PROPUESTA DE GUÍA METODOLÓGICA PARA EL MANEJO DE PROBLEMAS EN LA ESTANDARIZACION Y CALIDAD DE DATOS DE DIRECCIONES URBANAS EN COLOMBIA

DANIEL FELIPE RIVAS BURBANO

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN

TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

MAESTRÍA EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

MEDELLIN

2016

PROPUESTA DE GUÍA METODOLÓGICA PARA EL MANEJO DE PROBLEMAS EN LA ESTANDARIZACION Y CALIDAD DE DATOS DE DIRECCIONES URBANAS EN COLOMBIA

DANIEL FELIPE RIVAS BURBANO

Trabajo de grado para optar al título de

Magister en Tecnologías de la Información y Comunicación

Asesor

IVAN AMON URIBE

Magister en Ingeniería

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN

TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

MAESTRÍA EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

MEDELLIN

2016

DECLARACIÓN ORIGINALIDAD

"Declaro que esta tesis (o trabajo de grado) no ha sido presentada para optar a un título, ya sea en igual forma o con variaciones, en esta o cualquier otra universidad". Art. 82 Régimen Discente de Formación Avanzada, Universidad Pontificia Bolivariana.

FIRMA AUTOR (ES) _____

Medellín, 31 de marzo de 2016

A la memoria de mis padres, Orlando y Dalia Janeth...

AGRADECIMIENTOS

Mi más profundo agradecimiento a Dios, a mis padres y a la vida por darme el privilegio de estar vivo y tener el arte de orientarme hacia la excelencia personal, académica y profesional.

De igual manera, agradezco profundamente a la Universidad Pontificia Bolivariana por brindarme la oportunidad de cursar mis estudios de Maestría en Tecnologías de la Información y la Comunicación en los Programas de Formación Avanzada de Postgrados en la sede de Medellín.

En especial, mis más sinceros agradecimientos al Magister Iván Amón Uribe por tener la disposición incondicional de dirigir este trabajo y por estar presente en todo el desarrollo del mismo.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

La nomenclatura urbana de una población permite identificar y ubicar geográficamente aquellos elementos constitutivos de una ciudad, tales como son las edificaciones, lotes, vías y direcciones residenciales. Por diversos motivos, es frecuente que los registros de direcciones almacenadas por las organizaciones, tengan en su interior errores de calidad de datos. El presente proyecto pretende identificar la problemática de la nomenclatura urbana en Colombia y elaborar una guía metodológica para el manejo de problemas en las direcciones urbanas, con el propósito de mejorar la estandarización y calidad de datos de las direcciones residenciales en las fuentes de datos empresariales.

PALABRAS CLAVE: Nomenclatura urbana, Direcciones, Estandarización, Calidad de Datos.

ABSTRACT

Urban nomenclature allows identifying and geographically locates those constituent elements from a city such as buildings, lots, roads and urban addresses. Is very common that address stored by organizations have data quality issues. This project aims to identify the urban nomenclature issues in Colombia and develop a methodological guide to handle urban addresses problems in order to improve the standardization and urban address data quality on enterprise data sources.

KEY WORDS: Urban nomenclature, Addresses, Standardization, Address Matching, Data Quality.

1 INTRODUCCIÓN

En Colombia existen unas características dentro de la nomenclatura urbana que pueden ocasionar errores al momento de registrar direcciones residenciales en las bases de datos empresariales (Portafolio, 2011). Entre algunas de esas características presentes en las principales ciudades del país están las siguientes: errores de numeración de las edificaciones, problemas en la estandarización de la nomenclatura urbana y diferencias en el sentido y orientación de las vías. En una de las secciones de este trabajo, se describen en detalle esas características y un resumen de la nomenclatura urbana y vial nacional e internacional.

Por otro lado, a partir del trabajo elaborado por John Ballesteros, gerente de la empresa Gisco S.A.S[1] quien presentó la anatomía del proceso de geo codificación en municipios colombianos (Cely & Ballesteros, 2010), se propone la idea de describir la problemática de direcciones urbanas teniendo en cuenta las múltiples falencias presentes en la nomenclatura residencial de algunas ciudades del país. Partiendo de los antecedentes mencionados, este trabajo propone una guía metodológica para orientar a las empresas colombianas en el manejo de problemas en la estandarización y calidad de datos de las direcciones residenciales almacenadas en sus fuentes de datos y además, se realizó una revisión de algunas herramientas software de calidad de datos en direcciones residenciales para evaluar sus características funcionales y no funcionales utilizando direcciones residenciales colombianas e internacionales. La guía metodológica es aplicada a un caso de estudio con direcciones reales de la ciudad de Medellín y puede utilizarse como una herramienta para que las empresas logren organizar sus fuentes de datos para que tomen mejores decisiones en sus unidades de negocios.

La organización del trabajo es la siguiente: la sección 2 describe el Planteamiento del Problema, la sección 3 presenta los Objetivos General y Específicos, la sección 4 muestra el Marco Referencial y Estado del Arte, la sección 5 explica la Metodología empleada, la sección 6 detalla la Presentación y Análisis de Resultados, la sección 7 presenta las Conclusiones obtenidas, la sección 8 propone los Trabajos Futuros y finalmente, la sección 9 muestra las Referencias utilizadas.

[1] Gisco S.A.S es una empresa ubicada en la ciudad de Medellín que ofrece productos y servicios de Referenciación Geográfica, Cartografía y Mapas. Pagina Web: http://www.mygisco.com/home/

2 PLANTEAMIENTO DEL PROBLEMA

2.1 Problema

En Colombia, el relieve y la geomorfología están llenos de accidentes geográficos que ocasionan diferencias en el sentido y orientación de las vías principales (calles y carreras) de algunas de las principales ciudades del país. Además, la generación de otras vías alternas adicionales a las vías principales, tales como avenidas, transversales, circulares y pasajes, generan algunos inconvenientes en el momento de estandarizar la nomenclatura urbana a nivel nacional (Rosa, 2014), (Vargas & Horfan, 2013) y (Cely & Ballesteros, 2010).

Por otro lado, es frecuente observar diferentes problemas de calidad de datos que complican a las organizaciones a la hora de ubicar o hacer entregas a sus clientes, empleados, proveedores, haciendo que se pierda dinero y eficiencia organizacional (Amon, 2014). La empresa Gisco S.A.S lleva 14 años recorriendo el país palmo a palmo haciendo levantamiento en campo de información de las direcciones urbanas y orientación de las vías para ofrecer servicios de referenciación geográfica y es testigo de primera mano de las dificultades del sistema de nomenclatura en Colombia y de los múltiples problemas que presentan las organizaciones con sus datos de direcciones (Ballesteros, 2014).

Esta situación es corroborada en (Portafolio, 2011) cuando refiriéndose a los datos de las direcciones expresan que "…*es muy habitual que las empresas posean problemas de calidad de datos y problemas en la estandarización de las direcciones urbanas de sus clientes*". Cuando las direcciones tienen problemas de calidad de datos no es posible elaborar búsquedas acertadas de direcciones válidas en una ciudad. Además, el no tener direcciones confiables, influirá en la toma de decisiones en los procesos operativos de cada unidad de negocio de la empresa. En el caso específico de direcciones incorrectas, las empresas tendrán problemas en el momento de contactar clientes, la distribución de los productos no se realizará de una manera eficiente y se generará costos y re-procesos adicionales que afecten el sostenimiento y rentabilidad a largo plazo (Moore, 2007).

Como se evidencia con el estado del arte, la problemática de las direcciones urbanas en Colombia ha sido muy poco documentada. En este trabajo se propone la elaboración de un documento que guíe a las organizaciones para solucionar los problemas de calidad de datos de las direcciones que almacenan, con el fin de reducir los problemas que esto le acarrea al negocio.

2.2 Justificación

Como se indicó en el apartado anterior, la falta de calidad en los datos de direcciones acarrea dificultades y sobrecostos a las organizaciones. Aquellas direcciones urbanas que han sido debidamente estandarizadas serán el pilar fundamental para generar una base de conocimiento que pueda utilizarse en aplicaciones de SIG (Sistemas de Información Geográfica) y les permita a las empresas localizar aquellos clientes significativos para sus áreas de negocios (P. a. Zandbergen, 2011). Además de la estandarización, esto es, de que una dirección tenga la sintaxis correcta, es posible también que tenga problemas de exactitud, comenzando porque la dirección puede ni siquiera existir. En muchos otros países, se cuenta con bases de datos de direcciones que el gobierno provee con lo cual se puede verificar la existencia de una dirección, pero este no es el caso de Colombia.

A pesar de ser un problema que padecen muchas organizaciones del país, no se encontró evidencia de trabajos de investigación que aborden el asunto exhaustivamente y que orienten a las organizaciones colombianas para mejorar la calidad de sus direcciones. Por tal motivo, el aporte del presente proyecto está en identificar la problemática de la nomenclatura urbana en Colombia y elaborar una guía metodológica para el manejo de problemas en las direcciones que permita orientar a las empresas en la estandarización y calidad de datos de las direcciones residenciales que están almacenadas en sus fuentes de datos.

3 OBJETIVOS

3.1 Objetivo General

Proponer una Guía Metodológica para el manejo de problemas en la estandarización y calidad de datos de direcciones urbanas en Colombia

3.2 Objetivos Específicos

- Identificar la problemática de las direcciones y nomenclatura urbana en Colombia.
- Caracterizar herramientas software de calidad de datos para utilizarlas en la nomenclatura urbana colombiana.
- Construir una guía metodológica para depurar los problemas de las direcciones urbanas en Colombia.
- Realizar un caso de estudio de estandarización y calidad de datos para identificar posibles direcciones inválidas en la ciudad de Medellín.

4 MARCO REFERENCIAL

4.1 Marco contextual

El Instituto Geográfico Agustín Codazzi (IGAC) es la entidad encargada de elaborar el catastro nacional de las propiedades inmuebles y ha adoptado una nomenclatura numérica para la identificación de predios en las diferentes vías principales y alternas de las ciudades. Entre otras funciones, el IGAC también es la entidad encargada de producir los mapas oficiales y la cartografía del país (IGAC, 2005).

En Antioquia, la Dirección de Sistemas de Información y Catastro elaboró un manual de reconocimiento predial con el propósito de obtener el reconocimiento predial para los sectores urbano y rural, con el fin de unificar criterios en los procesos de formación, actualización y conservación catastral, teniendo en cuenta las características de las vías principales y alternas (Planeacion, 2010).

4.2 Marco conceptual

La *Nomenclatura Urbana* permite determinar las vías y edificaciones que pertenecen a una ciudad o población. A través de la nomenclatura, es posible ubicar y señalizar los accesos a las edificaciones y lotes identificando las vías próximas y adyacentes. La nomenclatura urbana está dividida en la nomenclatura vial y nomenclatura predial (domiciliaria). La *nomenclatura predial* describe la numeración de predios, residencias y domicilios. La placa predial la conforman dos valores numéricos separados por un guión (Ejemplo: 34 - 56). El primer valor hace referencia a la vía de menor denominación que delimita la cuadra del acceso al predio. El segundo valor corresponde a la distancia en metros entre la esquina formada por la intersección de la vía de menor denominación y la vía sobre la cual se encuentra el predio y el acceso principal del predio. De otra parte, *la nomenclatura vial* permite la identificación y numeración de la trama vial compuesta por calles, carreras, diagonales, transversales, avenidas, entre otras (Camacho & Tellez, 2009).

Ahora bien, con las direcciones se realizan procesos que apuntan a lograr direcciones de calidad.

