



**Intelligent**  
ENTERPRISE

Smarter, Faster, More Profitable

[About](#) | [Subscribe](#) | [Contact](#) | [Media Kit](#) | [Submissions](#) | [Back Issues](#) | [Site Map](#)

**Search**

more options

Feature

August 18, 2000 Volume 3 Number 13

# Payback

**It may not be visible, but poor data quality almost always has an impact on your bottom line**

By David Loshin

Last November, a New Jersey man admitted to circumventing computer fraud detection programs at two music-by-mail clubs: He used 1,630 aliases to buy CDs at special introductory rates, which he subsequently sold at flea markets at a 400-percent markup. Some of the culprit's methods included adding fictitious apartment numbers, unneeded direction abbreviations, and extra punctuation marks in his names and addresses. By exploiting these companies' inability to filter out bad data, he was able to take them for a ride for more than \$250,000.

This example illustrates how easy it is to recognize that poor data quality has a financial impact. But determining that impact is not always so simple. Without an economic framework for measuring the detrimental effect of poor quality data, arguing for an investment in a knowledge management or business intelligence strategy can be difficult. Traditionally, this process consists of little more than recounting anecdotes, hazy feelings about the difficulty of an implementation, or tales of customer dissatisfaction. Although these "measurements" superficially verify that "something" is amiss, they provide no objective information about the impact of the problem, nor do they imply a solution. This capability is all the more important when you consider that the evidence of such an impact may be difficult to correlate.

Common issues that could imply poor data quality include frequent system failures and service interruptions, a drop in productivity vs. volume, high employee turnover, high ratio of new business to continued business, increased customer service requirements, or customer attrition. The presence of any of these "tip-offs" can indicate opportunities to improve customer relations, optimize the production stream, and enhance employee satisfaction by analyzing and improving poor data quality. With a concrete means of doing so, you can indeed determine the extent to which bad information affects your bottom line and then take steps to address the problem. But first, it's important to know the main building

## Interact

Communities

### [IntelligentCRM](#)

Customer Relationship Management for Competitive Advantage  
Sponsored by [PeopleSoft](#)

### [IntelligentEAI](#)

E-Business Integration for Competitive Advantage  
Sponsored by [Level 8](#)

### [IntelligentERP](#)

Leveraging SAP™ for E-Business Solutions

### [IntelligentKM](#)

Knowledge Management for Competitive Advantage  
Sponsored by [Semio Corp.](#)

## Focus

Information Centers

### [Analytic Applications](#)

sponsored by WhiteLight Systems Inc.

### [Business Intelligence](#)

sponsored by Cognos Corp.

### [Data Integration](#)

### [Database](#)

[Featuring C.J. Date](#)

Sponsored by IBM Software

### [Data Warehousing](#)

sponsored by Acta Technology

### [E-Commerce](#)

### [Enterprise Development](#)

### [RealWare Awards](#)

### [Scalability](#)

## [Table of Contents](#)

blocks for constructing an economic model of data quality.

## Information Chains and Data Flows

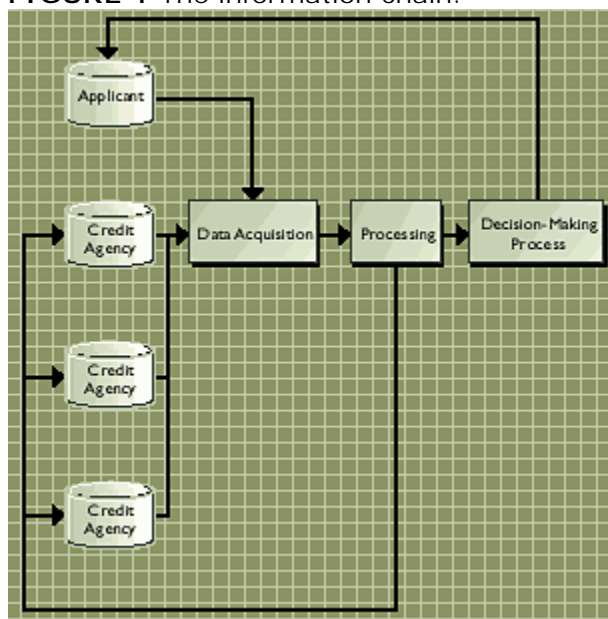
Most organizations use their data in operational as well as decision-making processing. Yet these processes are often treated as “black boxes” that take input data as “raw material” and then generate value-added information. This information in turn feeds into additional operational or decision-support processes.

