5.4. Describing Non-text Resources

Many of the principles and methods for resource description were developed for describing text resources in physical formats. Those principles have had to evolve to deal with different types of resources that people want to describe and organize, from paintings and statues to MP3s, JPEGs, and MPEGs.

Some descriptions for non-text resources are text-based, and are most often assigned by people. Other descriptions are in non-text formats are extracted algorithmically from the content of the non-text resource. These latter content-based resource descriptions capture intrinsic technical properties and in some domains are able to describe aboutness with some accuracy, thanks to breakthroughs in machine learning.

5.4.1. Describing Museum and Artistic Resources

The problems associated with describing multimedia resources are not all new. Museum curators have been grappling with them since they first started to collect, store, and describe artifacts hundreds of years ago. Many artifacts may represent the same work (think about shards of pottery that may once have been part of the same vase). The materials and forms do not convey semantics on their own. Without additional research and description, we know nothing about the vase; it does not come with any sort of title page or tag that connects it with a 9th-century Mayan settlement. Since museums can acquire large batches of artifacts all at once, they have to make decisions about which resources they can afford to describe and how much they can describe them.

German art historian Erwin Panofsky first codified one approach to these problems of description. In his classic Studies in Iconology, he defined three levels of description that can be applied to an artistic work or museum artifact. Figure 5.6, “Contrasting Descriptions for a Work of Art.” contrasts these three levels in the descriptions of a marble statue. It also shows the striking differences between the EXIF description in a digital photo of the statue and those created by people. 298[Mus]

> [298][Mus] (Panofsky 1972) proposes these three levels of description:
>
> *Primary subject matter*
>
> At this level, we describe the most basic elements of a work in a generic way that would be recognizable by anyone regardless of expertise or training. The painting The Last Supper, for example, might be described as “13 people having dinner.”
>
> *Secondary subject matter*
>
> Here, we introduce a level of basic cultural understanding into a description. Someone familiar with a common interpretation of the Bible, for example, could now see The Last Supper as representing Jesus surrounded by his disciples.
>
> *Intrinsic meaning or interpretation*
>
> At this level, context and deeper understanding come into play—including what the creator of the description knows about the situation in which the work was created. Why, for example, did this particular artist create this particular depiction of The Last Supper in this way? Panofsky

posited that professional art historians are needed here, because they are the ones with the education and background necessary to draw meaning from a work.

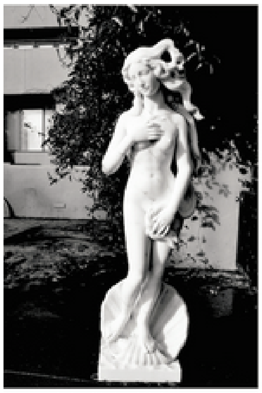
In other words, Panofsky saw the need for many different types of descriptors—including physical, cultural, and contextual —to work together when making a full description of an artifact.

Professionals who create descriptions of museum and artistic resources, architecture and other cultural works typically use the VRA Core from the Library of Congress, or the Getty Trust Categories for the Description of Works of Art (CDWA), a massive controlled vocabulary with 532 categories and subcategories. A CDWA-Lite has been developed to create a very small subset for use by non-specialists. 299[Mus]

[299][Mus] For CDWA, see (Harpring2009) at http://www.getty.edu/research/publications/electronic_publications/cdwa/.

For CDWA-Lite, see (Getty2006) at http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.pdf.

Figure 5.6. Contrasting Descriptions for a Work of Art.



EXIF Summary

Make	NIKON CORPORATION
Model	NIKON D90
Aperture	9
Exposure Time	1/320 (0.003125 sec)
Lens	ID AF-S DX VR Zoom-Nikkor 18-105mm f/3.5-5.6G ED
Focal Length	21.0 mm
Flash	Auto, Did not fire
File Size	4.7 MB
File Type	JPEG
Image Height	4288
Image Width	2848
Date & Time	2012:12:03 10:31:14

3 Levels

Primary
Marble statue of nude woman standing on a seashell.

