CHECKLIST 2022

Enhancing Data Integrity and Improving Decision-Making Through Geocoding and Data Enrichment

By David Loshin

precisely

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Enhancing Data Integrity and Improving Decision-Making Through Geocoding and Data Enrichment

By David Loshin

Organizational digital transformation initiatives are partially inspired by the desire to become more data-driven. Fundamentally, data-driven organizations use their data with business intelligence and machine learning to inform and improve their business decision-making. Not surprisingly, flawed, inaccurate, or inconsistent data sets impede an organization's ability to optimize the use of its data. To make certain that the beneficiaries of advanced analytics applications can make decisions confidently, an organization must institute practices for ensuring data integrity and trust.

There are two facets to data integrity assurance. One is rooted in fundamental data quality management—defining and complying with data quality and validity expectations. The second involves a sound strategy for acquiring hyper-accurate data to validate and verify the organization's data assets and enrich them to improve their value for analytics.

Seven ways to enhance your data's integrity:

1. Clarify the business questions
2. Streamline data enrichment using location information already in your business data
3. Standardize, verify, and validate addresses
4. Ensure data integrity with accurate geo-addressing
5. Employ a persistent, unique location identifier
6. Enrich your data
7. Leverage spatial analytics for added context

Location information is one domain that is nicely suited to data enrichment. Aside from being easily adaptable to data validation and verification, location information and spatial analytics provide context for decision-making processes. However, many organizations are still immature when it comes to high-quality location data for spatial analytics, and that suggests a need for accurate "geo-addressing," which includes both address verification and geocoding, to improve the analytics processes informing decision-making.

This Checklist Report provides guidance for enhancing data integrity through address verification, geocoding, and data enrichment. We consider the approach to soliciting data requirements from the business users, and then drill down into the concepts of addresses and locations. We discuss ensuring accuracy and consistency with geo-addressing and suggest that a persistent, unique identifier for locations can simplify data enrichment while improving precision and accuracy in matching and linkage. Finally, the report discusses how enriched data provides additional context and knowledge that contributes to more effective use of spatial analytics.

1 Clarify the business questions

As data volumes grow and data variety increases, organizations have expanded their investment in incorporating machine learning, AI, and other sophisticated analytics into their application environments. Although teams of data analysts and data scientists are positioned to leverage advanced analytics techniques, you cannot just throw technology at a problem and expect it to be solved. You must first understand the problem, be able to articulate what your expectations are for addressing the problem, and then determine what data sets are necessary to solve the problem.

Data analysts should not have to struggle to determine what information they need to ensure that business expectations can be met. Partner with your business clients to solicit their needs and clearly articulate the business questions. Although this may sound obvious, often the separation between the development teams and the business partners complicates the ability to assemble the right solution. Clarifying the business questions makes it simple to identify the most appropriate data source to deliver insights that address the business opportunities and inform decision-making.

Your process for clarifying the business questions should include these steps:

- **Engagement.** Reach out to the business clients and solicit their input about the business problems they face, why existing tools are insufficient, and what success metrics indicate that business challenges are being adequately addressed.

- **Clarification.** Identify who the decision makers are, what decisions they will be making, and what types of analyses will allow them to make good decisions. Review the use cases and clarify the specific business questions to be asked to address the business problems.

- **Data determination and integration**. Research the data needed to assemble a solution that facilitates the reporting, business intelligence, and analytics that inform the decision makers. Determine the best way to acquire and integrate that data to formulate a business-driven solution.

Once you take these steps, the data analyst can work with the business client to assemble a data environment and employ the right tools for accurate data, enrichment, and spatial analysis to address the business questions.

2 Streamline data enrichment using location information already in your business data

Businesses often conflate the concept of a "location" with an "address" and typically treat the two words as synonyms when solely focusing on ensuring the validity of pickup and delivery-point addresses. There are, however, numerous operational and analytics use cases that involve locations not necessarily bound to a street address, such as:

- Utility companies that manage large numbers of physical assets such as pipes, antennas, utility poles, wires, or fences need to keep track of the locations of these assets for ongoing monitoring and maintenance
- Mobile telephony companies log the locations from which mobile calls are initiated and track the mobile devices as they move to monitor connection quality and help determine where service gaps exist, requiring additional infrastructure
- Insurance companies analyze and then attempt to balance risk in coverage associated with the physical distribution of their customers across different geographic locations
- Community planners may install sensors in a variety of outfall locations to monitor water quality and volume of outflow across a region

In these examples, the locations don't need to be bound to deliverable addresses. Instead, they rely on a more general concept of a *geocode*. Although different types of codes are used to refer to locations, the most frequently used geocode employs a pair of latitude and longitude coordinates, where text addresses are mapped to latitude/longitude pairs through a process called *geocoding*.

Given a database of named locations and other points of interest, a geocoding service maps places to their location coordinates, including airport or railroad station codes, named points of interest, a street intersection, or place names. Alternatively, the service can perform the reciprocal operation of *reverse geocoding*: providing the nearest address or point of interest given a set of latitude and longitude coordinates. These processes allow you to link an address record to a geocoded location, opening the door for streamlined enrichment with additional location data attributes.

