

Identifying Duplicate Customers

Author: Jim Harris

Blogger-in-Chief, OCDQ Blog



About the Author, Jim Harris



[Jim Harris](#) is the Blogger-in-Chief at Obsessive-Compulsive Data Quality ([OCDQ](#)), an independent blog offering a vendor-neutral perspective on data quality and its related disciplines.

Jim is a recognized thought leader with almost 20 years of enterprise data management industry experience in data quality and its related disciplines.

Jim is an independent consultant, freelance writer, blogger, and regular contributor to [Information Management](#), [The Data Roundtable](#), and [Enterprise CIO Forum](#).

Jim is available for hire, offering a range of writing, speaking and consulting services. For more details visit his services page: <http://www.ocdqblog.com/services>

You can contact Jim at [OCDQ Blog](#), on twitter ([@ocdqblog](#)) or via his [Data Quality Pro profile](#).

About Data Quality Pro



Data Quality Pro is the global community marketplace for data quality professionals, products and project resources.

More than 5,000 members from all corners of the world use Data Quality Pro on a daily basis to:

- Study expert articles, guides and tutorials
- Learn about the latest products and technology innovations
- Connect with other professionals and peers
- Find specialist consultants and professionals for hire
- Discover new job and career opportunities

Membership is entirely free and takes less than 60 seconds to complete.

Find out more: <http://www.dataqualitypro.com>

Would you like to contribute to Data Quality Pro?

If you would like to be published on Data Quality Pro then [please contact us today](#) for more information.



Identifying Duplicate Customers by Jim Harris

Part 1 – The Symbiosis of Methodology and Technology

One of the most common data quality problems is the [identification of duplicate records](#), especially redundant information of the same customer throughout the enterprise.

The need for a solution to this specific problem is one of the primary reasons that companies invest in data quality software and services.

There are many [data quality vendors](#) to choose from and all of them offer viable solutions. Many of these solutions are driven by impressive technology using advanced mathematical techniques such as probabilistic record linkage theory, bi-partite graph matching algorithms, or my personal favorite, the redundant data capacitor, which makes identifying duplicate records possible using 1.21 gigawatts of electricity and a customized DeLorean DMC-12 accelerated to 88 miles per hour.

What is sometimes overlooked is that although technology provides the solution, what is being solved is a business problem.

What is sometimes overlooked is that although technology provides the solution, what is being solved is a business problem. Technology sometimes carries with it a dangerous conceit – that what works in the laboratory and the engineering department will work in the board room and the accounting department, that what is true for the mathematician and the computer scientist will be true for the business analyst and the data steward.

However, what truly determines that a duplicate customer has been identified is not what scientific techniques or mathematical models can justify, but what your business rules define as a duplicate customer.

My point is neither to discourage the purchase of data quality software and services, nor to try to convince you which data quality vendor I think provides the superior solution – especially since these types of opinions are usually biased by the practical limits of your personal experience and motivated by the kind folks who are currently paying your salary or hourly rate.

My goal in this guide of is to focus on [data quality methodology](#) and not data quality technology.

I believe that an effective methodology for implementing your business rules for identifying duplicate customers will help you maximize the time and effort as well as the subsequent return on whatever technology you invest in.

One of the recurring themes in this guide will be that the most significant challenge to solving this specific data quality problem is its highly subjective nature.

Data characteristics and their associated quality challenges are unique from company to company. Business rules can be different from project to project within the same company. Decision makers

on the same project can have widely varying perspectives. All of this points to the need for having an effective methodology.

Unsuccessful data quality projects are most often characterized by the business team meeting independently to define the requirements and the technical team meeting independently to write the specifications.

Typically, the technical team then follows the all too common mantra of "code it, test it, implement it into production, and declare victory" that leaves the business team frustrated with the resulting "solution."

Successful data quality projects are driven by an executive management mandate for business and technical teams to forge an ongoing and iterative collaboration throughout the entire project.

The business team usually owns the data and understands its meaning and use in the day to day operation of the enterprise and must partner with the technical team in defining the necessary data quality standards and processes.

During the business requirements phase of the project, some form of the following question will be asked:

"How do you define a duplicate customer?"

This is a critically important question – however, without an effective methodology, it can also prove to be a frustratingly difficult question. The participants in the requirements gathering process will most often respond with an answer that falls into one of the following two categories:

Category 1: "A duplicate customer is a duplicate customer."

In this category, the answer takes some form of stating that a duplicate customer occurs when the exact same information is repeated on multiple records, either within the same system or across multiple systems.

Sometimes, this answer is passive-aggressively provided by participants who doubt that such a problem could be prevalent in their systems. This "data denial" is not necessarily a matter of blissful ignorance, but is often a natural self-defense mechanism from the data owners on the business side and/or the process owners on the technical side. No one likes to feel blamed for causing or failing to fix the data quality problem. This is one of the many human dynamics that is missing from the relative clean room of the laboratory where the technology was developed. Your methodology must consider the human factor because it will be the people involved in the project, and not the technology itself, that will truly make the project successful.