Secondary
Statue made in 2005 by Lucio Carusi of Carrara, Italy, titled "Venus", made of local marble.

Interpretive
This is a 3d transformation of the 1486 painting by Italian painter Sondro Botticelli, titled "The Birth of Venus", now in the Uffizi Gallery in Florence. Carusi's Venus is substantially slimmer in proportions than Botticelli's because of changing notions of female beauty.

Descriptions for works of art can contrast a great deal, especially between those captured by a device like a digital camera and those created by people. Furthermore, the descriptions created by people differ according to the expertise of the creator and the amount of subjective interpretation applied in the description.

(Photo by R. Glushko. The statue, titled “Venus,” was made by Lucio Carusi, of Carrara, Italy, and is currently part of a private collection.)

5.4.2. Describing Images

Digital cameras, including those in cell phones, take millions of photos each day. Unlike the images in museums and galleries, most of these images receive few descriptions beyond those created by the device that made them. Nevertheless, a great many of them end up with some limited descriptions in Facebook, Instagram, Flickr, Picasa, DeviantArt, or others of the numerous places where people share images, or in professional image applications like Light Room. All of these sites provide some facilities for users to assign tags to images or arrange them in named groups.

Many different computational approaches have been used to describe or classify images. One approach uses the visual signature of an image extracted from low-level features like color, shape, texture, and luminosity, which are then used to distinguish significant regions and objects. Image similarity is computed to create categories of images that contain the same kinds of colors, objects, or settings, which makes it easy to find duplicate or modified images. 300[Com]

> [300][Com] See (Datta et al. 2008). The company Idée is developing a variety of image search algorithms, which use image signatures and measures of visual similarity to return photos similar to those a user asks to see.

For computers to identify specific objects or people in images, it is logically necessary to train them with images that are already identified. In 2005 Luis van Ahn devised a clever way to collect large amounts of labeled images with a web-based game called ESP that randomly paired people to suggest labels or tags for an image. The obvious choices were removed from contention, so a photo of a bird against a blue sky might already strike “bird” and “sky” from the set of acceptable words, leaving users to suggest words such as “flying” and “cloudless.” Van Ahn also invented the reCAPTCHA technique that presents images of text from old books being digitized, which improves the accuracy of the digitization while verifying that the user of a web site is a person and not a robot program. 301[Web]

> [301][Web] (von Ahn and Dabbish 2008).

However, if short text descriptions or low-level image properties are the only features available to train an image, otherwise irrelevant variations in the position, orientation, or illumination of objects in images will make it very difficult to distinguish objects that look similar, like a white wolf and the wolf-like white dog called a Samoyed. This problem can be addressed by using deep neural networks, which exploit the idea that low-level image features can be combined into many layers of higher-level ones; edges combine to form motifs or patterns, patterns combine to form parts of familiar objects, and parts combine to form complete objects. This hierarchical composition enables the highest-level representations to become insensitive to the lower-level variations that plague the other approaches.

