INFORMATION MANUFACTURING: THE ROAD TO DATABASE QUALITY

by Stephen E. Arnold
Database Consultant

The positive results of a technical enterprise are immediate. They are felt at once, as in the case of electricity or television. The negative effects, however, are long-term and are felt only with experience.

-Jacques Ellul [1]

You build quality in, not try to control it.

-PBS promotional clip for a program about quality [2]

One of the most significant technical achievements of the last 30 years is machine-readable information. Databases have become commonplace because the *benefits* of computer-assisted information access are immediate. Researchers with Ph.D.'s or a grade-school education can check for articles and books in seconds. Airlines, publishers, banks, hospitals, and virtually all other organizations know that databases enable innovation, serve constituencies, and improve efficiency.

Databases are constructs. People design them and then put together processes to manufacture information. Databases require a controlling idea, a record structure, raw material (content), and policies. Database publishers describe their activities as *building a database, making records,* or *producing citations.* It is important to remember that a system creates machine-readable files. A flawed system manufactures products that do not meet customer needs.

My use of this term "information manufacturing" refers to the process of creating machine-readable files from the moment an idea forms in the maker's mind to the feedback a user of the database provides.

In this article I want to formalize the concept of *information manufacturing.* To begin, I want to review the stages through which electronic publishing has moved and what direction I see it going. Then I want to explore the linkage of marketing and manufacturing machine-readable files. The end of this short excursion will bring us to an issue a few voices have raised. Questions about quality have been handled with silence. I believe discussions of data quality are meaningless unless information manufacturing finds itself under the microscope.

TRENDS IN DATABASE BUILDING

The number, diversity, and scope of databases are to me astonishing. There are text and numeric databases. There are hierarchical, flat, and relational databases. There are single-object and multi-object databases. There are public and private files. There are databases with terabytes of data and those with a few hundred records.

To see the sweep, I have to step back and try to view databases in a suitably broad context. I know that this approach blurs important distinctions for specific types of databases. The advantage gained

Four Stages In Information Manufacturing

Stage	Name	Main Feature
19651981	Invention	Development of original systems and procedures; emulation of print paradigms
1982- 1992	Duplication	Innovation of particular features; emulation and enhancement of existing machine-readable files
1992-l 997	Reconstruction	Databases with high value re-engineered for manufacturing in information factories using advanced systems and software; emphasis on replacing human effort with software and system processes.
1997-2001	Proliferation	Application of automated information manufacturing techniques to a broad, diverse range of products and services; embedded information in previously unintelligent products.

is an overview of the direction of movement in the “information manufacturing” industry.

My use of this term “information manufacturing” refers to the process of creating machine-readable files from the moment **an** idea forms in the maker’s mind to the feedback a user of the database provides. The concept of information manufacturing assumes that the database building is an ongoing activity. Although I will rely upon my experiences in commercial database publishing, I believe that these concepts I am articulating apply to other information manufacturers as well. To illustrate: a hospital capturing patient data for insurance claims is as much an information manufacturer as a commercial publisher like the Bureau of National Affairs. The common denominator is the application of computer technology to building or making machine-readable files.

The First Stage: Invention

The first stage of information manufacturing began in the 1960s. The U.S. government, scientists, and entrepreneurs applied computer technology to scientific information storage and retrieval. By 1980, information manufacturing had moved from bold experimentation when each solution was an innovation to technique, the application of procedures.

The earliest database builders had to write programs to get the results they desired. These programs had to be woven into a series of processes, some computer-assisted, some done by the innovators themselves. Each time a problem was solved, the innovators had other problems to solve. There were no tiles. But hardware, software, and systems evolved. Each advance presented new challenges.

The characteristics of this first stage of information manufacturing include:

- Little automation. Intensive “human” involvement in design and development of database systems, procedures, and records.

- Print models. Reliance on established models or metaphors for record structure (for example, machine-readable databases were electronic versions of printed indices).

- Experts only. Then slow but steady migration of the expertise needed to build (and use) a database from technical specialists to a broader audience.

In the period from 1970 to 1952, new databases and services started appearing. Commercial online services, specialized batch data base services, and a range of other resources developed to help people build, use, and maintain machine-readable files.

Information manufacturing processes grew more efficient as organizations gained database experience. The characteristics of databases designed and manufactured at this time include:

- Reliance on human intellect to select data for the file

- Manual data entry, first at the database producer’s facility and eventually in other locations to help contain costs

- Homogeneous records; that is, fixed-field length or standardization of the record structure

- Low-volume of production at start up with production increasing each year as funds or experience permitted.

