CUSTOMER
DATA PLATFORM
INSTITUTE

Redpoint

Published by:

Customer Data Platform Institute
231 Second Avenue
Milford, CT 06460
www.cdpinstitute.org

Sponsored by:

Redpoint Global
34 Washington Street, Suite 205
Wellesley, MA 02481
www.redpointglobal.com

Identity Resolution: See Your Customers Clearly

Identity resolution can be a bit of a mystery.

You visit a website for the first time and see ads for a product you researched elsewhere. You enter an unfamiliar hotel lobby and the doorman greets you by name. An insurance company asks you to identify addresses you lived at several years ago.

As a consumer, you might find those examples creepy or cool. As a marketer, you may find them exciting. In all cases, you're left to wonder exactly how the website, doorman, or insurer knew what they knew.

The answer in each case is the business was able to connect different bits of information related to the same person. The website used a third-party cookie to identify you and serve a targeted advertisement. The doorman was given pictures of high-value customers and their expected arrival times. The insurance company had a list of postal change-of-address records.

Identity Resolution Is Important

Making those connections is the role of identity resolution. The examples are a tiny fraction of identity resolution applications. A more comprehensive list would cover all business operations from customer acquisition through operational support.

- Marketing. Identity resolution can find customers across websites, email, mobile apps, game consoles, smart TVs, social networks, retail stores, and anywhere else they have a digital interaction. Once a customer is recognized, marketing systems can look up the profile to find information used to personalize the interaction. Ideally, the profile will combine information gathered in all channels and stored in a central location such as a Customer Data Platform. This profile can be supplemented with current context such as location and intent to deliver personalized offers, recommendations, and other treatments that increate profits and meet customer expectations for a quality experience. Identity resolution also enables marketers to measure the results of their efforts by correlating the treatments received by each customer with long-term behaviors such as purchases and retention.

- Operations, Service, and Support. Identifying customers and connecting to their profiles helps the company deliver its products and services more efficiently. Tasks such as ordering, payments, and returns are all streamlined when information can be read from existing records rather than re-entered separately for each transaction. Credit and return decisions can be based on customer history, enabling the company to fine-tune the balance between risk and reward on a personal level. Service systems can be given immediate access to the list of products a customer has purchased, product information such as common problems and solutions, and recent customer behaviors such as online searches of help pages. All of these items can be used to send inbound calls, chat requests, or email inquiries to the right agent and can give the agents guidance in solving the problem quickly and correctly. Other information such as loyalty status and churn predictions can further optimize decisions during each interaction. All of this information can be analyzed to understand product use and identify potential issues, find and remove friction in operational processes, assess agent performance, and measure the impact of customer experience.

- Advertising. Identity resolution can track individuals across websites, mobile apps, smart TVs, and elsewhere using cookies, device IDs, IP address, encrypted email addresses, or other shared

identifiers. While new privacy rules have limited some identity tracking, there are still many situations where people can be recognized and served advertising tailored to their interests, attributes, and history. This enables marketers to target both existing customers and non-customers. Non-customer targeting is often achieved through "look alike" models that find consumers who are similar to existing customers, regardless of what site those consumers are currently visiting. The same sort of tracking is also used to connect ad viewing with subsequent behaviors, such as purchases. Identity tracking also enables greater advertising efficiency through eliminating duplicates, suppressing existing customers and capping the number of ads sent to any one individual.

It's easy to list identity resolution applications. But in the excitement of building that list, don't forget that identity resolution isn't perfect. The quality of identity resolution available for each application will have a major impact on the value that application creates. Indeed, below some threshold, poor identity resolution can be worse than no identity resolution at all.

Identity Resolution is Complicated

Before looking at factors that impact identity resolution quality, let's define some fundamental concepts.

- Persistent keys vs transient identifiers. Identity resolution is the process of determining which identifiers refer to the same entity. An entity could be almost any real-world object, including a device, product, or building. In the context of customer data, it's usually a person, household, or business. Those kinds of entities have an independent existence that continues over time. While nothing lasts forever, we'll call them "persistent". The entities are also associated with attributes, or identifiers, such as telephone number, email or postal address, mobile device ID, or account ID. Those associations are not necessarily persistent: a person could get a new phone number, change their address, or buy a new mobile device. Some identifiers are clearly temporary, such as the number of your hotel room or airplane seat during a trip. Similarly, the same identifier may be associated with different entities at different times: someone else will have that hotel room tomorrow night; someone else may live in your old house after you move.

