Bad Data Costs More than you Think, Fortunately Solutions are Achievable

Scott Powell

RECORDMATCH, INC. | https://recordmatch.io/

Do you know what duplicate, incorrect, missing, inconsistent, and otherwise 'bad data' is costing you? Most organizations don't. The costs are high. Fortunately, improvements are achievable.

This whitepaper explores *bad data*, including:

- Costs and occurrence frequencies
- Causes & remediation strategies
- Data management best practices to consider when crafting a data management strategy
- Strategies to achieve ROI from data management projects

SUMMARY:

Poor quality or *bad data* is a source of financial and brand losses for most organizations today. On average, organizations spend 3-5 times more than they should due to bad data. Bad data results from many factors, including outdated information, conflicting information, missing information, and data entry errors. Technology can significantly improve data quality. When creating a data improvement strategy, consider interdependent best practices for disaster recovery, accessibility, and information security. To achieve the best results, follow proven program management practices to assess problems, define priorities, develop plans, implement, and evaluate & improve progress. Evaluate resource capabilities and experience when planning. Engage consultants to provide guidance, bandwidth, and expertise as needed.

THE COST OF BAD DATA

The 1992 Labovitz & Chang 1-10-100 rule estimates the relative costs of preventing bad data, the cost of correcting errors from bad data, and the cost of ignoring these data errors. In summary, the rule states for every $1 spent to prevent a data error, there is a relative $10 paid to fix the error (had it not been prevented), and a $100 for ignoring the mistake altogether. The cost ratio for these errors scale up and down relative to the complexity of the in-question data and error. For example, let's say it costs $5 in an associate's time to prevent a data error when processing a transaction. Had they not stopped the mistake, it would cost $50 to correct the operation when identified. Failure to identify the error (ignoring it) costs $500. Remember, the rule reflects averages and relative costs.

The 1-10-100 rule would overestimate costs if it were merely stating the price to correct the data. The rule reflects the cost of fixing the entire transaction. Imagine a data entry error at a government services agency. An associate enters an incorrect tax ID number or transaction amount. The cost to correct that error includes multiple actions: identifying the mistake, confirming it, reprocessing the transaction, providing amended services, and re-claiming or canceling services and benefits delivered in error. The 1-10 cost ratio is realistic.

Data error costs include: identification, data correction, transaction make-good, and brand damage costs.

Qualitative costs are excluded from the 1-10-100 rule. Poor experiences for stakeholders (citizens, customers, employees, and management) damages the brand and impacts future stakeholder actions, including renewals, employee turnover, funding. Data error costs include identification, data correction, transaction make-good, and brand damage costs.

The quantitative impact of individual data errors requires an understanding of the frequency of occurrence. Modern software has significantly improved user experiences and quality operations. Leading many to believe

data errors are low. Recent studies prove that data quality remains a challenge, and organizations experience higher operating costs.

- A 2019 IDC[1] study confirmed the high cost of data preparation and management vs. time spent actually using data; "data workers spend 80–90% of their time managing and preparing data and only 10–20% of their time performing analytics". 'Managing and preparing' includes the cost of correcting, deduplicating, and augmenting data after the initial data capture.
- A 2017 HBR[2] study found that only 3% of companies meet data quality standards and that 'on average, 47% of newly-created data records have at least one critical error, and a full quarter of the scores are below 30%, and half are below 57%'.
- A 2008 NIH study[3] of Clinical Trial Research Databases found error rates ranged from 2.3% to 26.9%.

Using the 1-10-100 rule and frequency statistics (26.9% and 47%), the cost of data error prevention is 3-5x lower than the price of correcting errors from bad data. The following table calculates the relative 'Repair vs. Prepare' for mitigating data errors during creation vs. repairing errors when identified. The error rates referenced in the study are in bold italics.

> … the cost of data error prevention is 3-5x lower than the cost of correcting errors from bad data.

