dun & bradstreet

Straight Talk on Matching

Why 100% Resolution Is Unrealistic – and May Be Counterproductive

dun & bradstreet

Executive Summary

There are many use cases that drive the need to properly resolve the identity of businesses, from risk mitigation and regulatory compliance to marketing and sales operations. Regardless of the use case, it's natural for users to expect their identity resolution solution to return a result for each record supplied.

However, for reasons that this whitepaper will explain, 100% identity resolution is unrealistic in most cases. Complete identity resolution – or as some would call it "matching" – across a full dataset is challenging due to factors that fall into three broad categories: the quality of data (whether it's input data or reference data); process factors including constraints of the matching algorithm and configuration specifications; and whether the underlying criteria support a match at all.

While not exhaustive, this framework for understanding unsuccessful attempts at identity resolution should allow most consumers of such technology to appreciate the challenges associated with matching, and understand the difficulty – and sometimes impossibility – of reaching the 100% mark. Finally, we propose a more productive lens for measuring success, by recasting the challenge in terms of effort versus value.

Some of the concepts of identity resolution may be demonstrated by considering a bookstore focused on young adult literature. The store has a highly curated selection of literature of specific interest to the young adult reader.

One day, a customer enters the store and asks the longtime owner where he can find *Harry Potter and the Sorcerer's Stone*.

The owner, in this case, acts as the identity resolution service, the resolution request is the title *Harry Potter and the Sorcerer's Stone*, and the output will be the book itself. One assumes that the owner is well-suited to answer, since she has owned the bookstore for a long time. Assuming further that the book is in the store, she ought to be able to quickly find it for the customer.

However, if the bookstore isn't limited to young adult literature, then it may have thousands of books rather than hundreds, and it becomes a more difficult task for the owner to find the book: Perhaps she doesn't know all the available titles, because there are too many of them; perhaps there is another book with a similar-sounding title; perhaps she doesn't know in what category the book can be found, which would aid her search.

In such cases, a request for *Harry Potter and the Sorcerer's Stone* by J.K. Rowling can be expected to provide a more precise answer – the owner now knows she is not looking for a book by any other author.

Of course, the strengths of the matching process must be considered. If the request is for a movie, the owner may be unable to assist. If there are multiple similar titles, additional refinement may be required. And if the owner is only willing to assist if customers get the title exactly right, the success rate will clearly suffer.

OVERVIEW OF IDENTITY RESOLUTION

Identity resolution is the process of matching a collection of data points about something – whether it's a person, a company, a product, or something else – against a trusted set of references, to identify the object that those data points truly represent, in order to establish a single cohesive view to use for marketing or other purposes. This can be a simple or complex process, dependent on various factors, including how wide the universe of potential entities is, how complete the information in the request is, and how capable the process – whether done by a person or a machine – is at interpreting and fulfilling a request.

The number of candidates in some universe of potential matches is perhaps the single most important factor in the degree of success one may have in trying to resolve an identity. While a large, complete universe is more likely to contain the right candidate, as the number of candidates increases, so too does the difficulty of determining the entity in that set that most appropriately matches a request.

One way to help reduce the impact of a large candidate pool is to make the request more complete. This can mute the effects of the high number of potential entities, since the likelihood of any particular entity meeting multiple criteria within the request is lower.

It is also important to keep in mind that there exists a significant difference between identity resolution – "matching" – and "searching."

HOW IDENTITY RESOLUTION DIFFERS FROM SEARCH

The concept of making a request more complete in order to specifically identify the correct entity speaks to the key differentiation between candidate retrieval – commonly known as "searching" – and identity resolution, which is often called "matching."

When performing a search, the objective is to return as many candidates as possible that meet the criteria of the request in some way; casting a wider net means that a greater number of results will be included. Because a search aims to retrieve all candidates that have the *potential* to be relevant, many of the results that are returned will be weak with, at best, tenuous ties to the request data.

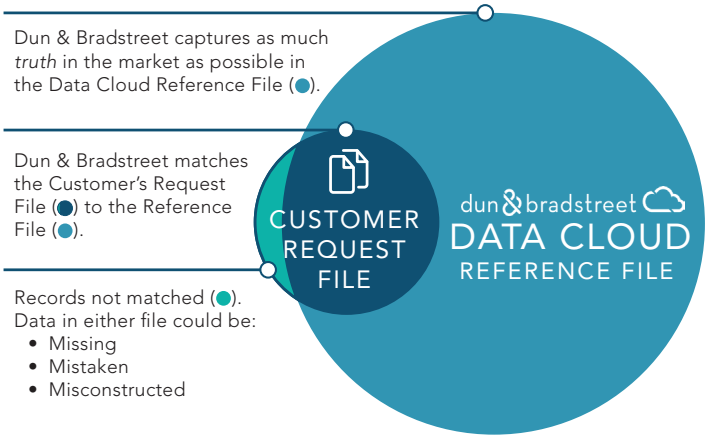

Figure 1. How Reference and Input Data Relate

Identity resolution, on the other hand, is focused on attempting to return a single right candidate that is, to an appropriate degree of certainty, the candidate that the requestor wants or needs to work with.

