCDigitalAssetManagers/>

Through early user interviews, it became clear that we needed to create two very different controlled vocabularies to enable searching in the DAM system. We were tempted by ambitious thoughts of creating these from scratch, but the amount of research involved was daunting. Keeping costs to a minimum was paramount, and therefore instead of beginning with a blank page we looked at storehouses of knowledge that already existed. In agency-wide systems and intra-departmental documents, we found what we needed. These were not formal taxonomies. Their creators did not refer to them as ‘controlled vocabularies’. What they provided, however, was essential raw material: lists of defined terms with established relationships, describing the data required in the DAM system.

BUT LET’S START FROM THE BEGINNING

The DAM solution was needed to manage thousands of files created by our agency. These include image files and layout documents (InDesign and Quark files). All were developed for marketing material devoted to specific brands, each brand addressing specific consumer needs.

The layout documents fall into a small range of types, such as layouts for single-page ads, or multiple-page brochures. These layout types fit into broader categories, such as direct-to-consumer material or outdoor advertising.

Users would need to search for files in many different ways, but there was a clear need to facilitate these two important searches:

1. What work had we done to address certain types of consumer needs?
2. What examples did we have of specific layout types and broader layout categories?

Adding metadata manually would be prohibitively time consuming, and therefore we needed two controlled vocabularies:

1. Consumer applications related to each brand we represent.
2. Layout categories containing subsets of specific layout types.

UTILIZING EXISTING EXPERTISE: WORKING WITH THE RIGHT PEOPLE AND SYSTEMS

Our agency is focused on pharmaceuticals, and therefore the consumer needs are, specifically, medical indications and medical applications. These applications, as addressed by our agency’s brands, are outlined in documents developed by and for our account teams. Containing beautifully detailed information, they covered more aspects of brand-specific data than were needed by our DAM system. The data were also written in prose form.

Our implementation team had system knowledge but not medical expertise. We needed partners who could translate the long-form information into term lists without losing accuracy. We found them in our Editorial department, each editor being assigned to a brand-specific team. The editors also possessed a critical skill: speed. They were able to groom the information without being swamped by the project.

Example:

Original data in prose form

(Brand Name) is indicated for candidemia and other forms of Candida infection (intra-abdominal abscess and peritonitis). (Brand Name) has not been studied in endocarditis, osteomyelitis and meningitis due to Candida.

Data groomed into search terms

Antifungal
Candida
Candidemia

With work proceeding on one front, we looked at defining layout types and their related layout categories. Users typically thought of each layout as belonging to one of about 20 common types, each within a broader category. We created a list of essential layout types and categories through user interviews. Layout documents were already categorized in the agency’s billing system but those data needed a lot of simplification. What worked for billing records wouldn’t necessarily be considered user-friendly in a searchable DAM system.

The billing system appeared to contain dozens of different types. A closer look revealed that there could be many different ways to describe one type of layout, such as

Table 1: Example, from our translation table

<i>Job type in billing system</i>	<i>Job category in billing system</i>	<i>Simplified type/Category</i>
Journal Ad Execution	Media Planning and Execution	Ad Journal Ads
Journal Ad Revisions/Resizing	Media Planning and Execution	Ad Journal Ads
Marketing Backgrounders	Global Country Marketing Communications	Backgrounder Sales Training
Competitive Backgrounders	Training Materials	Backgrounder Sales Training
Trial Backgrounder	Trial Commercialization	Backgrounder Sales Training
Competitive Backgrounders	Training Materials	Backgrounder Sales Training
Trial Backgrounder	Trial Commercialization	Backgrounder Sales Training
Backgrounders	Training Materials	Backgrounder Sales Training
Core Sales Brochure	Core Sales Materials	Brochure (Collateral Brochure) Core Collateral
Core Sales POA1 Brochure	Core Sales Materials POA1	Brochure (Collateral Brochure) Core Collateral
Core Sales POA2 Brochure	Core Sales Materials POA2	Brochure (Collateral Brochure) Core Collateral
DTP: Phys Brochure	DTP: Physician-Mediated Patient Ed	Brochure (Collateral Brochure) Core Collateral
In-Office Materials Brochure	In-Office Materials	Brochure (Collateral Brochure) Core Collateral
Mid-POA Brochure	Mid-POA Sales Materials	Brochure (Collateral Brochure) Core Collateral
Non-POA Brochure	Non-POA Sales Materials	Brochure (Collateral Brochure) Core Collateral
POA 1 Brochure	POA 1 Sales Materials	Brochure (Collateral Brochure) Core Collateral
POA 2 Brochure	POA 2 Sales Materials	Brochure (Collateral Brochure) Core Collateral

a direct-mail piece. Our DAM team created a translation table. This matched the dozens of layout types in the financial system to more user-friendly terms in the shorter list.

This time, we reached out to our Traffic department when we needed expert advice. As the team tasked with entering layout types into the billing system, they reviewed and refined our translation table (Table 1).