The achievements in this stage of information manufacturing were significant. Many of the most successful commercial files on which people rely date from this first stage. Examples: census data, fixed-field databases, and commercial databases of company financials, and citation (index-only and index’-and-abstract databases).

The Second Stage: Duplication

The second stage of database building extends from 1982 to the present. This ten-year interval may seem too large because the explosion of machine-readable files, technologies, and markets seems to demand finer distinctions.

My point is that database manufacturing volume accelerates rapidly in this stage. Innovation is anchored in moving information from print to electronic versions. Online services like Dialog Information Services and Mead Data Central expanded their offerings by several orders of magnitude: more databases, more technological power, more features. Databases penetrated virtually every organization in commercial and noncommercial sectors,

Why? First, the flow of expertise from technical specialists in database building to a person with little or no formal technical training was an important factor. Organizations of any size, almost anywhere in the world could build machine-readable files. Other key factors were the advent of personal computers, easv-to-use software, and market demanh. Thus, many forces energized the ebb and flow of information manufacturing in this second, familiar stage.

The databases created since 1982 built upon the machine-readable files in the market. Of course, new types of databases were invented. In 1984, the first videodisc appeared at an Information Industry Association meeting. This marvel combined sound, words and data. But products similar to those already available

→

from other publishers flowed from competing information factories. This contrasts with the first stage when the data pioneers were wrestling with how-tos, not consciously creating me-toos.

Since the end of the 1980s, trade shows featuring commercial databases have become predictable, featuring look-alikes, repositioned databases, technical gadgets, and price changes. A hot market-business information, for example-attracted scientific and technical publishers as well as new entrants. Database opportunity became synonymous with information saleable to business.

The characteristics of databases created in the duplication stage include:

- Incremental improvements of existing models; that is, examine other databases and add enhancements wanted by customers or users.
- Increasing reliance on machine-assistance for database creation; for example, optical character recognition instead of manual data entry and taking direct feeds of full-text materials.
- Greater throughput. Integration of personal computers into the manufacturing process and moving labor-intensive tasks to Korea, Ireland, and other countries where the cost of labor was less.
- Emergence of integrated manufacturing operations that create primary data, process data from other information producers, and distribute the data to specific markets; for example, West Publishing Co. and VNU's Disclosure and Inforum units. The greatest legacy of this stage was the foundation for the reconstruction stage we are now entering.

The Third Stage: Reconstruction

Let me define "reconstruction." The increasingly affordable computing, software, and graphics technologies are setting the stage for what promises to be a decade of manufacturing innovation. Files created in the next five to seven years will be rebuilds, not mere enhancements, of highly successful databases that now exist in digital form. What databases are candidates for reconstruction? They are files that cannot be changed to meet the needs of their customers quickly enough to prevent competitors from trying to woo the customers with a new and improved product. The successful reconstructions will entail a rethinking of the fundamental manufacturing policies and processes that underlie the original database. These are the policies and processes that make the manufacturing system difficult, even impossible, to change.

There are three connotations in my use of the term "reconstruction:"

First, new and advanced technologies will let a competitor duplicate, insofar as possible, known winners. Predicasts, Inc. has begun the job of evolving PROMT into a different construct. Humana, Inc. (Louisville, Kentucky1 has undertaken a similar effort in patient data across its dozens of hospitals.

> **The increasingly affordable computing, software, and graphics technologies are setting the stage for what promises to be a decade of manufacturing innovation.**

Second, reconstruction opens wide the doors of opportunity for organizations whose principal business may not be electronic publishing as we understand the market today. Examples:

(1) A seed company may include data of significant commercial value to farmers. Access to the data requires buying the seed company's products.

(2) A network software company bundles live data feeds with the operating environment.

Third, reconstruction creates opportunities to redefine product categories, particularly in the area of multiobject database manufacturing and enhanced user services. Examples:

(1) Does the purchaser of a word processing package know what database of correctly spelled words checks the author's text?

(2) Games on CD-ROMs can teach the player about anatomy. What product category does a game, scientific information, and images fit?

(3) Lotus 1-2-3 for Windows includes animated tutorials to teach users how to complete certain tasks in the program.

