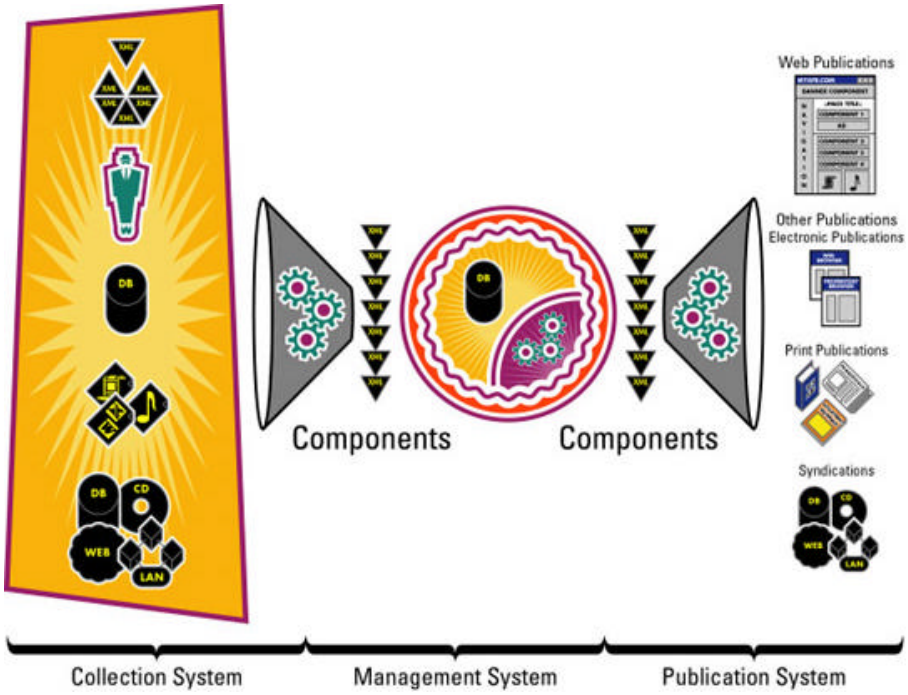
Defining Data, Information, and Content

A CM Domain White Paper

By Bob Boiko



This white paper is produced from the Content Management Domain which features the full text of the book "Content Management Bible," by Bob Boiko. Owners of the book may access the CM Domain at www.metatorial.com.

Table of Contents

Abstract

This paper contains the content of Chapter 1 of "Content Management Bible." It concerns the relationship between the terms in the title of the paper.

Computers were built to process data. Data consists of small snippets of information that have all the human meaning squeezed out of them. Today, people call on computers to process content. Like data, *content* is also information, but it retains its human meaning and context.

In this white paper I lay out one of the basic challenges of content management: Computers are designed to deal with data that's stripped of any context and independent meaning. Users want computers to deal with content, however, which is rich in context and meaning. How can you use the data technologies to manage and deliver very nondatalike content? This challenge isn't easy. If you err toward making your information too much like data, it looks mechanical and uninteresting to consumers. If you make your information too rich, varied, and context-laden, then you can't get a computer to automate its management.

The compromise, as you see in this white paper, is to wrap your information in a data container (known as *metadata*). The computer manages the data and the interesting, meaningful information goes along for the ride.

Content Is Not Data

Computers were first conceived as a way to perform computations that were too time-consuming or complex for humans. The model was (and to a large extent still is) as follows: If you can reduce a problem to a series of simple mechanical operations on numbers and logical entities (entities that are either true or false), it is amenable to solution by a computer

Computer professionals were either programmers or data input clerks. Programmers reduced problems to a series of mechanical operations according to the following simple maxim:

> *You input data; the computer processes it and then outputs it in a more useful form.*

Clerks took care of inputting the data. They sat in long rows and columns, typing long rows and columns of numbers as well as small phrases, such as first name/last name and street address. As time moved on, computer scientists invented databases (bases of data) to organize and hold vast quantities of these snippets.

As you may expect, some problems were better solved this way than others were. Thus, as computer technology developed, the use of computers moved naturally from science to manufacturing and finance, where numbers were still the main event. Today, of course, computers are in everything. But the part of everything that computers are in is the reducible part. The reducible part is the part where a finite set of very specific rules operating on numbers and logical entities can yield a useful result.

The idea of computers as data-processing machines runs deep and continues to this day as the main thing that computers do. On the other hand, everyone knows that most users want computers to do more than grind finely through mountains of snippets. Today, people want computers to sift through mountains of large, complete chunks (not snippets) of information and deliver the ones that they want most at that moment. In addition, people want computers to deliver information of the quality that they expect from more familiar sources of information, such as books, radio, TV, and film.