El proceso de *Address Standarization* (estandarización de una dirección), es un proceso en donde se prepara la dirección en un formato conocido corrigiendo los errores de escritura para estructurar y especificar una forma normalizada de escribir la dirección (Ranzijin, 2013).

El proceso conocido en inglés como *Address Matching* o en español como *concordancia de una dirección*, es un proceso que compara una dirección o una tabla de direcciones con los atributos de direcciones de un conjunto de datos de referencia para determinar si una dirección en particular está dentro de un rango de direcciones asociado con una característica del conjunto de datos de referencia. Si una dirección esta dentro del rango de características de la dirección, es considerada como una coincidencia (match) y la localización es recuperada (Farvacque-Viitkovic & Godin, 2005).

Los procesos anteriores no son lo mismo que la *Geo codificación* la cual es el proceso que determina las coordenadas geográficas (por ejemplo: latitud – longitud) a partir de la información como direcciones urbanas, puntos de interés, etc. Estas coordenadas geográficas luego pueden ser utilizadas para localizar elementos en un Sistema de Información Geográfico (Vargas & Horfan, 2013).

4.3 Estado del arte

El estado del arte fue dividido en 3 secciones teniendo en cuenta aquellos trabajos relacionados con los objetivos específicos propuestos:

4.3.1 Direcciones residenciales y la nomenclatura urbana en Colombia

El Banco Internacional para la Reconstrucción y el Desarrollo presenta un documento completo de Nomenclatura y Gestión Urbana que describe las experiencias de nomenclatura en varias ciudades de África (Farvacque-Viitkovic & Godin, 2005). Además propone un Manual de Nomenclatura y direcciones urbanas donde identifica las diferentes tareas y fases que debe tener la señalización y numeración de direcciones a través de un proceso de codificación de vías y sectorización de las ciudades.

El MEN (Ministerio de Educación Nacional) con el apoyo del IGAC busca elaborar el diseño e implementación de un Sistema de Información Geográfica del Sector Educativo, denominado SIG_MEN. El informe aborda cinco casos de nomenclaturas de direcciones basadas en: La estructura de la malla vial, Estructura Barrio-Manzana-Predio, Malla vial / Barrio-Manzana-Predio, Malla vial / Barrio, Sitios de interés (Camacho & Tellez, 2009). Este proyecto se encuentra en el marco del Convenio 269 de 2008 y busca especificar un modelo de estandarización de direcciones teniendo en cuenta la organización de los campos del código CUNU (DACD), la estandarización de abreviaturas de la Circular 300/01 (IGAC) y la Resolución 166/04 (MEN). Así mismo, el MEN en el anexo 2 de la resolución 166 de 2003, establece las abreviaturas de los nombres de las instituciones educativas y la nomenclatura correspondiente a la placa de los centros educativos.

En la propuesta de (Cely & Ballesteros, 2010) se genera una anatomía de geo codificación con el propósito de generar aplicaciones informáticas para la geo codificación de las direcciones en Colombia, denominados geo codificadores "tropicalizados". Utilizan como geo codificador a Address Locator de ArcGIS y como caso un municipio colombiano (Sabana Larga). El autor concluye que para localizar una dirección es necesario tener en la cuenta no solamente la dirección postal, sino también elementos como tipos de referencias directas o indirectas de lugares tales como construcción de nombres, códigos postales, códigos de área telefónica, etc.

En el estudio presentado por (Vargas & Horfan, 2013) el autor realiza un estudio empírico comparativo de los diferentes procesos de geo codificación que se realiza en la Alcaldía de Medellín para la georreferenciación de información en la ciudad de Medellín. Se analiza los métodos determinísticos y probabilísticos que permiten por medio de la estandarización y normalización de las direcciones un resultado enmarcado en un entorno espacial que cumpla con criterios de calidad tanto en el porcentaje de acierto de las direcciones encontradas como en la exactitud posicional del resultado. El proceso de geo codificación probabilístico es realizado con GEOCODING-ArcGIS y el de geo codificación determinístico con GEOCOD-Medellín. El estudio determina que los métodos determinísticos para la geo codificación de direcciones son más efectivos que los métodos probabilísticos.

4.3.2 Herramientas para calidad de datos en direcciones residenciales

Una medida generalmente aceptada en la industria para medir los líderes del mercado en tecnología son los estudios realizados por la firma Gartner (Friedman & Judah, 2013). En su cuadrante mágico de herramientas de calidad de datos, presenta las empresas líderes del mercado en herramientas informáticas para el manejo de calidad de datos y aborda el tema de detección de errores en campos de texto libre incluyendo direcciones. Estas herramientas informáticas permiten mejorar la calidad de los datos, ofrecen perfilamiento de datos, se adaptan a diferentes tipos de problemas en los datos de tipo numérico o texto y algunas poseen algoritmos de detección de duplicados no idénticos. Las herramientas que destaca la publicación son fuertes en la detección de diferentes tipos de problemas, pero en general cada fabricante ofrece una herramienta diferente para el manejo de calidad de datos en direcciones.

Por ejemplo, en el caso de los Estados Unidos, algunas herramientas validan una dirección para envíos de correspondencia y están certificadas por la oficina postal U.S.P.S., quien es la entidad encargada de las direcciones postales de todo el país y su información se encuentra previamente actualizada y validada. A partir de esta base de datos de direcciones, las herramientas pueden utilizarla para comparar y determinar si la dirección es válida o no, en caso de no ser exacta se puede aproximar y realizar la validación con un grado de probabilidad (Xu, Flexner, & Carvalho, 2012).

En Colombia, la entidad encargada es Código Postal que junto con 4-72 tiene disponible una página Web donde puede consultar y validarse las

direcciones residenciales de las principales ciudades del país a través de la búsqueda del código postal y utiliza la codificación vigente del DANE (Departamento Administrativo Nacional de Estadística). En la búsqueda se especifica el Departamento / Municipio / Centro Poblado y la dirección residencial en texto libre (MinTIC, 2014).

El proceso de validación de una dirección, conocida también como coincidencia de una dirección ó "Address Matching", se refiere a la tarea de búsqueda de una dirección dentro de una referencia válida y puede emplearse para la obtención de coordenadas de latitud/longitud en procesos de geo codificación (Wu & Rathswohl, 2010). Por ejemplo, para realizar una validación de direcciones eficaz se debe seguir un proceso que incluye: limpieza o corrección de errores en texto, segmentación, estandarización y búsqueda referencial. Algunas de estas fases son descritas en (Vargas & Horfan, 2013), donde se realiza un estudio empírico comparativo de los diferentes procesos de geo codificación realizados por la Alcaldía de Medellín para la georreferenciación de direcciones residenciales en la ciudad de Medellín. El estudio determina que los métodos determinísticos para la geo codificación de direcciones son más efectivos que los métodos probabilísticos.

El proceso de segmentación de direcciones, consiste en separar diferentes bloques dentro de una cadena de texto. Con programas que comparan el texto con el formato prestablecido no suelen tener buenos resultados, debido a problemas de bloques en desorden, espacios adicionales o separadores no estándar. Dentro de la búsqueda bibliográfica varios autores parecen coincidir que una buena manera para realizar la segmentación es mediante los Modelos Ocultos de Markov (HMM) como los descritos en (Viola & Narasimhan, 2005) y (Borkar, Deshmukh, & Sarawagi, 2001), ya que es una técnica posible de implementar computacionalmente, con buenos resultados y es resistente a varios de estos errores en las cadenas de texto.

Para enfrentar problemas de tipo ortográfico o tipográfico, se pueden utilizar técnicas para detección de errores en texto libre, algoritmos de "record linkage" en donde el texto de comparación es uno de los datos de la dirección (ciudad, barrio, cada bloque de la dirección), y dependiendo del dato que se quiera buscar, hay algoritmos más o menos efectivos. En los estudios de (Li, Wang, & Mei, 2010) y (Ranzijin, 2013), uno de los más recomendados es el algoritmo de distancia de edición ("edit distance"). La utilización de estas técnicas permite una buena preparación del texto para una posterior estandarización de las direcciones.

Algunas de las herramientas más utilizadas en la validación geográfica de direcciones, pueden trabajar de dos maneras principalmente: pueden utilizar la búsqueda en bases de datos completas con todas las direcciones disponibles o por medio de georreferenciación. El primero es un método poco escalable debido a los costos de actualización pero es muy eficaz, incluso si no se tiene alguna dirección con coincidencia exacta, se puede aproximar con un alto grado de confiabilidad si existe o no. El segundo ha tomado fuerza ya que aunque es un método que es menos preciso y requiere un buen número de recursos, no depende de recursos externos, sino que el mismo modelo pueda aplicarse en diferentes países. Por ejemplo, (Goldberg et al., 2013) sugieren algunas de las características que debe tener en la cuenta una herramienta de geo codificación: la escalabilidad del sistema, la integración con diferentes fuentes de información geográfica, la capacidad de estandarización y los algoritmos de comparación que contiene. En (Schootman et al., 2007) se evalúa la tasa de acierto de una de las herramientas más utilizadas para geo codificar direcciones y cómo la corrección de errores y la estandarización previa de estos elementos ayuda para mejorar el índice de acierto.

En el trabajo de (Yang, Bilaver, Hayes, & Goerge, 2004), utilizan una serie de direcciones para probar la efectividad de un conjunto de aplicaciones software en el proceso de geo codificación, concluyendo que los resultados son diferentes en cada herramienta y que cada fuente de información geográfica puede influir en esos resultados finales. En (Kravets & Hadden, 2007) explican algunas de las causas más comunes de errores en la geo codificación, donde uno de los factores importantes que influyen en la exactitud es el grado de urbanización del sitio. Por otro lado, en (Whitsel et al., 2006) describen la importancia de enriquecer la dirección con datos relacionados o “metadatos” para lograr una mejor aproximación y mayor certeza en el proceso de geo codificación.

El estudio propuesto en (P. A. Zandbergen, 2008) muestra que dependiendo de las características de la dirección puede variar el resultado, si bien el resultado de la geo codificación no puede ser exacto, puede llegar a tener una probabilidad de acierto cercana al 90% en el mejor de los casos. Esta probabilidad disminuye cuando la región objetivo presenta accidentes geográficos o zonas de difícil acceso terrestre o de baja urbanización. Dependiendo de estas características del terreno los resultados pueden llegar a nivel de calle, parcela o punto geográfico idealmente.

4.3.3 Manejo de problemas en direcciones residenciales

En la tesis de maestría de (Amon, 2010), el autor elabora una serie de guías metodológicas que apoyan a los analistas de datos en la selección de técnicas para depuración y limpieza de datos en cuanto a detección de duplicados, corrección de valores faltantes y detección de valores atípicos. Se describen las técnicas empleadas, posteriormente propone unos criterios y métricas de evaluación de éstas técnicas, luego realiza un diseño experimental de evaluación y finalmente interpreta los resultados obtenidos. En el trabajo no está descrito el manejo de errores sobre direcciones residenciales.

En la propuesta de (Batini & Cappiello, 2009), los autores realizan una descripción sistemática y comparativa de metodologías que apoyan la selección, personalización y aplicación de la evaluación de calidad de datos. Las metodologías son comparadas a través de varias dimensiones que incluyen las fases metodológicas, estrategias y técnicas, dimensiones de la calidad de datos, tipos de datos y finalmente los tipos de sistemas de información relacionados con cada metodología.

En el artículo elaborado por (Copano Ortiz, 2014), el autor describe las dificultades en la generación de un sistema integrado de direcciones residenciales en España. Toman como referencia un "Modelo de Direcciones de la Administración General del Estado" donde describen las características de un Modelo Único de Direcciones Normalizadas y Geo referenciadas de los inmuebles a nivel nacional. El autor propone una serie de consideraciones a tener en la cuenta para la captura de información de las direcciones residenciales y concluye que es necesaria una normativa específica de direcciones que determine cuáles son los organismos gubernamentales encargados de gestionar estas dificultades en las direcciones.