Other times, this answer is conservatively provided by participants who are concerned that being aggressive in identifying duplicate customers will negatively impact business decisions after duplicates are consolidated (either physically removed or logically linked). This answer is motivated by the fact that there is generally far greater concern about "false positives" than "false negatives" resulting from duplicate identification.

But what are false positives and negatives?

- False positives occur when a group of duplicates are identified that **do NOT represent the same customer**

Unsuccessful data quality projects are most often characterized by the business team meeting independently to define the requirements and the technical team meeting independently to write the specifications.

- False negatives occur when actual redundant representations of the same customer are **NOT identified**

Later, we will look at data examples that illustrate both of these scenarios and why the harsh reality is that they can and will occur regardless of the technology or methodology.

Category 2: "Isn't that what we are paying you to do for us?"

In this category, the answer takes some form of stating that identifying duplicate customers is either what the vendor's software is supposed to do clairvoyantly, or that the vendor's services team should just implement their proven methodology that worked for other clients in similar industries.

In the former case, it may be that the salesperson successfully "blinded them with science" to have such high expectations of the software. I am not trying to accuse salespeople of Machiavellian machinations (even though we have all encountered a few who would shamelessly sell their mother's soul to meet their quota) – as I stated earlier, all data quality vendors have viable solutions driven by impressive technology.

In the latter case, the participants may share my belief that it is the symbiosis of technology and methodology that leads to implementation success. However, the project team must still participate in the definition of the business rules and not simply send the vendor off to "do the voodoo that they do so well."

Both categories of responses (but especially Category 1) help emphasize the importance of my first recommendation – defining the business rules to identify duplicate customers can not be accomplished via a theoretical exercise.

Customer duplication is not a theoretical problem – it is a real business problem that negatively impacts the quality of decision critical enterprise information. Data-driven problems require data-driven solutions. Business rules are best illustrated by data examples. And I mean examples in the true definition of the word – real data from one or more of the project's actual data sources that exemplify the problem and not data metaphors that may meaningfully demonstrate the problem but are nonetheless fictional.

Therefore, it is highly recommended that before the requirements gathering phase, some preliminary analysis is performed on a representative sample of data from one or more of the project's actual data sources. This preparation of effective data examples will enable a far more productive discussion of the business rules.

...it is highly recommended that before the requirements gathering phase, some preliminary analysis is performed on a representative sample of data...

NOTES:



Part 2 – The Importance of Data Analysis

In Part 1 we explored why a symbiosis of technology and methodology is necessary when approaching this common data quality problem.

In addition, I also recommended performing a preliminary analysis on a representative sample of real project data in order to prepare effective examples for defining your business rules.

In Part 2, I will use data metaphors (i.e. *fictional* examples) to illustrate the importance of real data analysis as well as the highly subjective nature of this problem.

Specifically in this section, I will be focusing on the impact that false negatives can have on business rule definition.

For simplicity, the data metaphors will use the following three customer attributes:

1. **Customer Name** - only personal names
2. **Postal Address** - only United States address formats
3. **Tax ID** - for better fictional values, I used dates related to the customer name

False negatives can be caused when the greater concern about false positives motivates a cautious approach to duplicate identification.

This approach leads many projects to adopt a strategy allowing only exact matches. Therefore, let's begin by looking for duplicates where the exact same information is repeated on multiple records – meaning where all attributes are populated and have the same value.

Please refer to the example on the next page:

False negatives can be caused when the greater concern about false positives motivates a cautious approach to duplicate identification

Key	Customer Name	Postal Address	Tax ID
111	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
112	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
121	Tek Jansen	513 West 54th Street, New York, NY 10019	10262005
122	Tek Jansen	513 West 54th Street, New York, NY 10019	10262005
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
141	Joseph Heller	22 Catch Circle, Washington, DC 20004	19230501
142	Joseph Heller	22 Catch Circle, Washington, DC 20004	19230501
151	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
161	Biff Tannen	1809 Mason Street, Hill Valley, CA 94946	19370326
162	Biff Tannen	1809 Mason Street, Hill Valley, CA 94946	19370326

Would you argue that these are NOT duplicate customers?

Now you have an important choice – either you:

- **Take the blue pill** – stop reading and be content with implementing an exact match strategy and at least *some* of your duplicate customers will be identified
- **Take the red pill** – stay in Wonderland and I will show you just how deep the rabbit hole goes...

You have chosen wisely – so let’s get back to the future of identifying duplicate customers:

Key	Customer Name	Postal Address	Tax ID
111	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
112	Martin Seamus McFly	Twin Pines Mall, Hill Valley, CA 94942	11121955
113	Marty Calvin McFly	Lone Pine Mall, Hill Valley, CA 94941	11121955
114	McFly, Marty	Lone Pine Mall, Hill Valley, CA 94941	
115	McFly	Hill Valley, California	

Exact matching missed the last three records – do you think that all five are duplicates of the same customer?