The first step in understanding the effect of low data quality is to peer inside these black boxes and identify the steps through which information flows. Then, you can divide your system into a set of processing stages connected by directed information channels. (See sidebar, “Ebbs and Flows,”) Having broken the information-processing chain into a directed data flow, you now have a basis for determining the *cost-effect of low data quality*, or COLDQ. (See sidebar, “Six Steps to COLDQ,”)

An *information channel* is essentially a pipeline indicating the flow of information from one processing stage to another; a *directed information channel* additionally indicates the direction in which the data flows. (For instance, supplier data is delivered to an acquisition stage through an information channel.) I recommend that you create information channels of the directed variety because bidirectional communication may occur between any two points, and you want to be able to differentiate data flowing in one direction vs. another.

For example, any organization in the business of extending credit will have a process that, based on a set of inputs associated with an applicant, will determine whether the credit request is approved. Basically, this process consists of an information “chain” that provides a single yes-or-no answer based on the input values and some internal determination application. (See Figure 1)

**FIGURE 1** The information chain.



In this strategic process, data derives from the applicant as well as external agencies that collect data associated with credit worthiness. A data acquisition stage merges the data from the two sets of sources and forwards those records to a data processing stage, where the values of the records determine an answer. If the answer is "yes," it is packaged and forwarded to another department, where the applicant receives the details of the approval (for example, a credit card imprinted with the applicant's name).

An additional data packaging stage prepares a report detailing the credit application that is forwarded to an external data consumer, which may be the same data agent that originally supplied the process with data. Consider, however, if the applicant were to share a name with someone with a bad credit history. If the wrong data records for the applicant are forwarded by the external data suppliers, the resulting decision at the data processing stage could be an incorrect "no." If so, this poorly informed decision would subsequently affect the owner of the original application in the form of a denied mortgage or strategic business loan.

## Feeling the Impact

The assumption is that if no data quality problems are present, your data flows will operate smoothly. You can reasonably rely on decisions based on valid data, and an operation system will run smoothly as long as no invalid data items gum up the works.

Issues arise, however, when the information chains involve low quality data. The effects of low data quality propagate across systems, ultimately leading to poor decision making, tactical difficulties, increased costs, reduced revenues, and lowered customer satisfaction, among other things — ultimately rolling up to the company's bottom line. Conversely, improvements in data quality can reduce costs, increase revenues, streamline and enrich decision making, and improve customer satisfaction.

We can divide these impacts into *soft* impacts, which clearly have an effect on productivity but are hard to measure, and *hard* impacts, which have effects you can estimate and measure. (See sidebar, "Feeling the Impact," page 52.) But how do you measure a cost or a benefit? When calculating the total economic benefit of improved data quality vs. the economic detriment of poor data quality, you can quantify each impact directly against your company's bottom line. For simplicity's sake, I'll classify these measures into various categories:

- Cost increase — The degree to which poor data quality increases the cost of doing business
- Revenue decrease — Impact of poor data quality on current revenues
- Cost decrease — Measures how improved data quality can reduce costs
- Revenue increase — Indicates how improved data quality can increase revenues
- Delay — Measures any slowdown in productivity

- Speedup —The degree to which you can reduce a given process' cycle time
- Increase satisfaction — Indicates how much customer, employee, or shareholder satisfaction increases
- Decrease satisfaction — Indicates how much customer, employee, or shareholder satisfaction decreases.

You can accurately measure the economic impact of poor data quality in each of these categories. Although some impacts are difficult to tie to precise dollar amounts, you can at least estimate their monetary value at orders of magnitude.