En el trabajo de (Fang, Yu, & Zhao, 2010), los autores proponen un esquema de datos unificado de direcciones, analizando aquellas direcciones que son frecuentemente utilizadas en China. Este modelo describe y almacena todos los tipos de elementos de direcciones, tales como ciudades, municipios, poblaciones y sus relaciones. Basados en este esquema, las direcciones pasan por un proceso de división, estructuración y posterior búsqueda a partir de un conjunto de reglas y un algoritmo de Address Matching.

Los autores describen la secuencia de pasos empleada para la estandarización y búsqueda de la dirección, pero no son descritos los algoritmos empleados en el manejo de errores en las direcciones de entrada, ni los resultados alcanzados.

En el articulo de (Pérez Machado, 2008), el autor describe unas características del uso de la Geo codificación aplicados a las ciudades de São Paulo y Barcelona y realiza un listado de condiciones necesarias para la implementación de la geo codificación. Entre las condiciones sugieren que el proceso de separación de las direcciones en elementos específicos, permite tener mayor número de aciertos y menor cantidad de eventos ambiguos. Además, indica que en la experiencia práctica si una lista de direcciones es sometida previamente a un algoritmo de estandarización, el resultado obtenido en la geo codificación posterior alcanza el suceso en más del 95% de los casos.

Análisis del estado del arte

Como puede observarse en el estado del arte realizado, solo se detectaron tres trabajos realizados en Colombia relacionados con la problemática de las direcciones. El trabajo de (Vargas & Horfan, 2013) está orientado hacia el proceso de geo codificación para la georreferenciación de información en la ciudad de Medellín. En forma similar el trabajo de (Cely & Ballesteros, 2010) se orienta hacia la geo codificación haciendo un piloto con el municipio de Sabana Larga. Por último, el trabajo de (Amon, 2010) es genérico y las guías metodológicas presentadas están orientadas a la selección de las técnicas más adecuadas para detectar problemas de duplicados, valores faltantes y valores atípicos. No se detectaron trabajos orientados a la calidad de los datos de las direcciones para la realidad colombiana y mucho menos una guía que oriente a las organizaciones para solucionarlos.

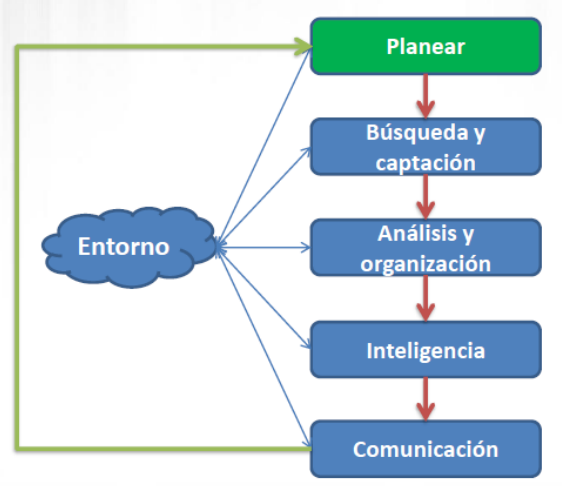
El caso colombiano se diferencia de otros países porque además de los accidentes geográficos que ocasionan anomalías en el sentido y orientación de las vías principales (calles y carreras), existen problemas en la nomenclatura urbana propuesta por el IGAC ya que en otras latitudes hay instituciones internacionales como la UPU (Universal Postal Union) en Europa y la USPS (United States Postal Service) de los Estados Unidos, que proveen estándares y bases de datos de direcciones postales y residenciales de una manera más completa, detallada y de acceso público.

5 METODOLOGÍA

La presente propuesta es desarrollada a partir de la metodología empleada en Vigilancia Tecnológica e Inteligencia Competitiva, conceptos descritos en la tesis de maestría de (Sanchez, 2002). La Vigilancia Tecnológica puede definirse como la búsqueda, detección, análisis y comunicación de informaciones orientadas a la toma de decisiones sobre amenazas y oportunidades externas en el ámbito de la ciencia y la tecnología (Jakobiak, 1992), mientras que la Inteligencia Competitiva va mas allá porque incluye a la Vigilancia Tecnológica y es una herramienta que les permite a las empresas captar información del exterior, analizarla y convertirla en conocimiento para tomar decisiones con menor riesgo y poder anticiparse a los cambios (Sanchez, 2002) y (Castillo, 2007). Los conceptos descritos anteriormente están enfocados principalmente a las empresas pero las fases que se describen alrededor de la Vigilancia Tecnológica e Inteligencia Competitiva son útiles para cumplir los objetivos y productos propuestos en la elaboración de este proyecto.

La Vigilancia Tecnológica nos permite realizar un proceso organizado, selectivo y sistemático para captar información sobre ciencia y tecnología, seleccionarla, analizarla, difundirla y comunicarla, para convertirla en conocimiento (Sánchez & Tamayo, 2014). En la gráfica siguiente se describen las fases de la metodología empleada en el proceso de Vigilancia Tecnológica (Figura 1).

Figura 1. Fases del proceso de Vigilancia Tecnológica



Fuente (Sánchez & Tamayo, 2014)

Planeación:

En esta primera fase se identificaron los factores críticos de vigilancia como los tópicos claves y las preguntas claves, traducidas en aquellas necesidades del ¿Por qué? ¿Para qué? ¿Para quién? ¿En qué?, se realiza el proceso investigativo.

Búsqueda y Captación:

Se realizó una "inmersión básica", con el propósito de identificar los conceptos y los términos, tanto los coloquiales como los técnicos de la temática de investigación. Es decir, selección de términos genéricos en inglés y español (estandarización de una dirección, nomenclatura urbana, address matching, address standarization, calidad de datos, geo codificación, entre otros). Además de la selección de términos secundarios (nomenclatura vial, nomenclatura domiciliaria, limpieza de datos, entre otros). Posteriormente, se realiza el proceso de búsqueda de fuentes de información confiables como libros, artículos, patentes, noticias, con el propósito de elaborar el estado del arte del tema de investigación.

Finalmente, se realizaron entrevistas a consultores expertos en la temática que describan la problemática de las direcciones urbanas en el país.

Análisis y Organización:

La documentación recolectada se administró a través de una aplicación de trabajo colaborativo utilizada en investigación y en la academia. En este caso particular, a través de la aplicación Mendeley se hizo una correcta gestión de referencias, búsqueda de palabras clave, así como la identificación de autores, instituciones y grupos de investigación, con el fin de conocer las redes de conocimiento que abordan investigaciones similares.

Inteligencia:

Creación de la propuesta de investigación del proyecto, construida a partir de la observación, el análisis y el tratamiento de la información recolectada en la fase anterior. Los datos e información analizados se transformaron en nuevo conocimiento, respondiendo al objetivo general y objetivos específicos formulados en la fundamentación del proyecto.

Con respecto al primer objetivo específico, la elaboración del estado del arte relacionado con la nomenclatura y estandarización de direcciones junto con las entrevistas realizadas a los consultores expertos, permitieron identificar la problemática de las direcciones y la nomenclatura urbana en Colombia.

En el segundo objetivo específico, una revisión bibliográfica de herramientas software para calidad de datos en direcciones, permitió realizar un proceso de selección de criterios y características de las herramientas software tales como accesibilidad, disponibilidad, escalabilidad, interoperabilidad, funcionalidad de corrección de errores, etc., que puedan utilizarse en la nomenclatura urbana colombiana.

En cuanto al tercer objetivo específico, la guía metodológica se construyó a partir de la revisión bibliográfica y los resultados obtenidos del proceso de estandarización y comparación de direcciones. La guía metodológica se expresó en forma de diagramas de flujo y recomendaciones que orientan a las empresas en la estandarización, calidad de datos y depuración de los problemas de las direcciones urbanas que están almacenadas en sus fuentes de datos.

Finalmente, en el cuarto objetivo específico se aplicó un proceso de limpieza de datos, detección de duplicados, valores atípicos incorrectos y demás técnicas de calidad de datos a una base de datos de direcciones de la ciudad de Medellín. Posteriormente, se realizó un proceso de estandarización de direcciones (adress standarization) y luego una identificación de posibles direcciones inválidas a través de una comparación de direcciones (address matching) en un conjunto de direcciones válidas de la ciudad de Medellín.

Comunicación:

La comunicación de la información se hará a partir de la presentación y divulgación de los resultados obtenidos en las fases anteriores y puede ser de manera escrita (informe, artículos, noticias) ó de manera oral (presentaciones, ponencias, congresos). De hecho se tiene planeado someter tres artículos a revistas científicas indexadas que son resultado del presente trabajo. El contenido de dichos artículos está inmerso en este documento en la sección 6 de presentación y análisis de resultados.

6 PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

La presentación y el análisis de los resultados es descrita manteniendo el orden de las tres secciones mencionadas en el estado del arte. Cada ítem aborda el cumplimiento de los objetivos específicos propuestos.

A continuación se detallan los resultados del primer objetivo específico: "*Identificar la problemática de las direcciones y nomenclatura urbana en Colombia".*

6.1 Direcciones residenciales y la nomenclatura urbana en Colombia: Particularidades y Diferencias

En el pasado, los habitantes de las grandes poblaciones optaban por colocar los nombres de próceres, santos o personas destacadas para distinguir las principales vías y localizaciones geográficas de sus ciudades. Esta convención, además de colocar puntos de referencia, orientar y guiar a la población, permitía recordar y rendir homenaje a aquellos personajes históricos propios de cada región.

Con el paso del tiempo, estas convenciones se volvieron insuficientes debido a la necesidad de ubicar y localizar puntos geográficos de una manera más precisa, por tal motivo fue necesario disponer de una nomenclatura que fuera fácil de utilizar y permitiera la identificación inmediata de vías y puntos geográficos. La nomenclatura urbana surge del crecimiento de las principales ciudades de Europa y Asia con el propósito de determinar la distribución de las propiedades y partiendo de algunas necesidades orden y seguridad pública. Es así, que entre los años de 1766 y 1768, empezó la primera numeración de predios en la comunidad de Crest (Francia) (Farvacque-Viitkovic & Godin, 2005).

En Colombia, la nomenclatura urbana es generada a partir de la identificación de vías descritas en los planes de ordenamiento territorial y la organización de la malla urbana propia de cada región. Las principales ciudades del país se caracterizan por tener una gran cantidad de accidentes geográficos que generan diferencias tanto en la nomenclatura como en el sentido y orientación de las vías principales y las vías alternas. En este escenario, es muy habitual que las empresas tengan problemas de calidad de datos y/o en la estandarización de direcciones residenciales de sus

clientes (Portafolio, 2011). Esta situación, puede afectar negativamente la toma de decisiones en las unidades operativas de la empresa.

Cuando estas direcciones tienen problemas de calidad de datos, no es posible elaborar búsquedas de direcciones válidas en las poblaciones y en el caso específico de direcciones incorrectas, las empresas tendrán problemas en el momento de contactar clientes, la distribución de los productos no se realizará de una manera eficiente y se generará costos y re-procesos adicionales que afecten el sostenimiento y la rentabilidad a largo plazo (Moore, 2007).

De acuerdo con lo anterior, es esencial para el entorno empresarial reconocer las características presentes en la nomenclatura urbana nacional que puedan ocasionar errores en el registro de sus direcciones residenciales. De esta manera, las empresas pueden tomar las medidas necesarias y realizar un proceso adecuado de validación, corrección y estandarización de direcciones residenciales que les permita tener búsquedas más precisas.