Because search and identity resolution both accept data as input and both return data as output, there is often confusion between the two, or a conflation of one with the other. However, the two remain related but separate functions, each with its own purpose and strengths.

PRECISION VS RECALL

Two concepts that may be helpful in understanding the difference between identity resolution and search are the related ideas of *precision* and *recall*.

Recall is a match algorithm's ability to return all relevant entities for a given search; in measuring recall, the number of irrelevant entities isn't considered. This means that if a request asks to identify a single specific entity, then any service that returns the right result will achieve 100% recall, even if that single, specific, correct result is hidden among millions of other, largely irrelevant results.

Precision, on the other hand, is the algorithm's ability to avoid irrelevant entities in the results returned to a requestor. Whether all of the *relevant* entities have been identified or not isn't considered when measuring precision, which means that if a match engine is asked to find companies based in Ohio, returning a single Ohio company and no others results in 100% precision, even though millions of other Ohio companies exist.

Search engines attempt to balance these two competing metrics to best fit their customers' needs. Some value recall over precision; others value precision over recall.

Identity resolution, on the other hand, attempts to achieve 100% precision and 100% recall on every request – generally a much more challenging task, given how data quality issues and other obstacles exist.

DATA QUALITY CONSIDERATIONS IN IDENTITY RESOLUTION

The first consideration to bear in mind as it relates to identity resolution is that of data quality. The quality of the data being used to perform the resolution is critical, both within the request and within the reference set.

But what is data quality? Data quality may be thought of as a measure of the consistency of structure, the completeness, and the accuracy of data that is used in evaluating choices or making decisions.

This paper will discuss data quality considerations for both the request and the reference dataset, beginning with the request.

REQUEST DATA QUALITY CONSIDERATIONS

Understanding the provenance of the request data is an important first step: Where did it come from? What goal do we hope to reach by matching the request data? The answers to these questions help to ensure that the process of matching the data will produce the intended results. If the request data is about a completely different universe than the match system is designed to work with, it will be difficult or impossible to get meaningful results.

For example, it would be a waste of time for a customer to walk into the small bookstore and ask to see their latest selection of lawnmowers. The bookstore isn't designed to help with that type of request. On the other hand, if a customer walks into a small bookstore and asks the owner to find the book *Harry Potter and the Sorcerer's Stone*, then, assuming that the owner stocks a book by that title, this request ought to be simple to fulfill.

However, if the customer's information is wrong, then it's reasonable to expect that the owner could hand them the wrong book. The book the customer gets in return might be a different book in the *Harry Potter* series. It might be a book unrelated to the character, but written by someone coincidentally named Harry Potter. The owner might not be able to hand the customer any book at all. Bad request data typically falls into one of three categories: missing, mistaken, or misconstructed.

Missing Data

Data that is simply not available in the request is referred to as "missing data." Missing data can occur for a number of reasons – it may not exist (for example, a person without a middle name), it may never have been known, or it may never have been recorded.

Mistaken Data

Data that doesn't reflect reality properly is known as "mistaken data." There are degrees of mistakes that can happen – from a value being completely wrong, such as writing down *Harry Potter and the Prisoner of Azkaban* when one ought to have written *Harry Potter and the Sorcerer's Stone*, to simply misspelling a value like *Hairry* instead of *Harry*.

Misconstructed Data

Data that is otherwise correct but is being used in an improper way or is called by an improper name (that is, in the wrong input field) is considered "misconstructed data." While the value is present, the information is wrong because the context in which the value is evaluated is wrong.

Consider the prior example. If the customer had entered the bookstore and asked for a book named *Harry Potter and the Sorcerer's Stone* by an author named Scholastic Book, it's possible that the owner would simply shrug. The customer would leave, unable to find the book, even though all the necessary information was there, just in the wrong context (Scholastic Book being the publisher, and not the author, of the title in question)

Missing, mistaken, and misconstructed data all affect the ability of an identity resolution algorithm to properly perform its task: Missing data may prevent its ability to isolate the proper candidate from among others; mistaken and misconstructed data may do the same, or alternatively cause it to identify the wrong candidate.

HOW DUN & BRADSTREET HANDLES REQUEST DATA QUALITY ISSUES

The evolution of Dun & Bradstreet's identity resolution service has included many innovations to deal with missing, mistaken, and misconstructed data. As a result, even when request data suffers from one or more of these maladies, we can still identify the proper entity.

Dun & Bradstreet attempts to work around missing data by considering as many combinations of the data points that are present in a request as possible. By doing so, it is often possible to identify the correct candidate in a way that doesn't require the missing data. For example, Dun & Bradstreet identity resolution service can accept a request to search for "ABC Company" in Ohio, even if the phone number and street address are missing, and if those data elements are reflected inside the Dun & Bradstreet Data Cloud, can successfully return the proper candidate to the requestor.