From manufacturing and finance, computers moved to the business desktop as the replacement for the typewriter and the paper-based spreadsheet. Then, three related developments occurred in the personal-computer industry. Together, the following breakthroughs set the stage for a major change in our expectations of what computers can, and are, to do:

✍✍ Digital media creation (images, sounds and video) became possible.

- Digital media output (color displays, sound cards, and video accelerators) became available.
- Consumer-oriented mass removable storage (CD-ROMs, in other words) became available and cheap.

These developments led to the meteoric rise of the multimedia industry. For the first time, you could create and cheaply deliver actual information and not just data. Soon, multimedia CD-ROMs proliferated, with everything from encyclopedias to full-motion games. You can now consider your computer an actual replacement for familiar information channels such as books, TV, and radio. What these traditional channels deliver is *content* and not data.

What the multimedia industry began, the Web is in the process of completing. Today, getting your content online isn't only possible, but it's often preferable to obtaining it in any other way. I now listen to more music and talk on my laptop computer and Personal Digital Assistant (PDA) than I do on my radio or stereo. Still, I'm usually frustrated whenever I go to the Web because I expect to quickly locate the content that I want and see it presented at least as well as - in the traditional channels. Unfortunately, that's not always what happens.

Note

If you've worked with digital content for a while, then you realize just how sticky a paradox this situation is. People expect access to prove easy and presentation to seem compelling. If they are, it's only because someone's put in an enormous amount of effort behind the scenes to make everything appear so easy and compelling. Making content natural is an unnaturally difficult endeavor.

Although users' needs and expectations changed, the guts of the computer didn't. Ten years ago, people came to computers to input, process, and output data. Today, most people come to find and consume content. At the base of all computer technology, however, is still the idea that you can reduce any problem to a set of simple instructions working on discrete and structured snippets.

Data and content are different, certainly, but that difference doesn't mean that they don't interact. In fact, innumerable transitions from one to the other occur every day. Moreover, from the standpoint of the computer system, content doesn't exist (only data exists. Today, users have few tools for dealing with content as content. Instead, you must treat it as data so that the computer can store, retrieve, and display it.

Consider, for example, a typical Web interaction where you may go to a music site. You browse a page that features a music CD that you like; you add it to your shopping cart and then pay for it. What you experience is a series of composed Web pages with information about music as well as some buttons and other controls that you use to buy it. All in all, the experience feels like a content-rich interaction. What happens behind the scenes, however, is a set of data-oriented computer programs exchanging data with a database.

Some of the data behind the scenes is very contentlike. A database stores, for example, the feature article with the artist's picture. The artist's name is in one field, and the text of the article is in another. A third field contains the picture of the artist. Some of the data is very datalike. It consists of numbers and other snippets that create an economic transaction between you and the record company. A database stores your credit card number, your order number, the quantity that you order, and the order price, for example, and uses them in calculations and other algorithms. Some of the data is in-between data and content. The song catalog contains song names, running time, price, and availability, which are all snippets of information that can look a whole lot like data as the site's database stores them, but they appear more like content as the site displays them (as shown in Figure 1).

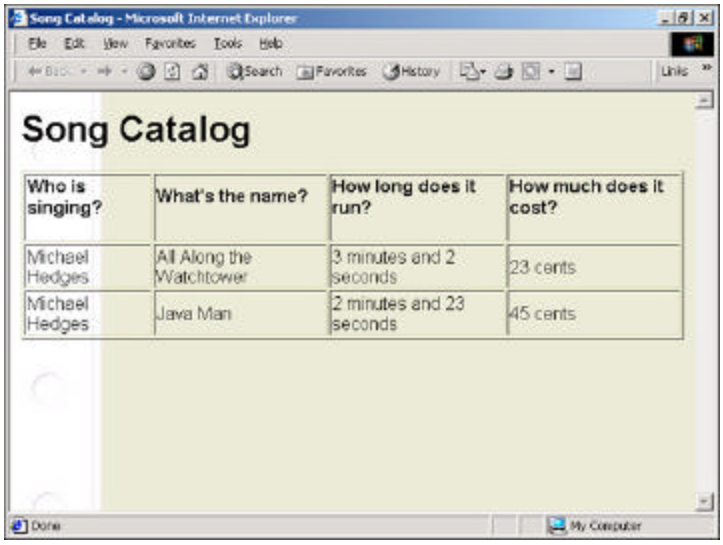


***Figure 1**: As a site stores content, it can look a lot like any other data. In displaying that content, however, it can't look like data if you want to hold your audience.*

In the database, it's all just data. On the page, the transaction data still looks a little like data; while the feature-article data looks nothing like data, and the catalog data can retain or lose as much of its data appearance as you want. On a well-designed page, however, visitors perceive it all as content.

So, from the user's perspective, information is all content, while from the computer programmer's perspective, it's all data. The trick to content management, in an age when the technology is still data-driven, is to use the data technologies to store and display content.