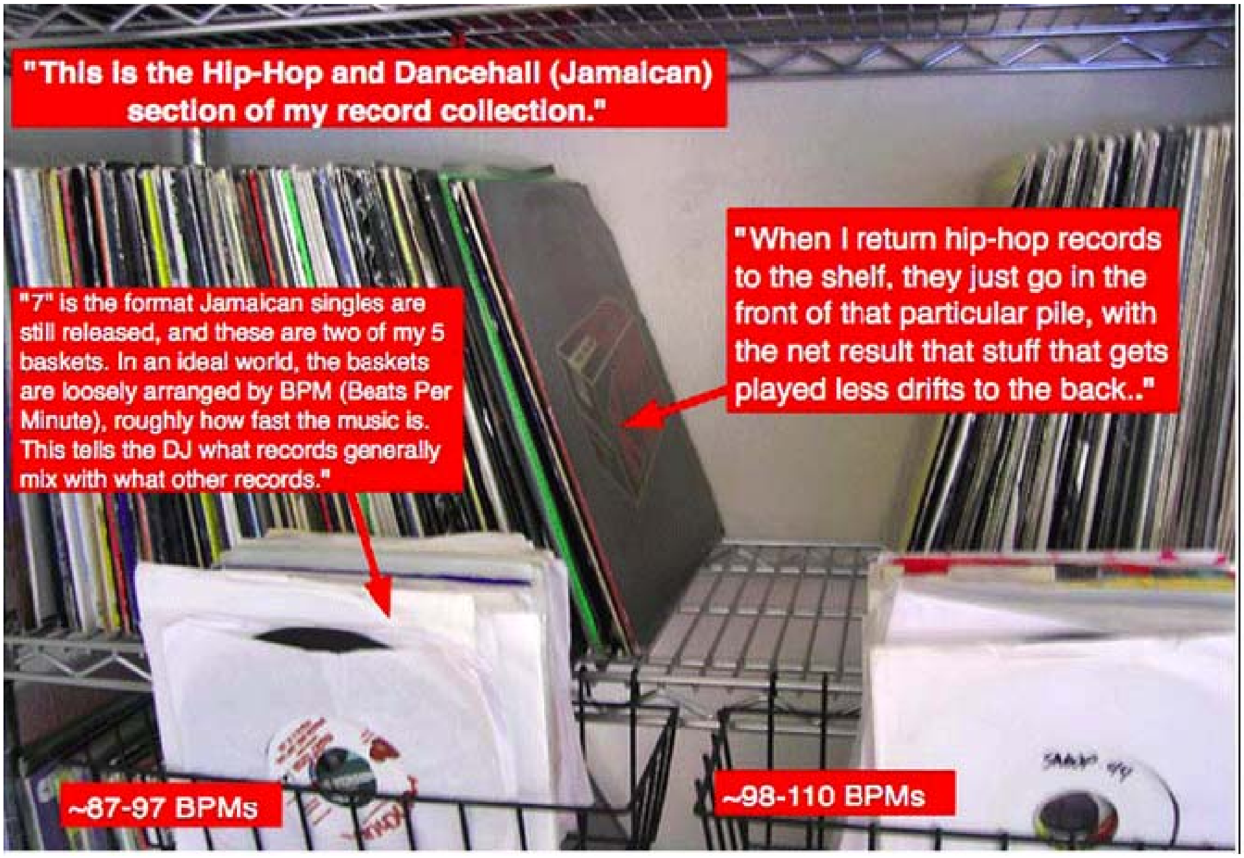
In 2012, when deep learning techniques were applied to a dataset of about a million images that contained a thousand different object categories, they reduced the error rate by half. This spectacular breakthrough, and the fact that the deep learning techniques that derive layers of features from the input data are completely general, rapidly caused deep learning to be applied to many other domains with high-dimensional data. Facebook uses deep learning to identify people in photos, Google uses it for speech recognition and language translation, and rapid captioning for images and video are on the horizon. Wearable computers might use it to layer useful information onto people's views of the world, creating real-time augmented reality. 302[Com]

[302][Com] The key idea that made deep learning possible is the use of “backpropagation” to adjust the weights on features by working backwards from the output (the object classification produced by the network) all the way back to the input. Mathematically-sophisticated readers can find a concise explanation and history of deep learning in (LeCun, Bengio, and Hinton 2015). LeCun and Hinton were part of research teams that independently invented backpropagation in the mid 1980s. Today, LeCun heads Facebook’s research group on artificial intelligence, and Hinton has a similar role at Google.