In a similar fashion, Dun & Bradstreet works around mistaken or misconstructed data by considering alternate combinations of data. When provided what might be considered a "business name," we check that value not only against business names, but against trade styles, executive names, and more. In this way, we're able to minimize, but not completely eliminate, the impact of mistaken or misconstructed data.

REFERENCE DATA QUALITY CONSIDERATIONS

In the prior example, a customer enters a small bookstore and asks the owner to find a book titled *Harry Potter and the Sorcerer's Stone*. Again, assuming that a book with that title is in stock, this request ought to be simple to fulfill.

This paper has already explored the possible outcomes that result from the customer's information being "wrong," whether that means missing, mistaken, or misconstructed; however, it's also possible that the bookstore owner's information – what can be called reference data – is wrong. In this case, it's again reasonable to expect that the owner could provide the wrong book, or none at all, to the customer. Bad reference data, like request data, typically falls into one of three categories: missing, mistaken, or misconstructed. Additionally, the cadence with which reference data is updated plays an important role in the overall utility and reliability of that data.

Missing Data

Missing reference data can occur for various reasons; the most typical cause of missing reference data is that the entity is too new to the world for the reference data to have been properly developed. In the example in this paper, assume an employee has recently put a new book on the shelf. If the owner hasn't yet had time to learn about the book – or perhaps hasn't yet noticed it – then it would be impossible for her to identify that book for the customer, even if it is in fact the title – *Harry Potter and the Sorcerer's Stone* – being requested.

Data may also be considered missing if there exists in its place a value that is meaningless – or so common as to be rendered meaningless. Consider a record where the book title is recorded as *NULL* or *Unknown Title*. In most situations, especially if there is a high presence of such values, these values don't really represent the title of a book; rather, they are placeholders but suggest that the true value is missing.

Mistaken Data

Reference data occasionally can be mistaken – that is, incorrect. The two primary causes of mistaken data are a recent change that has not yet been reflected in the data, and an improper interpretation or inaccurate recording of the information at hand.

Consider again the bookstore owner. She sees her employee putting a new book on the shelf, and she asks what its title is. He replies *Harry Potter and the Sorcerer's Stone* but she hears *Barry Slaughter and the Force of the Phone*. When later asked by the customer to find *Harry Potter and the Sorcerer's Stone*, she is unable to because the information she has – which is the best information for her to use to make the identification – is wrong with regard to the book's title.

It's important to note that there is a difference between incorrectness and legal accuracy. For example, it's possible for a legal document to reflect an incorrect name. In

such a case, while the data may be incorrect – perhaps the title is spelled *Hairry Potter,* with an "i" in the title character's first name – a reference data system would not necessarily be wrong to reflect that spelling. Even though readers might believe that the name ought to be spelled *Harry* and not *Hairry*, the reference data system is considered accurate since its source – the publisher's "official" information about the book – reflects the same apparent misspelling.

Curators of reference data are required to strike a balance between consistency with their trusted source, and accuracy with regard to the inherent truth. Understanding how a reference data curator deals with that friction is important in analyzing the impact of mistaken data on identity resolution success rates.

Misconstructed Data

Data may be present within a reference data provider's assets but not properly accessible due to improper construction of the data store.

Established and trusted reference data providers like Dun & Bradstreet have multiple safeguards and automated processes to help protect against these types of errors, but it is possible for issues to occur.

Occasionally, these types of issues arise because automated systems are unable to properly parse a given data value into two or more data elements as part of a standardization procedure. For example, a system might be asked to parse a free-text field into book titles, authors, and publishers. Without clear and consistent guidance for how to parse such a free-text field, a string such as "Harry Potter and the Sorcerer's Stone J.K. Rowling Scholastic" could be improperly parsed to suggest that "Harry Potter" is the name of the author, *and the Sorcerer's Stone* the title, and "J.K. Rowling Scholastic" the publisher – none of which are correct.

In the prior case of the bookstore owner, if the only information she is given for a book is the phrase "Harry Potter," it's possible she would interpret "Harry Potter" to mean that Harry Potter was the author's name, rather than the book's title. Asked, then, by the customer to find a book with a title similar to *Harry Potter*, the owner might not recognize that such a book exists, even though the data has been present all along.

Such issues can also happen when the extraction process that feeds into a reference data system improperly encodes the data – missing separators, terminators, or other important characters. High-quality match systems test for these situations and resolve them.

Update Cadence

A final aspect of reference data is the update cadence that the identity resolution provider has chosen for it.

The selection of an update cadence is a critical component of the success of identity resolution systems. Too slow a cadence results in stale data and potential unmatched requests; if the cadence is too fast the opportunity for quality checking of incoming data is lost. Additionally, faster updates can increase cost, both for the sourcing of data and for the processing required to integrate it into the reference data platform.