Content Is Information Put to Use

People in the computer world more or less agree to the definition of *data*. Data are the small snippets of information that people collect, join together in data records, and store in databases. The word *information*, on the other hand, contains all meanings, and no meaning, at the same time. You can rightly call anything, including data, *information*.

The word *information* holds a lot of meanings. For the purpose of this white paper, I use a pretty mundane definition. I take *information* to mean all the common forms of recorded communication, including the following:

- Text, such as articles, books, and news.
- Sound, such as music, conversations, and reading.
- Images, such as photographs and illustrations.
- Motion, such as video and animations.
- Computer files, such as spreadsheets, slide shows, and other proprietary files that you may want to find and use.

Before you ever see a piece of information, someone else has done a lot of work. That someone else has formed a mental image of a concept to communicate, and used creativity and intellect to craft words, sounds or images to suit the concept (thus crystallizing the concept). The person has then recorded the information in some presentable way. The author of the information pours a lot of personality and context into the information before anyone else ever sees it. So, unlike data, information doesn't naturally come in distinct little buckets, all displaying the same structure and behaving the same way. Information tends to flow continuously, as a conversation does, with no standard start, end, or attributes. You disrupt this continuity at your own peril. If you break up

information, then you run the risk of changing or losing the original intelligence and creativity that the author meant the information to express. If you break up information, then you run the risk, too, of losing track of, or disregarding, the assumptions the author made about the audience and the purpose of the information.

The now defunct ContentWatch organization (http://www.contentwatch.com/what.html) gave the following definition of content:

> *"What is 'content'? Raw information becomes content when it is given a usable form intended for one or more purposes. Increasingly, the value of content is based upon the combination of its primary usable form, along with its application, accessibility, usage, usefulness, brand recognition, and uniqueness."*

Information that passes casually around in the world isn't content. It becomes content after someone grabs it and tries to make some use of it. You grab and make use of information by adding a layer of data around it.

The step of adding data may seem like a small step (from information to its use (but it's not. By refocusing from the nature of information to its practical application, you open up a world of possibilities for applying the data perspective to information. The crux is that, although you can't treat information itself as data, you can treat information use that way. As you begin to use information, you wrap it in a set of simplifying assumptions (metadata) that enable you to use computers to manage that use. Humans are mandatory in creating the information and figuring out the simplifying assumptions that wrap the information, but after that, the computer is fine by itself in doling information out in a way that usually makes sense.

By wrapping information in data, a small action by a person can trigger a lot of work by the computer. Suppose, for example, that to simplify information management, you decide that you need to consider a piece of information as 1) new; 2) ready to publish; or 3) ready for deletion. By itself, no computer can decide which of these statuses to apply to a piece of information. By wrapping your information with a piece of metadata known as *status*, however, and by having a person set the status metadata, you can use a computer to perform a lot of work based on the status. The computer doesn't need to know anything about the information itself; it just needs to know what status a human is applying to the information, as the following list outlines:

- If you give a piece of content the status "new," the computer sends a standard e-mail message to a designated reviewer.
- If you give a piece of content the status "ready-to-publish," the computer outputs it to a designated Web page.
- If you give a piece of content the status "ready-for-deletion," the computer removes it from the Web page and deletes it from the system.

In this way, the computer can accomplish a lot of work as the result of a small amount of work by people.

The Web version of the Merriam-Webster dictionary (www.m-w.com) defines content in part as follows:

> *"1 a: something contained (usually used in plural [the jar's contents] [the drawer's contents b: the topics or matter treated in a written work [table of contents]..."*

This definition provides a nice angle on content (something that *something else* contains. By switching from information to content, you're switching from a consideration of a thing to its container. You're shifting the focus from the information itself to the metadata that surrounds it. The container for information is a set of categories and metadata that... well... contain the information. This additional data corrals and confines that information and packages it for use, reuse, repurposing, and redistribution.

If content is information that you put to use, the first question to ask is, "What use?" What is the purpose behind marshalling all this information? For such a simple question, an astounding number of content management projects go forward without an answer. Or, to state the situation more precisely, all projects have some purpose, but the purpose may have little to do with the content that the project involves. A question such as "Why are we creating this Web site?" all too often receives one of the following answers:

- Because we need to.
- Because our competitors have one.
- Because our CEO thinks that we need one.
- Because everyone's clamoring for one.

These answers initiate a project, but they don't *define* it. To define the project, you need to answer the following question, "What is the *purpose* of the content we're about to put together?" If you provide a solid and well-stated answer, it then leads naturally to how you need to organize the content to meet your goal. The key to a good purpose is that it's *specific* and *measurable*. Following are some examples of good purposes for an intranet:

- To provide a 24-hour turn-around in getting any new product data from the product groups to the field sales representatives.
- To provide a site with all articles from the identified sources that mention the identified competitive products.
- To show all the data on a pay-stub with a complete history for every employee.