Characteristics of information manufacturing in this stage will include:

- Greater dependence upon machine-generation of databases; for example, full-text and numeric databases built with minimal human input
- Introduction of new types of information constructs, which present data previously available in a format that restricted access or the inclusion of relevant "data objects" wanted and needed by users
- Proliferation of delivery options; for example, online, magnetic tape, direct broadcast, on-demand print, and others.

A growing awareness of the importance of "quality" in the machine-readable file will drive reconstruction. Greater familiarity with databases means more informed customers. Informed customers are better equipped to express their wants, needs, desires, and demands. William M. Bulkeley coined the phrase "reign of error" to express the importance of database quality [3].

Stage Four: Proliferation

Before addressing the issue of "quality," let me comment on the stage that I believe follows reconstruction. Once technologies allow existing information to be recast into more user-oriented forms, information manufacturing will be woven into the fabric of "things."

I am not sure electronic information will become a "consumer" product in the sense that household detergent is a "consumer product." In the proliferation stage, machine-readable databases will be an environmental factor. A purchaser will acquire a product because electronic information is embedded in the product and gives the product attributes that meet customer needs. Examples: (1) "Information" in the washing machine regulates water temperature, rinse cycle, even what detergent to add to the water. (2) An

→

automobile will include digital mapping and traveler information as part of the vehicle's infotainment module. Database options will be available: enhanced bed-and-breakfast listings. Upgrades will be databases that offer the buyer more "value" than the standard package.

"Value," however, underscores the importance of customer perception about databases. The products of an information manufacturing process are intangible until the "data" are shaped-by a query, a technology, or a human intelligence-into something tangible. The result may be a printed report, a display on a monitor, a CD-ROM, an interactive game, or some other deliverable artifact. At that moment, quality stops being an abstraction and becomes tangible. Data quality has moved from the shadows into the blaze of center stage.

THE PERCEPTION OF MACHINE-READABLE FILES

Marketing is a term applied to **a** wide range of functions that position the product and facilitate sales [4]. Marketing functions today range from advertising and public relations to product planning to warehousing and sales.

What is the link-this very moment— between information manufacturing, marketing, and nebulous concepts like "value" and "quality"? The question is at the core of justifying the existence of machine-readable files; for example:

- In banking: "We need to build a more effective database of our commercial customers. How much will this cost? How long will it take? What is the value of this investment to the bank?"

- In associations: "We have a wealth of information about our members, including the technical material each provides to us. How can we create a database of these resources? What will be the payoff for the members? What's the cost of building the system? What's its value?

- In publishing: "Our customers tell us they want charts and graphs as well as full text. How can we build a database that contains indices, full-text, and page images? What's the cost of creating this type of manufacturing operation? What's the value of this type of database today and over the next five years?"

These questions are obviously about making databases. As these examples show, marketing is secondary to the manufacturing process. The questions of value can be answered in part by costing out the hardware, software, and people needed to create the database. "Value" can also be determined by talking to potential customers and users of the service. Not surprisingly, customers talk about "quality" when asked about value. Ask, "Is this a good database for the money?" Hear, "Yes, but I wish the indexing were more consistent and I could find records with fewer spelling errors."

MARKETING'S RESPONSIBILITY

As people get more experience using electronic information, their expectations change. The fastest PC available in 1985 is a museum piece to a power user. The experienced user of information services can quickly form an assessment of a machine-readable file.

If manufacturing does not make the change, marketing has a difficult task convincing the customer about the value or quality of this particular database.

The judgment-whether it is correct or incorrect from the database producer's point of view-has a *direct impact* on the marketing of the product. A database can leave a good impression because the data answered the customer's question. What happens, however, when the information in the database is discovered to be incomplete or incorrect?

Yes, a high-performance system and an easy-to-use interface can influence the user's perception of the machine-readable file as well. There are dozens of other factors external to the database that influence a particular customer's view of an electronic information construct.

Let's assume that a customer reported a problem to the database producer's customer service hotline. The customer goes online and checks to see if the error has been corrected. When the customer uses the file and sees the change, the user perception of the database changes. The customer is likely to say, "The company took action and made work easier."

In summary, information marketing must deal with these tasks:

- Determining product needs (market research)
- Informing the user of the product availability (marketing communications)
- Supporting the user of the product (marketing support or customer service)
- Inputting user feedback to the organization (product development).

However, manufacturing is the central actor in this drama. If manufacturing does not make the change, marketing has a difficult task convincing the customer about the value or quality of this particular database.