  Because identifier associations are transient, it's often important to connect them with a persistent entity. This is what lets you know that this phone number and that email address belong to the same person. The set of these entity-to-identifier relationships is called an identity graph. Identity resolution systems usually assign their own identifier to each entity; this is separate from any external identifier, so it's available regardless of which other identifiers come and go. To accommodate identifiers that can belong to different entities over time, an identity resolution system will also often attach date ranges to the relationship or at least distinguish current identifiers from obsolete ones.

  While entities like a person are physically persistent, the identifier associated with an entity may be persistent or not. A non-persistent identifier is created by systems that rebuild their identity graph periodically, and assign new entity identifiers every time. This simplifies processing but also means that any data coded with identifiers from previous editions of the graph will no longer connect with the correct entity. By contrast, a persistent entity identifier is retained from one graph update to the next. This ensures that any record with the entity identifier will be linked correctly with the current identity graph, or directly with any other record holding the same entity identifier. This type of persistent entity identifier is what's usually meant by "persistent ID".

- Anonymous vs known. As just discussed, entities have an independent existence that is separate from any identifiers are associated with them. In practice, every entity in an identity resolution system is associated with at least one external identifier: otherwise, there would be no connection between that entity and anything else. There is, however, no guarantee that this identifier will provide enough information to locate the entity in the real world. Identifiers that do provide this information can be called "personal identifiers", although the legal meaning of that term varies with the jurisdiction. Identifiers that don't provide this information can be considered "anonymous". Traditional examples of anonymous identifiers include web browser cookies and device IDs. However, it turns out those can in fact often be connected with specific individuals, so privacy regulators are increasingly treating them as personal IDs. From the standpoint of identity resolution, an entity can be considered anonymous if it is associated only with identifiers that cannot themselves be used to locate or contact an individual. The entity becomes known once it is associated with an identifier that does connect it with an individual, such as email or telephone number.

- Matching methods. Identity resolution involves two major tasks: connecting identifiers with entities, and looking up connections after they are made. Making the connections usually involves matching different identifiers in some fashion. There are two main matching methods:

    - Deterministic matching uses connections that provided directly or indirectly by the entity. The simplest type of direct matching is an exact match: the email on a new account matches the email on an existing account, so they are linked to the same person. A second type relies on explicit connections. For example, if a customer provides an email address, postal address, and telephone number when she opens an account, then these three identifiers can all be linked to the account number and to the persistent ID associated with that number. If the customer later opens an email on a particular smartphone, the smartphone can also be linked to the email, bank account and persistent ID. This is still considered deterministic although additional controls are needed to avoid linking to a smartphone that does not truly belong to the customer, perhaps because the email was forwarded or the phone was borrowed.

    - Probabilistic matching uses data to estimate the chances that two identifiers relate to the same entity. One version of probabilistic matching uses behavior data, such tracking when a smartphone and tablet computer are frequently used at the same time in the same locations. There may be additional factors such as the type of content consumed or tasks performed. A different type of probabilistic matching compares data elements directly, such deciding whether "Rob Smith" and "Robert Smith" are the same person or "123 Main Street" and "132 Main Street" refer to the same address despite a typing error. This second type of probabilistic matching is used primarily for address data although it can sometimes match email addresses. Both types of probabilistic matching often rely on external services that collect more information than a company can gather on its own. Both also use complex algorithms to find as many correct matches as possible while minimizing false matches. Today, matching algorithms are often built by machine learning systems that are trained on sample data to estimate the likelihood that two sets of identifiers represent the same entity.

- Match levels. In most customer data applications, the entity being managed is an individual person. But in some situations, the relevant entity is business or household. These may have special identifiers, such as license numbers, which do not apply to individuals. Such situations require specialized matching rules to deal with the different identifiers, to distinguish different entities of the same type, and to connect individuals with the larger entities. Additional nuances include connecting

different levels of entities, such as parent companies and subsidiaries or branches; attaching the same individual to several entities; attaching several individuals to the same entity; and adding and removing individual-to-entity relationships as these change over time.