The abbreviation of first and middle names is a common challenge:

- Does a matching Tax ID guarantee that a variation is a duplicate?

- What about when Tax ID is missing?

An additional challenge that can occur with abbreviated first names:

Key	Customer Name	Postal Address	Tax ID
221	Tomas Kundera	245 Daniel Day-Lewis Drive, Kitch, NY 10022	19681985
222	Thomas Kundera	245 Daniel Day Louis Drive, Kitch, NY 10022	
223	T. Kundera	Daniel Day Lewis Drive, Kitch, NY 10022	19681985
231	Tereza Kundera	245 Daniel Day-Lewis Drive, Kitch, NY 10022	20061988
232	Teresa Kundera	245 Daniel Day Louis Drive, Kitch, NY 10022	
233	T. Kundera	Daniel Day Lewis Drive, Kitch, NY 10022	20061988

- Without Tax IDs, could you determine who Keys 223 & 233 should match?
- If both were missing Tax ID, would they then be considered duplicates of each other?

Name and address variations can combine to present additional challenges:

Key	Customer Name	Postal Address	Tax ID
151	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
153	Joseph Haydn	45 Farewell Symphony Stravenue, Austria, CO 80467	
154	Papa Haydn	Austria, CO 80467	17321809
241	Emmett Lathrop Brown, Ph.D.	1640 John F. Kennedy Drive, Hill Valley, CA 94942	11051955
242	Brown Emit Latrop	1640 Riverside Drive, Hill Valley, CA 94942	
243	Doc Brown	Hill Valley, California	11501955

- Is Key 153 an old postal address for the same customer? If postal validation confirms for Key 242 that "Riverside Drive" was renamed "John F. Kennedy Drive" would that make the name variation more acceptable?
- Do Keys 154 & 243 represent a possible nickname or another family member using the same Tax ID?
- Did you notice the transposed numbers in Tax ID on Key 243?

If so, did you give any partial credit? Marriages can be good for people but possibly bad for their data:

Key	Customer Name	Postal Address	Tax ID
251	Lorraine Baines	55 Enchantment Sea Park, Hill Valley, CA 94941	19382015
252	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94942	
253	Lorraine McFly	Hill Valley, California	19382015
261	Elinor Frost	1 Road Not Taken, Derry, NH 03038	18961938
262	Eleanor Rost	1 Road Not Taken, Derry, NH 03038	
263	Mrs. Robert Frost	122 Rockingham Road, Derry, NH 03038	18961938

- Did the hyphenated last name on Key 252 help you overcome the change of address and missing Tax ID?
- How do you know if Keys 261 and/or 262 are truly the same customer as Key 263?

In closing, please carefully consider the following pairs of records:

