Preparing Data for the Data Warehouse

Stephen M. Brown
Vice President, Client Services
Vality Technology
sbrown@vality.com

The ultimate success of a Data Warehouse is measured by the trust or degree of confidence which users have in the answers returned from queries, data mining and OLAP. The care and attention given to data transformation during initial loading and on-going maintenance directly determines the completeness and accuracy of the resulting answers. This presentation focuses on data preparation issues – the legacy data contaminants you will encounter, what can go wrong and how to ensure your data warehouse is stockpiled with good data.

If you are rewriting an operational application or feeding legacy data to a new information system, beware of the data contaminants described below. Without applying Data Re-engineering to defend against contaminants, you will subject users to erroneous and missing information about your organization's most important customers and business entities.

To build your new information system, you must migrate rarely audited legacy data to a heavily queried environment dependent on high accuracy. But legacy data, in its operational state, is not ready for the relational world. Organizations that blindly migrate legacy data risk problems arising from poor data quality, which cannot easily be corrected "after the fact." Since data contaminants, by nature, tend to cluster around your largest customers, even a small percentage of legacy data contamination will compound to invalidate information concerning your company's most prominent clients. From our experience, 80 percent of user queries will access that small area of your database suffering most from data ailments.

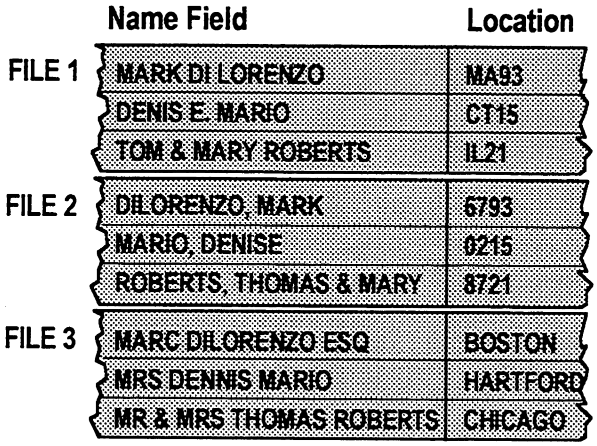
Your challenge: To gain an understanding of your data problems through an automated investigation of actual data values below the metadata surface. With this "data map," you can then determine and employ the best automated means for reconditioning and integrating your information.

The five big legacy data problems...

1. Lack of legacy standards: Multiple formats within disparate data files.

Legacy systems are disparate islands by themselves that often store data in very different grammatical structures and represent business entities in different ways, as the example at right shows. It is dangerous to assume that all location values actually represent locations. Even if they do contain location information, values in different files may possess different levels of granularity that preclude a one-to-one mapping to a desired standard.

For example, File 3 contains city values, while File 2 contains county codes, and File 1 contains a state abbreviation with census-tracking codes.

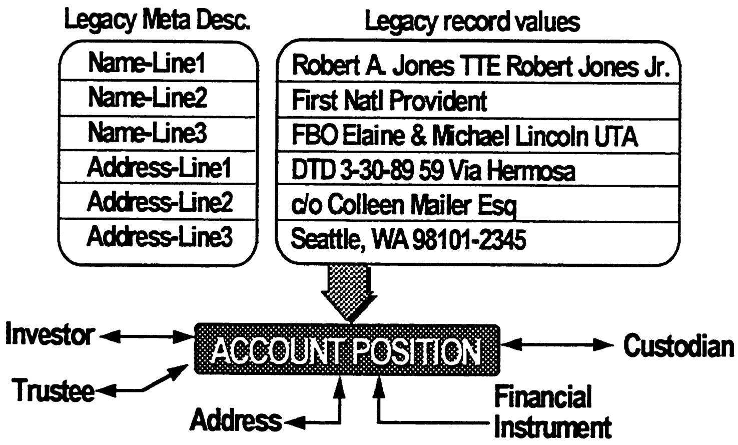


	Name Field	Location
FILE 1	MARK DI LORENZO	MA93
	DENIS E MARIO	CT15
	TOM & MARY ROBERTS	IL21
FILE 2	DI LORENZO, MARK	6793
	MARIO, DENISE	0215
	ROBERTS, THOMAS & MARY	8721
FILE 3	MARC DI LORENZO ESQ	BOSTON
	MRS DENNIS MARIO	HARTFORD
	MR & MRS THOMAS ROBERTS	CHICAGO