6.1.1 Conceptos Generales de Nomenclatura

La señalización adecuada de calles, carreras y direcciones permite una distribución eficiente de lotes y predios en los planes de ordenamiento territorial. De esta manera, el Estado tiene un recaudo de impuestos más eficiente a través de edificaciones, y una mejora en la gestión de los recursos y en las obras de inversión en las localidades urbanas. Las empresas públicas y privadas, tendrán un mejor funcionamiento de los servicios facilitando el pago de servicios públicos como agua, alcantarillado y electricidad. La información de las direcciones puede ser recolectada para generar una base de conocimiento que pueda utilizarse en aplicaciones para beneficio de la comunidad, como por ejemplo aplicaciones de SIG (Sistemas de Información Geográfica) y así, la población tiene la posibilidad de localizar, de una manera más sencilla, aquellos lugares significativos como clínicas, iglesias, colegios, estaciones de policía, universidades, etc. De acuerdo con lo anterior, resulta de vital importancia contar con una nomenclatura urbana organizada, estandarizada y actualizada para los diferentes actores de la sociedad (estado, empresas y población en general). A continuación, se describen algunos conceptos generales teniendo en cuenta sistemas de nomenclatura urbana utilizados en varias ciudades europeas, africanas, norte americanas y latinoamericanas (Farvacque-Viitkovic & Godin, 2005) y (Farvacque-Viitkovic & Chavez, 2011):

Nomenclatura urbana: La nomenclatura urbana es un elemento fundamental de orden y planeación que facilita la ubicación de los predios (viviendas, edificaciones, residencias, domicilios) y vías urbanas en una ciudad. Permite la señalización de las vías de acceso y la identificación de edificaciones con signos numéricos y alfanuméricos. La nomenclatura urbana se divide en dos partes: nomenclatura vial y nomenclatura predial o domiciliaria.

Nomenclatura Vial: Permite la identificación y numeración de la trama vial compuesta por calles, carreras, diagonales, transversales, avenidas, entre otras (Figura 2).

Nomenclatura Predial o Domiciliaria: permite la identificación y numeración de predios y se puede dividir en tres tipos: Numeración Secuencial, Numeración Métrica y Numeración Decamétrica (Figura 3).

Figura 2. Nomenclatura Vial en el mundo

Fuente https://commons.wikimedia.org/wiki/

Figura 3. Nomenclatura predial en el mundo

Fuente https://commons.wikimedia.org/wiki/

Numeración secuencial:

Los predios están numerados de manera secuencial y alternada, impares a la izquierda y pares a la derecha (Figura 4). La Tabla 1 presenta las ventajas y desventajas de este tipo de numeración.

Tabla 1. Ventajas y Desventajas del Sistema de Numeración Secuencial

Ventajas	**Desventajas**
- Permite numerar edificaciones aunque estén aisladas de la vía.	- Dificultad para recordar números de 3 y 4 cifras cuando las distancias son muy extensas.
- Es pertinente utilizar cuando hay un número determinado de edificaciones.	- La numeración puede no ser secuencial ocasionando confusión a los habitantes de otros lugares
- No es necesario colocar la palabra “bis” si existe una nueva edificación.	
- Saber la distancia facilita la labor de instalación y mantenimiento de los servicios públicos.	

Fuente: Elaboración propia

Figura 4. Sistema de Numeración Secuencial

Fuente: Elaboración propia

Numeración Métrica:

Corresponde a la distancia en metros desde la vía principal hasta la edificación y se mantiene la numeración impar al lado izquierdo y par al lado derecho (Figura 5).

Figura 5. Sistema de Numeración Métrica

Fuente: Elaboración propia

La Tabla 2 presenta las ventajas y desventajas de la numeración métrica.

Tabla 2. Ventajas y Desventajas del Sistema de Numeración Métrica

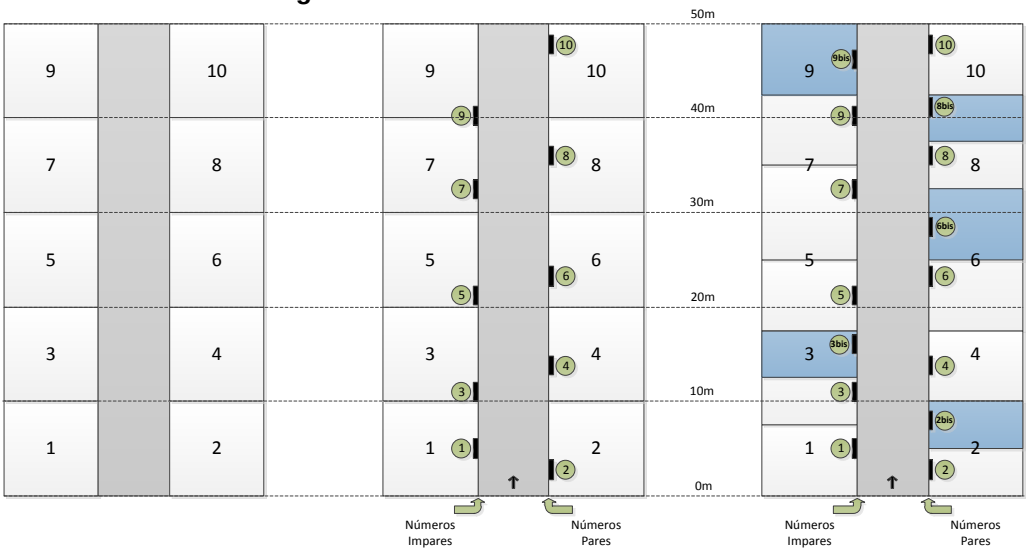
Ventajas	**Desventajas**
- Combina la facilidad de lectura de la numeración secuencial y la posibilidad de medir las distancias de la numeración métrica	- Este sistema de numeración es poco utilizado y está poco difundido, por tal motivo no hay un amplio conocimiento para aplicarlo.
- Saber la distancia facilita la labor de instalación y mantenimiento de los servicios públicos.	- Si existen dos edificaciones en el mismo segmento, la numeración se diferencia por letras. Por ejemplo: 4A y 4B.

Fuente: Elaboración propia

Numeración decamétrica:

La numeración de edificaciones se realiza teniendo en cuenta segmentos de distancias iguales (Figura 6). Por ejemplo, la vía se divide en tramos de 10 metros y la numeración es impar y secuencial en el lado izquierdo de la vía y la numeración es par en el lado derecho de la vía (Figura 6 y Tabla 3). La Tabla 4 presenta las ventajas y desventajas de la numeración decamétrica.

Figura 6. Sistema de numeración decamétrica



Fuente: Elaboración propia

Tabla 3. Sistema de numeración decamétrica

Segmento	Distancia (m)	Número lado izquierdo	Número lado derecho
1	0 – 10	1	2
2	10 – 20	3	4
3	20 – 30	5	6
4	30 – 40	7	8
5	40 – 50	9	10
6	50 – 60	11	12
7	60 – 70	13	14
8	70 – 80	15	16
9	80 – 90	17	18
10	90 – 100	19	20
11	100 -110	21	22

Fuente: Elaboración propia

Tabla 4. Ventajas y Desventajas del Sistema de Numeración Secuencial

Ventajas	**Desventajas**
- Combina la facilidad de lectura de la numeración secuencial y la posibilidad de medir las distancias de la numeración métrica	- Este sistema de numeración es poco utilizado y está poco difundido, por tal motivo no hay un amplio conocimiento para aplicarlo.
- Saber la distancia facilita la labor de instalación y mantenimiento de los servicios públicos.	- Si existen 2 edificaciones en el mismo segmento, la numeración se diferencia por letras. Por ejemplo: 4A y 4B.

Fuente: Elaboración propia

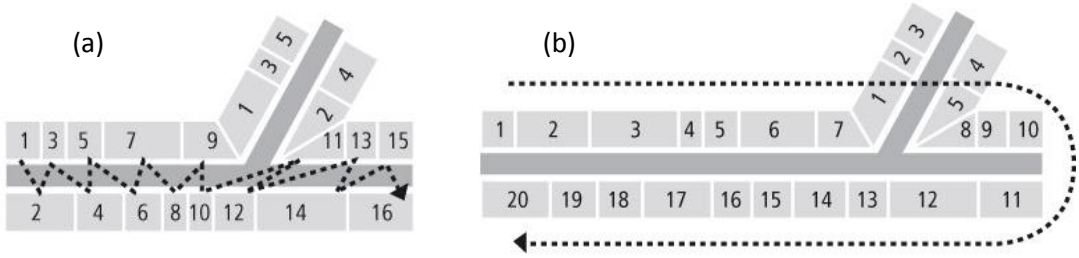
6.1.2 Historia de Nomenclatura Vial y Nomenclatura Residencial en el Mundo

En Europa existen algunas instituciones internacionales como la UPU (Universal Postal Union) quien tiene actividades encaminadas a proveer estándares de direcciones postales y residenciales. Por ejemplo, en Europa occidental y meridional la nomenclatura de predios se realiza utilizando números en orden ascendente, que pueden ser impares cuando los predios están ubicados en el lado izquierdo de la vía y números pares cuando los predios están ubicados en el lado derecho (Tantner, 2009).

En Australia y Nueva Zelanda se tiene el estándar AS/NZS 4819:2011 para la numeración de casas y predios que está basado en el mismo esquema europeo (Tantner, 2009) y (Newman & Haanen, 2011) (Figura 7a).

En algunas ciudades del Reino Unido la numeración se realiza de manera secuencial teniendo en cuenta el sentido de las manecillas del reloj. Generalmente, la numeración empieza en el lado izquierdo de la vía y continúa consecutivamente en el lado opuesto (Figura 7b).

Figura 7. (a) Nomenclatura Europea (b) Nomenclatura del Reino Unido



Fuente https://commons.wikimedia.org/wiki/)

En Japón las ciudades se dividen en pequeñas zonas numeradas, existen distritos (*ku*) que se dividen en barrios (*chome*), los cuales agrupan varias decenas de casas y forman bloques o manzanas. La numeración se realiza de acuerdo al bloque al que pertenecen y no en función de la calle. La numeración se hace según la fecha de su construcción; es decir, dos casas vecinas pueden no tener números consecutivos (Garcia, 2010) (Figura 8a).

En China la nomenclatura se caracteriza por ser como una brújula porque se utilizan los 4 puntos cardinales para identificar al país y sus principales ciudades. Esto se evidencia en los significados de los nombres: China (país

del medio), Beijing (capital del norte) y Nanjing (capital del sur). Las calles se subdividen en sectores (este, oeste, norte o sur según la orientación de la vía) y la numeración vuelve a empezar en cada sector. Por ejemplo, en Shanghái la “calle de Nankín” se subdivide en “Nanjing dong lu” o sector oeste de la calle de Nankín. Otras calles se pueden dividir en tres partes: dong lu, zhong lu, xi lu (sector este, sector central, sector oeste) (Niglio, 2014) (Figura 8b).