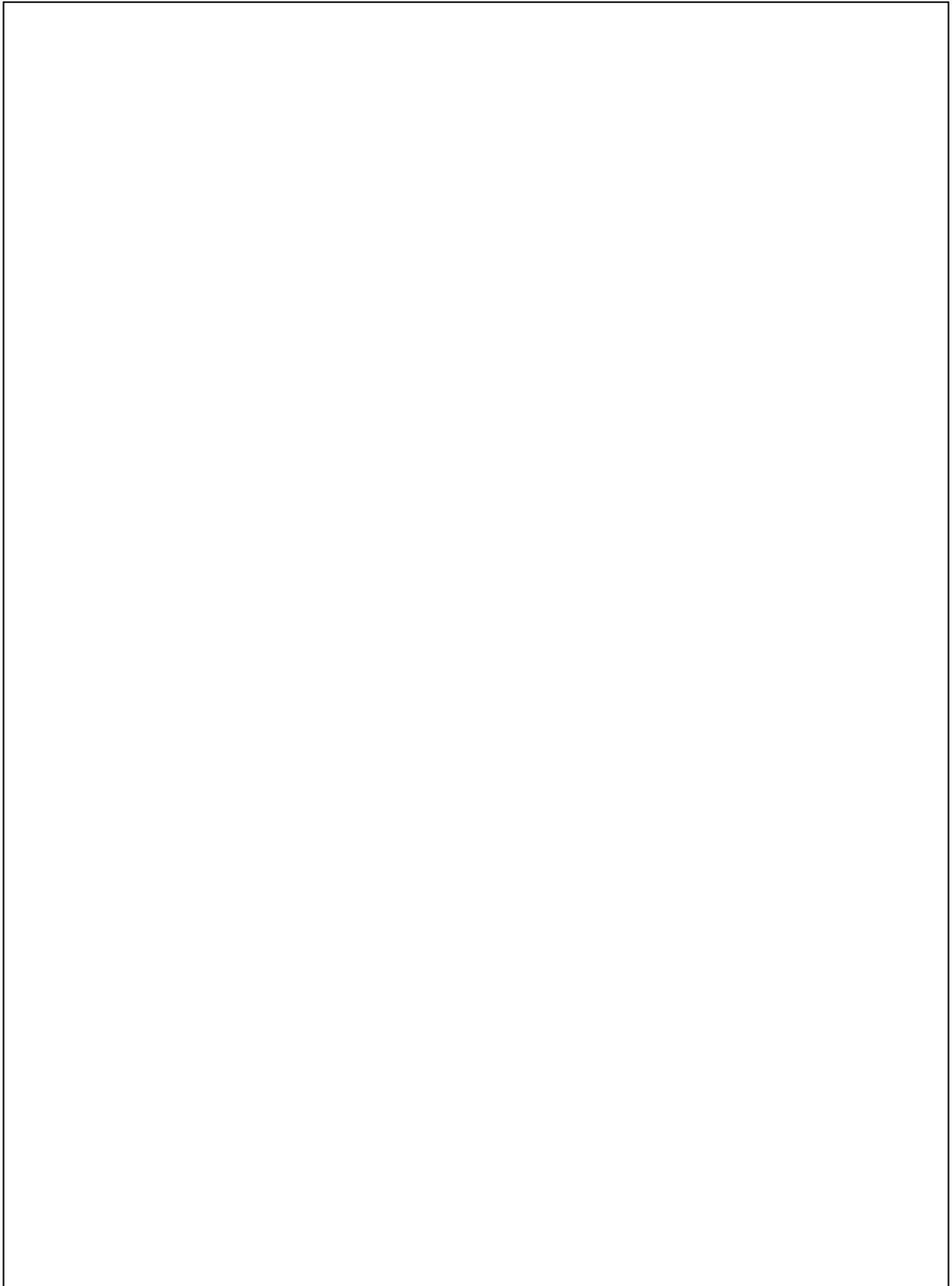
Key	Customer Name	Postal Address	Tax ID
271	William Shakespeare	12 Globe Theatre Road, Stratford-upon-Avon, NH 03576	26041564
272	Christopher Marlowe	12 Globe Theatre Road, Stratford-upon-Avon, NH 03576	26041564
281	Samuel Langhorne Clemens	11 Huckleberry Finn River, Tom Sawyer, MS 38967	30111835
282	Mark Twain	11 Huckleberry Finn River, Tom Sawyer, MS 38967	30111835

One of these pairs is a potential false negative caused by a pseudonym and the other is a potential false positive.

- Even if you know which is which, how do you define a business rule for this scenario?
- What other data metaphors can you think of that would illustrate the challenge of false negatives?

In Part 3 of this guide: We look at data metaphors that illustrate why some of the business rules that you just defined for resolving false negatives could result in creating the false positives that you started out trying to avoid.

NOTES:





Part 3 – False Positives and Business Rule Definition

In Part 2, data metaphors (i.e. *fictional* examples), illustrated the importance of using a detailed, interrogative analysis of real project data in your approach to identifying duplicate customers.

I explained how the greater concern about false positives could motivate you to take a cautious approach that can cause false negatives, especially when you restrict yourself to using only exact matching techniques.

In Part 3, additional data metaphors will illustrate the impact that false positives can have on business rule definition.

Again, for simplicity, the data metaphors will use the following three customer attributes:

1. Customer Name – only personal names.
2. Postal Address – only United States address formats.
3. Tax ID – for better fictional values, I used dates related to the customer name.

Your business rule adjustments for preventing false negatives can result in linking records of related (but not duplicated) customers. Sometimes, these false positives may reveal meaningful data relationships that are useful in other enterprise information initiatives.

Let's begin by looking at some false positives caused by a matching Tax ID and/or postal address:

Key	Customer Name	Postal Address	Tax ID
311	Gilbert A. Sullivan	571 H.M.S. Pinafore Plaza, Mikado, MI 48738	18711896
312	W.S. Gilbert	363 Pirates of Penzance, Patience, PA 19114	18711896
313	Arthur Sullivan	246 Princess Ida Island, Ruddigore, RI 02841	18711896
321	Abbott and Costello	First Baseman Boulevard, Cooperstown, NY 13326	
322	Bud Abbott	First Baseman Boulevard, Cooperstown, NY 13326	
323	Lou Costello	First Baseman Boulevard, Cooperstown, NY 13326	
331	David Starsky	93 Striped Tomato Ford, Bay City, CA 90731	19751979
332	Kenneth Hutchinson	93 Striped Tomato Ford, Bay City, CA 90731	19751979
333	Huggy Bear Brown	93 Striped Tomato Ford, Bay City, CA 90731	19751979

For Keys 312 & 313, do you think the matching Tax ID and similar name indicate possible duplication of Key 311 despite the different postal address?

For Keys 322 & 323, do you think the exact same postal address and similar name indicate possible duplication of Key 321 despite the missing Tax IDs?

What about Keys 331 – 333, where completely different names have the exact same postal address and Tax ID?

