# Metadata, Mon Amour

*author: Lev Manovich*

*year: 2002*

Metadata is the data about data: keywords assigned to an image in a media database, a number of words in a text file, the type of codec used to compress an audio file. Metadata is what allows computers to "see" and retrieve data, move it from place to place, compress it and expand it, connect data with other data, and so on.

The title of this chapter refers to the ongoing modern struggle between the visual data, i.e., images, and their creators and masters – the humans. The later want to control images: make new images which would precisely communicate the intended meanings and effects; yield the exact meanings contained in all the images already created by human cultures; and, more recently, automate these and all-over possible image operations by using computers. The former can be said to "resist" all these attempts. This struggle has intensified and became more important in a computer age – more important because the ease with which computers copy, modify, and transmit images allows humans to daily multiply the number of media records available.

Creating metadata is not, however, only the economic and industrial problem to be solved – it is also a new paradigm to "interface reality" and the human experience in new ways. This is already demonstrated by a number of successful art projects that focus on new ways to describe, organize and access large numbers of visual records. Importantly, these projects propose not only *new interfaces* but also *new types of images*, or, more generally, "records" of human individual and collective experience: film/video recordings embedded within virtual space (Sauter, *Invisible Shape of Things Past*; Fujihata, *Field-Work@Alsace*); photographs of people/objects organized into networks/maps based on their semantic similarity (Legrady, *Pockets Full of Memories*; Walitzky, *Focus*).

In summary, in terms of its creative and "generative" potential, "metadating the image" paradigm means following four related directions: (1) inventing new systems of image description and categorization; (2) inventing new interfaces to image collections; (3) inventing new kinds of images which go beyond such familiar types as "a still photograph" or a "digital video"; (4) approaching the new "super-human" scale of visual data available (images on the Web, web cam recordings, etc.) not as a problem but as a creative opportunity.

In short: new structure – new interface – new image – new scale.

## Description

Ancient and modern cultures developed rich and precise systems to describe oral and written communication: phonetics, syntax, semantics, pragmatics, rhetoric, poetics, narratology, and so on. Dictionaries and thesauruses help us to create new texts; the search engines and the ever present "find…" command in our software applications help us to locate the particular texts already created, or their parts; narratology and poetics provide us with concepts to describe the semantics and the formal structure of literary texts.

Paradoxically, while the role of visual communication has dramatically increased over the last two centuries, no similar descriptive systems and/or search tools were developed for images. While we do have some concepts such as Panofsky's iconography and iconology, or Pierce's index – symbol – icon, they do not approach the richness, the generality, and the precision of concepts available to describe the texts. While In the last four decades there have been many attempts to import concepts from literary theory and linguistics into art history and visual culture studies, these imported concepts have not been widely adopted.

Often the professionals working in some cultural field develop their own terms and taxonomies that are more precise than the terms used by the theorists studying the same field from the outside. In the case of images, there are a few professional practices we can look at – for instance, Hollywood cinematography or Bauhaus art education – but overall, the image taxonomies used in various contemporary professional fields are also quite

limited. Stock photography agencies divide millions of photographs into a few dozen categories, with names such as “joy,” “business,” and” achievement”. Graphic designers and their clients typically use even more limited range of categories to describe their projects: “clean,” “futuristic,” “corporate,” “conservative.”

In short, the way we usually deal with the problem of image description is to reduce the image to one or a few verbal labels (called “keywords” in software applications). In other words, we use natural languages (English, Spanish, Russian, etc.) as metalanguages for images.

Interestingly, when modern theorists have tried to address the questions of visual signification, they often ended up performing similar reduction. This tendency in modern thought even received a special label – “verbocentrism.” For instance, while Roland Barthes stimulated the interest in visual semiotics with his pioneering articles published in the late 1950s and early 1960s, he simultaneously strongly questioned the possibility of an autonomous visual language. In "Rhetoric of the Image" [1] Barthes investigated significations conveyed by the objects and their arrangement and in fact disregarded any contribution to meaning by the picture itself. [2] In *Elements of Semiotics* Barthes directly denied that a specifically visual language is possible: "It is true that objects, images and patterns of behavior can signify, and do so on a large scale, but never autonomously; every semiological system has its linguistic admixture." [3] And finally, in *The Fashion System* Barthes explicitly analyzed not clothes but "written clothes." [4]

While semioticians, art historians, and art critics were going back and forth between stating, a la Barthes, that images do not have meanings without a linguistic support and, on the contrary, searching for a unique pictorial language, these subtle debates concerning what happens *inside a single image* became now somewhat irrelevant. Computerization of media society introduced a new set of conceptual and practical challenges. Forget our inability to understand and describe how a single image may signify this or that – we now have to worry about more banal problems: how to organize, archive, filter and search billions and billions of images being stored on our laptops, network drives, memory cards, and so on.