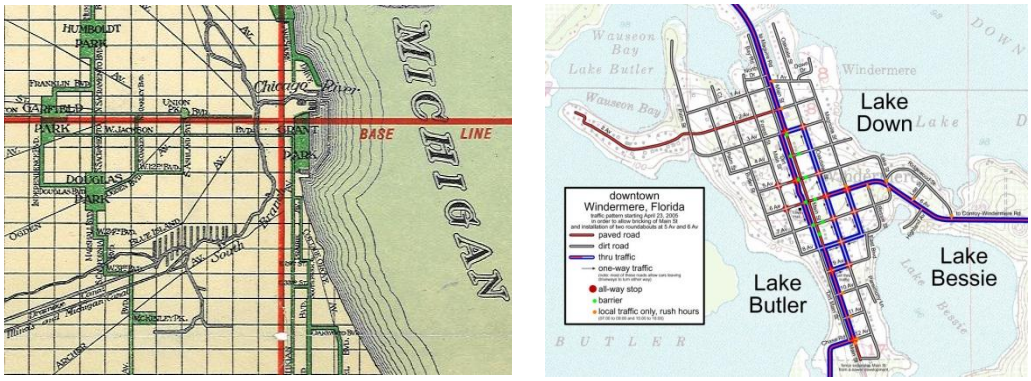
Figura 8. (a) Nomenclatura de Japón (b) Nomenclatura de China

Fuente: (Garcia, 2010) y https://commons.wikimedia.org/wiki/

En Corea del Sur se tienen dos sistemas de nomenclatura, el sistema oficial de nomenclatura llamado *Road Name Address System* que entró en vigencia en el 2011 y tiene la numeración de predios y vías de manera similar a los sistemas de nomenclatura de Estados Unidos, Canadá y Europa (Jae-un, 2013). La Nomenclatura es secuencial y alternada, los números impares van a la izquierda y los pares en la derecha. El sistema de Nomenclatura anterior es el Sistema Asiático del Este, que es el utilizado en Japón y Corea del Norte.

En Estados Unidos existen múltiples ciudades que fueron distribuidas geográficamente en forma de cuadrícula desde la época de la colonización británica. Philadelphia fue una de las primeras ciudades diseñada en forma de cuadrícula (Jackson, 1985), así como Boston, Atlanta, Chicago, Detroit, New York, Miami, Portland, San Francisco, Saint Luis, Washington (Figura 9). La organización de las ciudades en forma de cuadrícula viene históricamente de varias culturas como la babilónica, egipcia, romana y azteca. La numeración de las edificaciones comienza desde un punto inicial de partida, pueden aumentar de 100 en 100 entre cada bloque de la cuadrícula, y es común que tenga hasta cuatro o cinco dígitos. Los números de las edificaciones son pares e impares y están ubicados a cada lado de la vía.

Figura 9. Nomenclatura de Estados Unidos

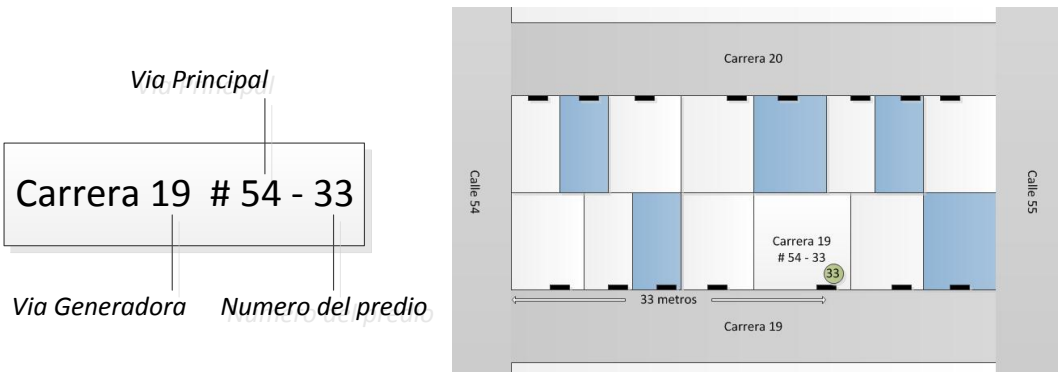


Fuente: (Wilberg, 1993) y https://commons.wikimedia.org/wiki/

6.1.3 Nomenclatura Vial y Nomenclatura Residencial en Colombia

La nomenclatura urbana colombiana se genera a partir de la identificación de vías descritas en los planes de ordenamiento territorial y la organización de la malla urbana propia de cada región. La placa predial está conformada por dos valores numéricos separados por un guión. Por ejemplo en la siguiente dirección: Carrera 19 # 54 – 33, el primer valor (en este caso 54) está constituido por la vía de menor denominación que delimita la cuadra sobre la cual se encuentra el acceso al predio. El segundo valor será el correspondiente a la distancia en metros (33 metros) entre la esquina formada por la intersección de la vía de menor denominación y la vía sobre la cual se encuentra el predio (en este caso 19) y el acceso principal del predio (Figura 10) (Camacho & Tellez, 2009).

Figura 10. Numeración en Colombia



Fuente: Elaboración propia

Se conoce como Vía Principal a la vía sobre la cual está ubicado el acceso principal al predio y como *Vía Generadora* al eje vial que tiene intersección con la vía principal donde está ubicado el predio. La vía generadora figura en la placa del predio y va acompañada de otro número que representa la distancia aproximada en metros desde el eje generador.

Hace varias décadas, en el año de 1932 en Santafé de Bogotá, el Concejo de Bogotá estableció los parámetros técnicos para la asignación de la nomenclatura para el Distrito a través del Acuerdo 7 de 1932. Este modelo de nomenclatura fue el que empezó a adoptarse en la mayoría de ciudades del territorio colombiano.

Asimismo, para el año de 1934 en la ciudad de Medellín, la Sociedad de Mejoras Públicas propuso ante el Concejo Municipal una nomenclatura numérica para Medellín similar a la de Bogotá, a través del acuerdo 253 del primero de diciembre de 1934, por el cual se adopta el plan general de nomenclatura de la ciudad (Olano & Morales, 2006).

El IGAC (Instituto Geográfico Agustín Codazzi) es la entidad encargada de elaborar el catastro nacional de las propiedades inmuebles y ha adoptado la nomenclatura numérica para la identificación de predios en las diferentes vías principales y alternas de las ciudades. El Instituto Geográfico Agustín Codazzi a través de la Circular 300 de 2001 propone las especificaciones requeridas para unificar la captura y trascripción de datos que conforman la información catastral. Este documento se elaboró teniendo en cuenta el documento "Procedimientos Generales para Codificación Catastral" publicado en 1989 dentro del proceso de sistematización de la información alfanumérica catastral (IGAC, 2005).

En Antioquia, la Dirección de Sistemas de Información y Catastro elaboró un manual de reconocimiento predial con el propósito de obtener el reconocimiento predial para los sectores urbano y rural, con el fin de unificar criterios en los procesos de formación, actualización y conservación catastral, teniendo en cuenta las características de las vías principales y alternas descritas anteriormente (Planeación, 2010).

Sentido y Orientación de vías en Colombia:

Colombia se caracteriza por tener una gran cantidad de accidentes geográficos que generan diferencias en el sentido y orientación de las vías principales (Calles y Carreras) y las vías alternas (Avenidas, Transversales, Circulares y Pasajes) en las principales ciudades del país. Estos accidentes generaron algunas diferencias en el momento de estandarizar la nomenclatura urbana a nivel nacional (Rosa, 2014), (Vargas & Horfan, 2013) y (Cely & Ballesteros, 2010). A continuación, se presenta una revisión de la orientación de las vías urbanas en varias ciudades del país, utilizando imágenes que fueron tomadas de la aplicación de escritorio de Google Earth.

En Bogotá, Medellín y Villavicencio, las Calles aumentan en sentido norte y las Carreras en sentido occidente. Las Diagonales atraviesan las Calles y las Transversales atraviesan las Carreras. Además, en la ciudad de Medellín ocurre la particularidad que existe una vía alterna denominada Circular que no existe en las otras ciudades del país (Figura 11a y Figura 11b).

En Bucaramanga y Santa Marta, las Carreras aumentan en sentido oriental y las Calles disminuyen en sentido sur. Además, las Diagonales atraviesan las Calles y las Transversales atraviesan las Carreras (Figura 12a y 12b).

En Ibagué y Pasto, las Carreras aumentan en sentido norte y las Calles aumentan en sentido oriente. Además, las Diagonales atraviesan las Carreras y las Transversales atraviesan las Calles (Figura 13a y 13b).

En Cali y Manizales, las Carreras aumentan en sentido sur y las Calles aumentan en sentido oriente. Además, las Diagonales atraviesan las Carreras y las Transversales atraviesan las Calles (Figura 14a y 14b).

En Barranquilla, las Carreras aumentan en sentido norte y las Calles aumentan en sentido occidente. Además, las Diagonales atraviesan las Carreras y las Transversales atraviesan las Calles mientras que en Cartagena las Carreras aumentan hacia el oriente y las Calles aumentan en sentido norte. Las Diagonales atraviesan las Calles y las Transversales atraviesan las Carreras (Figura 15a y 15b).

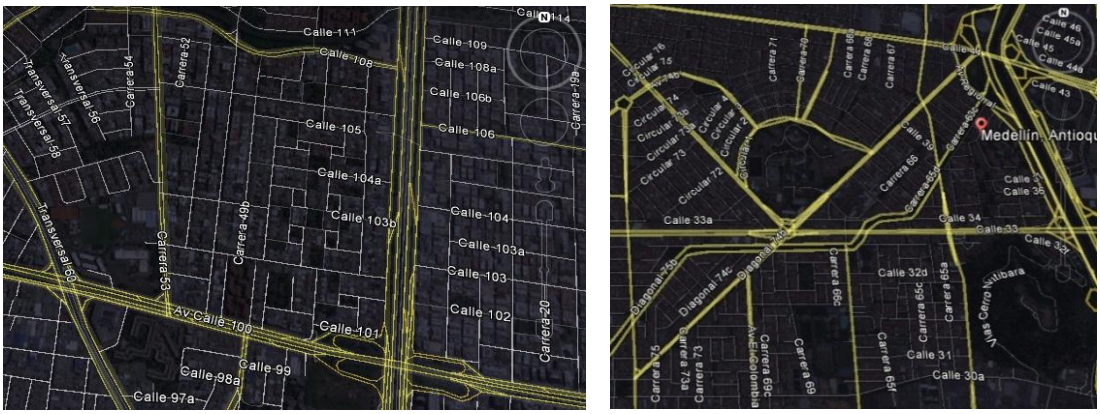
En Pereira, las Carreras aumentan en sentido sur y las Calles aumentan en sentido occidente. Además, las Transversales atraviesan las Calles y no existen las vías Diagonales mientras que en Armenia, las Carreras aumentan en sentido norte y las Calles aumentan en sentido sur. Además, las

Diagonales atraviesan las Carreras y las Transversales atraviesan las Calles (Figura 16a y 16b).

La ciudad de Cúcuta, tiene un desorden más generalizado porque tiene problemas de duplicidad y discontinuidad en su nomenclatura. Las Carreras se denominan Avenidas y van en múltiples sentidos dependiendo del cuadrante de la ciudad donde se encuentre. La Avenida 0 y la Calle 0 ó Diagonal Santander dividen la ciudad en 4 cuadrantes. El IGAC junto con la Asociación de Empresas Unidas Prestadoras de Servicios Públicos para el Área Metropolitana de Cúcuta, la Cámara de Comercio y la Alcaldía de Cúcuta, elaboraron un estudio con el fin de cambiar el 80% de la nomenclatura vial de la ciudad (Nieto, 2015; Torres, 2015).

