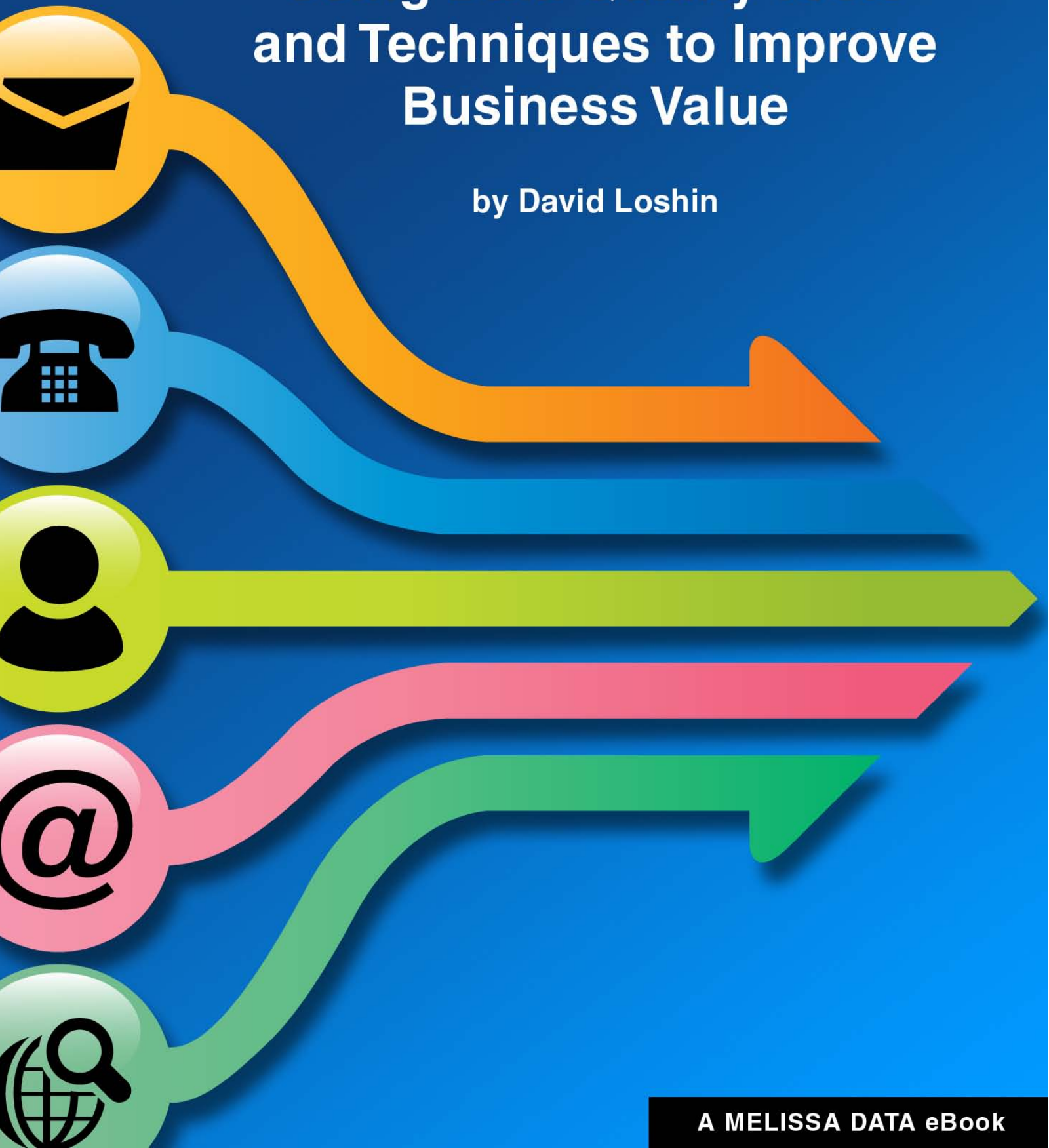
A Data Quality Primer:

Using Data Quality Tools and Techniques to Improve Business Value

by David Loshin



A MELISSA DATA eBook

Table of Contents

Foreword

Over the past two decades, the conventional wisdom on data quality has seemed to drastically change. As decision support systems, data warehousing, and business intelligence triggered greater scrutiny of the data used to measure and monitor corporate performance, a changing attitude has gradually altered our perception of what is meant by "quality information."

Instead of focusing on specific uses of data in the contexts of how data sets support operation of transactional systems, we have started to consider data reuse and repurposing, and noting the data's inherent value, which goes beyond its use to make functional applications work. And as opposed to a knee-jerk reaction to data errors, the industry now focuses on evaluating conformance to business rules that are indicative of a data set's fitness for its (potentially numerous and varied) purposes.

Yet the fundamental aspects of data quality improvement have generally remained the same, and center on a virtuous cycle:

1) Evaluate data to identify any critical errors or issues that are impacting the business
2) Assess the severity of the errors and prioritize their remediation
3) Develop and deploy mitigation strategies
4) Measure improvement to the business
5) Go back to step 1)

All well and good, but we still rely on tools and techniques for each one of these steps. We need tools to evaluate when data errors exist, tools for evaluating the severity of the problem, tools for eliminating the root causes and correcting data, and tools for inspection and monitoring. In this ebook, we focus on part of the challenge: understanding when data values are or are not valid and correct; how data values can be made correct; and how data cleansing services can be integrated into the environment. We describe five key aspects of data quality management:

- **Data cleansing**
- **Address data quality**
- **Address standardization**
- **Data enhancement**
- **Record linkage and matching**

We then look at some of the practical aspects of introducing proactive data quality management into the organization.

CHAPTER 01

The Theory of Data Cleansing

1.1 Data Cleansing and Simple Business Rules

Data cleansing combines the definition of business rules in concert with software designed to execute those rules. Yet there are some idiosyncrasies associated with building an effective business rules set for data standardization and particularly, data cleansing. At first blush, the process seems relatively straightforward: We have a data value in a character string that we believe to be incorrect and we'd like to use the automated transformative capability of a business rule to correct that incorrect string into a correct one.

Here is a simple example: For address correction, we'd like to expand out the abbreviations for the street type such as "road," "street," "avenue," etc.). For the road type of "STREET," we might have rules such as:

- STR is transformed into STREET
- ST is transformed into STREET
- St is transformed into STREET
- St. is transformed into STREET
- Str is transformed into STREET
- Str. is transformed into STREET

And so on. The approach that would be taken is to integrate these rules into a data cleansing rules engine, and then present our strings to be corrected through the engine. To continue the example (and if we also included a rule that upper-cases all letters), the string "1250 Main Str." might be transformed into "1250 MAIN STREET" and provided back to the calling routine. Seems simple, no?

Of course it is. And *simplistic* as well, since the same transformation might happen when presenting this street name as well: "St. Charles St," which would be changed into "STREET CHARLES STREET" when using that same set of rules. Because the rule is so basic, there are no controls over how, where, and when the rule is applied. We'd have to have more rules and a bit more control to effectively transform and *correctly* correct the data.

1.2 Designing More Effective Data Cleansing Rules

After considering a simple data transformation and cleansing rule that was to be used to standardize a representation of a street type., we found that an uncontrolled application of the rule made changes where we didn't really want a change to happen.

There are two reasons why applying the rule led to an undesired result. The first issue is context: even though we want to map the string "St." to "STREET" when it appears at the end of an address, that same string "St." appearing earlier in the street name and before a proper noun is more likely to be an abbreviation for "Saint," not "Street." The obvious way to fix that is to introduce a new rule that maps "St." to the word "SAINT."