Under the name field, Is DENISE MARIO the same individual as DENIS E MARIO or did the "e" in Denise's name simply move over through a keying error to change her gender? (And what if Denise Mario opted on occasion to use her maiden name -- Denise Mathews. That would certainly complicate your effort to identify customers. But that's a matching and consolidation problem better exemplified by "Legacy Contaminant #4: The Anomalies Nightmare.")

2. **Legacy information buried and floating within free-form text fields.**

As implied by the diagram, your legacy data contains robust and complex relationships that remain hidden and "floating" within free-form text. You must be able to unlock the true identity of critical business entities buried in legacy data in order to preserve roles and relationships needed for your relational world.



Legacy Meta Desc.	Legacy record values
Name-Line1	Robert A. Jones TTE Robert Jones Jr.
Name-Line2	First Natl Provident
Name-Line3	FBO Elaine & Michael Lincoln UTA
Address-Line1	DTD 3-30-89 59 Via Hermosa
Address-Line2	c/o Colleen Mailer Esq
Address-Line3	Seattle, WA 98101-2345

This example depicts a registration label for a trust account. To determine who the customer is, you need to extract the meaning of information buried in the fields. "Hidden treasures" include: FBO (for the benefit of), TTE (Trustee), UTA (Under trustee account), c/o (Care of), and DTD (dated).

"Floating domains" complicate your ability to locate and extract relationships from free-form text. For example, a data value that appears on Address Line 1 in one record might appear on Line 2 in the next record, and might get split between Line 2 and 3 in yet another record. To build accurate consolidated views of key business entities and their relationships across accounts, you must effectively break down and establish links to related values *before* the information can be mapped to your relational tables.

In the example below, "care of" values appear in various locations within name and address fields, complicating your effort to identify and parse information into separate entities.

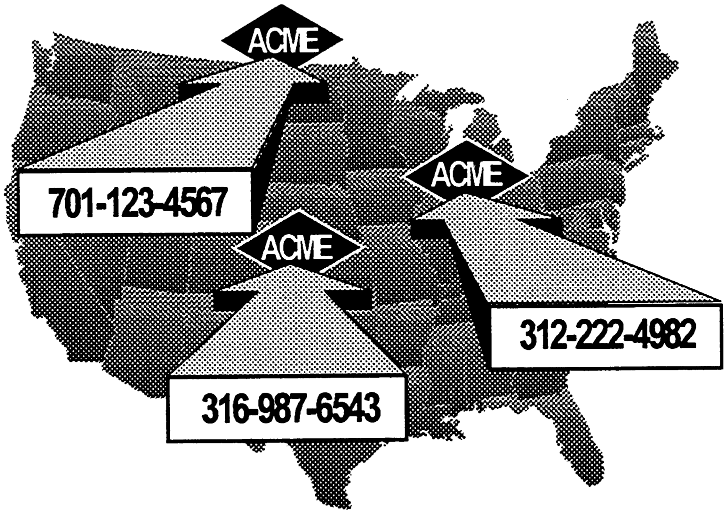
NAME FIELD 1	NAME FIELD 2	STREET
POWELL, KIMBERLY		PO BOX 2345 C/O RAY WHITE
EVANGELICAL METHOD	1ST CHURCH	C/O CHELSEY BARTLETT
ENSMINGER, J DAVID	D/B/A J D & E INC C/O MRS. ENS	2999 S UNIVERSAL BLVD

Conventional data scrubbers that employ look-up tables and name-and-address dictionaries cannot possibly contain listings for every conceivable variant in anomaly, field location, and line break. Because of the uniqueness of your data and business rules, rigid look-up-table approaches cannot meet your needs.