WHY CHANGES AREN'T MADE

At this point, you may be asking yourself, "What a ridiculous example. Information manufacturers should fix errors! They do, don't they?" My advice: do some checking. Ask around. Scrutinize databases as a consumer of the data, not a neutral intermediary.

The reasons for ignoring user needs and wants vary by organization, of course. For our purposes, we will assume that the information manufacturing "system" cannot accommodate the change. Typical reasons for not responding include:

- Programming cost is beyond the organization's resources
- Priority of the change is too low to warrant investment
- Technically impossible in the present manufacturing "plant"
- Return on investment does not meet organization's target
- Copyright or other legal issues block the change.

It makes little difference if the unresponsive provider of electronic information has the world's most superb marketing engine. The key is the ability of the information manufacturer to deliver what the customer wants. In practical terms, a machine-readable file can lose user and customer support if enhancements are not made. A competitor can enter the market and take customers.

If an information manufacturer does not respond to customer needs, is it likely the customer will seek an alternative? No one seeks an alternative under these conditions: (1) no acceptable options exist, (2) no one knows the data are flawed, or (3) one believes some information is better than no information.

Marketing does not make a successful machine-readable file. Manufacturing does. To state the obvious: the information in the machine-readable file must be accurate, and the overall experience must conform to the customer's expectations of fair value, ease-of-use, and other subjective factors.

MANUFACTURING DEFINES EXCELLENCE

Peter Drucker offered these observations in his essay *The Emerging Theory of Manufacturing:*

[A] plant must be redesigned from the end backwards and managed as an integrated flow. . . . few companies have enough knowledge about what goes on in their plants to run them as systems. . . . As soon as we define manufacturing as the process that converts things into economic satisfactions, it becomes clear that producing does not stop when the product leaves the factory [5].

The customer's needs become the inputs for changing the manufacturing processes to deliver products the customer will buy. Henry Ford, an authentic American genius, allegedly said in response to a question about the choice of paint for a Model T: "Any color so long as it is black." Customers today are unlikely to have their choices limited.

As we enter the Reconstruction Stage, options will present themselves to users and customers. Gross distinctions like basic functionality or subject (discipline) will blur and electronic constructs will be evaluated on such attributes as:

- Price
- Accuracy (correctness of "spelling," factual precision)
- Presentation (appearance, usability)
- Trade offs (completeness versus cost, ease-of-use versus accuracy).

We are entering a period when marketing will be increasingly dependent upon manufacturing to deliver what the customer wants. Organizations that have the ability to build machine-readable files that meet needs and solve problems will then have the difficult job of differentiating their databases from others.

Despite the tendency to place great emphasis on marketing, managers of information factories must be skilled "information engineers" and expert manufacturers if the organization's products are to gain wider customer and user acceptance.

In the United States, there are hundreds, if not thousands, of niche markets. Individuals in these markets use a variety of techniques and

→

technologies to form virtual communities of interest. News and information about a new machine-readable file can reach individuals in a niche rapidly without much support from a marketing department.

Skirmishes will be fought between informatidn manufacturers who are making files that follow the assembly line model of minor customization and organizations that can create customized information constructs.

SIGNS OF CHANGE

One of the most visible signs of the advent of the reconstruction stage is the existence of two types of information manufacturers.

(1) Manufacturers using information factories constructed on policies, systems, hardware, and software from the 1982-1992 period or earlier. Some in this category are integrating newer technologies into their existing systems. Examples: (a) Dow Jones News/Retrieval's experiments with parallel processing computers. (b) The Whole Earth 'Lectronic Link's virtual communities and their user-created databases.

(2) Manufacturers using advanced technologies, which will become the foundation for the reconstruction stage beginning to gather momentum. Examples: (a) Voyager Corporation's multimedia discs for Macintosh computers. (b) Indices built by agent software operating on Internet nodes.

Talking about what *will* happen is risky, of course. I want to point out that the principal difference between these two movements is mass production with some individualization opposed to information manufacturing specifically tailored to the individual's needs and requirements.

Skirmishes will be fought between information manufacturers who are making files that follow the assembly line model of minor customization and organizations that can create customized information constructs. The likely winners will *not* be the combatants. Integrators able to resolve the differences will capture the customer's loyalty. Likely characteristics of an integrator will be:

- Ability to use data from other information manufacturers and reconfigure it to meet the needs of particular markets. Example: a producer of software tools, that reside in a network operating system.
- Lower costs for such features as indexing, machine translation, and formatting. Example: software handles these functions. When high-cost human labor is required, it is for high-value added tasks.
- Flexibility in assembling the information objects needed to provide the machine-readable file the market or the customer wants.