3 Standardize, verify, and validate addresses

Addresses in data provide the link for enriching that data with location information. However, addresses themselves can be tricky. An address is typically structured as a multifield representation of a postal delivery location. Some countries (such as the U.S.) have well-defined standards describing the structure and semantics of all address fields. Most countries are part of the Universal Postal Union (UPU), a United

Nations agency acting as the postal sector's primary forum for international cooperation. You can look up any country and see the addressing standards and requirements, with sample addresses and their respective structures.

However, even when there is an existing standard, it does not mean that every organization using addresses will be familiar with or comply with that standard. In cases where the organization uses the standard, people unfamiliar with the standard will still create addresses using nonstandard values or allow components of the address to be assigned to the wrong fields. Address values are often incorrect or assigned to the wrong field, resulting in false positives and false negatives when matching addresses, which in turn leads to incorrect insights.

Despite the efforts by collaborative organizations such as the UPU to identify national organizations that might provide postal standards, postal addresses are complicated. Parsing, standardizing, and verifying addresses across different countries is complex because each country has a different address format. Your enterprise can employ tools and techniques that standardize addresses while improving their trustworthiness.

- **Address standardization.** This process parses the values in each address field and transforms nonstandard values to conform to the address standard, such as transforming the word "Street" to its standard representation of "ST" in the United States or "Saint" to its standard representation of "ST" in France.

- **Address verification.** This process checks a standard address against an authoritative source to ensure the address can receive mail.

- **Address validation.** This process checks to see if the address is valid within the defined standard. For example, validation will check if a combination of a building number and a street name are valid within the number range for that street. Validation is available in countries that can certify the correct address. For example, in the United States, there is a method called CASS certification for address validation.

Because address standardization is tightly coupled with address data quality, look for vendor solutions that offer data quality capabilities built into their geocoding solutions.

4 Ensure data integrity with accurate geo-addressing

Geographic data integrity is critical for empowering downstream data consumers and analysts using spatial analytics to answer their business questions. However, processes linking business data sets to the wrong source reference data can misinform and lead to impaired decisions.

Address standardization, verification, and validation provide one level of precision and allow for accurate matching against a high-quality reference data source. Both precision and accuracy are critical for a number of use cases, such as:

- Differentiating between points of interest at a specific location (e.g., the front door of a building versus the center of the rooftop or the location of a loading dock)

- Identifying the location of non-addressable assets (such as a cell tower or a windmill)
- Monitoring how weather events such as hurricanes move across regions to determine whether certain properties are in the path of a storm

Two criteria are necessary to ensure the highest quality data enrichment: processes that accurately parse address data and geocoding reference data containing the most accurate, up-to-date information at the precise location. In other words, the need for hyper-accurate location coordinates implies a need for geocoding/reverse geocoding processes that accentuate the core capabilities for location data integrity: accuracy (ensuring correctness of the location information) and currency (ensuring that the location information is up to date).

Many organizations, however, underestimate the significant amount of work it takes to maintain address accuracy and currency, especially when those addresses are subject to updates from a variety of sources. Even the best parsing and matching algorithms will produce low-quality data if the reference data set contains incorrect latitude/longitude coordinates, if the reference data is outdated, or if there are false positive matches. That suggests that when you need hyper-accurate location coordinates, look for partners that can help implement a solution with precise address parsing, high-quality record matching that can link records about the same entities from multiple sources, and access to a trustworthy reference data source with current and correct location data.

5 Employ a persistent, unique location identifier

Data sets created or updated using manual data entry processes are prone to introduced errors and inconsistency. Messy or nonstandard addresses, alternate names, multiple data sources, and variations in table structure combine to make it a challenge to enrich data across an organization's systems. Assigning a unique identifier to a standardized address streamlines joining data sets together as well as data enrichment.

Benefits of using a persistent, unique identifier for location data include:

- **Providing a rapid lookup.** All data sets containing information about a location can use the unique identifier as an index for querying.
- **Improve data privacy.** A unique location identifier can mask personally identifiable information (PII) associated with residential addresses by referring to the identifier instead of sharing sensitive location information.
- **Streamline enrichment.** Joining data sets using a unique location identifier simplifies the data enrichment process for augmentation with location attributes.
- **Empower data analysts and data scientists.** Mapping multiple enriched data sets to specific locations streamlines the augmentation of features that can be used for reporting and more advanced analytics.

6 Enrich your data

Collecting and verifying location information is a prelude to effectively leveraging that information to address business challenges. More specifically, spatial data enrichment is a process of augmenting existing data sets with additional characteristics associated with your organization's location data.