3. Legacy myopia: Multiple account numbers block a consolidated view.

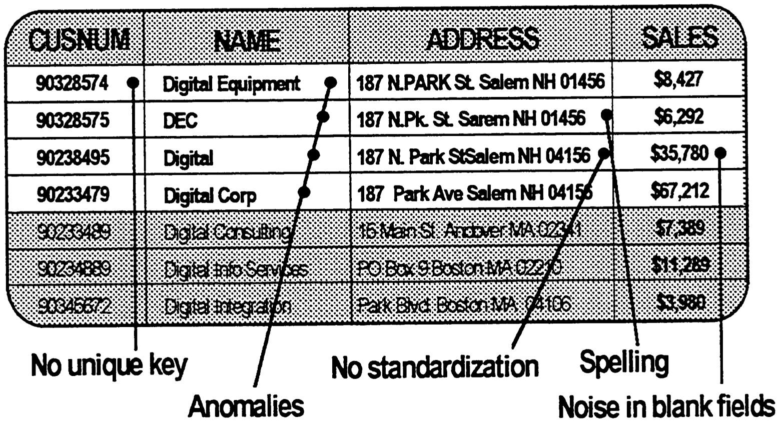
Acme Co. has telephones in three locations. But legacy data refers to three accounts under different numbers. How will your business identify the customers when account numbers block your view?

Similarly, you may need to relate a business entity (i.e., a customer) across different lines of business. Or you may need a consolidated view of subsidiaries under a particular parent firm, say United Technologies. To identify its subsidiaries -- Otis Elevator, Pratt & Whitney, Hamilton Standard, and Carrier -- you will have to employ a custom-tunable matching engine that can evaluate any and all attributes within records in order to locate matching and related entities.



4. The anomalies nightmare: Complex matching and consolidation.

Business entities can be represented in a wide variety of ways. One of Vality Technology's clients uncovered more than *one hundred* anomalies for their customer, Digital Equipment Corp. Locating anomalies from among millions of records and multiple files is difficult.
But your responses to important queries will be wrong without entity integrity; you must either relate all "instances" of Digital or condense all instances to a single representation defined by one and only one set of attributes. Without entity integrity, a query seeking sales from each customer that exceeds $75,000, say, would omit Digital!

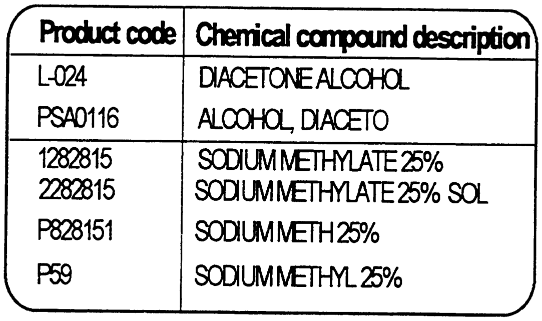


CUSNUM	NAME	ADDRESS	SALES
90328574	Digital Equipment	187 N.PARK St. Salem NH 01456	$8,427
90328575	DEC	187 N.Pk. St. Sarem NH 01456	$6,292
90238495	Digital	187 N. Park StSalem NH 04156	$35,780
90233479	Digital Corp	187 Park Ave Salem NH 04156	$67,212
90233489	Digital Consulting	15 Main St. Andover MA 02341	$7,389
90234889	Digital Info Services	PO Box 9 Boston MA 02210	$11,289
90345672	Digital Integration	Park Blvd. Boston MA 04106	$3,980

In the example below, inconsistent use of middle and first names makes consolidation difficult. Anomalies in Name and Address values abet the problem.