5.4.3. Describing Music

(Written by Graham Freeman for the 3rd Professional Edition of TDO)

A DJ Describes and Organizes Music


"This is the Hip-Hop and Dancehall (Jamaican) section of my record collection."
"7" is the format Jamaican singles are still released, and these are two of my 5 baskets. In an ideal world, the baskets are loosely arranged by BPM (Beats Per Minute), roughly how fast the music is. This tells the DJ what records generally mix with what other records."
"When I return hip-hop records to the shelf, they just go in the front of that particular pile, with the net result that stuff that gets played less drifts to the back.."
~87-97 BPMs
~98-110 BPMs

Casual music fans might describe their music using the names of the songs or performers and might organize it according to genres like “Pop,” “Rock,” or “Classical.” A professional DJ, however, emphasizes different properties, especially the beats per minute of the music.

This annotated photo shows a portion of the music collection of noted DJ “Kid Kameleon” (http://kidkameleon.com/).

(Photo and annotation by Matt Earp. Used with permission.)

Some parts of describing a song are not that different from describing text: You might want to pull out the name of the singer and/or the songwriter, the length of the song, or the name of the album on which it appears. But what if you wanted to describe the actual content of the song? You could write out the lyrics, but describing the music itself requires a different approach.

Describing music presents challenges quite different from those involved in describing texts or images. Poems and paintings are tangible things that we can look at and contemplate, while the aural nature of music means that it is a fleeting phenomenon that can only be experienced in the performative moment. Even musical scores and recordings, while as much tangible things as paintings and poems, are merely containers that hold the potential for musical experience and not the music itself. Most contemporary popular music is in the form of songs, in which texts are set to a melody and supported by instrumental harmonies. If we want to categorize or describe such music by its lyrical content, we can still rely on methods for describing texts. But if we want to describe the music itself, we need to take a somewhat different approach.

Several people and companies working in multimedia have explored different processes for how songs are described. On the heavily technological side, software applications such as Shazam and Midomi can create a content-based audio fingerprint from a snippet of music. Audio fingerprinting renders a digital description of a piece of music, which a computer can then interpret and compare to other digital descriptions in a library. 303[Com]

[303][Com] (Cano et al. 2005).

On the face of it, contemporary music streaming services represent the apex of music classification and description. Pandora, for example, employs trained musicologists to listen to the music and then categorize the genres and musical materials according to a highly controlled musical vocabulary. The resulting algorithm, the “Music Genome,” can essentially learn to define a listener’s musical tastes by means of this musical tagging, and can then use that information to suggest other music with similar characteristics. 304[Com]

[304][Com] (Walker 2009).

But musicians have been thinking about how to describe music for centuries, and while the Music Genome certainly brims with complexity, it pales in comparison to the sophistication of the much older “pen-and-paper” methods from which it derives. Ethnomusicology (loosely defined as the study of global musical practices in their social contexts) has arguably made greater strides towards comprehensive descriptions of musical resources than any other field of musicological study. Since the late 19th century, ethnomusicologists have created complex methods of notation and stylistic taxonomies to capture and categorize the music of both Western and non-Western cultures.

Hungarian composer and scholar Béla Bartók collected and transcribed thousands of Eastern European folk songs to which he applied a complex classification system to group them into “families” derived from melodic archetypes. More recently, American ethnomusicologist Alan Lomax’s Cantometrics project classified songs collected from around the word according to 37 style factors in an effort to create a common controlled vocabulary that would facilitate cross-cultural comparison and analysis. 305[LIS]

[305][LIS] Bartók's method for transcribing and categorizing each tune into families was as follows:

1. All tunes end on the note "g" for ease of comparison;
2. Tunes are divided and categorized according to the number of lines;
3. Tunes are classified according to the placement of the final note of various tune lines with the final note indicated by figures;
4. Sub-groups are categorized according to the number of syllables to each tune line;
5. Tunes are categorized according to their melodic compass with the lowest and highest note of each tune labeled.