Of course, the questions of visual semiotics and hermeneutics still matter – but they need to be re-calibrated. *The cultural unit is no longer a single image, but a large scale structured or unstructured (such as the Web) image database*. This shift becomes clearly visible if we compare how visual epistemology works in *Blow-Up* (Antonioni, 1966), *Blade Runner* (Scott, 1982), and *Minority Report* (Spielberg, 2002). The protagonists of the first two films are looking for truth within a single photographic image. Panning and zooming into this image reveals new information about reality: the killer hiding in the bushes, the identity of a replicant. In contrast, the protagonist of *Minority Report* is looking for truth *outside a single image*: he works by matching the image of a future murder to numerous images of the city contained in a database to identify the location of the murder. The message is clear: by itself, a single image is useless – it only acquires significance in relation to a larger database.

## Structure

How did computer scientists and the image industries respond to the dramatic increase in the amount of media data available? The response has been to gradually shift towards more structured ways to organize and describe this data. The industries are moving from HTML to XML to Semantic Web; from MPEG-1 to MPEG-4 to MPEG-7; from "flat" lens-based images to "layered" image composites to discrete 3D computer generated spaces. [5] In all these cases the shift is from a "low-level" metadata (the fonts used in a PDF file, the resolution and compression settings of a digital video file) to a "high-level" metadata that describes the structure of a media composition and ultimately its semantics.

This gradual shift occurs in two complementary ways. One involves adding metadata to all the media data *already* accumulated during the last hundred or fifty years of media society. Slides, photographs, recordings of television programs, typewritten records stored in numerous archives, state, university, and corporate libraries – all of these are being digitized and stored in computer databases with the metadata usually entered manually. (Often the reports on these efforts read as though they came from fiction by Borges or Lem: for instance, as I write this, hundreds of thousands of slides in an art collection at my

university library are being digitized and logged; the recent report proudly announced that the speed of the process has reached 12,500 slides a month.)

The second is to assure that any media data generated *in the future* – from a page of text on the Web to an image snapped by a cell phone camera to a TV show – will contain "high-level" metadata. This involves implementing various structured media formats such as already mentioned MPEG-4 and MPEG7 that I will focus on here as my examples. [6] The designers of MPEG-4 describe it as "the content representation standard for multimedia information search, filtering, management and processing." MPEG-4 standard is based on the concept of a media composition that consists of a number of a media objects of various types, from video and audio to 3D models and facial expressions, and the information on how these objects are combined. MPEG-4 provides an abstract language to describe such a composition.

MPEG-7 represents the next logical step in a gradual transition towards structured media data that comes with machine and code readable descriptions of its structure and contents. MPEG-7 is defined as "a standard for describing the multimedia content data that supports some degree of interpretation of the information's meaning, which can be passed onto, or accessed by, a device or a computer code." It is worth quoting the longer passage from the ISO/IEC document describing the standard as it explains well the importance of the last part of this definition:

*More and more audiovisual information is available from many sources around the world. The information may be represented in various forms of media, such as still pictures, graphics, 3D models, audio, speech, and video. Audiovisual information plays an important role in our society, be it recorded in such media as film or magnetic tape or originating, in real time, from some audio or visual sensors and be it analogue or, increasingly, digital. While audio and visual information used to be consumed directly by the human being, there is an increasing number of cases where the audiovisual information is created, exchanged, retrieved, and re-used by computational systems. This may be the case for such scenarios as image understanding (surveillance, intelligent vision, smart cameras, etc.) and media conversion (speech to text, picture to speech, speech to picture, etc.). Other scenarios are information retrieval (quickly and*

*efficiently searching for various types of multimedia documents of interest to the user) and filtering in a stream of audiovisual content description (to receive only those multimedia data items which satisfy the user's preferences)…*

Audiovisual sources will play an increasingly pervasive role in our lives, and there will be a growing need to have these sources processed further. This makes it necessary to develop forms of audiovisual information representation that go beyond the simple waveform or sample-based, compression-based (such as MPEG-1 and MPEG-2) or even objects-based (such as MPEG-4) representations. Forms of representation that allow some degree of interpretation of the information's meaning are necessary. These forms can be passed onto, or accessed by, a device or a computer code.