As I write this, I know of no fully operational information factory or integrator operating along these principles.

THE END AS BEGINNING: QUALITY

An information factory produces individual information a record at a time. (Other units of a database are *data sets* and *time series*, in the jargon of numeric file manufacturers.)

Some of the attributes customers recognize in electronic information are timeliness (real-time updates) and brand identity (Dow Jones). Other "visible" factors are data consistency; that is, elements of the record are in a predictable "place" and "style" as other records from the file. The packages have to be reliable; that is, they have to operate to user expectations or past experience suggests they will. The packages have to be affordable; that is, they have to be priced so the customer can afford to buy them or sees the value of having the data. There is a subjective aspect as well: the information packages have to be usable; that is, the information should not create confusions, disappointment, *or* frustration.

Quality is electronic publishing's golden idol. I know of no information manufacturer who would argue for the manufacture of flawed information.

But What Is ***Quality*?** This is a question with many answers. Toyota Motors defines quality as products that conform to the specification. Items that exceed allowed variances are, therefore, poor quality.

Can electronic information be measured like a Lexus' fender? When does a database have quality? The answer depends upon who asks the question, the knowledge of the person evaluating the database and its records, and the use of the electronic information.

At the May 1992 Workshop on Instruction in Library Use, representatives of North American academic libraries had several workshops from which to choose. One of these workshops allowed the 30 attendees to offer their opinions about the different challenges electronic products pose. A surprising number of North American librarians in these sessions voiced concern about software features, local control of documentation, and functionality of CD-ROM and online public access catalog products [6]. This was a small sample of the attendees and the comments were made in an open discussion.

Can electronic information be measured like a Lexus' fender? When does a database have quality?

As I listened, I thought the group demonstrated consensus on this point: electronic publishers (database producers) seemed unable or unwilling to make product-related changes. The publishers, particularly those with CD-ROMs, found it easy to say, "Changes are coming." Database producers seemed to make partial changes or none. Why are CD-ROM publishers resisting customer pleas for software features?

Consider online:

- Why are traditional timesharing services unwilling to expand electronic mail and user-created database services?

When information manufacturing is recognized as the central issue, change will be possible and reconstruction will gain momentum.

- Why are large text timesharing companies unable to accommodate user requests for data formatted for specific word processors?
- Why are databases unable to provide charts and graphs in graphic and "table" formats?
- Why are information manufacturers blaming online services for limiting their ability to fix data errors?
- Why are online services saying database quality is the responsibility of the manufacturer?

The reasons are rooted in information manufacturing processes and policies. The majority of electronic publishers, therefore, wisely try to steer a middle course. Certain features and functions are added if they can be provided at "reasonable" cost, in a "reasonable" period of time, and without necessitating an overhaul of the complete information manufacturing "plant."

But as users and customer get "smarter" about machine-readable files, they want more. When an information manufacturer promises but cannot deliver, sales can be lost. As a librarian from the American West might say, "Big hat, no cattle." Quality is rooted in manufacturing. When information manufacturing is recognized as the central issue, change will be possible and reconstruction will gain momentum.

Marketing talks, manufacturing delivers, customers experience. The only road to database quality is through the factory, an information factory.

REFERENCES

[1] Jacques Ellul, *The Technological Bluff*, translated by Geoffrey W. Bromiley (Grand Rapids, Michigan: William B. Eerdmans Publishing Co., 1990), p. 73.

[2] Aired on PBS stations, 30 May *1992.*

[3] The *Wall Street Journal,* 26 May 1992, p. B6 in the occasional column "Information Age."

[4] Stephen E. Arnold, "Marketing Electronic Information: Theory, Practice, and Challenges, 1980-1990," in *Annual Review of Information Science and Technology,* Martha E. Williams, editor (Elsevier Science Publishers B.V., 1991), pp. 87-144.

[5] Peter F. Drucker, *Managing for the Future: The 1990s and Beyond* (New York: Truman Talley Books, 1992) p. 314.

[6] A more detailed description of this conference appeared in *Library Monitor,* June 1992.

Communications to the author should be addressed to Stephen E. Arnold, P.O. Box 300, Harrod's Creek, KY 40027; 502/228-3820; Fax 502/228-0548; CompuServe-76060,325.