Update cadences may vary for different portions of the reference data; certain critical data may be updated in near-real time, while other, less critical data may be updated weekly, monthly, or quarterly. The differences in update cadence may mean that the reference data, while appropriately up-to-date for the vast majority of use cases and requests, will occasionally lack a data point that could be used to fulfill a specific request. In such situations, it is possible that the request would fail to match. Understanding the cadences with which identity resolution providers update their reference data should be an important consideration in choosing a provider.

HOW DUN & BRADSTREET HANDLES REFERENCE DATA QUALITY ISSUES

Dun & Bradstreet data goes through a series of cleaning, parsing, and standardization steps that aim for the highest level of matchability to request data. When request data is received, that data is processed through the same steps in order to align as closely as possible to Dun & Bradstreet's reference data.

Some of these steps include the removal of null or null-equivalent values (like *Title Unknown*) to reduce the possibility of improper matching across records, and ensuring that the proper number of fields is received in each input file.

Additionally, customers may request new investigations into companies within the Dun & Bradstreet Data Cloud; if a customer identifies a potential discrepancy with the data, Dun & Bradstreet is able to follow established procedures to research the business and update or confirm the information in question.

SUMMARY

Data quality, both in the request and in reference data, plays a huge role in the success of identity resolution. Taking proactive steps to enhance data quality prior to requesting identity resolution will improve the result quality. Similarly, choosing a provider with a demonstrated commitment to reference data quality will also enhance match quality.

Dun & Bradstreet enforces rigorous quality checks during data ingestion, applying critical cleansing, parsing, and standardization rules, and by curating data at all times from dependable sources through reliable channels. The DUNSRight process, pictured in Figure 2, is the centerpiece of these efforts.

PROCESS FACTORS IN IDENTITY RESOLUTION

A second consideration to keep in mind when evaluating the success rates of identity resolution systems is the process used to perform the resolution. Any identity resolution system is built to balance speed with accuracy. The design choices that result from this balancing act impact the ability of the system to quickly and accurately identify entities based on a request.

Some of the most common design factors that define identity resolution systems are:

- The definition of minimal search criteria
- Consistency between request and reference data
- The degree of imprecision – or "fuzziness" – permitted between the request and reference data
- The ability to determine the context or meaning of a match request
- The algorithm used to determine closeness of match

We will now briefly explore each of these common design factors and the impact these choices can have on identity resolution outcomes.

MINIMAL SEARCH CRITERIA

Any identity resolution system, in order to consider a request, requires that request to meet some set of minimum criteria. If a request doesn't fulfill the minimum criteria, then it will not be considered by the system and will not result in a match.

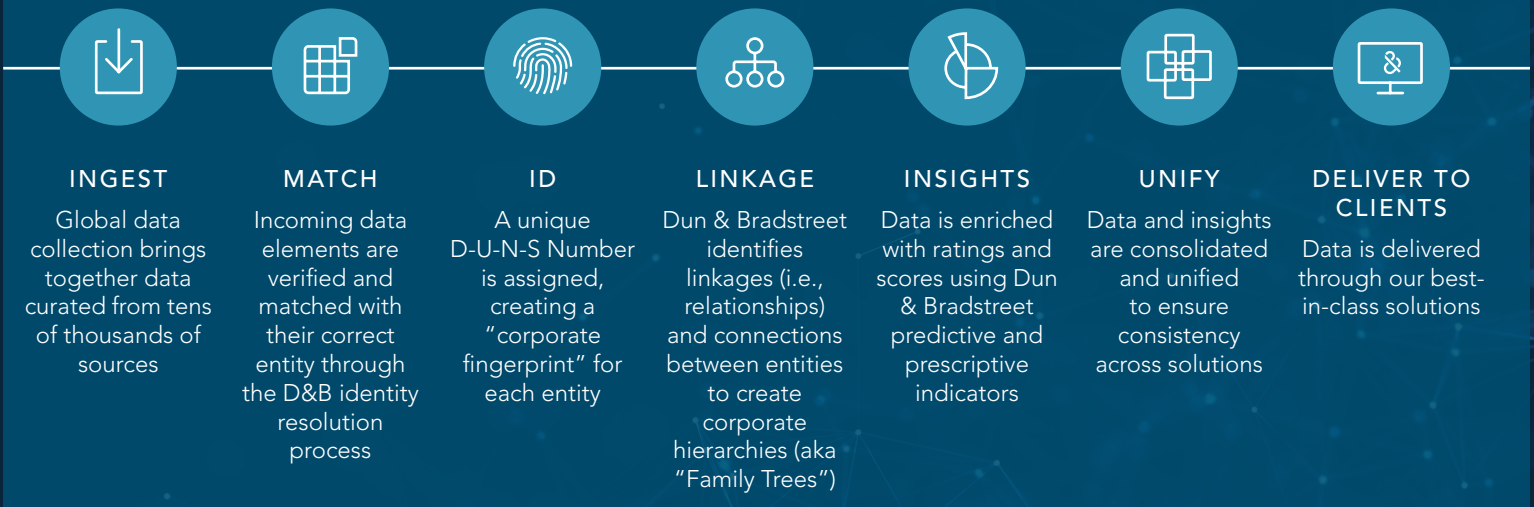

Figure 2. The DUNSRight Process

Returning to our example, consider again the bookstore owner who has been asked to find a book. The bookstore owner may insist that any request include certain information before she will try to find a book – perhaps she requires at least the title and the author, or the book's International Standard Book Number (ISBN).