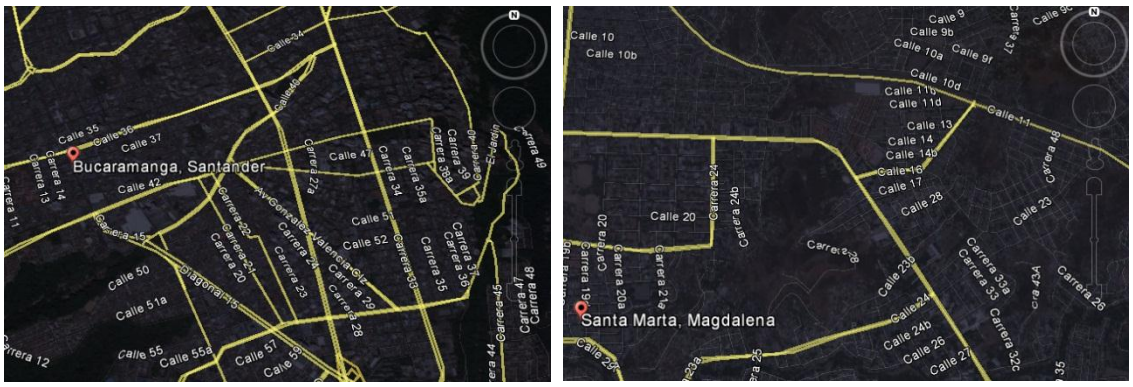
En Popayán las Carreras aumentan en sentido Occidente, las Calles aumentan hacia el Sur con la particularidad de que la ciudad ha crecido hacia el norte, por tal motivo a partir de la Calle Primera, se coloca la palabra Norte. Además, las Diagonales atraviesan las Carreras y las Transversales atraviesan las Calles (Figura 17a y 17b).

Figura 11. Orientación de vías de (a) Bogotá y (b) Medellín



Fuente https://www.google.com/earth/

Figura 12. Orientación de vías de (a) Bucaramanga y (b) Santa Marta



Fuente https://www.google.com/earth/

Figura 13. Orientación de vías de (a) Ibagué y (b) Pasto

Fuente https://www.google.com/earth/

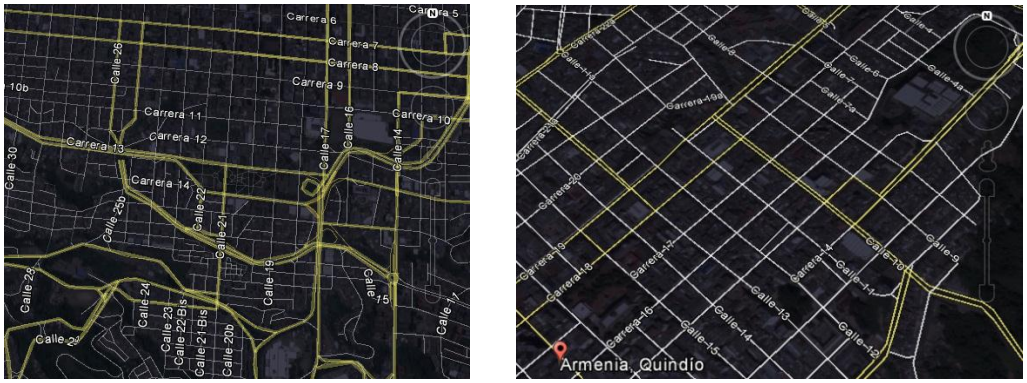
Figura 14. Orientación de vías de (a) Cali y (b) Manizales

Fuente: https://www.google.com/earth/

Figura 15. Orientación de vías de (a) Barranquilla y (b) Cartagena

Fuente: https://www.google.com/earth/

Figura 16. Orientación de vías de (a) Pereira y (b) Armenia



Fuente: https://www.google.com/earth/

Figura 17. Orientación de vías de (a) Cúcuta y (b) Popayán

Fuente https://www.google.com/earth/

Como puede observarse, los accidentes geográficos del país y la falta de planeación y coordinación, han generado algunas diferencias al momento de estandarizar la nomenclatura urbana a nivel nacional e influyen en gran medida en que la orientación de las vías principales (Carreras y Calles) y las vías alternas (Avenidas, Transversales, Circulares y Pasajes) varíe entre las diferentes ciudades colombianas.

La ciudad de Cúcuta, es la que tiene un desorden más generalizado porque tiene problemas de duplicidad y discontinuidad en su nomenclatura, pero en general en las principales ciudades del país, es evidente que la nomenclatura decamétrica corresponde al sistema de nomenclatura estándar. Además, existen diferentes entes territoriales quienes han propuesto un marco legal para la definición de nomenclatura urbana en Colombia, entre ellos encontramos al Departamento Administrativo de Catastro Distrital, Departamento Administrativo Nacional de Estadísticas y el Instituto Geográfico Agustín Codazzi en su dependencia de Subdirección de Catastro Nacional (Camacho & Tellez, 2009).

6.2 Herramientas para calidad de datos en direcciones residenciales

A continuación, se presentan los resultados del segundo objetivo específico: "*Caracterizar herramientas software de calidad de datos para utilizarlas en la nomenclatura urbana colombiana".*

6.2.1 Definición de características funcionales y no funcionales

Dentro de la búsqueda bibliográfica, se identificaron unos pasos fundamentales en el proceso de validación de una dirección y esto se tomó como punto de partida para la definición de las características y criterios de evaluación para cada una de las herramientas. Considerando los aportes de (Padrón Torres, 2007), (Vargas & Horfan, 2013) y (Goldberg et al., 2013), puede hablarse de algunos pasos principales en la validación de direcciones: Segmentación, Corrección de errores en texto, Estandarización o Normalización y Búsqueda referencial (en bases de datos, proceso de geo codificación u otros).

Para la definición de las características que servirían como criterios de comparación de las herramientas, se tuvo en la cuenta algunas

consideraciones iniciales: los pasos que realiza para validar una dirección respecto a los descritos anteriormente y la capacidad de resolver direcciones en varios países, incluyendo las direcciones colombianas. Finalmente, las características seleccionadas fueron las siguientes:

- **Accesibilidad:** Es moderadamente sencilla de encontrar, de probar, instalar, si se tiene un demo o servicio web donde se pueda acceder.

- **Segmentación:** Es capaz de segmentar correctamente cada uno de los bloques de un formato de dirección.

- **Corrección de errores en texto:** Es capaz de corregir errores en texto libre específicos en el caso de direcciones residenciales.

- **Estandarización:** Tiene una secuencia de pasos para estandarizar o normalizar direcciones, utiliza técnicas de corrección de textos, segmentación, etc.

- **Búsqueda Referencial:** Utiliza algún motor de búsqueda referencial para validar la dirección luego de tenerla estandarizada.

- **Soporte geográfico:** Variedad de direcciones de países o regiones que pueda validar.

- **Disponibilidad:** Qué tan difícil es de adquirir, tiene muchos requisitos para la instalación o el despliegue y qué tanta dependencia se tiene del proveedor para el servicio.

- **Interoperabilidad:** Se puede integrar con otras herramientas para validar direcciones, integración a través de servicios web.

- **Escalabilidad:** Qué tanta cantidad de registros soporta y si tiene tiempos de respuesta aceptables a medida que crece el número de registros.

- **Conectividad:** A que fuentes de información puede conectarse (bases de datos, archivos planos).

- **Licenciamiento:** Tipos de licencias que tiene disponibles por la utilización de la aplicación o el servicio.

El listado anterior fue dividido en dos partes: características funcionales o directamente relacionadas a la funcionalidad (Corrección de errores en texto, Estandarización, Segmentación, Búsqueda referencial y Soporte Geográfico) y no funcionales o que hacen parte de características propias del software (Accesibilidad, Conectividad, Escalabilidad, Interoperabilidad, Disponibilidad y Licenciamiento).

6.2.2 Búsqueda y selección de herramientas software

Tras la revisión de fuentes de información y la consulta de algunas empresas locales que tienen un nivel de madurez alto en gobierno de datos, fueron identificadas algunas herramientas líderes del mercado para el manejo de calidad de datos en direcciones (Friedman & Judah, 2013).

AddressDoctor: Es la solución para calidad de datos en direcciones de la empresa Informática, uno de los líderes en esta especialidad (Informatica, 2014).

Google Maps (API): No es precisamente una herramienta de calidad de datos pero su amplia capacidad, permite utilizarla en otros servicios diferentes a los de localización (Svennerberg, 2010).

SmartyStreets: Empresa enfocada específicamente en direcciones, con su producto en línea del mismo nombre (SmartyStreets, 2015).

Servinformación: Empresa dedicada a diferentes servicios de información geográfica en Colombia, siendo uno de los pocos proveedores de Geo codificación en el País (Servinformacion, 2015).

Trillium Software, Global Locator: Producto de Geo localización de la empresa Trillium, otro de los líderes en temáticas de calidad de datos (Company, 2015).

Human Inference, Data Cleaner: Producto Originalmente Open Source y gratuito, ahora es un Open Source comercial y recientemente añadió funciones para validación de direcciones (Team, 2015).

Github, Open Refine: Proyecto originario de Google, denominado Google Refine que fue abandonado y luego retomado por Github, un grupo de

colaboradores de software de código abierto, quienes lo renombraron como Open Refine (GitHub, 2015).

SqlPower DQGuru: Empresa dedicada a soluciones de gestión de bases de datos e inteligencia de negocios con software de código abierto y gratuito, dentro su portafolio cuentan con una solución de calidad de datos DQGuru que tiene algunas funcionalidades de validación de direcciones (SQLPower, 2015).

Para realizar una preselección de herramientas que reduzca el conjunto a evaluar, se tomaron tres características de las mencionadas anteriormente seleccionadas por su importancia en el manejo de direcciones: Accesibilidad, Estandarización y Búsqueda Referencial. Con base en una tabla de calificación (Tabla 5), se calificó a todas las herramientas para posteriormente seleccionar a aquellas que obtuvieran los mayores puntajes (Tabla 6). Cabe aclarar que los puntajes son subjetivos de acuerdo con la apreciación del autor de este trabajo.

Tabla 5. Tabla de Calificación de Características

Valor	Calificación
4 a 5	Alto
3 a 3.9	Medio
2 a 2.9	Bajo
1 a 1.9	Muy Bajo

Fuente propia

Tabla 6. Ponderación de las herramientas candidatas

Herramienta	Accesibilidad	Estandarización	Búsqueda referencial	Total
AddressDoctor	3	5	5	13
Google Maps API	5	3	5	13
Data Cleaner	4	4	4	12
Servinformación	3	4	3	10
Global Locator	1	3	5	9
Open Refine	5	3	1	9
SmartyStreets	2	4	3	9
SQLPower, DQGuru	5	2	1	8

Fuente: Elaboración propia

Las herramientas que obtuvieron los valores más altos fueron: AddressDoctor, Google Maps API, Data Cleaner y Servinformación. A continuación, se detallan los aspectos más relevantes de las herramientas preseleccionadas.

AddressDoctor: Esta es una herramienta con amplia funcionalidad y documentación, posee una serie de soluciones que aplican en diferentes escenarios para asegurar la calidad de datos en una dirección. Mediante el "Data Quality Center" realiza las operaciones en línea y en segundo plano, y con una cuenta gratuita puede realizarse un número limitado de transacciones (Informatica, 2014).

Adicionalmente, permite enriquecer la dirección con datos relacionados a la ubicación, por ejemplo, coordenadas de latitud y longitud vía georreferenciación. Esa función está disponible solo dentro de la cobertura, no aplica en ciertos países y subregiones. El formato con que aparece la lista de candidatos es un formato de envío de correo postal, incluso el mismo programa permite imprimirlo de esta forma. Posee una gran cantidad de países que se encuentran dentro de la cobertura de validación; más allá de que los campos pedidos son predefinidos, en el campo de dirección tiene la versatilidad de organizar el texto y segmentarlo para realizar la búsqueda adecuadamente. Para poder utilizar la funcionalidad de validación, es necesario diligenciar algunos parámetros simples como la ciudad o población y el país. Puede seleccionarse un país por defecto, si es que todos los registros están dentro del mismo y tiene cobertura parcial en 240 países.