Frequency of Data Errors	Work Units		Work Unit Costs		Cost of Data Management	Repair Cost vs. Prepare Cost
	Good Data	Bad Data	Prepare Costs	Repair Costs		
Data Error Prevention Model						
0%	100	0	$100	$0	$100	0%
Error Correction Model						
10%	90	10	$0	$100	$100	100%
20%	80	20	$0	$200	$200	200%
27%	***73***	***27***	***$0***	***$270***	***$270***	***270%***
47%	***53***	***47***	***$0***	***$470***	***$470***	***470%***
50%	50	50	$0	$500	$500	500%
75%	25	75	$0	$750	$750	750%
100%	0	100	$0	$1,000	$1,000	1000%

Organizations can achieve operational ROI from data error prevention. Data quality ROI is easily invisible when data management is in the IT department budget, and the operations budget absorbs the operating costs of correcting transaction errors from bad data.

Unfortunately, many organizations today just keep paying higher costs because they are invisible in operating budgets. Ignoring unknown waste is expensive.

---

[1] Enabling Organizations with Trusted Data: Data Governance Driven by Data Quality

August 2019, Written by: Stewart Bond, Research Director, Data Integration and Integrity Software

[2] "Only 3% of Companies' Data Meets Basic Quality Standards", Harvard Business Review by Tadhg Nagle , Thomas C. Redman and David Sammon; September 11, 2017

[3] Analysis of Data Errors in Clinical Research Databases; National Institute of Health, by Saveli I. Goldberg, PhD,a Andrzej Niemierko, PhD,a,d and Alexander Turchin, MD, MSb

CAUSES & REMEDIATION

There are many causes of incorrect data. If it were as simple as data entry, the problem would be much easier to solve. In addition to data entry errors, there are duplicate records for the same individual, conflicting records, siloed databases with different data for the same individual, stale and outdated data that is no longer accurate, and incomplete records.

Type	Description	Remediation Methods
Data Entry	Data entry errors occur when users enter incorrect or inaccurate values. For example, a name may be recorded as "Steve" in place of "Steven" or a city is registered as "Pittsburgh" when the individual lives in "Carnegie." Data entry errors include incorrect data from mis-spellings, accidental transcription, and intentional or unintentional false values.	Change workflows to improve data capture: ● Verify data against existing records during entry ● Prompt users to verify net-new records and differences to existing records ● Automatically validate addresses to postal standards
Duplicate Records	Multiple record errors are when two or more records exist for the same thing in a database. Duplicates frequently occur when various roles identify and enter data, such as CRM systems, security master files, SKU master files, and similar situations. Duplicate records often happen when operations are consolidated across departments or during company mergers.	Use data review and validation tools to identify duplicates: ● Rules engines with logical conditions ● Fuzzy matching ● Prompt users to manage edge cases Sustainability Note: Technology solutions need support adjustments over time.
Conflicting Records	Conflicting records occur within duplicate files when the information is contradictory. That is, the records are materially different. For example, the addresses have different streets, house numbers, or city names.	Conflicting records require higher investigation than typical duplicates as the correct information needs to be determined. This process need not be 100% manual: ● Maintain date-time metadata to support time assisted validation ● Keep confidence hierarchies to compare sources ● Compare records to 'gold' records Use these validations to prompt users during data entry and assist the resolution of records within databases.
Incomplete Records	Incomplete records have missing material information. Incompletes can occur when users skip fields, and when multiple data sources are combined, and required data fields are bypassed.	Incomplete records are mitigated differently during data entry and when identified within the database. ● Require data capture user interfaces to capture all required fields ● Use rules engines to identify missing data within databases ● Augment missing data from alternative sources ● Generate workflows to engage customers/citizens to update data records
Siloed Database	Siloed database errors occur when there are two or more separate databases with overlapping information without any cross-database validation, preventing the identification and remediation of data errors.	Fix siloed database errors with a "gold" data store (comprehensive database) that combines records from multiple sources. The gold datastore does not replace the source databases. Instead, it serves to maintain data validation rules, gold records, push remediated data back to functional databases, and assist users with new data entry.