And that introduces the second issue: order of execution. Let's say we did have that second rule, so our rule set now looks like this:

- St is transformed into SAINT
- St. is transformed into SAINT

- STR is transformed into STREET
- ST is transformed into STREET
- St is transformed into STREET
- St. is transformed into STREET
- Str is transformed into STREET
- Str. is transformed into STREET

Applying that rule set to our string "St. Charles St" yields "SAINT CHARLES SAINT" which is also clearly incorrect. The reason is that we have conflicting rules, in which the same input maps to two different outputs, ordering the rules won't work either, since the same thing will happen (just with the other output values). To address this challenge, we need an approach that accommodates both context and order of execution.

1.3 Adding Context to Data Quality Rules

We can look at a formal summarization of the challenge of conflicting data quality rules. We have two rules, R1 and R2, and the same input X:

- R1: Transform string X into string Y1
- R2: Transform string X into string Y2

It is easy to see this conflict using the simple examples in my previous posts, but in fact, as your data cleansing rule set grows, the potential for introducing conflicting rules not only grows, the ability to find them diminishes.

There are a couple of approaches for addressing this challenge. The first is greater differentiation in defining the cleansing rule through the use of contextual cues. In our example, we might look at these conflicts:

1. St is transformed into SAINT
2. St. is transformed into SAINT

3. St is transformed into STREET
4. St. is transformed into STREET

and introduce contextual constraints:

1. St is transformed into SAINT at the beginning of a street name
2. St. is transformed into SAINT at the beginning of a street name
3. St is transformed into STREET at the end of a street name
4. St. is transformed into STREET at the end of a street name

This approach somewhat addresses the problem in some cases, but becomes an issue again when there are new contexts, such as a string like “Trevor St. Lawrence St.” which would necessitate yet another contextual rule.

1.4 Adding Context and Ordering to Data Quality Rules

One approach to resolving data cleansing rule conflicts is the introduction of contextual constraints for application of the rules. This could help in differentiating the application of rules, but could grow to be complex quickly. There is a second approach that could be used, which is adjusting the rule set somewhat to ensure distinction of abbreviation and then phasing the application of rules.

The idea is that if we have two rules that share the same input but have different outputs, using the form:

- R1: Transform string X into string Y1
- R2: Transform string X into string Y2

then a modification to the rule set to break that conflict could work if we first correct all

instances of one type of context-dependent inputs into a modified form and then apply modified rules during a second pass. Here is another stab at modifying our sample rules into two passes. Here is pass 1:

1. St is transformed into __STREET__ at the end of a street name
2. St. is transformed into __STREET__ at the end of a street name

Here is pass 2:

1. St is transformed into SAINT
2. St. is transformed into SAINT
3. __STREET__ is transformed into STREET

We used the more predictable contextual rules for the first pass and changed the flagged items into some token that would probably never appear as a placeholder for the next pass. Hopefully, we will have filtered out all of the strictly context-dependent instances in the first pass, allowing us to loosen the constraint for the instances in which the context is less predictable (thereby transforming "Trevor St. Lawrence St." into "TREVOR ST. LAWRENCE __STREET__" after pass 1 and into "TREVOR SAINT LAWRENCE STREET" after pass 2).

This is just one way to approach the challenge, yet there are other ideas that can be applied. The first step is to look at the ways your data cleansing tools define rules as a way to consider the options, and we will explore this in greater detail in the following sections.

CHAPTER

02 Address Quality

2.1 Characterizing the Quality of Address Data

One aspect of managing the quality of master address and location data involves reviewing a lot of the existing documentation that has been collected from a number of different operational systems, as well as reviewing the business processes to see where location data is either created, modified, or read. There are likely to be many references to operations or transformations performed on addresses, mostly with the intent of improving the quality of the address.

Curiously, there are often a number of different terms used to refer to these different transformations: validation; verification; standardization; cleansing; correction. But what do all these things mean? And why are these different terms used if they mean the same thing?

The first step in exploring the answer to this question is reflecting back on the nature of deliverable addresses. When an item is sent to an addressed location, there are some core concepts that need to be right:

1) The item must be directed to a specific recipient party (either an individual or an organization).
2) The address must be a deliverable address.
3) The intended recipient must be associated with the deliverable address.

In addition, there are certain incentives provided to senders when the addresses are completely aligned with the Postal Standard, adding one more concept:

4) The delivery address must conform to the USPS® standard.

These directives provide us with some material with which to work for differentiating the different terms used for postal data quality.

2.2 Translating Expectations into Quality Directives

After considering the variety of terms used in describing address quality, we introduced a set of core concepts that needed to be correct to provide the best benefits for accurate parcel delivery. Let's look at these more carefully:

1) The item must be directed to a specific recipient party (either an individual or an organization).
2) The address must be a deliverable address.
3) The intended recipient must be associated with the deliverable address.
4) The delivery address must conform to the USPS standard.

Together these concepts have implications for address quality, and we can start with the first 3 concepts. The first concept implies a direct connection between entities: the sender and the recipient. The corresponding business rule is relatively subtle – it suggests that the recipient must be identifiable to the sender. Concept #2 is a bit more direct: the address must be a deliverable address. This means that the address must carry enough information to enable a carrier to locate the address as a prelude to delivery. Concept #3 establishes a direct dependence between the recipient and the addressed location, implying awareness of that connection.

Together we can infer more discrete assertions:

- The address must be accurately mappable to a real location.
- The address must contain enough information to ensure delivery.
- The recipient must be a recognized entity.
- The recipient must be connected to the address.

Our next steps are to figure out what these assertions really mean in terms of transforming a provided address into a complete, validated, and standardized address.

2.3 Postal Standards and Address Quality – Take 1

The USPS Postal Standard (Publication 28) provides at least some of the specifications we need for address quality. For example,

> "The Postal Service defines a *complete address* as one that has all the address elements necessary to allow an exact match with the current Postal Service ZIP+4 and City State files to obtain the finest level of ZIP+4 and delivery point codes for the delivery address."

The next paragraph provides some additional details:

"A *standardized address* is one that is fully spelled out, abbreviated by using the Postal Service standard abbreviations (shown in this publication) or as shown in the current Postal Service ZIP+4 file."

A large part of the remainder of the document guides what is valid and what is not valid, as well as the postal standard abbreviations (as mentioned in the definition of *standardized*). So an address must be complete, which by definition implies that it can be matched with current Postal Service

ZIP + 4® and City State files. This match is to obtain the ZIP + 4, so the implication is that **verification** means that a complete address matches the USPS files and has the correct ZIP + 4. The address components must be consistent with the postal standard in terms of valid and invalid values. For example, a street address cannot have a number that is outside the range of recognized numbers (that is, if the USPS file says that Main Street goes from 1-104, an address with 109 Main St is invalid). So **validation** means that the street address is consistent with what is documented by the USPS files.

Standardization is also defined by the above reference: it is spelled out, and uses the USPS standard abbreviations.

In turn, the process for address quality would be to:

3) Ensure the address is complete.
4) Ensure that the address values are valid by checking it against the USPS files.
5) Verify the address's ZIP + 4 by matching against the USPS files.
6) Standardize the address according to the USPS standardized abbreviations.

2.4 Address Quality – Take 2

We have dealt with some of our core address quality concepts, but not this one:

The intended recipient must be associated with the deliverable address.

The problem here is no longer *address quality* but rather *address correctness*. The address may be complete, all the elements may be valid, the ZIP + 4 is the right one, and all values conform to standardized abbreviations… and still be **incorrect** if the recipient is not associated with the address!

This is the bigger challenge with address data quality, since address correctness or accuracy is a factor of real-world events that are not necessarily synchronized with your databases. Some level of control is again served by the Postal Service through the NCOA^{Link}® (National Change of Address) dataset that is licensed to tools providers. Checking against the NCOA^{Link} data set will notify you if an entity linked to a location

has self-reported a change of address, and this accommodates a large portion of the address correction issues. However, there are estimates about the percentage of people that moved, and I recall reading a Census Bureau press release about their 2009-2010 statistics noting that 12.5% of the population moved over the year.