- Golden record. As identity resolution systems sift through input data to determine what will be considered a match, they also assess the quality and currency of each item. This information is often consolidated to produce a "golden record" that contains what the system believes to be the most accurate version of each element. Factors include completeness ("John Jacob Smith" is more complete than "J. Smith", correctness ("123 Main St" may be the actual address, not "132 Main St"), compliance with standards ("Robert" is a standard name while "Bobby" is a nickname), and currency (distinguish an old address from a current address). The purpose of the golden record is to provide other systems with the best available version of the entity data and to ensure all systems use the same data. Some definitions of golden record limit its contents to entity attributes, but others include transactions and derived values such as segment assignments and predictive model scores. Note that while the golden record will usually pick a single "best" value for each attribute, there may be multiple attributes that are valid, such as multiple email addresses or device IDs. Different values may be best for different purposes, such as distinguishing a business and personal email address. Matching processes need access to all values of each attribute, not just the best one.

Identity Results Can Vary

All identity resolution systems are governed by the same fundamental concepts. But the performance of these systems varies considerably. Key factors include:

- Update vs rebuild frequency: Each new identifier must be ingested into the identity resolution system and either matched with an existing entity record or used to create a new one. This process might update the existing identify graph by looking for matches with existing entities, or it might rebuild the graph by re-examining all identifier-to-entity relationships. A rebuild could result in assigning existing identifiers to a different entity, in merging two entity records into a single entity, or in splitting one entity into two different entities. A rebuild requires much more processing than an update, so systems often do updates more often than rebuilds. In particular, real-time updates for each new piece of data are fairly common while real-time rebuilds are not. However, the frequency of updates and rebuilds has a direct impact on the accuracy of identity resolution, since more current data enables the system to adjust more quickly to new information.

  Update and rebuild frequency are particularly important when the identity resolution system maintains a golden record, since any changes to that record will impact other systems that use it. If the golden record includes transactions and other non-identity data, quick updates have even greater impact because they will directly influence audience selections and personalized recommendations based on that data.

- Data quality. Identity resolution is highly dependent on the quality of data fed into the system. Identifiers that are missing or captured incorrectly due to input errors will result in missed matches, false matches, or both. Data captured for a different original purchase often doesn't meet the quality standards needed for effective identity matching. For example, a building address may contain errors that the local post office will overlook but will confuse a matching algorithm. Similarly, a valid phone number may not be needed to set up a new loyalty account but a missing or invalid one might prevent

the identity resolution system from matching that account with other records. Although the quality of the inputs is generally beyond control of the identity resolution itself, it is something the system can monitor and help guide users to improve. In addition, some identity resolution systems can make limited quality improvements on data after it's loaded, such as standardizing the formats of dates, phone numbers, and addresses.

- Reference data. Identity resolution can often be improved by applying external data. This might include postal files that determine whether an address is valid and report address changes when a person has moved. Other reference data might include connections between different identifiers, such as linking email address to postal address, or associating one device ID with another. This information is often gathered by specialist firms that have built vast data collection networks which yield much more complete results than a company's own internal sources. Some reference data is based on deterministic matches while other reference data relies heavily on probabilistic techniques.

- Accuracy tuning. Identity resolution always involves a trade-off between finding correct matches and avoiding false matches. Even deterministic methods involve some decisions such as how to treat apparent data entry errors, how many times a relationship must appear before it's accepted and how long to treat an inactive identifier as valid. Probabilistic methods involve much more complicated algorithms and many more judgements. Different applications require a different balance between missed vs false matches: in financial transactions, a false match could have severe consequences, such as crediting a deposit to the wrong account or approving a loan to an unqualified borrower. By contrast, a false match for an advertising program might only mean one less customer sees an offer. Some identity resolution systems address this directly by allowing users to set different matching rules for different purposes. Identity resolution systems vary significantly in how they build their matching rules, how easily users can understand the rules, and how much control users have over the rules that apply in any given situation.

Privacy Rules Have an Impact

Every marketer is aware of Google's plans to block third-party cookies in Chrome, and recognizes that will make it much harder to understand who sees which advertisements. Other constraints include third-party cookie blocking on Apply Safari and other browsers; Apple and Google Android reducing consumer tracking via device IDs; and proliferating legal rules starting with the European Union's GDPR. Each of these changes has an impact on identity resolution processes and results. Major considerations include:

- Limits on collection. Some identifiers are no longer available, either entirely (e.g. third-party cookies blocked by a browser) or unless the user agrees (e.g. device IDs requiring consent). There can also be limits on location and other information which is not a personal identifier but can resolve identity in combination with other data. Additional rules apply to particular types of data, such as health information, and to data about classes of people, such as children. Identity management processes need to be adjusted to work without these elements or to only use them where allowed.