A common challenge is the same family name and the exact same postal address:

Key	Customer Name	Postal Address	Tax ID
411	Elizabeth Barrett Browning	43 Portuguese Sonnets, Counting Ways, IL 62650	18061861
412	Robert Browning	43 Portuguese Sonnets, Counting Ways, IL 62650	18061861
413	Robert Wiedemann Barrett Browning	43 Portuguese Sonnets, Counting Ways, IL 62650	18061861
421	Felix Mendelssohn	46 East Elijah Expressway, Oratorio, OR 97289	18090203
422	Abraham Mendelssohn	46 East Elijah Expressway, Oratorio, OR 97289	18090203
423	Moses Mendelssohn	46 East Elijah Expressway, Oratorio, OR 97289	18090203
431	Horatio Alger Jr.	One American Dream Avenue, Revere, MA 02151	
432	Horatio Alger Sr.	One American Dream Avenue, Revere, MA 02151	
433	Horatio Alger	One American Dream Avenue, Revere, MA 02151	
441	Patrick Thames	1831 London Bridge, Lake Havasu City, AZ 86403	
442	Patricia Thames	1831 London Bridge, Lake Havasu City, AZ 86403	
443	Pat Thames	1831 London Bridge, Lake Havasu City, AZ 86403	

- Do you think Key 413 a duplicate of Key 412 or a son named after both of his parents?
- Do you think Keys 421 – 423 are duplicates caused by multiple pseudonyms or a son, father and grandfather living in the same house?
- Without Tax IDs, Keys 431 & 432 can only be differentiated by generation (i.e. "Jr." indicating Junior and "Sr." indicating Senior), however do you think Key 433 a duplicate for either of them?
- Keys 441 & 442 might only be differentiated by gender, but what about Key 443?

An additional complexity that can occur with families at the exact same postal address:

Key	Customer Name	Postal Address	Tax ID
511	Peter and Lois Griffin	31 Spooner Street, Quahog, RI 02903	19990131
512	Lois and Stewie Griffin	31 Spooner Street, Quahog, RI 02903	20020214
513	Stewie and Brian Griffin	31 Spooner Street, Quahog, RI 02903	20050501

You may find it useful to first split the compound customer names into separate records:

Key	Customer Name	Postal Address	Tax ID
511-a	Peter Griffin	31 Spooner Street, Quahog, RI 02903	19990131
511-b	Lois Griffin	31 Spooner Street, Quahog, RI 02903	19990131
512-a	Lois Griffin	31 Spooner Street, Quahog, RI 02903	20020214
512-b	Stewie Griffin	31 Spooner Street, Quahog, RI 02903	20020214
513-a	Stewie Griffin	31 Spooner Street, Quahog, RI 02903	20050501
513-b	Brian Griffin	31 Spooner Street, Quahog, RI 02903	20050501

Performing this split reveals two potential pairs (511-b/512-a & 512-b/513-a) with the exact same name and postal address but completely different Tax IDs – are these duplicates?

How many customers do you think are represented by Keys 511 – 513?

Keys 411 – 513 are also examples of a non-duplicate data relationship commonly referred to as a *family household*, where multiple distinct customers are linked for having the same family name and the same postal address. This relationship is useful in marketing programs that target family units (e.g. vacation packages, mobile phone plans) or that target the head of a household (i.e. customers making purchasing decisions).

A common family name and street name can combine to present an additional challenge:

Key	Customer Name	Postal Address	Tax ID
611	Agent Smith	1999 Main Street, Touthville City, Irgendein Land	
612	Dudley Smith	1987 Main Street, Touthville City, Irgendein Land	
613	Jefferson Smith	1939 Main Street, Touthville City, Irgendein Land	
614	Hannibal Smith	1983 Main Street, Touthville City, Irgendein Land	
615	Winston Smith	1984 Main Street, Touthville City, Irgendein Land	
616	Sarah Jane Smith	1973 Main Street, Touthville City, Irgendein Land	
617	Doctor Zachary Smith	1965 Main Street, Touthville City, Irgendein Land	

Do you think that any are duplicates of the same customer(s) or relate the same family household(s)?

In closing, please carefully consider the following groups of records:

Key	Customer Name	Postal Address	Tax ID
711	Shawn Spencer	Pineapple Place Apartments, Santa Barbara, CA 93121	
712	Burton Guster	Pineapple Place Apartments, Santa Barbara, CA 93121	
713	Carlton Lassiter	Pineapple Place Apartments, Santa Barbara, CA 93121	
714	Juliet O'Hara	Pineapple Place Apartments, Santa Barbara, CA 93121	
721	F. Scott Fitzgerald	Literary Luxury Lofts, Littera, LA 70116	
722	James Joyce	Literary Luxury Lofts, Littera, LA 70116	
723	Jay Gatsby	Literary Luxury Lofts, Littera, LA 70116	
724	Stephen Dedalus	Literary Luxury Lofts, Littera, LA 70116	
731	Dr. Leonard Leakey Hofstadter, Ph.D.	California Institute of Technology, Pasadena, CA 91125	
732	Dr. Sheldon Cooper, Ph.D.	California Institute of Technology, Pasadena, CA 91125	
733	Mr. Howard Wolowitz, M.Eng.	CALTECH, Pasadena, CA 91125	
734	Rajnish Koothrappali, Ph.D.	California Institute of Technology	
741	Jack Carter	Global Dynamics, Eureka, OR 97086	
742	Allison Blake	Global Dynamics Headquarters, Eureka, OR 97086	
743	Henry Deacon	Global Dynamics Research Division, Eureka, OR 97086	
744	Douglas Fargo	Global Dynamics Cryogenics Division, Eureka, OR 97086	