Not all changes propagate to the NCOALink file at the right time, and it may take a while before all consumers of that data actually synch up with the NCOALink dataset. Even if you do a quarterly review, if we trust that 12.5% statistic, then there is a pretty good chance that by the end of the quarter you will still have a 3-4% inaccuracy rate for mapping entities to locations.

And there are other considerations that are not incorporated into this calculation. For example:

- Individuals change jobs and therefore change business addresses
- Third-party data vendors incorrectly link individuals to locations
- Miskeyed data
- Purposely incorrect data
- Propagation of legacy addresses overwriting updated addresses

This a small sample of challenges. But what it means is that there are many aspects of assessing and assuring the quality and correctness of addresses, and it may be worth reviewing the ways that your organization verify, validated, standardize, and correct location data!

2.5 Non-USPS Addresses: Verifying Addresses Off the Beaten Path

In many smaller and rural communities, including village and mountainous regions, the USPS does not deliver directly to homes and businesses. Instead, the USPS assigns PO boxes to every address in the community and only delivers mail to the PO Box™. Trying to verify actual street addresses using only the USPS database would return these unknown, non-USPS addresses as "bad." There are more than 5 million of these so-called "bad," non-USPS addresses.

The inability to verify non-USPS addresses can be a problem for many businesses. For instance, while the USPS does not deliver packages to these non-USPS addresses, UPS, FedEx and other shippers might. But most order fulfillment and contact center systems are tied to USPS data, which wouldn't be able to verify a non-USPS address as "shippable" – resulting in a lost order.

Cell phone providers often require a verifiable street address (not a PO Box) for account activation. Subprime lenders and other credit providers want a verified address to help ensure a person applying for a loan has some piece of valid information – and that they are reachable at a physical address. And, verifiable addresses are useful for political parties and civic organizations for pre-election campaigns. With traditional methods (tied solely to USPS address data) verification in some rural areas would be impossible.

Fortunately, there is a way to verify these non-USPS addresses. A few data vendors are able to aggregate contact data from multiple sources to identify and verify these non-USPS addresses.

Vendors offering multisourced data precludes the need to have contracts with third party shipping providers for the additional reference data – resulting in lower costs.

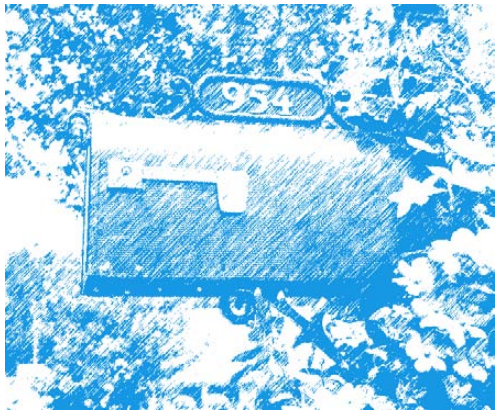
Some vendors also offer the option of geocoding these non-USPS addresses to the rooftop level. Geocoding is useful to identify and pinpoint the exact latitude and longitude coordinates for each non-USPS address, helping businesses make informed decisions about risk exposure, tax jurisdictions, sales clusters, marketing segmentation, logistics, and much more.

CHAPTER

03 Address Standardization

3.1 The Importance of Standards for Addresses and Locations

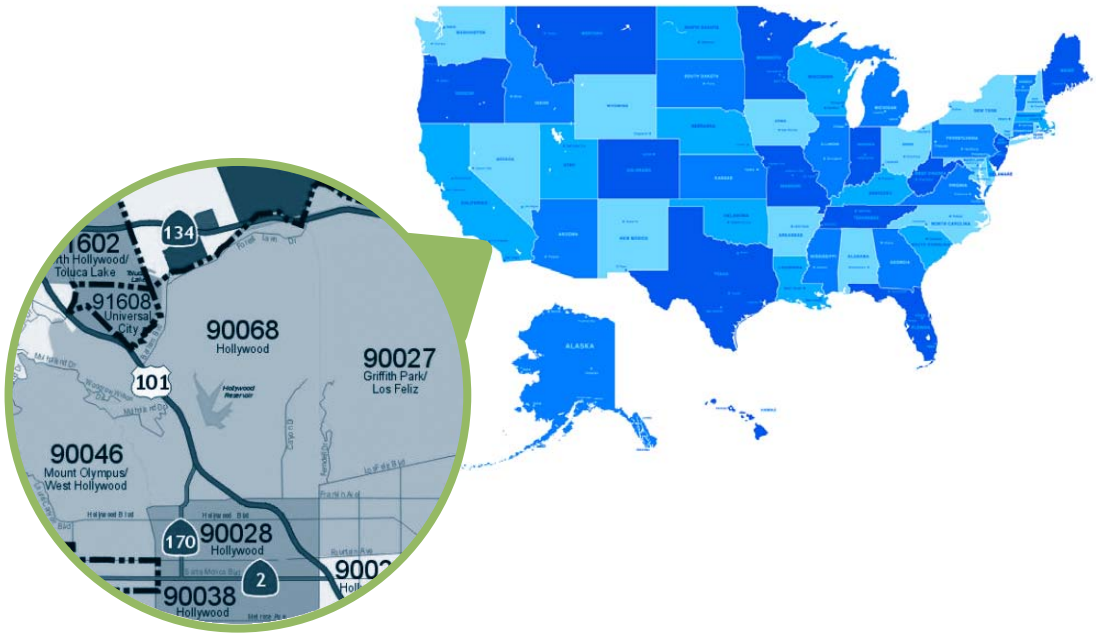
Here is a simple scenario, followed by what should be a simple question: We have an item that we'd like to have delivered to a specific individual at a particular location, and we'd plan to engage an agent to deliver the item on my behalf. How can we communicate to the agent where the item is to be delivered? From the modern day perspective, it should be obvious: you only need to provide the street address and you expect that the agent will be able to figure it out on his own.



However, this begs the question, specifically because we never think about the existing framework of standards that has evolved around any communication system (whether that is for physical package delivery, connecting telephone calls, or sending emails). We expect that the delivery agent will be able to figure out how to get to a location because the standard address format contains a hierarchical breakdown for refining the location at finer levels of precision. In the US, an address contains a street name and number, as well as a city, state, and a postal code. The refinement can begin with the state, then

resolve down to the city, and then the postal code. By the time you have resolved down to that level, one is likely to be able to easily find the street, and the specific location is determined using the street number.

This process works in the US because there is a postal standard, and in fact, the driving force behind addressing standards is the need for accuracy in delivery. Following the standard ensures that anyone reading an address has enough information to be able to find the addressed location. In the US, there is actually a very comprehensive standard, called **"Publication 28,"** and it describes all the gory details. And next, we will look at some interesting aspects of addressing, address standards, and the concept of "location."



3.2 The Need and the Mechanics of Address Standardization

In the last section, we discussed the topic of address standards, and conjectured that the existence of a standard for addresses not only simplifies the processes of delivery, it helps to ensure delivery accuracy. Ultimately, delivery accuracy saves money, since it reduces the amount of effort to find the location and it eliminates rework and extra costs of *failed delivery*.

This is all well and good as long as you use the standard. The problem occurs when, for some reason, the address does not conform to the standard. If the address is slightly malformed (e.g. it is missing a postal code), the chances are still good that the location can be identified. If the address has serious problems (e.g. the street number is missing, there is no street, the postal code is inconsistent with the city and state, or other components are missing), resolving the location becomes much more difficult (and therefore, costly).