Dun & Bradstreet requires a business name for all identity resolution requests; more information is permitted (and encouraged), but without a business name the system will not attempt to find a match.

CONSISTENCY BETWEEN REQUEST AND REFERENCE

Identity resolution systems are most effective when the data used to request a match is aligned consistently with the reference data against which matches are made, as closely as possible. The degree to which the match algorithm ensures this consistency has a significant impact on match success.

Consistency has myriad aspects but some of the most typical approaches to ensuring consistency include:

- Language or character set consistency
- Capitalization (case consistency)
- Consistency among variations

Language or Character Set Consistency

Particularly when dealing with data of a global nature, the variety of languages and characters sets that exists around the world may come into play. Matching algorithms that attempt to resolve identities in such environments need to be prepared to deal with these sorts of regional differences.

A customer asking for *Harry Potter y la Piedra Filosofal* is likely to be frustrated by the shop owner's inability to locate the book; however, if the reference data (the owner's list of available titles) is in English, then an appropriate way to approach the challenge would be to attempt to translate requested titles into English prior to trying to find the match. Assuming the shop owner can recognize the request as having been made in Spanish, and can translate from Spanish to the English equivalent – *Harry Potter and the Sorcerer's Stone* – the book may be located.

Other situations may be more challenging, especially when dealing with different character sets. Dun & Bradstreet addresses some of these issues through its exposure of multiple matching engines, including one that accepts some Asian languages that utilize non-Roman character sets. Directing match requests to the appropriate engine can ensure consistency of language and character set between the request and the reference data.

Capitalization

Other aspects of consistency are much simpler; capitalization is one of these. By doing something as easy as transforming any request and reference data to a similar case (upper- or lowercase), match algorithms reduce their complexity and improve outcomes.

Consider again a request for *Harry Potter and the Sorcerer's Stone*. If a customer writes down a request for *harry potter and the sorcerer's stone*, it's – to most reasonable observers – identical to a request for *Harry Potter and the Sorcerer's Stone* or for that matter *HARRY POTTER AND THE SORCERER'S STONE* or *hArRy POTter aNd The sORcerER's StoNe*. There are thousands of ways the letters in that title could be selected as upper or lower case.

Dun & Bradstreet simplifies matching by standardizing requests to a single case and by maintaining reference data in a similar fashion.

Consistency Among Variations

Beyond capitalization, there are other ways in which consistency is helpful in improving match success.

There are often words or phrases that may be represented in several different ways even within a particular language. For example, the word "and" may be represented by itself, or by an ampersand (&) or by the plus sign (+). A robust matching system should attempt to address these types of variations and ensure that each results in an equally successful match attempt.

Another example of variations is ordinal street names, for example First Street, Second Street, Third Street. Each of these could additionally be represented as 1st Street, 2nd Street, or 3rd Street. Even the word "Street" may be shortened to "St," another example where having a consistent representation of such variations

is important. Robust match engines are designed to identify and appropriately standardize these types of values, in order to optimize the ability to match even when the exact same words are not used.

Dun & Bradstreet match engines test for these types of variations and standardize appropriately, improving match rates and simplifying the process for users.

IMPRECISION (FUZZINESS)

One of the primary reasons that identity resolution systems exist is because there is ambiguity between different entities, and a need to be able to quickly and efficiently disambiguate one entity from another.

Accordingly, most identity resolution systems allow for a certain degree of imprecision between the request data and the reference data to which they match. This imprecision is also called "fuzziness," and the various mechanisms by which imprecision tolerance is implemented can affect the match rates for identity resolution.

Consider again our example; this time, assume that the bookstore owner has been asked for a book entitled *Harry Potter* – but no book with exactly that title exists in her store. However, the owner is aware that a book titled *Harry Potter and the Sorcerer's Stone* is in her store, and is able to surmise that despite the difference, there is a probability that this may be the book that her customer is seeking. In this case, the owner has implemented a "fuzzy matching" technique to accommodate the imprecision of the customer's request. However, there are seven books in the Harry Potter series, and many other ancillary books on the topic. The bookstore owner's "fuzzy matching" contains a certain amount of risk that her choice will be incorrect.

Some common "fuzzy matching" techniques are shown in Figure 3. In each of these cases, a typical implementation would measure the differences between the elements and score the comparison on a scale, perhaps 0 to 100. A zero score would indicate no meaningful relationship between the elements, while a score of 100 would indicate a very strong match; a score in between those two extremes suggests a partial match, which can be appropriately integrated by the algorithm into its results.