MPEG-7 and similar schemes call for the inclusion of high-level metadata along with the media data that will enable computers to automatically process this data in a variety of data. But where would this metadata come from? I have briefly discussed above our overall tendency to describe images in terms of verbal labels. Can computers at least generate such labels automatically? Or maybe they would even finally allow us to describe image with more precision than natural languages?

Computerization creates a promise that images that traditionally resisted the human attempts to adequately describe them will be finally conquered. After all, we now easily find out that a particular digital image contains so many pixels and so many colors; we can also generate a histogram (in Photoshop 7.0 it is a command found under "image" menu) that shows up how frequently each value appears in the image; etc. In short, by turning an image into a mathematical object digital computers gave us a *new metalanguage for images – numbers*. Building on such simple statistics, a computer can also tease out some indications of image structure and semantics – for instance, it can easily automatically find most edges in photograph and sometimes even segment it into parts corresponding to individual objects.

Yet this promise may be only the illusion. The metadata provided by a image database software I use to organize my digital photos (iView MediaPro 1.1) tells me all kinds of

technical details such as what aperture my digital camera used to snap this or that image – but nothing about the image content (in technical terms, this is typical "low-level" metadata). Visual search engines that can deal with the queries such as "find all images which have a picture of X" or "find all images similar in composition to this one" are still in their infancy. More generally, after almost fifty years of research, computer vision systems still can only recognize objects in photographs or video when they know what these objects would be beforehand – presented with an arbitrary image, they become "blind."

In short, while computerization made the image acquisition, storage, manipulation, and transmission much more efficient than before, it did not help us much in dealing with its side effects – how to more efficiently describe and access the vast quantities of digital images being generated by digital cameras and scanners, by the endless "digital archives" and "digital libraries" projects around the world, by the sensors and the museums. Although standards such as MPEG-7 would allow computers to automatically process visual data based on metadata, there still remains a basic and very time-consuming task: entering this metadata. In other words, computers can help us but only after we help them first by feeding image descriptions.

**Scale**

The constantly growing quantities of media data which are already available in numerous public and private various archives and databases or which can be generated on purpose (by storing all access logs of a Web site, by continuously recording the output of some sensors or video cameras, and so on) represents not only the problem to be solved (if it can be solved at all) but also a unique artistic opportunity. [7] This unique opportunity can be summed up as the shift from "sampling" to "complete recording."

One of the most basic principles of narrative arts is what in computer culture called "compression." A drama, a novel, a film, a narrative painting or a photograph compresses weeks, years, decades, and even centuries of human existence into a number of essential scenes (or, in the case of narrative images, even a single scene). Non-essential is stripped away; essential is recorded. Why? Narrative arts have been always limited by the capacities

of the receiver (i.e., a human being) and of storage media. Throughout history, the first capacity remained more or less the same: today the time we will devote to the reception of a single narrative may range from 15 seconds (a TV commercial) to two hours (a feature film) to forty hours (the average time spend by a player on a new computer game) to maybe hundreds of hours (following a TV series or soap opera). But the capacity of storage media recently changed dramatically. Instead of 10 minutes that can fit on a standard film roll or two hours that can fit on a DV tape, a digital server can hold practically unlimited amount of audio-visual recordings. The same applies for audio only, or for text.

In short, if both traditional narrative arts and modern media technologies are based on sampling reality, that is, representing/recording only small fragments of human experience, digital recording and storage technologies greatly expand how much can be represented/recorded. This applies to granularity of time, the granularity of visual experience, and also to what can be called "social granularity" (i.e., representation of one's relationships with other human beings).

In regards to time, it is now possible to record, store and index years of digital video. By this I don't mean simply video libraries of stock footage or movies on demand systems – I am thinking of recording/representing the experiences of the individuals: for instance, the POV of single person as she goes through her life, the POVs of a number of people, etc. Although it presents combined experiences of many people rather than the detailed account of a single person's life, the work by Spielberg's Shoah Foundation is a relevant here as it shows what can be done with the new scale in video recording and indexing. The Shoah Foundation assembled and now makes accessible massive amount of video interviews with the Holocaust survivors: it would take one person forty years to watch all the video material, stored on Foundation's computer servers.