Notice that these goals are pretty specific, so quite a few of them may prove necessary to motivate a large intranet. As you organize your goals, you organize the content behind them. Your process is complete after all your individual goals fit together into a coherent whole and you can ultimately summarize them under a simple, single statement such as "Full support for the field" or "Zero unanswered employee questions."

Content Is Information Plus Data

There is something human and intuitive about information that makes treating it simply as data quite impossible. With data, what you see is what you get. With information, much of what it conveys isn't actually in the information but, rather, in the mind of the person creating or consuming it. Information lives within a wider world of connotation, context, and interpretation that make it fundamentally not amenable to the data-processing model. In fact, the concept of data was created specifically to remove these subjective qualities from information so that computers could manage it with strict, logical precision.

So, can computers ever really manage information? I believe that they can (albeit poorly by human standards). Until a new computer processing model appears, you can put methodologies, processes, and procedures in place to "handle" information by using data-age tools. Rather than reducing the information to mere data, you can capture whole, meaningful chunks of information and wrap it in descriptive data that computers can then read and act on.

Metadata makes the context and meaning of information explicit enough that a computer can handle it. Adding data to information (to metadata, that is) helps split the difference between keeping the information whole and enabling data techniques to effectively manage it. The data that you add to information is a way of making the context, connotation, and interpretation explicit. More important, metadata can make explicit the kind of mind that you expect to interpret the information. By adding a piece of metadata known as *audience type* to each chunk of information that you produce, for example, you can make explicit any implicit assumption about toward whom you're directing the information. Then, a computer can perform the simple task of finding information based on who's on the other side of the terminal. Of course, not all tasks are this

simple, but the concept is always the same. You tag a large chunk of information with the data that the computer needs to access so that it can figure out what to do with that information.

Content, therefore, is information that you tag with data so that a computer can organize and systematize its collection, management, and publishing. Such a system, a *content management system*, is successful if it can apply the data methodologies without squashing the interest and meaning of the information along the way.

Until computers (or some newer technology) can handle content directly, you must -figure out how to use the data technologies to collect and deliver content. Using data technologies to handle content is a central theme of this white paper.

From Data to Content and Back

From the perspective of this white paper, which attempts to reconcile the data and content perspectives, what is data and what is content depends mainly on how you create, manage, and bring each type of information onto a page, as the following list describes:

- Transaction information are datalike snippets that you typically use to track the processes of buying and selling goods. For transaction information and other very datalike sources, you don't generally use a content management system (CMS). Usually, they are managed as part of a traditional data-processing system. The templating system of a CMS, however, often does manage the displaying of such information on the -publication page.
- Article information is the sort of information that is text-heavy and has some sort of editorial process behind it. You create and manage article information, and other very contentlike information, most often using a CMS. In fact, without this sort of information, the CMS has little reason to exist. The CMS is also generally responsible for displaying this sort of information.
- Catalog information, which is information that might appear in a directory or product listing, can go either way. You sometimes create and manage it by using a CMS, and sometimes a separate data-processing application handles it. Obviously, if an organization already has a full-featured catalog system, tying your CMS to it is the easiest way to go. On the other hand, if you need rich media and lots of text, along with the part numbers and prices in your catalog, then including the catalog as part of the overall CMS may make the most sense.

The purpose of content management isn't to turn all data into content. The purpose of content management is to oversee the creation and management of rich, editorially intensive information and to manage the integration of this information with existing data systems. The CMS must, in all cases, carry the responsibility for the creation of the final publication. If this publication includes both data and content, then the CMS is there to ensure that the right data and content appear in the right places and that the publication appears unified and coherent to the end consumer.

Summary

Data was invented because it was much easier to deal with than information. It's small, simple, and all its relationships are clearly known (or else ignored). Data makes writing computer programs possible. Information is large, complex, and rife with relationships that are important to its meaning but impossible for a computer to decipher.

Here are some points to keep in mind as you hear and discuss the terms data, information, and content:

- Content (at least for the purposes of this white paper) is a compromise between the usefulness of data and the richness of information. Content is rich information that you wrap in simple data. The data that surround the information (metadata) is a simplified version of the context and meaning of the information.
- In a content management system, the computer manages information indirectly through data. The content compromise says, "I can't get the computer to understand and manage information, so I must simplify the problem. I must create a set of data that represents my best guess of the important aspects of this information. Then, I can use the computer's data capabilities to manage my information via the simplified data."
- Someday, someone may invent computers that can deal directly with information. I'm not holding my breath, however, because to do so requires cracking some of computer science's hardest problems, such as artificial intelligence and natural language processing. In the meantime, everyone must make do with the current dumb but very strong beasts. (If someone does finally invent these new machines, they may need to change their name from computers to something less mechanical.)