It is not difficult to see the parallels of this method with the Pandora algorithm, as well as the greater level of descriptive detail afforded by Bartók's method. See (Bartók 1981).

Every folk song collection contains several examples of the same song performed at various times by the same singer or by many different singers. Often these different songs (or "variants") are so drastically different that we begin to ask the question: "At what point does a variant become a completely new piece of music?" Up until the beginning of the twentieth-century, many scholars believed that variants were simply poor performances by folk singers who were attempting to recreate a pristine, archetypical version of the song.

It wasn't until Australian collector Percy Grainger suggested that variants represented a vital and dynamic performance practice among folk singers that the idea of variants as flawed archetypes gave way to one in which all performances were unique entities unto themselves that possess what Wittgenstein would eventually refer to as "family resemblances" with one another. (Grainger 1908)

More recently, Newsweek magazine compiled a list of 60 different versions of Leonard Cohen's "Hallelujah," many of which differ so drastically from Cohen's original as to seem to be completely different songs with only the most rudimentary family resemblances. Does "Hallelujah" as a "work" even exist anymore? Or is it simply an idea, a potential for music that only exists during each varied performance? (http://www.newsweek.com/60-versions-leonard-cohens-hallelujah-ranked-303580).

On a more granular level, musicians are endlessly innovative in finding ways to categorize, describe, and analyze not simply large-scale musical genres, but the notes themselves. In the accompanying photo showing the record collection of professional DJ "Kid Kameleon," we see that the records are arranged not simply by genre, but also by beats-per-minute (BPM). For Kid Kameleon, these records represent the resources of his musical creative process, and arranging them by BPM allows him to pull exactly the correct musical material he needs to keep the music flowing during a performance. His classification system is therefore a taxonomy that moves from the broad strokes of genre down to the fine grains of specific arrangements of notes and rhythms. This photo is not simply a picture of a record collection: it is a visual representation of an artist's creative process. 306[LIS]

[306][LIS] This method of organizing musical resources for ready access (physically and cognitively) is one that has both an illustrious past and a fascinating future. Musicologist Robert Gjerdingen has studied the way in which composers in 18th century Naples learned their art by

studying an organized system of musical schemata that could be expanded, strung together, and varied to create an endless series of pleasing compositions in the galant style of the period (Gjerdingen 2007). A current approach to this same idea can be found in the work of composer David Cope, whose Experiments in Musical Intelligence software ("Emmy" and the next-generation "Emily Howell") can analyze existing music, break its musical resources down into identifiable schema, and then recombine those schema to create a musical output in the style of the original musical input (Cope 2001). Emmy can recombine these elements in millions of different ways to produce compelling, convincing, and somewhat unnerving works in the style of any composer whose music has been fed to her. Different though they may all seem, 18th century Neapolitans, Kid Kameleon, and Emmy all represent a creative process dependent on the input, description, organization, recombination, and output of musical resources.

5.4.4. Describing Video

Video is yet another resource domain where work to create resource descriptions to make search more effective is ongoing. Video analytics techniques can segment a video into shorter clips described according to their color, direction of motion, size of objects, and other characteristics. Identifying anomalous events and faces of people in video has obvious applications in security and surveillance.307[Com] Identifying specific content details about a video currently takes a significant amount of human intervention, though it is possible that image signature-matching algorithms will take over in the future because they would enable automated ad placement in videos and television.308[Bus]

[307][Com] (Regazzoni et al. 2010) introduce a special issue in IEEE Signal Processing on visual analytics.

[308][Bus] One organization that sees a future in assembling better descriptions of video content is the United States' National Football League (NFL), whose vast library of clips can not only be used to gather plays for highlight reels and specials but can also be monetized by pointing out when key advertisers' products appear on film. Currently, labeling the video requires a person to watch the scenes and tag elements of each frame, but once those tags have been created and sequenced along with the video, they can be more easily searched in computerized, automated ways (Buhrmester 2007).