**Table 1** The risks of poor data quality.

Costs	Associated with...
Detection	A data quality problem that provokes a system error or processing failure, invoking a separate process to track down the problem
Correction	Correcting a problem as well as the restarting of any failed processes or activities, including the time associated with the activity that failed and any extraneous employee activity
Rollback	"Undoing" work that has already been done
Rework	All work performed before the successful run took place
Prevention	The design, implementation, and integration of new processes for identifying data quality problems and preventing operational failure because of unexpected data problems
Warranty	Fixing data quality problems as well as compensating the customer for damages
Reduction	A customer deciding to do less business with your organization in reaction to your data quality problem
Attrition	A customer completely ceasing business with your organization in reaction to poor data quality
Blockading	Complete customer dissatisfaction that causes other potential customers to decide against doing business with your organization in the first place

Data quality problems may arise at the acquisition stage, during any number of internal data creation stages, or at the time of data packaging. We can classify these problems into cost categories: *detection*, *correction*, *rollback*, *rework*, *prevention*, and *warranty*. Furthermore, as errors propagate across the enterprise and to the customer base, the risk of customer dissatisfaction can increase, possibly leading to lost revenues. The risk areas here include *reduction*, *attrition*, and *blockading*. (See Table 1, page 54.) For the most part, all these areas represent only those associated with operational activities, and do not even touch on impacts involving tactical or strategic processes, although reduction, attrition, and blockading are certainly oriented toward strategic activities.

## Putting it All Together

Now that we have all the pieces in our economic model, let's look at the

actual steps involved in building one. The result will be a *data quality scorecard* that summarizes the overall cost associated with low data quality and helps you identify the best opportunities for improvement.

- **Map the information chain.** The first step in the analysis is to map your information chain and data flows (as I described them earlier) in order to understand how information flows across your organization. Using this chain, you can locate the sources of any potential problems.
- **Identify data flow.** Given an information chain, the next step is to determine what data your system is using, and through which information channel source and target points the data passes. If possible, you should detail the record or message structure so that you can directly associate any error conditions with the specific data set in which the error occurs.
- **Interview employees.** You should interview all involved employees to assess the internal impact of flawed data. Next, aggregate and sum the time all employees devote to data quality issues for each stage in the information chain.
- **Interview customers.** To determine the impact of decreased customer revenue, talk to current and former customers to understand the reasons for any decrease in business or attrition and blockading.


**Table 2** Matrix for classifying data quality problems.

Problem and Location	DO Activities	Economic Factor	Estimated Economic Impact	Units Affected	Total
Data Acquisition	Detection	100	\$0.02	20,000,000.00	\$4,000
	Correction	100	\$0.04	20,000,000.00	\$7,000
Malformed Addresses	Detection	1	\$3.00	20,000.00	\$60,000
	Correction	1	\$0.20	20,000.00	\$4,000
Data Processing Stage 2 — Unknown ID	Prevention	1			\$50,000

- **Isolate flawed data.** Next, annotate the information chain with the results of your interviews. Note any source of a data flaw at each point where a data set is sent, received, or manipulated, along with a list of the activities to which you can attribute those flaws.
- **Identify the impact domain.** Now that you have an information chain annotated with the list of data flaws and the activities associated with each of those flaws, it's time to start attributing the flaws and activities to impact domains by building a matrix that classifies each data quality problem. (See Table 2.) The first axis identifies the problem and its location in the information chain, the second axis represents the activities

## Rate This Article

Let us know what you think.

Rating:  

Comments:

associated with each problem, and the third axis denotes the impact areas for each activity. In each cell in this matrix, you should insert the estimated cost associated with that impact, using the economic measures described earlier. If no estimate can be made, at least indicate the impact's order of magnitude.

- **Aggregate the total.** Now you can superimpose the matrix onto a spreadsheet and build an aggregation model. You can tally and summarize the costs in different ways and use them as input to the improvement stage.

- **Identify opportunities for improvement.** Finally, use the model to look for the biggest "areas of pain." Having categorized the location and impacts of your data quality problems, the next logical step is to find the best opportunities for improvement — where you can get the biggest value with the smallest investment. You can do so by adjusting your spreadsheet model to include the costs of the improvement projects and offsetting them by the value the improvement project yields. The result is an environment for calculating ROI for improvement project implementation. For any suggested improvement, add the cost of designing and implementing the improvement to the model, along with a time frame for implementation. Each improvement must correspond to the elimination of at least one cost impact. Consequently, you can calculate ROI based on the decrease in costs (or alternatively, increase in revenues) vs. the cost associated with the improvement project.