- Limits on use. Many privacy regulations limit what companies can do with the data they are still allowed to collect. Some limits are based on gaining consumer consent to particular uses, such as California's rule to allow consumers to opt out of sharing their data with other companies. Other regulations specify the types of uses that are permitted without consent, such as GDPR rules that allow processing to fulfill a contract or conduct normal business activities ("legitimate interest"). The

same identity resolution process might be permitted in one situation and forbidden in another depending on the purpose involved.

- Privacy-enhancing techniques. Data clean rooms, homomorphic encryption, pseudonymization, email hashing, differential privacy, and other techniques are increasingly applied to enable companies to process customer data without exposing personal identities. Many of these techniques involve some form of identity resolution to create unified profiles which are then anonymized. Others may allow some types of identity resolution after the data is anonymized.

As privacy rules become more stringent, achieving identity resolution within privacy constraints will become more important.

Match Identity Resolution Methods to Your Needs

Identity resolution becomes more complicated – and more important – every day. This makes selecting the right tools an ever-greater challenge. Here are some helpful tips to guide your choice:

- Different applications rely on different techniques. There are many different kinds of identity resolution, including variations in entity definitions, data sources, match methods, and outputs. Learn which techniques are appropriate for which purposes so you can ensure you focus on what' necessary.

- Define your needs and identify your gaps. Your company probably has some identity resolution capabilities in place, if only for operational tasks. Take some time to understand how those systems work, how they might be improved, and what impact those improvements might have on business results. Then expand your vision to assess what new applications identity resolution could make possible, the value of those applications, and changes in existing identity resolution systems to achieve this value. This lets you focus on filling the gaps that matter.

- Compare internal and external solutions. Identity resolution involves many components. Some, such as data quality, might already exist in your company or be easily developed by extending existing systems. Others, such as advanced matching algorithms, persistent identity management, and comprehensive reference data, are almost surely best purchased from outside specialists. Be aware that external options exist for nearly all identity resolution features and consider those as options when planning to enhance your system.

- Test early and often. The complexity of identity resolution systems makes it almost impossible to predict the exact impact of changes in any one component. It's essential to test new data sources, quality processes, match algorithms, update rules, and other elements to see how they impact the results of different applications. Reference data sources in particular must be evaluated closely and reviewed frequently as their own quality can vary over time.

What Next?

Identity resolution doesn't have to be a black box. Take the time to understand how identity resolution works in general and at your company in particular. Then develop a vision for how it should work, including current and future applications. As you work to realize that vision, keep in mind the ultimate goal: to future-proof your business and gain full value from your customer data investments.

About Redpoint Global

With the Redpoint CDP, innovative companies are transforming their customer experiences across the enterprise and driving higher revenue. Redpoint's solutions provide a remarkably unified, single point of control where all customer data is connected and every customer touchpoint intelligently orchestrated. Delivering the perfect customer experience – more engaging, highly personalized moments, relevant next-best actions, and tangible ROI—this is how leading marketers lead markets. To learn more, visit www.redpointglobal.com.

Contact:
Redpoint Global
888 Worcester Street, Suite 200
Wellesley, MA 02482
www.redpointglobal.com
contact.us@redpointglobal.com

About the CDP Institute

The Customer Data Platform Institute educates marketers and marketing technologists about customer data management. The mission of the Institute is to provide vendor-neutral information about issues, methods, and technologies for creating unified, persistent customer databases. Activities include publishing of educational materials, news about industry developments, best practice guides and benchmarks, directories of industry vendors, and consulting on related issues.

The Institute is managed by Raab Associates, a consultancy specializing in marketing technology and analysis. Raab Associates identified the Customer Data Platform category in 2013. Funding is provided by a consortium of CDP vendors.

Contact:
Customer Data Platform Institute
231 Second Avenue
Milford, CT 06460
www.cdpinstitute.org
info@cdpinstitute.org