Do you think that any are duplicates of the same customer(s)?

Keys 711 – 744 (as well as Keys 321 – 513) are also examples of a non-duplicate data relationship commonly referred to as a *geographic household*, where multiple distinct customers are linked for having the same postal address. This data relationship is useful in mass mailing programs that benefit from the cost savings of eliminating redundant deliveries to the same postal address.

- What other data metaphors can you think of that would illustrate the challenge of false positives?
- What other meaningful data relationships can you think of that may be revealed by false positives?

In Part 4 we will discuss recommendations for documenting your business rules as well as setting realistic expectations about the first iteration of application development and guidelines for the necessary collaboration of the business and technical teams throughout the entire project.

NOTES:



Part 4 – Business Rule Documentation and Application Development Expectations

So far in this guide, we have discussed:

- Why a symbiosis of technology and methodology is necessary when approaching the common data quality problem of identifying duplicate customers
- How performing a preliminary analysis on a representative sample of real project data prepares effective examples for discussion.
- Why using a detailed, interrogative analysis of those examples is imperative for defining your business rules
- How both false negatives and false positives illustrate the highly subjective nature of this problem

Now we discuss recommendations for documenting your business rules as well as setting realistic expectations about application development and guidelines for the necessary collaboration of the business and technical teams throughout the entire project.

Business Rule Documentation

The goal of a business requirements document (BRD) is to provide clear definitions of business problem statements that include associated solution criteria. Although your project's BRD will obviously contain other necessary material, here are a few recommendations for documenting your business rules for identifying duplicate customers:

1. **Include data examples** – parts 2 and 3 of this guide illustrated the effectiveness of examples for facilitating discussion. They should also be included in the documentation. Data examples convey business rules far better than either concise (but esoteric) statements, or detailed (but verbose) pages of attempted explanation.
2. **Accentuate the negative** – although it may sound counterintuitive, it is simply easier to explain something when you don't like it. Recall your answers to the questions in parts 2 and 3 of this guide. When you looked at records that you believed should be considered duplicates, did you feel the need to justify your decision with an elaborate explanation? Compare that with your reaction when you looked at records that you believed should NOT be considered duplicates. This effect is known as "negativity bias" where bad evokes a stronger reaction than good in the human mind – just compare an insult and a compliment, which one do you remember more often? Therefore, focus on documenting the rules that identify what is NOT a duplicate customer.
3. **Avoid technology bias** – it is often easier to define your business rules *before* vendor evaluation. Knowing how the vendor's software works can sometimes cause a "framing

effect” where rules are defined in terms of software functionality, framing them as a technical problem instead of a business problem. Remember that all data quality vendors have viable solutions driven by impressive technology. Therefore, focus on stating the problem and solution criteria in business terms.

Application Development Expectations

Too many data quality initiatives fail because of lofty expectations, unmanaged scope creep, and the unrealistic perspective that problems can be permanently “fixed” as opposed to needing eternal vigilance.

Here are a few recommendations for setting realistic expectations for application development:

1. **Plan for multiple iterations** – in order to be successful, application development must always be understood as an iterative process. ROI will be achieved by targeting well defined objectives that can deliver small incremental returns that will build momentum to larger success over time. Projects are easy to get started, even easier to end in failure and often lack the decency of failing quickly. Just like any complex problem, there is no fast and easy solution for data quality.
2. **Prepare for more reviews** – review of preliminary data analysis was used to help discuss and document your business rules. Additional reviews are necessary during application development in order to refine the matching criteria before implementation. Also, most implementations will include logic for identifying scenarios of uncertainty that require manual review.
3. **Focus on the data** – every vendor’s software has some way to rank match results (e.g. numeric probabilities, weighted percentages, confidence levels). Ranking is often used as a primary method in differentiating the three possible result categories: (1) automatic matches, (2) automatic non-matches, and (3) potential matches requiring manual review. Although this functionality is necessary, it can sometimes be a distraction when reviewers become obsessed with ranking to the point that they actually ignore whether or not records have been properly categorized. First and foremost, focus on the data (e.g. are the “automatic matches” truly duplicates?). Modifying matching criteria is where science meets art. Perform trending analysis on the effects caused by changing criteria to guard against doing more harm than good. I have used software from most of the Gartner Data Quality Magic Quadrant and many of the so-called niche vendors. Without exception, I have always been able to obtain the desired results by staying focused on the data.
4. **Perfection is impossible** – it doesn’t matter if your vendor’s match algorithms are deterministic, probabilistic, or even supercalifragilistic. The harsh reality is that false negatives and false positives can be reduced, but never eliminated. A relentless quest to find and fix every one of them is a self-defeating cause. Although this is easy to accept in theory, it is notoriously difficult to accept in practice. For example, let’s imagine that your project is processing one billion records and that exhaustive analysis has determined that the results are correct 99.99999% of the time, meaning that incorrect results occur in only 0.00001% of the total data population. Now, imagine conducting a review by explaining the statistics but providing **only** the 100 exception records. Do not underestimate the difficulty that the human mind has with large numbers. Also, don’t forget about the effect of negativity bias. A focus on exceptions can undermine confidence and prevent acceptance of an overwhelmingly successful (but not completely perfect) implementation.