	NAME FIELD	STREET	CITY/STATE
File 1	GRIFFITH, CARRIE EILEEN	3834 SOUTH V ST	FT SMITH, AR
File 2	GRIFFITH, CARRIE	3835 (sic) SO V STREET	FORT SMITH, AR
File 3	GRIFITH (sic), EILEEN	3834 S V STREET	FORTSMITH, AA (sic)

In addition, anomalies are not limited to names and addresses. You must find an automated means to investigate, standardize, transform, and integrate *all* your information: from loan numbers to part descriptions and -- in this case of a major manufacturer -- chemical compounds:



Product code	Chemical compound description
L-024	DIACETONE ALCOHOL
PSA0116	ALCOHOL, DIACETO
1282815	SODIUM METHYLATE 25%
2282815	SODIUM METHYLATE 25% SOL
P828151	SODIUM METH 25%
P59	SODIUM METHYL 25%

5. Legacy data surprises in individual fields: Data values that stray from their field descriptions and business rules.

What you see at the metadata surface does not accurately represent what exists beneath the surface.. An automated data investigation will reveal sundry exceptions and noise that fail to jive with meta field descriptions and business rules. Here are some examples:

- Commercial names improperly mixed with personal names.
- Foreign names mixed with domestic names.
- Relationships such as “doing business as” and “c/o” that spill over into address fields.
- Name fields that contain unexpected relationships and location information.
- Directions (e.g., “Bear left at fork”) in address fields.
- Extraneous noise (e.g., an 8-character string within a 10-digit SSN field).
- Addresses with missing area codes, lot numbers, etc.
- Information that has been truncated.
- Part descriptions that call for part numbers.
- Inconsistent use of white space, special characters, and field boundaries. For example, sometimes a second name field contains an address. Sometimes a word breaks at the end of a line and continues onto the next.
- “Care of” values that “float” through name and street address fields, making it impossible to predict their location from record to record.
- Names within address fields. Missing values.
- Abbreviations.

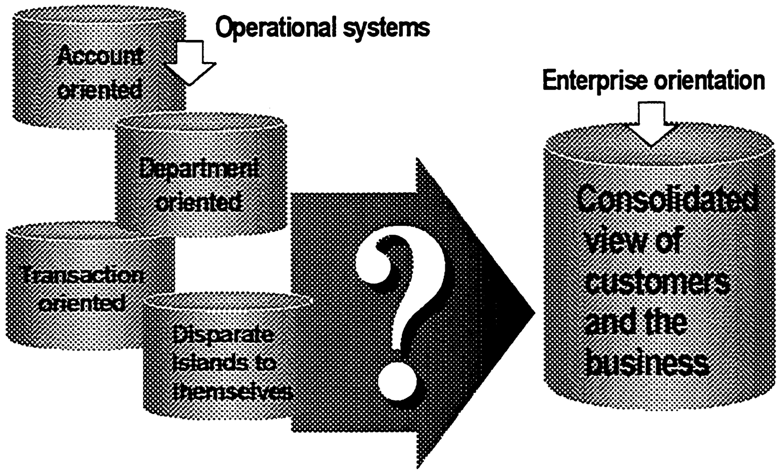
For organizations moving data from multiple operational systems to a new data warehouse or client/server environment, adhering to any of the following “excuses” could lead to a data disaster. Outlined here are actual reasons that we’ve heard from Fortune 500 companies, for avoiding Data Re-engineering. They opted to avoid investigating, reconditioning, and consolidating data from multiple sources before migrating the information to new relational databases.

1. "The new data will be as good as the old, and the old data seems to work just fine."

Your old data works fine in your operational systems, but will likely break down in your new information systems. Aging legacy systems were designed as transaction processors, not as information engines that could satisfy queries about the business and enable cross-department and cross-product marketing. Disparate data within legacy systems have account-number and department orientations. These characteristics, which may be benign in an operational setting, can become cancerous if migrated blindly to an information system.