Clearly, if a system is more tolerant of imprecision with any of these techniques, there will be more opportunity to identify the right candidate, but also more possibilities that must be considered, which impacts speed and performance.

TECHNIQUE	DESCRIPTION	EXAMPLES
String Similarity Matching	Two strings are tested for how closely they match to one another	• "All the king's men" vs "All of the kings' men" • "TGI Friday" vs "T.G.I. Friday's"
Phonetic Matching	Two strings are rewritten as how they might sound, and compared for similarity	• "Smith" vs "Smythe" • "There's a bad moon on the rise" vs "There's a bathroom on the right"
Acronym Matching	Two strings are evaluated to determine if one represents a likely acronym for the other	• "UTEP" vs "University of Texas El Paso" • "IBM" vs "International Business Machines" • "UD Arena" vs "University of Dayton Arena"
Geolocation Proximity Matching	Two locations are converted to latitude and longitude and evaluated for how close they are to one another	• Burbank vs Los Angeles • Boca Raton vs Delray Beach • Dallas vs Fort Worth

Figure 3. Fuzzy Matching Strategies

At Dun & Bradstreet, the team responsible for identity resolution is constantly reevaluating techniques and tolerance thresholds for "fuzzy matching," implementing new approaches and adjusting others. As a result, Dun & Bradstreet is able to set and meet aggressive expectations around imprecision tolerance.

Nonetheless, it is possible that some possible matches will not be made because they fall outside the bounds of predefined Dun & Bradstreet tolerance settings.

CONTEXT AND MEANING

The ability of an identity resolution system to return the proper match for a given request is dependent in part on its ability to assess the context or meaning of the request being made. Doing so enables the system to provide a match that is most likely to fulfill the needs of the requestor.

The most common way in which this is accomplished is for the system to impute additional information from the request, based on a set of rules or a machine

learning algorithm. Once the additional information has been imputed, the system uses it either to broaden its candidate retrieval or to refine its scoring.

Consider the bookstore example once again; suppose now that a patron enters and asks the owner for help in finding a book called *Dolores Umbridge*. While no such title exists, the bookstore owner has enough context to identify that the customer is actually trying to find *Harry Potter and the Order of the Phoenix*. Similarly, the owner of some other bookstore might have a customer ask for a book he thinks is titled *The Joy of Woodworking*. Even if that book doesn't exist, the owner can put the request in context and suggest *The Joy of Carpentry* or *The Joy of Woodcrafting*. Both concepts are similar to the original request.

At Dun & Bradstreet, the data in identity resolution requests is parsed and then probed for additional imputable information. Business names may have keywords (e.g., "construction") that suggest that the most correct match would be in a particular industry; other words or phrases may be brand names for well-known companies; a provided phone number's area code may suggest a particular geographic region. This additional data is used to broaden the search and refine the scoring that ultimately identifies the best match for the request.

SCORING ALGORITHMS

Every identity resolution system requires a mechanism for determining the best candidate among several candidate entities in order to report back the match for a given request. Simply put, a scoring algorithm is required to sort the candidates from best to worst.

Scoring algorithms can range from simple to complex, and the specific scoring algorithm that is implemented will vary from system to system. Regardless, scoring algorithms typically rely on the following to generate a score:

- Similarity between request data and reference data points
- Source(s) of reference data points
- Recency of reference data points
- Uniqueness of reference data points relative to request data

Once a score is generated, the candidate entities are ranked according to their scores, and the highest-scoring candidate is selected as the match for the request.

In some cases, however, identity resolution systems implement a minimum match score. In these situations, if the highest-scoring candidate doesn't reach this threshold, a match won't be returned for the request. This is another way in which valid request data, and valid reference data, may seem to align as a match but do not.

Some providers, including Dun & Bradstreet, permit users to configure some or all of the scoring algorithm, in order to prefer certain classes of candidate entities over others.

Dun & Bradstreet's approach includes supplying the proprietary Match Grade® String (MGS), calculated based on similarity for each of 11 critical data points. The MGS is also mapped to a confidence code (CC)

ranging from 0 to 10. We regularly assess these mappings between MGS and CC to align different MGS values with the proper CC values and to make the confidence code as simple and straightforward to use as possible. Nonetheless, it remains possible to adjust the configuration to meet particular needs, and even when this has been done, certain matches will still not rise above the matching threshold.

SUMMARY OF PROCESS FACTORS

The matching process is the core of identity resolution and therefore its particular implementation and efficacy have a direct impact on the success of the overall use case.

Even the most advanced processes, however, have constraints that can lead to match request failures. Certainly, higher-quality match processes lead to higher-quality results, but one cannot assume that higher match rates necessarily equal better results: It is possible to have high match rates but for the matches to be of low quality.

Dun & Bradstreet has been refining its identity resolution processes for decades; nonetheless, it would be misguided to expect 100% match success. Instead, Dun & Bradstreet considers the business end game and optimizes match performance to produce the results that best serve the use case.