Team Collaboration

As I explained in Part 1 of this guide, successful projects are driven by an executive management mandate for business and technical (i.e. IT) teams to forge an ongoing collaboration. Here are a few recommendations for fostering that collaboration:

1. **Provide leadership** – not only does the project require an executive sponsor to provide oversight and arbitrate any issues of organization politics, but the business and IT must each designate a team leader for the initiative. Choose these leaders wisely. The best choice is not necessarily those with the most seniority or authority. You must choose leaders who know how to listen well, foster open communication without bias, seek mutual understanding on difficult issues, and truly believe it is the people involved that make projects successful. Your team leaders should also collectively meet with the executive sponsor on a regular basis in order to demonstrate to the entire project team that collaboration is an imperative to be taken seriously.
2. **Formalize the relationship** – consider creating a service level agreement (SLA) where the business views IT as a supplier and IT views the business as a customer. However, there is no need to get the lawyers involved. My point is that this internal strategic partnership should be viewed no differently than an external one. Remember that you are formalizing a relationship based on mutual trust and cooperation.
3. **Share ideas**– foster an environment in which a diversity of viewpoints is freely shared without prejudice. For example, the business often has practical insight on application development tasks, and IT often has a pragmatic view about business processes. Consider including everyone as optional invitees to meetings. You may be pleasantly surprised at how often people not only attend but also make meaningful contributions. Remember that you are all in this together.

In Part 5 of this guide: We will discuss topics related to duplicate consolidation, including physical removal vs. logical linkage, techniques for creating a "best of breed" representative record for duplicates, and consolidation vs. cross population where the representative record is used to update duplicates with a consistent representation of the highest quality data available.

NOTES:



Part 5 – Record Consolidation and Creating “Best of Breed” Records

So far in this guide, we have discussed:

- Why a symbiosis of technology and methodology is necessary when approaching the common data quality problem of identifying duplicate customers
- How performing a preliminary analysis on a representative sample of real project data prepares effective examples for discussion
- Why using a detailed, interrogative analysis of those examples is imperative for defining your business rules
- How false negatives and false positives illustrate the highly subjective nature of this problem
- How to document your business rules for identifying duplicate customers
- How to set realistic expectations about application development
- How to foster a collaboration of the business and technical teams during the entire project

In this part, the fifth and final part in the guide, we will discuss topics related to duplicate consolidation, including techniques for creating a “best of breed” representative record for duplicates, physical removal vs. logical linkage, and consolidation vs. cross population.

E Pluribus Unum – Out of Many, One

We will use the following data metaphors from Part 2 of this guide:

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	
134	T.S. Eliot	Wasteland, Texas	26091888
211	J.D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
212	Jerome D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	
213	J. David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
214	Jerome Salinger	Holden Caulfield Highway, Agerstown, PA 19102	
215	David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
151	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
153	Joseph Haydn	45 Farewell Symphony Stravenue, Austria, CO 80467	
154	Papa Haydn	Austria, CO 80467	17321809
251	Lorraine Baines	55 Enchantment Sea Park, Hill Valley, CA 94941	19382015
252	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	
253	Lorraine McFly	Hill Valley, California	19382015

Consolidation evaluates groups of identified duplicate records and creates one representative record for each group.

Creating a "Best of Breed" Representative Record

Typically, consolidation creates the representative (i.e. "best of breed") record using one of two techniques:

1. **Record Level Consolidation** – choosing one complete record from within the group
2. **Field Level Consolidation** – constructing fields from potentially different records from within the group

The business rules for performing consolidation can vary as much as the business rules for identifying duplicate customers. The selection criteria are usually fairly straightforward, however complexity can be caused when multiple criteria with nested levels of tie-breakers are needed to choose or construct the "best of breed" data for the group.