Some typical data contaminants:

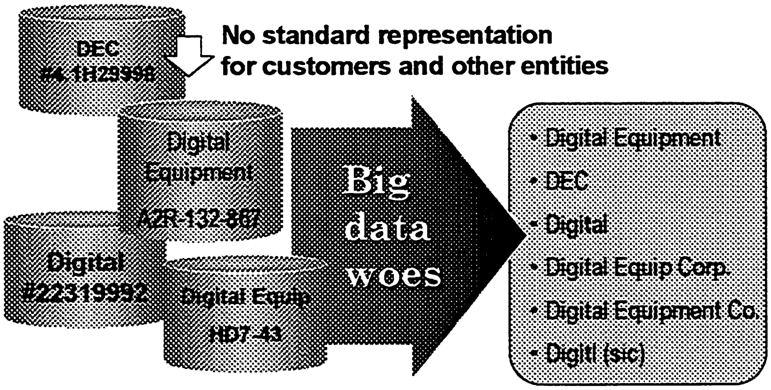
- Important entities, attributes, and relationships hidden and floating in text fields. ("Bob and Mary Fine, DBA Fine Foods").
- Data values that stray from their field descriptions and business rules.
- Anomalies and multiple account numbers for the same entity.



How will you re-engineer legacy data, saddled by account and department orientations, into a consolidated business view?

Without reconditioning and consolidating legacy data, you cannot gain an accurate, consolidated view of customers and the business.

Good data + good data = bad data. Even legacy files that, individually, have adequate data quality for their original purpose, become laden with data problems when merged. Due to the lack of standards and business rules between systems, as well as "creative" data-entry practices and keying errors, the integration of multiple legacy files creates uncommon keys in related records. Without identifying and linking relationships in entities across legacy inputs, the resulting lack of "entity integrity" will expose your systems to erroneous information and generate wrong answers to queries.

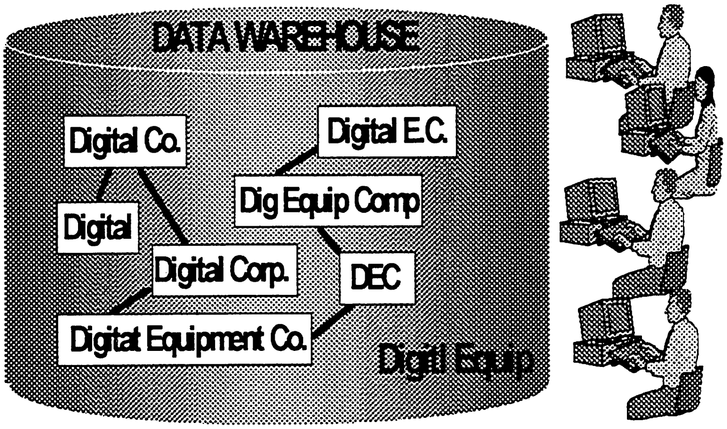

No standard representation for customers and other entities

Even "legacy" data files of good quality in the old world become drenched with contaminants when merged.

For example, a user queries a customer information system to determine fourth-quarter sales from Bristol-Myers Squibb Company. The answer might be a strewn among records for: Bristol-Myers, Bristol Myer, Bristol Meyer, Bristol-Meyers, Brystol Myers, Bristol-Myers Squib, Bristol Myers Squibb Corp, Bristol Myer Squibb Co., etc. If your IS organization is unwilling or unable to make these "fuzzy joins"-- joins without unique keys -- credibility in your new system will be lost quickly.

2. "We have keys to join the data most of the time."

For an enterprise information system critical to improving your decision support, enhancing customer service and enabling new marketing initiatives, "most of the time" is not good enough. Since anomalies and uncommon keys, by nature, tend to gather around your largest customers, even a small percentage of legacy data contamination will compound to invalidate information concerning your company's most prominent clients. From our experience, 80 percent of your queries will access data relating to 20 percent of your customers -- exactly where your data contaminants reside.



THE COMPOUNDING EFFECT OF LEGACY DATA CONTAMINATION:

"Bad" data tends to relate to your most important customers and business entities – the information most commonly accessed for queries. Here, all "instances" of Digital Equipment Corp. are related except for one anomaly. Just this one data "error" will render all queries about Digital sales inaccurate.

3. "If necessary, we'll clean up the data after we've populated the new system, after the pilot."