Google Maps API: Esta es una librería propietaria de Google, la cual permite realizar peticiones de geo codificación a los servidores de mapas. La extensa funcionalidad de esta librería hace que pueda utilizarse con diferentes propósitos. Tiene un número determinado de peticiones que se detallan en su acuerdo de servicio y la librería puede utilizarse para validar direcciones, aunque debe realizarse mediante las especificaciones del servicio web. Para la realización de las pruebas, se utilizó el proyecto de código abierto jgeocoder que se conecta al servicio de mapas para estandarizar direcciones (Svennerberg, 2010).

Data Cleaner: Es una de las herramientas de código abierto con más fuerza para trabajar en calidad de datos y cuenta con dos versiones, una comercial y otra de la comunidad. Permite realizar diferentes operaciones en los datos, tiene un buen número de funcionalidades de análisis y transformaciones para

realizar correcciones de diferentes problemas en los datos. Mediante un complemento “EasyDQ”, el software realiza tareas de limpieza de direcciones, tiene buena cobertura geográfica y realiza operaciones de segmentación y corrección de la dirección, aunque en los resultados no es muy claro con las modificaciones que realizó sobre el conjunto de datos original. Para utilizar la función de limpieza de direcciones, es necesario que en las columnas de la fuente de datos estén ciertos campos complementarios de la ubicación (Team, 2015).

Servinformación, Georreferenciador: Este software está enfocado en convertir direcciones de una base de datos en una ubicación geográfica y está basado en las mallas viales de cada municipio, en donde identifica cada uno de los componentes de la dirección y lo traduce a coordenadas. Este software tiene comunicación con algunos sistemas de información geográfica por medio de exportación de archivos planos y en formatos de gráficos. Al trabajar con información de la malla vial que es construida y almacenada de manera propietaria, hace que solo se pueda realizar la georreferenciación sobre las ciudades que estén disponibles en el listado de cobertura. Para realizar la búsqueda sobre alguna población es necesario adquirir la licencia de esa población (Servinformacion, 2015).

El programa está diseñado para trabajar con las particularidades del sistema de nomenclatura colombiano, tiene la definición completa de cada tipo de componente de la dirección (calle, carrera, avenida, etc.) y realiza una muy buena estandarización gracias a su lista de alias o sinónimos para cada componente, adaptado al lenguaje comúnmente utilizado en el país para escribir una dirección, además permite añadir un alias nuevo mediante el mismo programa, lo que mejora su flexibilidad para la segmentación y en donde puede ayudar al usuario final (Servinformacion, 2015).

Para las cuatro herramientas preseleccionadas, se hizo la misma evaluación para todas las características funcionales: Corrección de errores en texto, Estandarización, Búsqueda Referencial, Segmentación y Soporte Geográfico. En la Tabla 7 se presentan las puntuaciones asignadas para cada una de las herramientas seleccionadas.

Aunque se aclara nuevamente que los puntajes son subjetivos de acuerdo con la apreciación del autor de este trabajo, la herramienta Informática AddressDoctor y “Georreferenciador” de Servinformación resultan bastante completas en su funcionalidad, no tanto las herramientas “Data Cleaner” y “Google Maps API”.

Tabla 7. Tabla de evaluación de características funcionales

Herramienta	Segmentación	Corrección	Estandarización	Búsqueda referencial	Soporte Geográfico
AddressDoctor	4	1	5	5	4
Google Maps API	1	1	3	5	5
Data Cleaner	1	5	4	4	3
Servinformación, Georreferenciador	4	1	4	3	2

Fuente: Elaboración propia

En la revisión de las características no funcionales, se presentan aquellas que cumplen con las opciones deseables en un software de calidad de datos, estas permiten ver que tan completa es la herramienta y que tan amigable es con el usuario final (Tabla 8).

Tabla 8. Tabla de evaluación de características no funcionales

Herramienta	Accesi-bilidad	Conecti-vidad	Escala-bilidad	Interope-rabilidad	Disponi-bilidad	Licencia-miento
AddressDoctor	3	5	5	5	4	Por transacciones
Google Maps API	5	1	3	2	4	Por transacciones
Data Cleaner	4	2	5	5	5	Comercial 1 año / comunidad
Servinformación, Georreferenciador	3	1	5	2	4	Por transacciones

Fuente: Elaboración propia

En esta revisión, las herramientas Data Cleaner y AddressDoctor fueron las que obtuvieron mejores resultados debido a su amplia funcionalidad e integración con diferentes fuentes de datos, plataformas y otras soluciones, el software AddressDoctor adicional a su amplia funcionalidad en el manejo, validación y corrección de direcciones, posee además opciones deseables para mejorar la calidad de los datos de direcciones en los diferentes pasos, desde la captura, hasta la verificación en bases de datos y su presentación en diferentes formatos deseados.

6.2.3 Evaluación de las herramientas de calidad de datos

Mediante un caso de evaluación se realizan pruebas con un listado pequeño de direcciones locales válidas, se evalúa cada una de las herramientas contra un listado de ejemplo de 20 direcciones existentes y con problemas de calidad de datos ingresados a propósito, con el objetivo de verificar la

funcionalidad de cada una de las herramientas en el escenario de Colombia y con problemas típicos de diferentes tipos, con direcciones principalmente de las ciudades de Bogotá, Medellín y Cali. En dos de las herramientas seleccionadas no se obtuvieron resultados en esta prueba.

En la herramienta Data Cleaner, que utiliza el complemento EasyDQ y según su documentación realiza limpieza de direcciones de 240 países, no se obtuvo un resultado para estas consultas del experimento preliminar, se realiza otro experimento con direcciones de Estados Unidos y en esta devuelve el mismo conjunto de datos como salida y no da alguna pista de que verificación realizó o que problemas corrigió, por esto no fue posible obtener resultados de la prueba, el software no es muy claro en los resultados que entrega. No se encuentra documentación detallada sobre la funcionalidad, resultados o problemas frecuentes en este complemento, así como en las otras funcionalidades de la herramienta original. En esta herramienta no fue posible realizar la prueba completa ya que aun completando todos los datos requeridos, la herramienta presentaba mensajes de errores de ejecución.

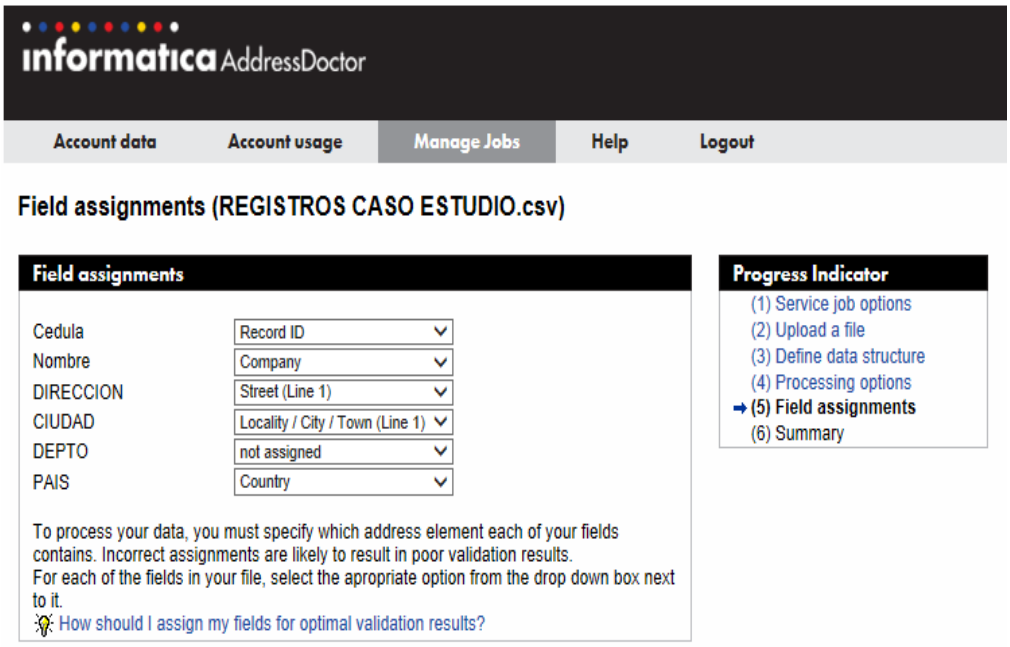
En el software Google Maps API, no se encontró una interfaz de usuario aceptable para poder utilizar el servicio adecuadamente. Tiene una potente funcionalidad pero no se encontró un cliente que realice la validación de direcciones con las características esperadas en una herramienta de calidad de datos. El proyecto de código abierto “jgeocoder” puede realizar validaciones solo en Estados Unidos y el proyecto “GeoGoogle” no tiene la suficiente documentación sobre cómo utilizarlo. Este último puede ser una opción desde el desarrollo de una nueva herramienta aunque dependa de las restricciones de uso de la plataforma. A pesar de su muy buena funcionalidad que realiza validación de direcciones antes de una petición interactiva en la aplicación de mapas, en general no se encuentra un cliente lo suficientemente bueno para ser una opción de validación de direcciones ni de la misma empresa ni externo que utilice la librería de Google Maps.

En la herramienta de Servinformación se realiza pruebas con los 20 registros, el programa posee una funcionalidad específica para geo codificar direcciones de algunas ciudades principales de Colombia. Los resultados de la prueba son descriptivos ya que entrega un dato de “dirección traducida” en donde se puede comparar la dirección original y ver la interpretación que realiza la herramienta. Identifica e interpreta correctamente los componentes de la dirección en Colombia (“Calle”, “Carrera”, “Avenida”) y también otorga

una clasificación de la calidad del proceso de geo codificación (Excelente, Buena, Deficiente, etc.), esto permite definir la limpieza, por ejemplo se podría sustituir las direcciones con clasificación “excelente” y excluir o realizar el esfuerzo por volver a recuperar las direcciones clasificadas como “deficiente”. Esta herramienta puede llegar a ser una buena opción si se va a trabajar con direcciones locales que se encuentren dentro de la cobertura. Por restricciones de licenciamiento no fue posible realizar la prueba con los registros deseados y con los problemas de calidad de datos insertados en el listado y no se puede verificar a fondo sus capacidades en otros escenarios.

En la revisión de la herramienta AddressDoctor, se realiza la prueba preliminar de manera exitosa y los resultados que entrega son claros y de fácil interpretación.

Figura 18. Ingreso de direcciones en AddressDoctor.