Data enrichment can be used to enhance location information and set the stage for more sophisticated analyses that can improve customer experiences, optimize operations, reduce insurance risk, streamline marketing, identify food deserts, or inform analysis of communicable disease transmission. Enrichment can incorporate relatively static data such as boundaries, points of interest, average incomes, home prices, and other property attributes, as well as more dynamic data such as traffic, weather, and footfall data.

The process of data enrichment consists of three stages:

- **Data requirements assessment.** During this stage, data analysts engage with data consumers to understand what data is needed for their specific use cases. For example, an initiative to improve traffic safety might need neighborhood location data detailing the types of streets, historical data about traffic accidents, and accessibility data such as locations of accessible pedestrian signals.

- **Selection of data sources.** In this second stage, data professionals must identify potential data providers, which can include public or open data sets (such as those published by municipalities), primary data sources (such as government agencies, universities, or private companies), aggregators (that collect data from multiple sources and add value to the collected data), or even emerging sources such as data marketplaces managed by trusted third parties.

- **Record matching and linkage.** The third stage of enrichment matches identifying information in your organization's data with records from the provided data sources. This process also establishes a link between those records and augments your organization's data with additional attributes. For example, you could match address locations to a U.S. Census block and augment your organization's data with census variables such as median income and education level.

These stages comprise an iterative process—as the data consumers implement their use cases, they will incrementally identify additional data needs, require additional data sources, and go through the linkage process each time. Streamline these iterations by partnering with a vendor whose products enable enrichment using a broad array of attributes across multiple data sets while simplifying the linkage process using a persistent, unique location identifier.

7 Leverage spatial analytics for added context

Although map-based visualizations provide an effective means for communication, location intelligence and spatial analytics encompass a much wider array of practices that extend beyond the map. Once your organization's data sets are enriched with additional attributes, there are several types of analyses that may require querying and aggregating based on a variety of dimensions (such as distance or region geometry) that provide spatial insight, such as (but not limited to):

- **Map-based visualizations,** such as map-based overlays of flood zones on top of residential areas or regions not currently served by a physical retail location
- **Aggregate analyses,** such as the number of individuals living within a particular distance from a company's physical retail locations or the average duration of a mobile call's connections to different cell towers
- **Measurement analyses,** such as the distances that members of low-income communities must travel to a supermarket or average walking distances within a set of housing developments to public transportation stops
- **Optimization analyses,** such as calculating an area for a municipality's social services or finding the best route for a circulator bus

Even though maps are useful for visually conveying analytics results, they are only one part of a more comprehensive end-user presentation such as an interactive operational dashboard. Data analysts can append the results of spatial analyses to augment existing business data sets. Business analysts can then run queries or apply filters to the data, which would simultaneously update the map view and other visualizations displayed on the dashboard. Examples might include filters associated with weather events, the service area for public transportation, or queries to assess the impacts of a failure of a power grid segment.

The benefits of spatial analytics are best accomplished using enriched data. High-quality location data enriched using a persistent, unique location identifier will improve spatial analytics and intelligently inform the desired results.

Afterword: Power data-driven decision-making

Data enrichment is a core competency for any organization that wants to leverage location data to provide a broader context for spatial analytics. However, the level of effort associated with continued maintenance and upkeep of reference location data can be overwhelming, especially when there is a need for subject matter expertise to ensure accuracy and consistency, and as the number of data sources containing (or impacting) location data continues to grow.

Engage with a vendor or service provider with domain expertise, processes, and scalable platforms to ensure quality data, especially when it comes to location data. Establishing a relationship with a data integrity provider that can validate and verify location data as well as provide persistent, unique location identifiers can ensure organizational data quality, improve data linkage and enrichment, and allow your business clients to focus on their business and know they are making decisions using trusted data.

About our sponsor

precisely

Precisely is a global leader in data integrity, providing accuracy and consistency in data for 12,000 customers in more than 100 countries, including 99 of the *Fortune* 100. Precisely's data integration, data quality, data governance, location intelligence, and data enrichment products power better business decisions to create better outcomes. Learn more at www.precisely.com.

About the author

David Loshin,
president of Knowledge Integrity, Inc., (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices as the author of numerous books and papers on data management, including *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* and *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is a frequently invited speaker at conferences, web seminars, and sponsored websites and channels. David is also the program director for the Master of Information Management program at the University of Maryland's College of Information Studies.

David can be reached at loshin@knowledge-integrity.com.

About TDWI Research

TDWI Research provides industry-leading research and advice for data and analytics professionals worldwide. TDWI Research focuses on modern data management, analytics, and data science approaches and teams up with industry thought leaders and practitioners to deliver a broad and deep understanding of business and technical challenges surrounding the deployment and use of data and analytics. TDWI Research offers in-depth research reports, commentary, assessments, inquiry services, topical conferences, and strategic planning services to user and vendor organizations.

About TDWI Checklist Reports

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

tdwi

Transforming Data
With Intelligence™

A Division of 1105 Media
6300 Canoga Avenue, Suite 1150
Woodland Hills, CA 91367

E info@tdwi.org

tdwi.org

Email requests or feedback to info@tdwi.org.

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.