The examples of new finer visual granularity are provided by projects of Luc Courchesne and Jeffrey Shaw which both aim at continuous 360 o moving image recordings of visual reality. [8] One of Shaw's custom systems which he called Panosurround Camera uses 21 DV cameras mounted on a sphere. The recordings are stitched together using custom software resulting in a 360o moving image with a resolution of 6000 x 4000 pixels. [9]

Finally, the example of new "social granularity" is provided by the popular computer game *The Sims*. This game that is better referred to as "social simulator" models ongoing relationship dynamics between a number of characters. Although the relationship model itself can hardly compete with the modeling of human psychology in modern narrative fiction, since *The Sims* is not a static representation of selected moments in the characters' lives but a dynamic simulation running in real time, we can at any time choose to follow any of the characters. While the rest of the characters are off-screen, they continue to "live" and change. In short, just as with the new granularity of time and the new granularity of visual experience, the social universe no longer needs to be sampled but can be modeled as *one continuum*.

Together, these new abilities open up vast new vistas for aesthetic experimentation. They give us an unprecedented opportunity to address one of the key goals of art – a representation of reality and the human social and subjective experience of it – in new ways. In other words, what for the industry and computer science are difficult questions which need urgent solutions instead should be viewed as possibilities to play with. For instance, if it already possible to record and store practically unlimited number of still and moving images of one's existence, what kind of interface can we use to organize and navigate these images? Or, given that we now can use database software to classify, link, and retrieve images and image sequences along with other media, how can a database structure be used to represent the life of a modern city, the history of a place, etc. In short, behind the problem of visual metadata that became more urgent because of the new scale of media data available there is an exciting promise – the promise to rethink the nature of representation.

## Re-inventing media

Has the revolution in the scale of available storage been accompanied by the new ideas about how such media recording may function? It is not hard to see that most of the commercial and academic research into new structures and interfaces for organizing and accessing media data takes for granted commercially supported media formats and media conventions the way they exist today– photographs, consumer video, professional

television programs, and the like. For example, when ISO/IEC document which specifies MPEG-7 standard talks about various types of media that can be supported by this standard, the list include not only such "general" types as video and 3D models, but also more particular ones such as "talking heads" (an obvious reference to television and industrial video convention). Given that most of this research is geared towards *existing* applications by the industry, government agencies, and the military, this orientation towards media formats and conventions the way they exist today can be expected. However, some research projects are trying to re-invent media formats and their uses beyond what exists today. These projects come from different research paradigms that are not tied in to broadcasting and commercial video production industries the way MPEG community is.

Since the beginning of the 1990s, working within the paradigms of Computer Augmented Reality, Ubiquitous Computing, and Software Agents at places such as MIT Media Lab and Xerox Park, computers scientists advanced the notion of a computer as an unobtrusive but omni-present device which automatically records and indexes all inter-personal communications and other user's activities. A typical early scenario envisioned in the early 1990s involved microphones and video cameras situated in the business office which record everything taking place, along with indexing software which makes possible a quick search through the years' worth of recordings. More recently the paradigm has expanded to include capturing and indexing all kinds of experiences of many people. For instance, a DARPA-sponsored research project at Carnegie-Mellon University called Experience-on-Demand which begun in 1997 aims to "developed tools, techniques, and systems that allow users to capture complete records of personal experience and to share them in collaborative settings." [10] A report on the project from 2000 summarizes the new ideas being pursued as follows:

*Capture and abstraction of personal experience in audio and video as a form of personal memory.*

*Collaboration through shared composite views and information spanning location and time.*

*Synthesis of personal experience data across multiple sources.*

*Video and audio abstraction at variable information densities.*

*Information visualizations from temporal and spatial perspectives.*

*Visual and audio information filtering, "understanding," and event alerting. [11]*

Given that a regular email program already automatically keeps a copy of all send and received emails, and allows to sort and search through these emails, and that a typical mailing list archive Web site similarly allow to search through years of dialogs between many people, we can see that in the course of text communication this paradigm has already been realized. However, the difficulties of segmenting and indexing audio and visual media already discussed above are what delays realization of these ideas in practice in relation to other media. But the recording in mass itself is already can be easily achieved: all is takes is an inexpensive Web cam and a large hard drive.

What is important in this paradigm –- and this applies for computer media in general – is that *storage media became active*. That is, the operations of searching, sorting, filtering, indexing, and classifying which before were the strict domain of human intelligence, become automated. A human viewer no longer needs to go through hundreds of hours of video surveillance to locate the part where something happens – a software program can do this automatically, and much more quickly. Similarly, a human listener no longer needs to go through years of audio recordings to locate the important conversation with a particular person – software can do this quickly. It can also locate all other conversations with the same person, or other conversations where his name was mentioned, and so on.