There are two ways to try to deal with this problem. The first is to bite the bullet and treat each non-standard address as an exception, forcing the delivery agent to deal with it. The other approach attempts to fix the problem earlier in the process by trying to transform a non-standard address into one that conforms to the standard. Address standardization is actually not that difficult, especially when you have access to a good standard. At the highest level, the process is to first determine where the address does not conform to the standard, then to standardize the parts that did not conform.

Remember that an address captures the incremental knowledge to resolve the location. We can use this fact, plus the information provided in the standard, to consider ways to fix non-standard addresses. Each component has its specific place inside the address, and there are standards for abbreviations (such as ST for "street," or AVE for "avenue") as well as for common terms (such as ATTN for "attention"). One can define a set of rules to check if the address has all the right pieces, if they are in the right place, and if they use the officially-sanctioned abbreviations. You can also use rules to move parts around, to map commonly-used terms to the standard ones, and use lookup tables to fill in the blanks when data is missing. So in many cases, it is straightforward to rely on tools and methods to automatically transform non-standard addresses into standardized ones.

3.3 Techniques for Address Standardization

We've discussed the value and importance of address standardization as an integral component of both transactional and analytical applications, especially when seeking levels of accuracy associated with the concept of *location*, which in some cases goes beyond the concept of "address." But, knowing that with some degree of precision, we

can map locations to their nearest geocoded location, let's think about aspects of a more general challenge: ensuring resolution of a provided, descriptive address to an actual known address.

Let's clarify this a little. When we discuss a "provided, descriptive address," we refer to what an individual has presented as an address. And, while another individual might be able to infer enough meaning from a presented address to make a delivery, the address might have misspellings, errors, or other variations that might prevent it from being adequately mapped to a specific geocoded location.

Aside from the other benefits we have already considered, transforming the address into a standard form will simplify the geocoding process. That transformation process leverages a few straightforward ideas, namely:

1. There is a representative model for "standardized" addresses with its accompanying formats, syntax, acceptable value lists, and rules.
2. An application is able to scan a non-standardized (or what I called a "provided descriptive") address, differentiate between the parts that are good and the ones that do not meet the standard.
3. There is a way to map the non-standard parts into standard ones.

In fact, all three of these ideas are doable, and over the next set of postings, let's look at each one of these in greater detail.

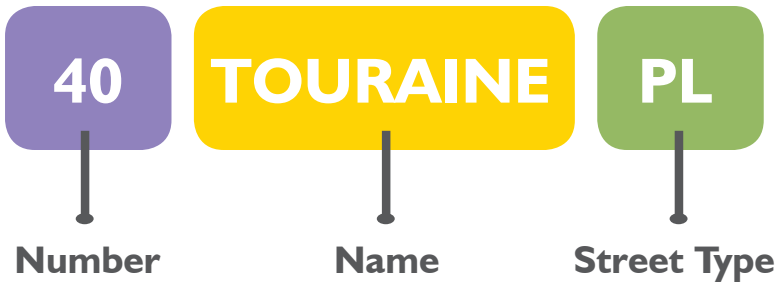
3.4 Formats, Syntax, and Content

One great thing about having a standard representation for data is that it becomes easy to see whether any value does or does not meet the standard. Let's use a simple example: we can say that a street address has to have three parts – a number, a name, and a "street type." We can further specify our example standard with these constraints:

- The number must be a positive integer number
- The name must have one, and only one word
- The street type must be one of the following: RD, ST, AV, PL, or CT

While there are many streets with names that span more than one word, and there are a lot more types of streets, this experiment is to demonstrate how we can use the standard to determine if an address is valid or not by comparing it against the defined format, syntax, and content characteristics, such as:

- The address string must have three components to it (format)
- The first component has to only have characters that are digits 0-9 (syntax)
- The first character of the first component cannot be a '0' (syntax)
- The third component must be of length 2 (format)
- The third component has to have one of the valid street types (content)



In other words, we are refining the rules for validity into ones that we can test. If the first component of the address has any characters other than digits, it is not a valid address, and if the last component of the address is "AVE" the string is not valid, since the length of that component is 3, not 2.

3.5 Business Rules for Standardization: Bringing it all Together

While we have been talking in the last few sections about checking whether a data value observes the standard (and is therefore a valid value), the real challenge in standardization is in determining (1) that a value does not meet the standard and then (2) taking the right actions to modify it so that it does meet the standard. That process, strangely enough, is called "standardization," and it extends the tokenization and parsing

to recognize both valid tokens and common patterns for *invalid* ones, and that is where the power of standardization lies.

Here is the basic idea: when you recognize a token value to be a known error, you can define a business rule to map it to a corrected version. The example we have used is a simple address standard:

- The number must be a positive integer number
- The name must have one, and only one word
- The street type must be one of the following: RD, ST, AV, PL, or CT

And deriving these additional expectations:

The address string must have three components to it (format)

- The first component has to only have characters that are digits 0-9 (syntax)
- The first character of the first component cannot be a '0' (syntax)
- The third component must be of length 2 (format)
- The third component has to have one of the valid street types (content)

The next step would be to consider the variations from the expected values. A good example might look at the third token, namely the street type, and presume the types of errors that could happen and how they'd be corrected, such as:

Possible errors	Standard
Rd, Road, Raod, rd	RD
Street, STR	ST
Avenue, AVE, avenue, abenue, avenoo	AV
Place, PLC	PL
CRT, Court, court	CT

In this example, we see some variant abbreviations, fully-spelled out words, a finger flub (such as when the typist hit the b key instead of the v in "abenue"), and a transposition ("Raod" instead of "Road", also very frequent).

Different types of formats and patterns can be subjected to different kinds of rules. The first token has to be an integer, but perhaps some OCR reader mis-translated what it scanned into a character instead of a number, so we might see O instead of 0, A instead of 4, S instead of 8,) instead of 9, etc. That means that part of the standardization process looks for non-digits and then applies rules that traverse through a string and convert according to the defined mappings (A becomes 4, for example).

For the second token, the challenge is when more than three words appear. One set of rules might take all tokens between the first and the last and concatenate them together into a single word. Another approach is to scan the tokens and pluck out the one that most closely matches one of the street types and move that to the end.

So these are the basic ideas for standardization: defining the formats and patterns; determining the tokenization rules; parse the data and recognize valid tokens and invalid tokens; define rules for mapping invalid tokens to valid ones; and potentially rearrange tokens into the corrected version. In reality, there are many more challenges, opportunities, and subtleties, but this description (at least) provides a high level view of the general process.

3.6 Tokenization and Parsing

To summarize, the data values stored within data elements carry specific meaning within the context of the business uses of the modeled concepts, so to be able to standardize an address, the first step is identifying those chunks of information that are embedded in the values. This means breaking out each of the chunks of a data value that carry the meaning, and in the standardization biz, each of those chunks is called a *token*. A token is representative of all of the character strings used for a particular purpose. In our example, we have three tokens – the number, name, and type.

Token categories can be further refined based on the value domain, such as our street type, with its listed, valid values. This distinction and recognition process starts by *parsing* the tokens and then rearranging the strings that mapped to those tokens through a process called *standardization*. The process of parsing is intended to achieve two goals – to validate the correctness of the string, or to identify what parts of the string need to

be corrected and standardized. We rely on metadata to guide parsing, and parsing tools use format and syntax patterns as part of the analysis.

We would define a set of data element types and patterns that correspond to each token type and the parsing algorithm matches data against the patterns and maps them to the expected tokens in the string. These tokens are then analyzed against the patterns to determine their element types. Comparing data fields that are expected to have a pattern, such as our initial numeric token or the third street type token, enables a measurement of conformance to defined structure patterns. This can be applied in many other scenarios as well, such as telephone numbers, person names, product code numbers, etc. Once the tokens are segregated and reviewed, as long as all tokens are valid and are in the right place, the string is valid.