Consider a simple direct mail campaign. Any mailing list is bound to contain errors such as entity duplication, incorrect addresses, or invalid addresses. Different kinds of costs are associated with each of these types of errors, depending on the application. Mailings sent to invalid addresses have a cost associated with the postage and handling of the material, as well as the correction costs associated with updating the mailing list database, although the detection costs are relatively low. Incorrect addresses are detectable as well (they will likely be returned), but a high number of undeliverable packages will severely reduce the campaign's response rate.

Duplicate entries are not detectable as such unless the recipient reports the mistake — and the cost incurred is also associated with the postage. Sometimes the cost is greater, however. I recently received two mailings from the same organization, one addressed to David Loshin ("Dear David ..."), the other addressed to Loshin David ("Dear Loshin ..."). I would have dismissed this as a minor error and ignored it except for one simple fact: the letters were sent by the local Direct Marketing Association. You would expect that I could trust an organization promoting the direct marketing industry to at least perform straightforward duplicate elimination correctly.

## Know Your Costs



If you would like us to respond, please enter your e-mail address below:

In this article I have proposed a framework for characterizing and understanding the costs associated with low data quality. By calculating a real cost associated with bad information, instead of propagating anecdotes, the enterprise that is truly interested in improving its information resource can prioritize and fund data quality improvement projects based on a solid return on investment model.

EBBS AND FLOWS	FEELING THE IMPACT	SIX STEPS TO COLDQ
<p><b>Understanding the directed data flow process is crucial for identifying problems</b></p>	<p><b>The practical effects of poor data quality can ripple across your organization</b></p>	<ul style="list-style-type: none"> <li>• Understand how information flows across your organization</li> <li>• Interview employees and customers to understand data quality issues</li> <li>• Locate the areas where data quality problems arise</li> <li>• Identify the impact associated with each instance of poor data quality</li> <li>• Characterize and aggregate the economic impact</li> <li>• Identify opportunities for improvement.</li> </ul>
<ul style="list-style-type: none"> <li>• Data supply — data suppliers forward information into the system</li> </ul>	Soft Impacts	
<ul style="list-style-type: none"> <li>• Data acquisition — data is accepted from external suppliers and injects it into the system</li> </ul>	Difficulty in decision making	
<ul style="list-style-type: none"> <li>• Data creation — internal to the system; data may be generated and forwarded to another processing stage</li> </ul>	Time delays in operation	
<ul style="list-style-type: none"> <li>• Data processing — any stage that accepts input and generates output (as well as generating side effects)</li> </ul>	Organizational mistrust	
<ul style="list-style-type: none"> <li>• Data packaging — any point where information is collated, aggregated, and summarized for reporting</li> </ul>	Lowered ability to effectively compete	
<ul style="list-style-type: none"> <li>• Decision making — the point at which human interaction is required</li> </ul>	Data ownership conflicts	
	Lowered employee satisfaction	
	Hard Impacts	
	Customer attrition	
	Costs attributed to error detection	
	Costs attributed to error rework	
	Costs attributed to prevention of errors	
	Costs associated with customer service	

- Decision implementation — where the decision made at a decision-making stage is executed, which may affect other processing stages or a data delivery
  - Costs associated with fixing customer problems
  - Costs associated with enterprisewide data inconsistency
  - Costs attributable to delays in processing
- Data delivery — where packaged information is delivered to a known data consumer
- Data consumption — the exit stage of the system.

**David Loshin** ([loshin@knowledge-integrity.com](mailto:loshin@knowledge-integrity.com)) is president and CTO of Knowledge Integrity Inc., a consulting firm specializing in knowledge management and data mining. He is the author of Enterprise Knowledge Management: The Data Quality Approach (Morgan Kaufmann, 2000).

---

**Links:** [Home](#) | [Subscribe](#) | [About](#) | [Media Kit](#) | [Contacts](#) | [Edit Cal](#) | [Submissions](#) | [Archives](#) | [White Papers](#)

**Communities:** [IntelligentERP](#) | [IntelligentEAI](#) | [IntelligentCRM](#) | [IntelligentKM](#)

---