INHERENT TRUTH AND IDENTITY RESOLUTION

A third consideration surrounding the success of an identity resolution request is the inherent truth within the universe of discourse.

If a request is made to identify an entity that simply does not exist, then a proper identity resolution system will – correctly – return an indication that it cannot find the entity being requested.

Returning to our example, if the customer asks the bookstore owner to find "Harry Potter" in the store, and there simply is not a book with that title – not just no book with *Harry Potter* in the title, but no misspelled *Hairry Potter*, no misheard *Barry Slaughter*, no book written by Harry Potter, no book with the words "Harry Potter" together in the title – then she should reply that she can't find it, and that would be a proper response.

No match process improvement or data quality improvement can force an entity to exist that simply does not exist. It is imperative that requestors of identity be prepared to consider the possibility that a failed identity request is an indication that the requested entity simply doesn't exist, and have processes in place to properly handle such a situation.

One possible approach is to use an identity resolution failure as a trigger to add an entity to the universe with the information provided. While effective at avoiding multiple failures for the same set of input, it risks the introduction of "noisy" data into the reference dataset, and eliminates the valuable feedback that is implicit in the nonresolution of a request. Data governance, additionally, becomes much more challenging due to the introduction of a much less structured and curated source (to wit, the failed requests).

Another possible approach for dealing with request failures is to extend the universe of discourse. To use our example again, the customer, their request in the young adult bookstore having failed, may broaden their search to include other bookstores that the owner operates, or bookstores in other towns, or libraries, or a few friends who are known bibliophiles.

Dun & Bradstreet makes judicious use of both mitigating approaches. With permission of the requestor, Dun & Bradstreet may record the request data for failed requests and use that information to identify opportunities for future investigations into new businesses, adding them to the universe once they have been properly verified. In this way, a failed request can trigger the introduction of a new entity, but only after it has undergone rigorous vetting.

Additionally, Dun & Bradstreet is constantly expanding its breadth and depth of reference data, including additional countries and sources of data for countries with established reference datasets. This gives identity resolution clients a greater chance of finding the business they're looking for.

Nonetheless, there will remain instances of business locations that do exist but are not in the Dun & Bradstreet Data Cloud. For some of these instances, the omission is by design, as Dun & Bradstreet does not assign a D-U-N-S Number, for example, to stand-alone ATM locations or vending machines, nor to oil rigs or windmills. For others, the omission is a result of

processes and policies intended to safeguard the reliability of the information in the Data Cloud as a whole; as mentioned previously, Dun & Bradstreet rigorously vets new additions to its Data Cloud. Depending on the business, this process can take time, and therefore a relatively new business entity may not be immediately locatable within the Dun & Bradstreet Data Cloud.

While Dun & Bradstreet continually works to capture and offer data about businesses around the world, it's also important to consider that when identity requests fail to match, it's possible that the business either never existed or has ceased to exist. Large numbers of failed identity requests may represent an opportunity to perform a thorough purge of data that may have grown outdated and, as a result, outlived its utility. While challenging, these efforts are often rewarding for consumers of the data.

OPTIMIZING OPTIMIZATION

Since optimizing identity resolution is dependent on so many variables, not all of which are fully under a requestor's control, it is not uncommon to struggle with understanding when such optimization should be considered complete – in other words, when is it "done"?

As seen in Figure 4, our historical data suggests that a median customer has approximately 92% of their match requests fulfilled; customers at or above the 90th percentile may achieve 97.5% or more matches.

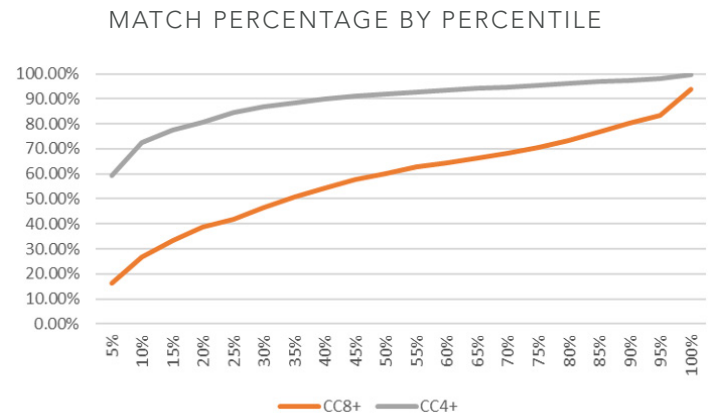

Figure 4. Match Performance Statistics by Percentile

Obviously, such metrics rely heavily on the quality of the request data submitted. As requests include more comprehensive data, the rate of fulfillment increases.

Of course, the ultimate answer to the question of when to scale back optimization efforts is dependent on an unknown quantity – the fulfillment rate that can actually be achieved by any provider. Since this target is unknown, it may be best to attempt to modulate optimization experimentally.

As with any experiment, it is important to establish baselines and then understand the impacts that changing variables have on that baseline. One approach to conducting such an experiment would be to determine the level of fulfillment that is achievable with a given provider upon a set of request data – setting the baseline – and then apply one or more of the recommended techniques in the next section before testing the fulfillment rate again.