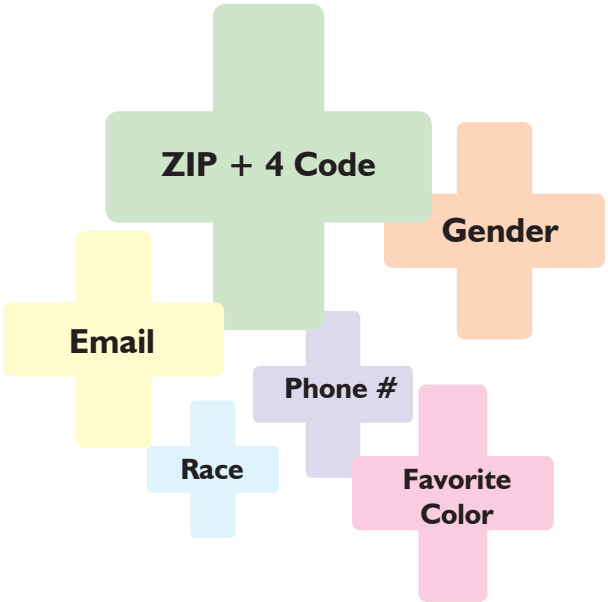
CHAPTER

04 Data Enhancement

4.1 Increasing Data Utility and Gaining Business Value from Data Enhancement

Most business applications are originally designed to serve a specific purpose, and consequently, the amount of data either collected or created by any specific application is typically just enough to get the specific job done. In this case, the data is utilized for the specific intent, and we'd say that the "degree of utility" is limited to that single business application.

On the other hand, businesses often use data created by one application to support another application. As a simple example, customer location data (such as ZIP™ codes) that is collected at many retail points of sale, is used later by the retail business to analyze customer profiles and characteristics by geographical region. In these kinds of scenarios, the degree of utility of the data is increased, since the data values are used for more than one purpose.



In fact, data sets are constantly targeted for repurposing, but one challenge that emerges is that sometimes the data that is collected or created is not of sufficient quality for the secondary uses. Errors, missing values, misfielded data, or any number of other data flaws detract from its potential utility. Fortunately, many of these data flaws are easily addressed through data enhancement, which (informally) is a process that adds information to a data set to improve its potential utility.

There are a number of different ways that data sets can be enhanced, including adapting values to meet defined standards, applying data corrections, and adding additional attributes. In this section, we will look at scenarios in which specific data enhancements help businesses to improve value along a variety of value drivers.

4.2 Data Enhancement for Operational Purposes

The concept of data enhancement is a collection of methods for adding information to a data set to increase its utility, and suggested that there are a number of scenarios where enhancement adds business value. Let's review two examples that show ways that data enhancement can be incorporated into operational scenarios.

We can start with a very common use of enhancement: postal standardization and address correction. A delivery address describes a specific location to which an item can be delivered. In the United States, it is usually composed of a street name, a street number, a city name, a state identifier, and a postal code. When executing a sales transaction, you would probably want to make sure that you have a valid delivery address to ensure that the purchased products can be sent to the customer. So, although there is a wide variety of ways that people could assemble a delivery address, the address data can be submitted to an address validation and correction enhancement process to ensure proper delivery.

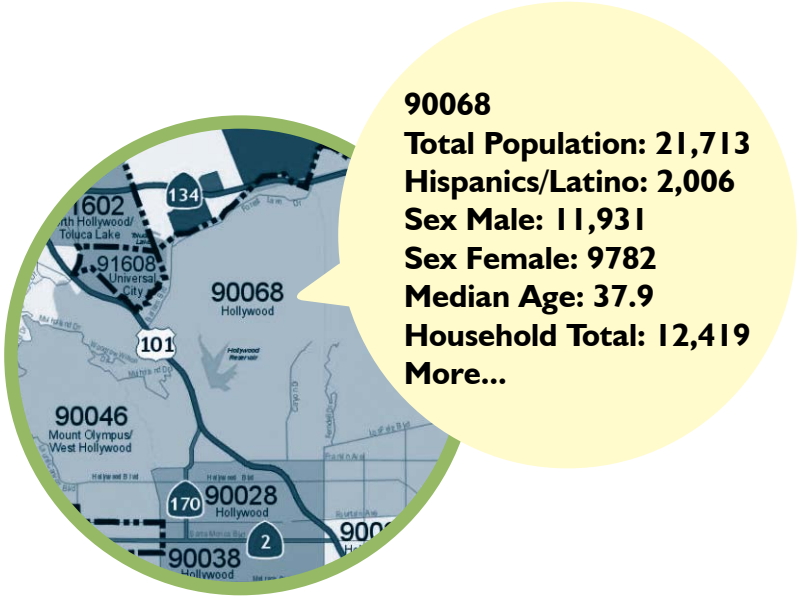
Another common example involves individual's names, which can appear in data records in different ways: first name followed by last name; last name with a comma, followed by first name; with or without titles such as "Mr." or "Professor;" different generational suffixes… In a recent conversation with some colleagues at the US Census Bureau, they shared with me that they have over 1000 different patterns for ways that names can

appear in data. Again, a data standardization and enhancement process can parse out the key components of a person's name, fill in the blanks (if necessary) through lookups in master data tables, and reorganize those components into a format so that a customer's identity can be established for verification purposes when providing customer service at an inbound call center.

4.3 Data Enhancement for Analytical Purposes

The value of data enhancement is not limited to insertion in specific operational workflows, because enhancement is often done to provide additional detail for reporting and analysis purposes. And in these cases, enhancement goes beyond data standardization and correction; instead, the enhancement process can add more information by linking one data set to another. The appended data can augment an analytical process to include extra information in generated report and interactive visualizations.

As an example, consider collecting ZIP Code values at a point of sale. A retail company can take sales data that includes this geographic data element and then enhance the data with **demographic profiles provided by the US Census Bureau** to look for correlation between purchasing patterns and documented demographics about the specific locations (including sex, age, race, Hispanic or Latino origin, household relationship, household type, group quarters population, housing occupancy, and housing tenure).



Geographic data enhancement also adds value for analysis. Given a pair of addresses, an enhancement process

can evaluate different types of distances (direct distance and driving distance are two examples) between those two points. This can be useful in a number of analytical applications, such as site location planning, which compares properties based on a variety of criteria (possibly including the median driving distance for local customers for a bank branch, or average driving time for delivering pizza to frequent customers).

There are many data aggregators who can supply demographic and behavior data that can enhance your customer data sets. And, you can use your own company's data for enhancement as well, such as your own product sales by region used to develop your own customer segmentation data.

Standardizing names and addresses is the first step, and linking those records to the reference data collections allows direct linkage based on specific criteria, ranging from gross-level linkage (say, at the county level) down to specific enhancement at the individual level (such as the names of the magazine to which a customer subscribes). These qualitative enhancements augment the business intelligence and analytics processes to help companies make more sales, increase revenues, and improve profitability.

4.4 Integrating Analytical Results with Operational Activities

We have looked at using enhancement for operational purposes, as well as analytical purposes, but there are ways that we can merge the two into a hybrid: using enhanced data for analytics, whose results are incorporated into operation activities using the same types of enhancements.

We have already begun to lay the groundwork for more interesting uses of enhancement, beginning with the use of enhancement for person names, perhaps in a customer support capacity, in order to verify identity. At the same time, we can examine the use of analytics that use customer enhanced data for customer profiling and geo/demo/psychographic segmentation. These two processes can be combined to provide even more effective recommendations to improve customer support, or even drive additional sales.

Consider this scenario: call origination data (such as telephone number) is provided when callers reach out to an inbound call center. Location data associated with the originating telephone number is used as a key to look up geographic/demographic data, which is used to enhance the inbound caller's record with segmentation data nominally associated with the individual. As the call center representative walks through specific scripts provided to help the caller, the enhanced profile information is used to adjust offers in real time based on previously calculated statistics.