Outdated Information	Outdated information was correct originally, then degraded to incorrect over time. Changes in postal standards and missed updates cause outdated information errors. For example, a zip code extension is added or changed, additional children are born, or other life events.	Remediate outdated information errors with periodic evaluation and comparison to standards and new data sources: ● Prompt users to engage individuals during engagements when the records meet date thresholds ● Compare multiple sources of information to identify discrepancies ● Compare postal information to current standards

ADDITIONAL BEST PRACTICES

When evaluating the data quality and solution strategies, it is the right time to review Disaster Recovery, Accessibility, and Security considerations.

DISASTER RECOVERY

"**Disaster Recovery** involves a set of policies, tools, and procedures to enable the recovery or continuation of vital technology infrastructure and systems following a natural or human-induced disaster. Disaster recovery focuses on the IT or technology systems supporting critical business functions,[1] as opposed to business continuity, which involves keeping all essential aspects of a business functioning despite significant disruptive events. Disaster recovery can be considered a subset of business continuity.[2][3]"[4]

Treating data quality and Disaster Recovery (DR) as mutually exclusive will negatively impact both. Independent initiatives introduce data quality issues if the DR process brakes data remediation processes within production systems. As an example, a DR process replicates an application's primary database to an alternative location. Excluding the replicated database from data, quality introduces opportunities for errors to creep in from divergent data. Costs are compound during actual DR events.

ACCESSIBILITY

Data strategies must consider data access management. Data accessibility is essential to realizing the value of technology investments, and it is an increasing priority of governance policies and auditors. Many databases contain sensitive information, especially in government, medical, and financial services institutions. Businesses have customer and financial information that could damage the company if made public.

Data quality strategies must account for access control by role, organization, and use case. Consider data sources when augmenting content to ensure unintended cross-referencing does not occur. For example, court records contain criminal histories that are helpful to police agencies but must be kept separate from housing authorities.

Achieve accessibility through the maintenance of meta-data within consolidated databases. Meta-data enables rules engines to comply with data privacy parameters when cross-referencing and combining information. For

[4] https://en.wikipedia.org/wiki/Disaster_recovery

example, name and address records are used across a hospital system while keeping medical histories need-to-know confidential. Accessibility strategies must account for human and machine access control at the individual element level.

INFORMATION SECURITY

In addition to accessibility, data strategies must address the fundamental security of information. Historical assumptions that on-premise data is fundamentally more secure than cloud data should be questions. Security is an always moving target, and depending on your business, you may be required to meet regulatory standards.

Is it realistic to believe that on-premise network ecosystems have higher security investments and expertise than the major cloud providers? Cloud providers must consider security as a primary part of their value proposition. While many on-premise networks are cost centers, evaluate who can cost-effectively manage robust security at scale?

Many industries have required regulatory standards. Markets may require them of vendors to compete.

- Do you need to meet HIPAA, SJIS, or other standards?
- Does GDPR or CCPA impact your data management practices?
- Will you be pursuing ISO 27001 certification as your customers require it?

The best practice is to define all standards required, those that may be needed, and future standards. Update your security policy equipped with the right information. If you need help, engage consultants to help define security policy, best practices, and implementation strategies. Combine requirements and standards with best practices for data encryption and access control.

GETTING STARTED

Today's technology landscape is challenging. Today's businesses have state of the art SaaS and 'green screen' mainframes and everything in-between in their data operations. Improving operations can appear insurmountable as too often 'the cure is worse than the disease.' In this environment, many organizations struggle with data quality, accessibility, and efficiency challenges.

Improving digital operations and stakeholder benefits doesn't have to be a long, painful remediation process. Follow simple steps to achieve measurable gains:

- Evaluate your needs: Assess the current environment, identify pain points for users, and the frequent errors, quantify costs from mistakes.
- Define and prioritize goals: Define achievable goals and set priorities. Everything cannot be #1 priority.
- Create a plan: Develop a remediation plan. Focus on achievable actions and continuity of operations.
- Implement & evaluate: Act to implement the plan, maintain governance process to maintain progress, manage risk, and improve the process.

Most importantly, use qualified resources and enable them to succeed. If needed, retain consultants to assist. Consultants can support anywhere from 100% outsourcing to management and oversight best practices. Doing things as right as possible (or reasonable) the first time is the best investment.