To refer to the famous story by Borges, not only can computers make maps as big or larger than the territory, but they can also be used to make new types of maps impossible before. Instead of compressing reality to what the author considers the essential moments, very large chunks on everyday life can be recorded, and then put under the control of software. I imagine for instance a "novel" which consists of complete email archives of thousands of characters, plus a special interface that the reader will use to interact with this information. Or a narrative "film" in which a computer programs assembles shot by shot in real time, pulling from the huge archive of surveillance video, old digitized films, Web cam

transmissions, and other media sources. (From this perspective, Godard's *History of Cinema* represents an important step towards such database cinema. Godard treats the whole history of cinema as his source material, traversing this database back and forth, as though a virtual camera flying over a landscape made from old media.)

As this essay has tried to suggest, "metadating the image" paradigm can be looked at as a problem to be solved or as a unique creative opportunity to pursue. This paradigm points toward four directions for artistic research – new structure / new interface / new image / new scale – which are interrelated. New *scale* in the quantity of media available makes it difficult to use this data efficiently without automation. The automation – that is, processing of media by computers – requires new *structured* media formats such as MPEG-7 that include *metadata describing the semantics* of the data. The same change in scale calls for new *interfaces* that would allow human users to navigate and access media collections efficiently. But since the interface can be approached not just as a tool but also as a cultural form – a mechanism to "interface reality" as well as to construct new reality – working on such new interfaces to media becomes an important task for media/software arts. (While new media artists have extensively critiqued existing software interfaces in general and developed many particular alternatives, surprisingly little energy has been spend so far thinking on how we can interface image and other media collections in new ways.) Finally, along with creating new structures and new interfaces to existing media forms, both researchers and artists are also working on new media forms including new forms of visual media – new *images* which by themselves already "interface reality" in new ways.

## References:

[1] Barthes, Ronald, trans. (1964). "Rhetoric of the Image." *Image -Music - Text*. Ed. Stephen Heath. New York: Hill and Wang, 1977. 32-51.

[2] Sonesson, Göran. *Pictorial Concepts. Inquiries into the Semiotic Heritage and its Relevance for the Analysis of the Visual World*. Lund, Sweden: Lund University Press, 1989. Page 127.

[3] Barthes, Ronald, trans. (1964). *Elements of Semiology*. New York: Hill and Wang, 1968. Page 10.

[4] Barthes, Ronald, trans. (1967) *The Fashion System*. New York: Hill and Wang, 1983.

[5] For a detailed discussion of compositing in terms of this shift, see the section "From Image Streams to Modular Media" in my *The Language of New Media* (MIT Press, 2001).

[6] MPEG-4 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group), the committee that also developed the successful standards known as MPEG-1 (1992) and MPEG-2 (1994). Version 1 of MPEG-4 was approved in1998, and version 2 in 1999. All quotations in this section are from http://mpeg.telecomitalialab.com/standards/.

[7] This and the following section use the material from my article "Reality Media" published as "Old Media as New Media: Cinema" in *The New Media Book*, edited by Dan Harries (London: BFI Publishing, 2002).

[8] For Courchesne's Panoscope project, see http://www.din.umontreal.ca/courchesne/; For Jeffrey Shaw's projects, see http://www.jeffrey-shaw.net. Both discuss their projects in relation to previous strategies of "experience representation" in panorama, painting, and cinema in *New Screen Media: Cinema/Art/Narrative*, edited by Martin Rieser and Andrea Zapp (London: BFI and Karlsruhe: ZKM, 2001).

[9] Private communication between Shaw and the author, July 4, 2002.

[10] http://www.informedia.cs.cmu.edu/. For more information on the project, see Howard D. Wactlar et al., "Experience-on-Demand: Capturing, Integrating, and Communicating Experiences Across People, Time, and Space," http://www.informedia.cs.cmu.edu/eod/; see also Howard D. Wactlar et al., "Informedia Video Information Summarization and Demonstration Testbed Project Description," http://www.informedia.cs.cmu.edu/arda-vace/. Both of these research projects were conducted at Carnegie-Mellon University; dozens of similar projects are going on at universities and industry research labs around the world.

[11] http://www.informedia.cs.cmu.edu/eod/EODforWeb/eodquad00d.pdf.