As a more direct example, complaints about dropped mobile calls might lead into a script to recommend upgrading equipment. Based on the caller's enhanced location and geographic profile, historical measures of accepted offers coupled with connectivity statistics in the given area can be mined. The results can then be fed into the call center application, which provides specific suggestions for the call center rep to offer a particular type of phone that is best suited for maintaining connections in the customer's location, at a promotional price that the customer is likely to accept.

With a little bit of thought, we can come up with many types of these hybrid scenarios – ones where data enhancement is used for both analytical and operational purposes, including data standardization and cleansing, all with a focus towards improved business functions.

CHAPTER 05

Record Linkage and Matching

5.1 Distributed Data and Distributed Information

You might not realize how broad your electronic footprint really is. Do you have any idea how many data sets contain information about any specific individual? These days, any interaction you have with any organization is likely to be documented electronically. And, for those curious enough to read the fine print of the "privacy" policies, you might not be surprised to find that many of those organizations managing information about you are sharing that information with others.

Profile Informaiton

What is your gender? M or F
What is your age?
What is your highest level of education?
What is your annual household income?
Do you have any children? If so, how many?
Are you a local resident?
How many vocations do you take a year?
Are they domestic or international?

Contact Informaiton

Last name
First name
Street Address
City, State, ZIP
Country
Phone Number
Email Address

Actually, this is not a new phenomenon; this has been going on for many years by data aggregator

companies who just love to collect data and turn it into salable products. The easiest example to share is that of the mailing list company with the reference database that can be segmented across numerous geographic dimensions (in incremental precision such as state, county, town, ZIP Code, ZIP + 4, street name, etc.) as well as demographic dimensions such as number of cars owned, favorite leisure activities, or household income.

And, any time you fill out some form or respond to some survey or another, more information about you is captured. Remember that registration card you filled out for the toaster you bought? The survey you filled out to get that free subscription? Didn't you subscribe to some magazine about fishing and other outdoors activities? How about that contest you entered at the county fair?

Actually, you are not the only one collecting your information. Did you buy a house? Home sales are reported to the state and the data is made available, including address, sales price, and often the amount of your mortgage. Wedding announcements, birth announcements, obituaries log life cycle events.

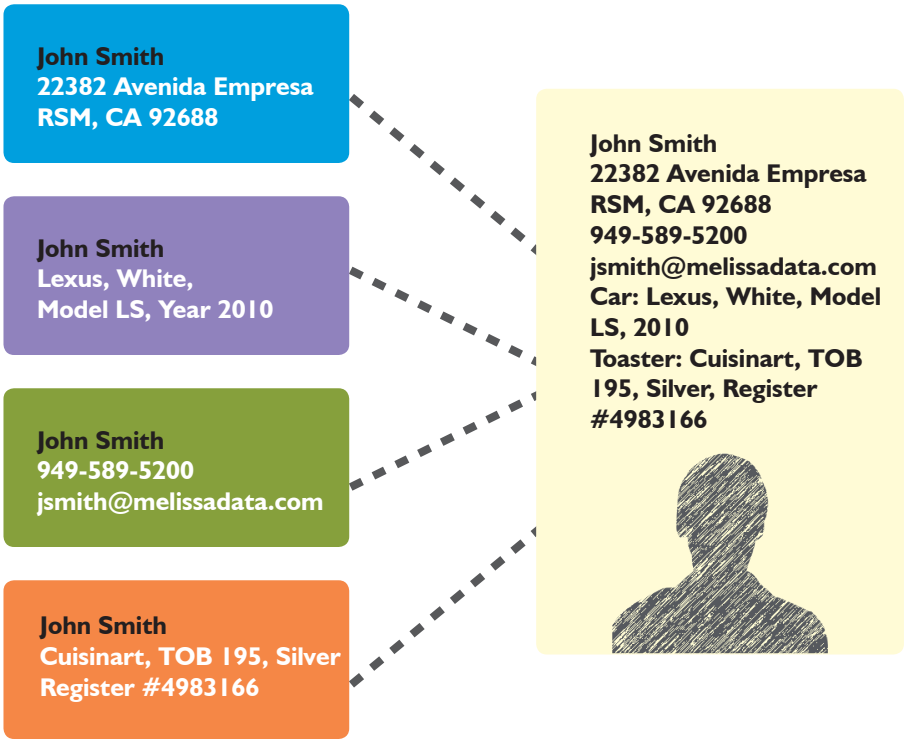
Every single one of these artifacts captures more than just some information about an individual – it also captures the time and place where that information is captured, sometimes with accurate precision (such as the time of an online order) or less precision (such as the day the contest entry was collected from the box. There are many distributed sources of information about customers, and each individual piece of collected data holds a little bit of value. But when these distributed pieces of data are merged together, they can be used to reconstruct an incredibly insightful profile of the customer. In this section we will examine exactly how this process works.

5.2 What is Record Linkage?

Many distributed pieces of data about a single individual can be combined together to form a deep profile about that individual. But how are different data records from disparate data sets combined to formulate insightful profiles?

The answer lies in the ability to collect the different pieces of data that belong to a single individual and then glom them together. For example, let's presume the existence of a record in one data set that has a person's address, a record in another data set that has that person's telephone number, a third record that has that person's registration number for a toaster, another with the person's car year, make, and model, etc. As long as you can find all the records that are associated with each person and connect them together, you could collect all the interesting information together and create a single representative profile. That profile is then suitable for use in list generation, but is also used for more comprehensive analytics such as segmentation, clustering analysis, and classification.

The way these records are connected together is through a process called "record linkage." This process searches through one or more data sets looking for records that refer to the same unique entity based on identifying characteristics that can be used to distinguish one entity from all others, such as names, addresses, or telephone numbers.



When two records are found to share the same pieces of identifying information, you might assume that those records can be linked together.

It sounds simple, but unfortunately, there are a number of challenges with linking records across more than one data set, such as:

- The records from the different data sets don't share the same identifying attributes (one might have phone number but the other one does not).
- The values in one data set use a different structure or format than the data in another data set (such as using hyphens for social security numbers in one data set but not in the other).
- The values in one data set are slightly different than the ones in the other data set (such as using nicknames instead of given names).
- One data set has the values broken out into separate data elements while the other does not (such as titles and name suffixes).

Luckily, there are numerous software products that are designed to address these discrepancies, which can simplify the record linkage process. If you recall, you may begin to see how parsing and standardization start to fit in. These tools will parse and standardize the values prior to attempting to compare for the purposes of linkage, and that alleviates some of the noted challenges.

5.3 Record Linkage and Data Enhancement

We've considered how the distribution of information about entities and the use of record linkage to find corresponding data records in different data sets can be linked together. Record linkage can be used for a number of processes that we bundle under the concept of "data enhancement," which we'll use to describe any methods for improving the value and usefulness of information. There are three different types of enhancement to review in greater detail:

- **Data cleansing** – The first type of enhancement is relatively straightforward: our idea is to link records together for the purpose of cleansing the data, or making it more suitable for use. Often, one data set may have a more trustworthy representation of an entity, or we may have more than one data set, each potentially containing overlapping data elements such as birth date, address, telephone number. By linking two different records, you can compare the corresponding values, find those that are of better quality (e.g. more complete or more current values) and update the "delinquent" record with the higher quality values.

- **Enrichment** – Existing records for entities (such as people or products) can be matched against other data sets with additional reference information. For example, you might want to match your customer data with a credit bureau's data and enrich your own data set with each individual's credit ratings.
- **Merge/Purge** – Duplicate records entered into one data set often plague the business in attempting to actively manage customer accounts. Applying the record linkage methodology to the records in a *single* data set helps find multiple records that refer to the same individual. These records can be presented to a data analyst to review and determine the surviving record and updating the record with the highest quality values.