Unfortunately, post-migration cleanups come too late. You may have already destroyed the credibility of the new system and lost the financial backing of your sponsor, who doubts your ability to recover. Even if you still have the funds to proceed, after-the-fact fixups are often expensive, complex and fraught with risk. What if mending the data requires expanding the data model because you failed to anticipate attributes hidden in free-form text fields? The cost of this kind of extensive system reworking can easily exceed one quarter of your initial implementation cost.

By failing to re-engineer data prior to migration, you will be taxing your data propagation programs, which lack the means to properly map complex relationships (attribute values whose meaning and target destinations depend on correlating values in other fields or records). In addition, data propagation programs lack the logic to parse mixed and floating domains hidden in free-form fields, and to normalize and map these attributes to their appropriate relational domains.

Furthermore, it may be impossible to fix the data after the move. If your data-propagation strategies were too simplistic, you may have misclassified attributes extracted from mixed-domain fields. You will be unable to reverse engineer the mistake back to the point where you can perform the parsing correctly, if the original data sources were not retained. You cannot move forward, because the data is wrong. And you cannot go back, because the bridges were burned behind you.

4. "We're going to fix the data at the point of entry with GUIs and better edit processes."

There are four flaws to this reasoning. First, never underestimate the creativity of data-entry operators, who connive clever ways to circumvent edit rules in order to store unplanned and unanticipated attributes.

Second, a company's business practices and policies will always evolve faster than the application designed to capture them. Tomorrow your CEO wants to change sales compensation strategies. Is your business going to wait for the database and GUI redesign? No. You're going to see that your product continues to ship and that your salespeople continue to get their commissions. But in so doing, you will be entering data values that have strayed from their metadata labels!

Third, there are categories of data whose values and formats are beyond your control and for which no standards exist. For example, customers can choose any way they want to represent themselves. A customer calls up one week to order merchandise and gives his name as Richard Simmons. In subsequent transactions he offers his name as Dick Simmons and Dr. Richard K. Simmons. You can't force a standard on the customer.

Fourth, while it's a good intention to fix the data at the point of entry from this day forward, what will you do about the huge volumes of existing records awaiting the legacy-to-relational bulldozer? A company with millions of customer records needs an automated means to investigate, recondition, and consolidate the information to attain domain and entity integrity. Domain integrity exists if a data value is valid for its range and it correctly relates to other values within a record. For example, the first name, Richard, is a valid first name and it properly correlates to the gender value: male. Entity integrity exists if an entity (i.e., an individual, business or location) is clearly identified by one and only one set of attributes.

Actual (and typical) data sample from a Vality client...

KENTUCKY FRIED CHCKN
KENTUCKY FRIED CHKN
KENTUCKY FRIED
KENTUCKY FRIED CHIC
KENTUCKY FRIED CHIK
KENTUCKY FRIED CHKN
KENTUCKY FRIED 01
KENTUCKY, FIRED CH
KFC CORP
KFC CORP X270-915
KFC CORP OF DESHTSPR
KFC CORP OF POLLYS INC

How will you correctly identify and consolidate anomalies among millions of records?

5. "The users will never agree to change their data."

Will they also disagree about the need for high data quality and accurate answers to queries about their business?

In any event, you can still attain accuracy without changing the data. Through foreign keys, synonym tables and cross-record linkages, you can preserve the original legacy values, yet still produce accurate answers to queries that require a consolidated view. Users may not understand that in a relational system, you can relate many records to represent complex relationships without destroying, eliminating, or changing any of the original data values. You do not need to "dedupe" or standardize "DEC", "Digital" and "Digitl (sic) Equipment" into one surviving record. You can give users an accurate consolidated view simply by linking the records after their relationships have been determined.

Unless you take responsibility for the actual data values that provide the footprint for your business, you may never get the results you're seeking. You're spending huge amounts of money on new information systems. Can your business afford to be misinformed?

This is not to say that the burden is solely on IS shoulders. Your business sponsors must communicate a strong message that the goal is to better understand business practices and trends across departments and lines of business -- currently hidden within those disparate legacy files.

Don't let these excuses prevent your organization from utilizing information as a strategic asset. The health of your company depends on the health of your data.