Although not a comprehensive list, here are some of the most common selection criteria:

Completeness	For record level consolidation, this usually means selecting the record with the highest number of populated fields. For field level consolidation, this usually means selecting the fields with the longest values.
Frequency	More common in field level consolidation where the most frequently occurring value is selected for a given field. However, it can be used in record level consolidation to select the record that has the most frequently occurring combination of values across fields. Either way, the assumption is that the most frequently occurring value indicates preferred information.
Recency	For record level consolidation, this usually means selecting the record most recently updated. For field level consolidation, this usually means selecting the field value from the record most recently updated. Either way, the assumption is that the most recent update contains reliable information.
Source	More common in record level consolidation to select the record that originated in a preferred source system. However, it can be used in field level consolidation to select the value for a given field from the record that originated in a preferred source system.

Examples of Record Level Consolidation:

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	
134	T.S. Eliot	Wasteland, Texas	26091888
151	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
153	Joseph Haydn	45 Farewell Symphony Stravenue, Austria, CO 80467	
154	Papa Haydn	Austria, CO 80467	17321809

Choosing a record based on the most frequently occurring combination of values across fields:

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
151	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809

Examples of Field Level Consolidation:

Key	Customer Name	Postal Address	Tax ID
211	J.D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
212	Jerome D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	
213	J. David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
214	Jerome Salinger	Holden Caulfield Highway, Agerstown, PA 19102	
215	David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
251	Lorraine Baines	55 Enchantment Sea Park, Hill Valley, CA 94941	19382015
252	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	
253	Lorraine McFly	Hill Valley, California	19382015

Constructing a record based on selecting the fields with the most complete (longest) values:

Key	Customer Name	Postal Address	Tax ID
211	Jerome David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
252	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	19382015

"Frankenstein Consolidation"

Field level consolidation is sometimes referred to as "Frankenstein consolidation" since constructing fields from different records within the group can assemble an "unnatural data monster" by creating an invalid combination of field values. This concern typically makes record level consolidation the far more common consolidation technique.

Physical Removal vs. Logical Linkage

Typically, consolidation is implemented using one of two techniques:

1. **Physical Removal** – where the group of identified duplicate records is replaced with the representative record, meaning either the duplicate records are actually deleted from the source system or simply excluded from the target system. Physical removal is most commonly associated with record level consolidation.
2. **Logical Linkage** – where the group of identified duplicate records are updated with a reference identifier field whose value points to the identifier field value of the representative record, which is differentiated by either not having a reference identifier field value or with an additional indicator field.

Consolidation vs. Cross Population

An alternative strategy in consolidation is cross population, where the representative record is used to update the identified duplicate records with the "best of breed" data. Cross population is most commonly associated with field level consolidation.

Typically, cross population is implemented using one of two techniques:

1. **Fill in the blanks** – values from the representative record are used to update only the unpopulated fields in the records in the group.
2. **Create consistent values** – the representative record is used to update all fields in all records in the group to create a single consistent representation of the highest quality data available.

Applying "fill in the blanks" cross population using the above representative records:

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	T.S. Eliot	1917 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
134	T.S. Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
211	J.D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
212	Jerome D. Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
213	J. David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
214	Jerome Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
215	David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
151	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
153	Joseph Haydn	45 Farewell Symphony Stravenue, Austria, CO 80467	17321809
154	Papa Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
251	Lorraine Baines	55 Enchantment Sea Park, Hill Valley, CA 94941	19382015
252	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	19382015
253	Lorraine McFly	85 George Douglas Heights, Hill Valley, CA 94941	19382015

Applying "create consistent values" cross population using the above representative records:

Key	Customer Name	Postal Address	Tax ID
131	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
132	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
133	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
134	Thomas Stearns Eliot	1915 J. Alfred Prufrock Lane, Wasteland, TX 79526	26091888
211	Jerome David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
212	Jerome David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
213	Jerome David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
214	Jerome David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
215	Jerome David Salinger	Holden Caulfield Highway, Agerstown, PA 19102	19191951
151	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
152	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
153	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
154	Franz Joseph Haydn	94 Surprise Symphony Stravenue, Austria, CO 80467	17321809
251	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	19382015
252	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	19382015
253	Lorraine Baines-McFly	85 George Douglas Heights, Hill Valley, CA 94941	19382015

Discover More Guides on Data Quality Pro

There are a wealth of additional guides and tutorials to help you with this topic on Data Quality Pro.

Visit Data Quality Pro for more details: <http://dataqualitypro.com>



Discover More on OCDQ Blog

Jim Harris has written extensively on this topic over on the popular OCDQ Blog:

<http://www.ocdqblog.com>

You can contact Jim at [OCDQ Blog](#), on twitter ([@ocdqblog](#)) or at his [Data Quality Pro profile](#).

Hire Jim Harris for Writing, Speaking or Consulting Assignments

Did you find this guide useful and informative?

Why not hire Jim to write or speak on similar topics for your business or carry out a consulting assignment within your organization?

Please review [Jim Harris' services page](#) for more details.

Please Share This Guide on Twitter

<http://clicktotweet.com/oW9eS>