There are many variations on these themes. For example, merge/purge can be used for combining customer data sets after a corporate acquisition; enrichment can be used to institute a taxonomic hierarchy for customer classification and segmentation. Loosening the matching rules for merge/purge can help with a process called "householding," which attempts to identify individuals with some shared characteristics (such as "living in the same house").

5.4 Inferred Knowledge and Customer Intelligence through Matching and Linkage

One of the most interesting byproducts of record linkage is the ability to infer explicit facts about individuals that are obfuscated as a result of distribution of data. As an example, consider these records, taken from different data sets:

A:

David	Loshin	301-754-6350	1163 Kersey Rd	Silver Spring	MD	20902

B:

Knowledge Integrity, Inc	1163 Kersey Rd	Silver Spring	MD	20902

C:

H David	Lotion	1163 Kersey Rd	Silver Spring	MD	20902

D:

Knowledge Integrity, Inc.	301	7546350	7546351	MD	20902

We could establish a relationship between record A and records B and C because they share the same street address. We could establish a relationship between record B and record D because the company names are the same. Therefore, by transitivity, we can infer a relationship between "David Loshin" and the company "Knowledge Integrity, Inc" (A links to B, B links to D, therefore A links to D). However, none of these records alone explicitly shows the relationship between "David Loshin" and "Knowledge Integrity, Inc" – that is inferred knowledge.

You can probably see the opportunity here – basically, by merging a number of data sets together, you can enrich all the records as a byproduct of exposed transitive relationships. This provides us with one more valuable type of enhancements that record linkage provides. And this is particularly valuable, since the exposure of embedded knowledge can, in turn, contribute to our other enhancement techniques for cleansing, enrichment, and merge/purge.

5.5 Entities and Their Characteristics

How can you tell if two records refer to the same person (or company, or other type of organization)? In previous sections, we have looked at how data quality techniques such as parsing and standardization help in normalizing the data values within different records so that the records can be compared. But what is being compared?

A simplistic view might suggest that when looking at two records, comparing the corresponding values is the best way to start. For example, we might compare the corresponding names, telephone numbers, street addresses – stuff that usually appears in records representing customers, residences, patients, etc.

But the simple concept belies a much more complex question about the attributes used to describe the individual as well as differentiate pairs of individuals. Much of this issue revolves around the approaches taken for determining what characteristics are being managed within a representative record, the motivation for including those characteristics, and importantly, are those data elements used solely as "attribution" (or additional description of the entity involved) or are they used for "distinction" (to help in unique identification).

More to the point: what are the core data elements necessary for determining the uniqueness of a record? We often take for granted the fact that our relational models presume one, and only one record per entity, and that there might be business impacts should more than one entry exist for each individual. Yet individual "entities" may exist in multiple data sets, even in different contexts. Some characteristics are part and parcel of each entity, while others describe the entity playing a particular role.

5.6 The Challenge of Identifying Information

We have already started to look at the process of determining which characteristics are used to uniquely differentiate between any pair of records within a data set. The same question is relevant when attempting to match a pair of records as well, once they are determined to represent the same entity. We can call these "identifying attributes," and the values contained therein can be referred to as "identifying information."

Let's look at an example for customer data integration: What data element values do we compare when trying to link two records together? Let's start with the obvious ones, namely first and last names. Of course, we all know that there are certain names that are relatively common (think: "John Smith"). But even if you have an uncommon name, you might be surprised. For example, if you type in my name ("David Loshin") at Google, you will find entries for me, but you will also find entries for a dentist in Seattle and a professor.

Apparently, first and last names are not enough identifying information for distinction. Perhaps there is another attribute we can use? You probably know that I have written some books, (see http:\\dataqualitybook.com), so maybe that is an additional attribute

to be used. But if you go to Amazon and do a search for "David Loshin," you will find me, but it turns out the professor has also written a book.

Even an uncommon name such as mine still finds multiple hits, and while attempting to add more identifying information can reduce that number of hits, a poorly selected set of attributes may still not provide the right amount of distinction. It may take a number of iterations to review a proposed set of identifying attributes, determine their completeness, density, and accuracy before settling on a core set of identifying characteristics to be used for comparison.

One more thing to think about, though. Once you get to the point where you are pretty confident that those attributes are enough for differentiation, there is one last monkey wrench in the works: Even if you had the absolute set of identifying attributes, there is no guarantee that the values themselves are exact matches!

5.7 Approximate Matching

Actually, my first name is not David – that is really my middle name, but it is the given name my parents used when talking to me. This has actually led to a lot of confusion over the years, especially when confronted with a form asking for me "first name" and my "last name." For official forms (like my driver's license) I use my real first name as my "first name," but for non-official forms I often just use David. The result is that there is inconsistency in my own representation in records across different data systems.

If we were to rely solely on an exact data element-to-data element match of values to determine record duplication, the variation in use of my first or middle name would prevent

> **"[Approximate Matching] allows for two values to be compared with a numeric score that indicates the degree to which the values are similar."**

two records from linking. In turn, you can extrapolate and see that any variations across systems of what *should* be the same values will prevent an exact match, leading to inadvertent duplication.

Fortunately, we can again rely on data quality techniques. We have our stand-bys of parsing and standardization, which can be enhanced through the use of transformation rules to map abbreviations, acronyms, and common misspellings to their standard representations – an example might be mapping "INC" and "INC." and "Inc" and "inc" and "inc." and "incorp" and "incorp." and "incorporated" all to a standard form of "Inc."

We can add to this another tool: approximate matching. This matching technique allows for two values to be compared with a numeric score that indicates the degree to which the values are similar. An example might compare my last name "Loshin" with the word "lotion" and suggest that while the two values are not strict alphabetic matches, they do match *phonetically*. There are a number of techniques used for approximate matching of values, such as comparing the set of characters, the number of transposed, inserted, or omitted letters, different kinds of forward and backward phonetic scoring, as well as other more complex algorithms.

In turn, we can apply this approximate matching to the entire set of corresponding identifying attributes and weigh each score based on the differentiation factor associated with each attribute. For example, a combination of first name and last name might provide greater differentiation than a birth date, since there is a relatively limited number of dates on which an individual can be born (maximum 366 per year).

By applying a weighted approximate match to pairs of records, we can finesse the occurrence of variations in the data element values that might prevent direct matching from working.

5.8 Modeling Issues and Entity Inheritance

In our last section we looked at matching and record linkage and how approximate matching could be used to improve the organization's view of "customer centricity." Data quality tools such as parsing, standardization, and business-rule based record linkage and similarity scoring can help in assessing the similarity between two records. The result of the similarity analysis is a score that can be used to advise about the likelihood of two records referring to the same real-life individual or organization.

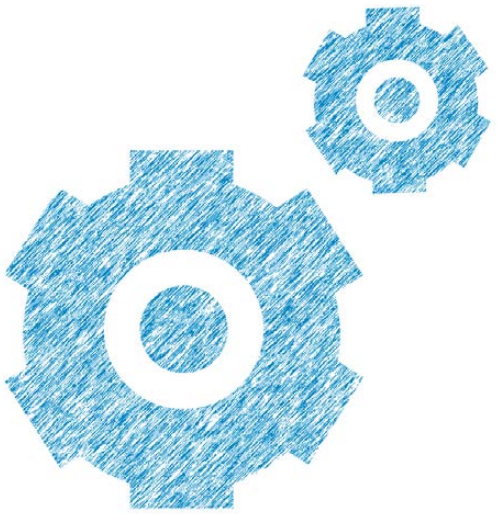
One last thought: this approach is largely a "data-centric" activity. What I mean is that it looks at and compares two records regardless of where those records came from. They might have come from the same data set (as part of a duplicate analysis), or from different data sets (for consolidation or general linkage). But it does not take into consideration whether one data set models "customer" data and another models "employee" data. While you may link a customer record with an employee record based on a similarity analysis of a set of corresponding data attributes, the contexts are slightly different.