RECOMMENDATIONS FOR GETTING THE MOST OUT OF IDENTITY RESOLUTION

This paper has explored the typical scenarios that lead to failed identity resolution requests, as well as some of the ways Dun & Bradstreet has accommodated some of these situations. Having dispelled the notion that a 100% match rate is a desirable goal in most business processes, we can now present these 10 recommendations for evaluating and optimizing match performance:

1. Don't expect 100% match at scale. Unless datasets are small enough to curate manually, some requests will fail to match, and some will return imperfect matches.

2. Focus on value. Direct optimization efforts into identity resolution requests that offer the greatest business impact, and provide as much data as possible in these requests. Prioritizing by spend or revenue data is one approach to achieving proper focus. In other cases, prioritizing records that represent strategic advantages even with limited revenue or spend may be the more appropriate approach.

3. Know the limits. Some match processes have very broad capabilities and others are intended to have a narrow focus. Understand the reference universe prior to attempting to match data against it.

4. Understand the match process. Understand the specific process's minimum match criteria, prepare request data to meet them, and consider delaying requests for records that do not fulfill them until more data points have been gathered.

5. Configure the match system. World-class match engines offer tuning configurations that allow customers to align matching behavior with specific expectations and needs. Take advantage of these capabilities.

6. Fix known problems. Discuss and resolve systemic data quality issues with the upstream source. If a source regularly supplies bad city names, for example, work with them to fix the problem before allowing the data to move forward.

7. Learn the lingo. Understand the metadata returned with match output and what it implies. Examples of this include the Dun & Bradstreet Confidence Code, Match Grade String (MGS) and Match Data Profile.

8. Have a backup plan. Establish a secondary protocol for addressing identity requests that do not meet match acceptance criteria. This could include a manual review, an automated second pass with modified configurations, or more. Consider reattempting the identity resolution for non-matching requests a few days, weeks, or months later.

9. Watch for trends. Track success rates over time, both to quantify improvements versus historical baselines and to monitor for performance issues.

10. Enlist help. Dun & Bradstreet has a team of identity resolution experts who specialize in optimizing match performance for specific use cases and business goals. These experts can be engaged through the Dun & Bradstreet Client Director.

CONCLUSION

Identity resolution is a critical workflow component for many organizations. The downstream impacts of missing or improper resolution can be significant. As such, consumers of identity resolution services naturally look to maximize their match rates.

However, an increase in match rates does not necessarily equate to an increase in match quality, and, depending on a customer's specific use case, can actually be counterproductive. Rather than aiming for 100% match rates, organizations should focus on optimizing match performance – finding the appropriate balance between recall (finding any match) and precision (finding the single correct match) for their purposes.

This goal can be brought into closer reach through request data quality improvements, the selection of a provider with high reference data quality and sophisticated match processes, and an acknowledgment that the inherent truth may not support a match under all circumstances.

Dun & Bradstreet is well positioned to help organizations with these challenges. We have provided critical information about customers and suppliers for over 177 years, enabling better decision-making and more successful outcomes.

The Dun & Bradstreet team is constantly evolving its industry-leading expertise in identity resolution, and making those improvements available to customers. The Dun & Bradstreet Data Cloud offers the world's largest set of business decisioning data and analytical insights, providing insights on hundreds of millions of businesses and other commercial entities across the globe. We source data from tens of thousands of sources, tens of millions of websites, and crowdsourcing and validating initiatives. We continuously monitor our vast number of sources for changes that impact information in the Dun & Bradstreet Data Cloud, verify changes, and update the Data Cloud accordingly.

Additionally, the Data Cloud offers the deepest and richest insights into relationships of all types among companies, identifying millions of relationships that can inform decision-making. We leverage information from our global sources, along with proprietary capabilities, to discover and curate millions of business-to-business relationships. These relationships can include corporate hierarchies, ultimate beneficial ownership, alternative-type relationships, historical ownership, and analytically derived connections. We continuously monitor the dynamic changes to these relationships, including corporate actions such as mergers, acquisitions, and divestitures and make relevant updates.

AUTHORS

GEORGE C. L'HEUREUX, JR., is Principal Consultant – Data Strategy at Dun & Bradstreet. Contact George at lheureuxg@dnb.com.

CECILIA PETIT is Principal Consultant – Data Strategy at Dun & Bradstreet. Contact Cecilia at petitc@dnb.com.

About Dun & Bradstreet

Dun & Bradstreet, a leading global provider of business decisioning data and analytics, enables companies around the world to improve their business performance. Dun & Bradstreet's Data Cloud fuels solutions and delivers insights that empower customers to accelerate revenue, lower cost, mitigate risk, and transform their businesses. Since 1841, companies of every size have relied on Dun & Bradstreet to help them manage risk and reveal opportunity.

dnb.com | Twitter: @DunBradstreet