MARRYING GROOMED DATA TO EAGER ASSETS

We had the data. We had the assets. How would we join the two?

Volatility of the information not only added to the puzzle but also led to a solution. Consumer applications for a given brand can change over time. The product knowledge repositories, groomed by the Editorial team, would require periodic updating so that our DAM system would always pull updated information. Our IT partners created an

application in Sharepoint to meet this need. Information in the product knowledge repositories was imported into the Sharepoint. Account teams can easily update the information within this application, and the DAM system can grab the data from it.

To define a trigger for grabbing the data, we laid out the necessary series of events. Each asset required a brand association before the right consumer application data could be targeted. Upon ingestion into the DAM, the files are tagged with the name of the correct brand. This tells our DAM system to look to the Sharepoint application and access the correct consumer application information.

Next, we needed to tag layout documents in our DAM system with types and categories. Our IT partners could tell our DAM system to query our financial system, but adding a separate translation table to that chain would be problematic. Our agency's Traffic department, with their mastery of our billing system, were

able to identify fields within it that could house the translated terms and be accessed by a web services application.

From there, marrying the data to the files was relatively easy, thanks to the well-established processes in our agency. Each layout document is held to a rigid naming convention. This makes it easy, by analyzing the filename, to tell a layout document from another kind of file. Upon ingestion into the DAM, the system looks at each filename. If the name indicates a layout document, the financial system is referenced. When a match is found for the filename, its corresponding layout type, translated for usability, is also located and migrated into the DAM system.

Pulling controlled vocabularies from pre-existing knowledge repositories has done more than save us hours of labor while enabling crucial searchability. It has brought systems and teams together under a common language, defined by users but refined by DAM experts.

A TALE OF TWO SEARCHES: THE IMPACT ON REUSE

Case 1: Consumer needs vocabulary

This vocabulary was most needed by our agency's business development teams to make the agency's work history more accessible and searchable.

When preparing to pitch work to a new client, past agency work that addresses similar consumer needs is gathered as part of an agency portfolio. To do this, teams had relied on personal familiarity with agency history. Long-term employees, and other staff who had worked on agency portfolio projects, had become the keepers of this detailed knowledge of agency history.

Some team members had unknowingly demonstrated that a DAM solution was needed for this work. Pitch development often occurs under tight turnaround time. In user interviews, our DAM implementation team learned that several individuals had gathered personal collections of files to help with fast assembly of pitch work. Each was a gathering of greatest hits of our work, and provided proof that users needed a centralized searchable repository.

These personal collections have been imported to the new DAM system. These,

and all agency work from the last 5 years and onward, will be tagged with metadata about consumer needs. When work is needed, which will show that the agency has experience addressing a specific need such as cardiology or dermatology, the DAM system will show all files associated with prior brands addressing the same. Of course, once our work history is accessible and searchable, it can be leveraged agency-wide for other a variety of uses, including tracking product history and providing an educational resource for new employees.

These metadata will apply to all file types: image files, layout documents, Microsoft Office documents and so on. Search results can be refined to one specific type of file such as a digital image, or a type of layout document such as brochures or ads. This brings us to

Case 2: Job types and job categories vocabulary

Also needed to help serve up the agency's work history, these metadata help teams find more specific kinds of work. Job types and categories apply exclusively to layout documents (InDesign files, Quark files and pdfs exported from them). Several teams, for a variety of reasons, need to recall prior work based on layout types. A typical case would be a team needing to recall a brochure made for a specific client in a known year.

In the past, searchers relied on a search of filenames that may or may not describe the layout. The filenames consist of two parts: the rigidly maintained alpha-numeric job code mentioned above followed by a brief user-generated description of the layout. This second part usually refers to the layout type, but could be abbreviated by the user creating the name, or modified for other reasons. Our previous DAM system's folder structure allowed for navigation into folders based on client and year. After that, barring any additional information and with no predictable way for a job type to be entered in the filename, a user could be left to flip through many pages of small thumbnails to find the needed file.

Now all layout documents will be tagged with predictably entered job type metadata. A dictionary document imported into the DAM can further help guide users through common variations of types. (A user looking for a Visual

Aid may try searching for VisAid, VIs Aid or Visual Aid.) Accounting for these variations, users will still be searching based on common language, because the DAM system grabs metadata translated from the billing system. That billing system is where a project's life begins, and determines the type under which a layout is categorized for its entire life cycle. There is very little risk that an 'Ad' will be tagged as a 'Convention Panel'.

Another benefit of this vocabulary is the pairing of job types with broader job categories. Capturing work based on these categories is tremendously beneficial for business development teams and others who are

developing portfolios of agency work. As with the consumer needs information, this kind of insight into agency work was not previously formalized, and was owned by long-term employees and others well versed in agency history.

The new controlled vocabularies make our DAM system an instrumental tool for finding assets to generate new business and save production time. Just as importantly, they draw from language created and used by subject matter experts, and reflect categorizations in other agency-wide systems. We are reaping the rewards of approaching DAM as an enterprise-wide team effort.