A match across the two data sets is a bit of a hybrid: we have matched the *individual* but one playing different roles. That introduces a different kind of question: Are the identifying attributes associated with the "customer" or the individual acting in the role of "customer"? The same question applies for individual vs. employee. And finally, are there attributes of the roles that each individual plays that can be used for unique identification within the role context? The answers to these questions become important when matching and linkage are integrated as part and parcel of a business application (such as the consolidation of data being imported into a business intelligence framework).

CHAPTER

06 Data Quality Control

6.1 Reactivity vs. Proactivity



So far we have looked at technical approaches to data quality and the use of data quality tools to parse, standardize, and cleanse data. Finally, it is time to look at harnessing the power of these tools and techniques to support a data quality management program. Most organizations are relatively immature when it comes to addressing data quality issues. Some typical behaviors in an immature organization include:

- Few or no well-defined processes for evaluating the severity or root causes of data issues;
- Little or no coordination among those investigating data errors;
- Evaluating the same issues multiple times; and
- Correcting the same errors multiple times.

These are all manifestations of a more insidious problem: knee-jerk reactivity, which presumes that addressing the symptoms solves the problem. But in reality, applying these bandages to open wounds is merely a temporary fix. This suggests that

incremental maturation of data quality processes involves transitioning from a reactive environment to one that operates within the context of a series of policies and controls.

The manifestations of immaturity listed here are some fertile areas for improvement, namely:

- Defining processes for evaluating data errors when they are identified;
- Instituting methods for coordinating those evaluations; and
- Applying corrections once, and only once.

As a byproduct of coordinating evaluation, your team will be less inclined to evaluate the same issues multiple times! In this section, we will look at ideas for each of these suggestions.

6.2 Standardizing your Approach to Monitoring the Quality of Data

We have considered three techniques for maturing your organizational approach to data quality management. The first recommendation was defining processes for evaluating errors when they are identified. These types of processes actually involve a few key techniques:

1) An approach to specifying data validity rules that can be used to determine whether a data instance or record has an error. This is more of a discipline that can be guided by formal representations of business or data rules. Often, metadata management tools and data profiling tools have repositories for capturing defined rules, leading to our next technique…

2) A method for applying those rules to data. This often will take advantage of the operational aspects of a data profiling or monitoring tool to validate a data instance against a set of rules. It may also incorporate parsing and standardization rules to identify known error patterns.

3) A means for reporting errors to a data analyst or steward. Some data analysis and profiling can be configured to automatically notify a data steward when a data validity rule is violated. In other situations, the results of applying the validation rule

can be accumulated in a repository, and a front-end reporting tool is used to provide visualization and notification of errors.

4) An inventory of actions to take when specific errors occur. As your team becomes more knowledgeable about the types of errors that can occur, you will also become accustomed to the methods employed for analysis and remediation.

In time, the repeated use of tools and the corresponding actions for remediation can be evolved into standardized methods, which can be documented, published, and used as the basis for training data quality analysts.

6.3 Data Quality Incident Management

The previous step in our transition from uncontrolled reactivity to being proactively engaged in managing data quality involved defining processes for identifying and evaluating data errors using standardized methods. Providing well-defined processes to data stewards and data quality analysts helps reduce any confusion around the appropriate steps to take when those commonly-occurring data failures are discovered in process.

But in many cases there is still an issue of coordination. While standardizing the approaches to monitoring for data validity helps reduce the effort and complexity of analyzing and remediating issues, there is still the situation that the same error may impact multiple data consumers; if each data consumer reports an issue to one of the data stewards, you have many stewards investigating the same problem. So this is where our second suggestion comes in: instituting methods for coordinating those evaluations.

This is an area in which the data management world can learn lessons from our friends in desktop or network support, who rely on incident management systems for the reporting, tracking, and management of issues. Data consumers impacted by a data error can report the flaw in the incident management and tracking system, which can assign a unique identifier to the logged issue and then route it to a specific data steward. However, by carefully structuring the ways that errors are described when reported provides hierarchies and organization in a way that facilitates assignment of issues to those stewards with the greatest corresponding experience. In other words, issues

can be grouped to reduce the amount of replicated effort. In turn, an incident tracking system for data quality issues also provides entry points for tracking the status of the issues – whether the root cause has been identified, if a correction has been performed, or if further evaluation is being performed.

6.4 Achieving "Proactivity" ?

Standardizing the approaches and methods used for reviewing data errors, performing root cause analysis, and designing and applying corrective or remedial measures all help ratchet an organization's data quality maturity up a notch or two. This is particularly effective when fixing the processes that allow data errors to be introduced in the first place totally eliminates the errors altogether.

In the cases where the root cause is not feasibly addressed, we still have another standardized approach: defining data validity rules that can be incorporated into probe points in the processes to monitor compliance with expectations, and alert a data steward as early as possible when invalid data is recognized. This certainly reduces the "reactive culture" we previously discussed, and governing the data stewardship activities by combining automated inspection tools such as data profiling, automated data correction and cleansing tools, and incident management reduces replicated analysis efforts as well as repetitive fixes applied at different places and times. In fact, many organizations consider this level of maturity as being proactive in data quality management because you are anticipating the need to address issues that you already know about.

However, to truly be *proactive,* you'd have to go beyond anticipating what you know. In this light, we might say that instituting controls supporting inspection, monitoring, and notifications is less about being not proactive and more about being reactive much earlier in the process. To really be proactive, perhaps it might be more worthwhile to attempt to anticipate the types of errors that *you don't already know.* Instead of only using profiling tools to look for existing patterns and errors, you might use these analytical tools to understand the methods and channels through which any types of potential errors could occur and attempt to control the introduction of flawed data before it ever leads to any material impact!

<<

What's Next?

In this ebook we have examined some of the core concepts necessary for cleansing data: parsing; standardization; business; data correction; and data enhancement. In turn, we examined how approximate searching and matching can be used to link records together and facilitate the same cleansing and enhancement techniques discussed. And lastly, we looked at some of the drivers and requirements for improving the processes surrounding managing data quality.

Yet there are many more topics that can be discussed in terms of improving data quality for business value. We only touched upon aspects of data quality assessment, and there are many opportunities to review ways to develop and deploy "Data Quality as a Service," especially as cloud computing becomes even more prevalent. Many organizations are looking at drastically increasing their consumption of information with "big data" analytics programs, and at the same time, people are exploring many different ways to reuse and repurpose data for both operational and strategic benefit.

We began the ebook by noting the drastic changes in the attitudes toward data quality and data quality management over the past twenty years. We can only anticipate that more changes lie ahead, and we look forward to sharing more ideas in the future!

About the Author.

David Loshin, president of Knowledge Integrity, Inc. **(www.knowledge-integrity.com)**, is a recognized thought leader, TDWI instructor, and expert consultant in the areas of data management and business intelligence. David is a prolific author on business intelligence best practices, including numerous books and papers on data management such as *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at **www.dataqualitybook.com**. David is a frequent speaker at conferences, Web seminars, and sponsored Web sites and channels such as **www.b-eye-network.com**. His best-selling book, *Master Data Management*, has been endorsed by data management industry leaders, and his MDM insights can be reviewed at **www.mdmbook.com.**

MELISSA DATA®
Your Partner in Data Quality

About Melissa Data Corp.

Melissa Data **(www.MelissaData.com)** provides powerful, yet affordable contact data verification solutions and consulting services for any size organization. Melissa Data's data quality software, plug-ins, data enhancement services, and developer tools verify, standardize, consolidate, enhance and update U.S., Canadian, and global contact data. More than 5,000 companies rely on Melissa Data to gain a complete, accurate, and trusted view of critical information assets.