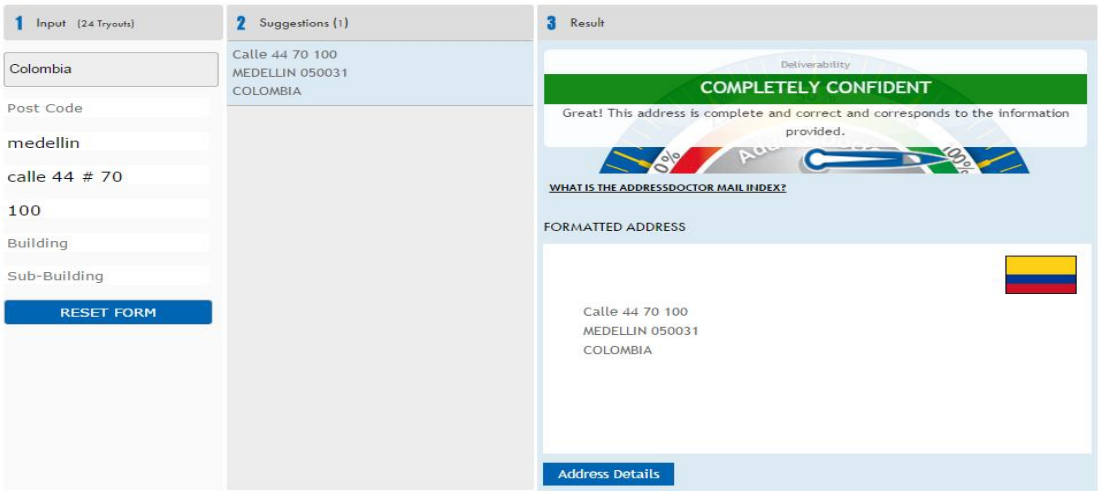
Fuente (Informatica, 2014)

Presenta los resultados consolidados y con las observaciones del procedimiento, también entrega como salida el conjunto de datos original con las clasificaciones, las correcciones y los datos adicionales que el usuario solicitó en la configuración del proceso, tuvo un comportamiento satisfactorio con las direcciones del país, claro que es necesario una prueba más sólida como la que se realiza al final de este trabajo. Para poder utilizar la funcionalidad es necesario el diligenciamiento del formulario de definición de atributos, estos los puede tomar de diferentes fuentes de datos siempre que estén en formato tabular y contenga los atributos obligatorios de “dirección”,

“ciudad” y “país”, también se puede agregar otros atributos adicionales de localización como “departamento”, “localidad” y “barrio”, que pueden ayudar a mejorar los resultados (Figura 18).

Se encuentran también algunos aspectos a mejorar: presenta dificultades con la configuración de las columnas, es muy rígido en la conformación del formato de dirección, errores tipográficos, tiene dificultades con la forma de escritura de los componentes y no es tan efectivo identificando diferentes formas de escritura. La cobertura de la geo codificación, solo está disponible en algunos países. Después de realizadas las pruebas con las opciones disponibles se encontró que la herramienta AddressDoctor cumple con los criterios fundamentales para realizar validación de direcciones, además de ser amigable con el usuario y con unas cualidades no funcionales que ayudan a ser una buena opción para trabajar con direcciones residenciales con una buena cobertura mundial.

Figura 19. Ejemplo de validación de una dirección con AddressDoctor.



Fuente (Informatica, 2014)

Ninguna de las herramientas estudiadas es lo suficientemente buena en la corrección de errores en texto, no soportan problemas como bloques en desorden, datos sobrantes, separadores no estándar y todos los esfuerzos previos por mejorar la calidad de los datos en la captura inicial influyen en el resultado final de la validación. Las funciones de limpieza de direcciones que soporten búsqueda referencial todavía no están lo suficientemente desarrolladas para las características de una dirección de Colombia, algunas de ellas pueden mejorar su calidad en algunos casos en áreas urbanas registradas, pero todavía no cubren la mayor parte del territorio.

Finalmente, después del proceso de evaluación, se sugiere el uso de una combinación de herramientas en el proceso de limpieza de texto, segmentación, corrección de errores, validación y enriquecimiento de las direcciones residenciales (Figura 20).

Figura 20. Herramientas sugeridas para el proceso de validación y estandarización de direcciones.

Fuente: Elaboración propia

6.3 Guía Metodológica para el manejo de errores en la estandarización de direcciones residenciales colombianas

A continuación se presentan los resultados del tercer objetivo específico: "*Construir una guía metodológica para depurar los problemas de direcciones urbanas en Colombia"*. Para la construcción de la guía, es necesario tener en la cuenta la estructura de las direcciones residenciales en Colombia.

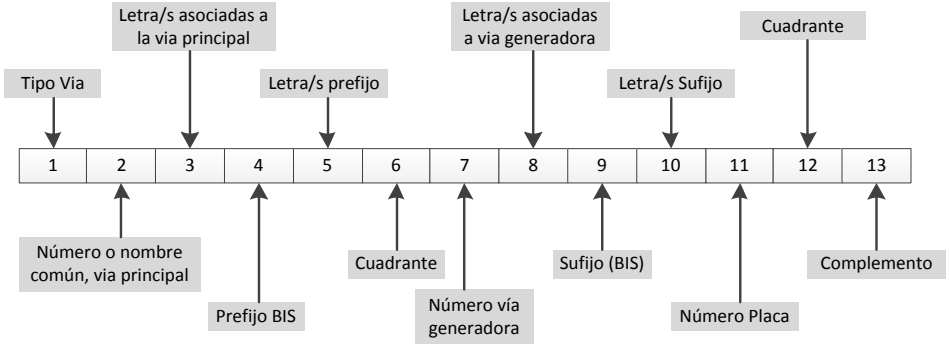
6.3.1 Estructura de Direcciones Residenciales en Colombia

En Colombia pueden existir diversas maneras de representar una dirección residencial y esto se ve reflejado en el informe realizado por el MEN (Ministerio de Educación Nacional) junto con el IGAC (Instituto Geográfico Agustín Codazzi) (Camacho & Tellez, 2009). En el informe puntualizan cinco tipos de direcciones residenciales basadas en las estructuras de: Malla Vial, Barrio-Manzana-Predio, Malla Vial/Barrio-Manzana-Predio, Malla Vial/Barrio y Sitios de interés. Esta representación de las direcciones está fundamentada en: la organización de campos del CUNU (Código Unificado de Nomenclatura Urbana) elaborado por el DACD (Departamento Administrativo de Catastro Distrital) de Santafé de Bogotá, la estandarización de abreviaturas de la Circular 300/01 propuesta por el IGAC y en la Resolución 166/04 del MEN. A continuación, se describen las diferentes representaciones de direcciones residenciales en el país.

Estructura Malla Vial

Está conformada por una vía principal, vía generadora y un número de placa. Este es el sistema más utilizado a nivel nacional (Figura 21).

Figura 21. Nomenclatura de la estructura de Malla Vial

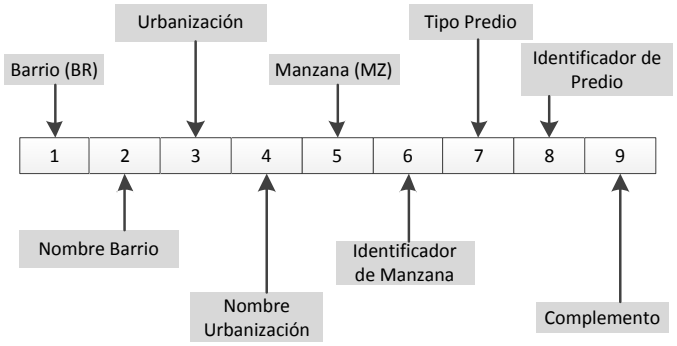


Fuente: Elaboración propia

Estructura Barrio-Manzana-Predio

Esta nomenclatura puede ser un complemento de la Nomenclatura Malla Vial/Barrio y está definida por un nombre de barrio, un código de manzana y un código de predio, aunque en la práctica pueden darse otras variaciones (Figura 22).

Figura 22. Nomenclatura de la estructura de Barrio-Manzana-Predio

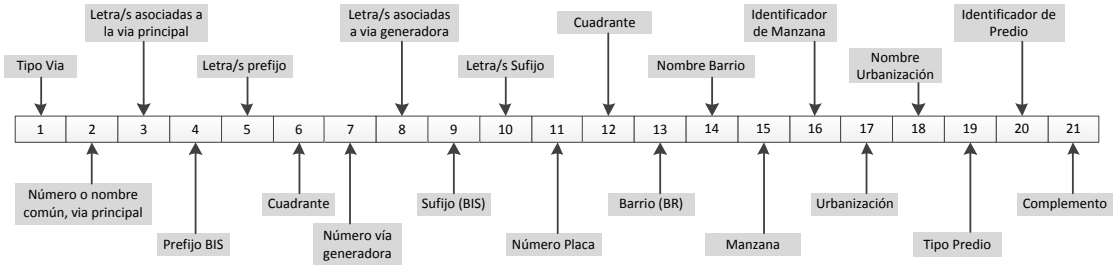


Fuente: Elaboración propia

Estructura Malla Vial/Barrio-Manzana-Predio

Este caso ocurre cuando se combinan el caso 1 y el caso 2 descritos anteriormente (Figura 23).

Figura 23. Nomenclatura de la estructura de Malla Vial / Barrio-Manzana-Predio

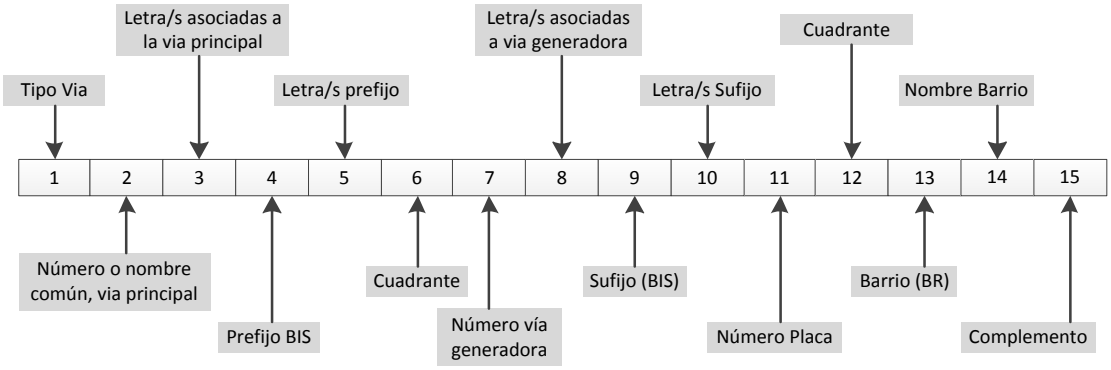


Fuente: Elaboración propia

Estructura Malla Vial/Barrio

Esta nomenclatura introduce el barrio para diferenciar dos o más direcciones iguales en una misma población. Es por esto que se agrega la abreviatura de barrio (BR) y el nombre del barrio (Figura 24).

Figura 24. Nomenclatura de la estructura de Malla Vial / Barrio

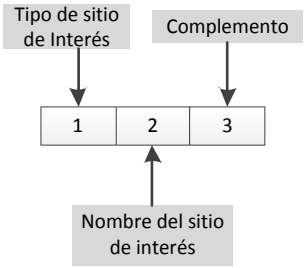


Fuente: Elaboración propia

Estructura Sitios de Interés

Este tipo de nomenclatura es empleada cuando quiere ubicarse un sitio de interés (parques, estadios, centros comerciales, etc.) (Figura 25).

Figura 25. Nomenclatura de la estructura de Malla Vial / Barrio



Fuente: Elaboración propia

6.3.2 Propuesta de Guía Metodológica

Existen empresas colombianas que organizan sus direcciones adoptando los estándares internos que emplean los organismos gubernamentales. Por ejemplo, la DIAN tiene una página Web para estandarizar las direcciones a nivel nacional con una codificación empleada por la misma organización (DIAN, 2014). La página presenta las nomenclaturas más usadas en el ingreso de direcciones residenciales de algunos servicios ofrecidos como el RUT (Registro Único Tributario), el cual permite identificar, ubicar y clasificar a las personas y entidades que tengan la calidad de contribuyentes declarantes del impuesto de renta en el país.

De igual manera, Código Postal Colombia junto con 4-72 tiene disponible una página Web donde puede consultar y validarse las direcciones residenciales de las principales ciudades del país a través de la búsqueda del código postal utilizando la codificación del DANE (Departamento Administrativo Nacional de Estadística) (MinTIC, 2014).

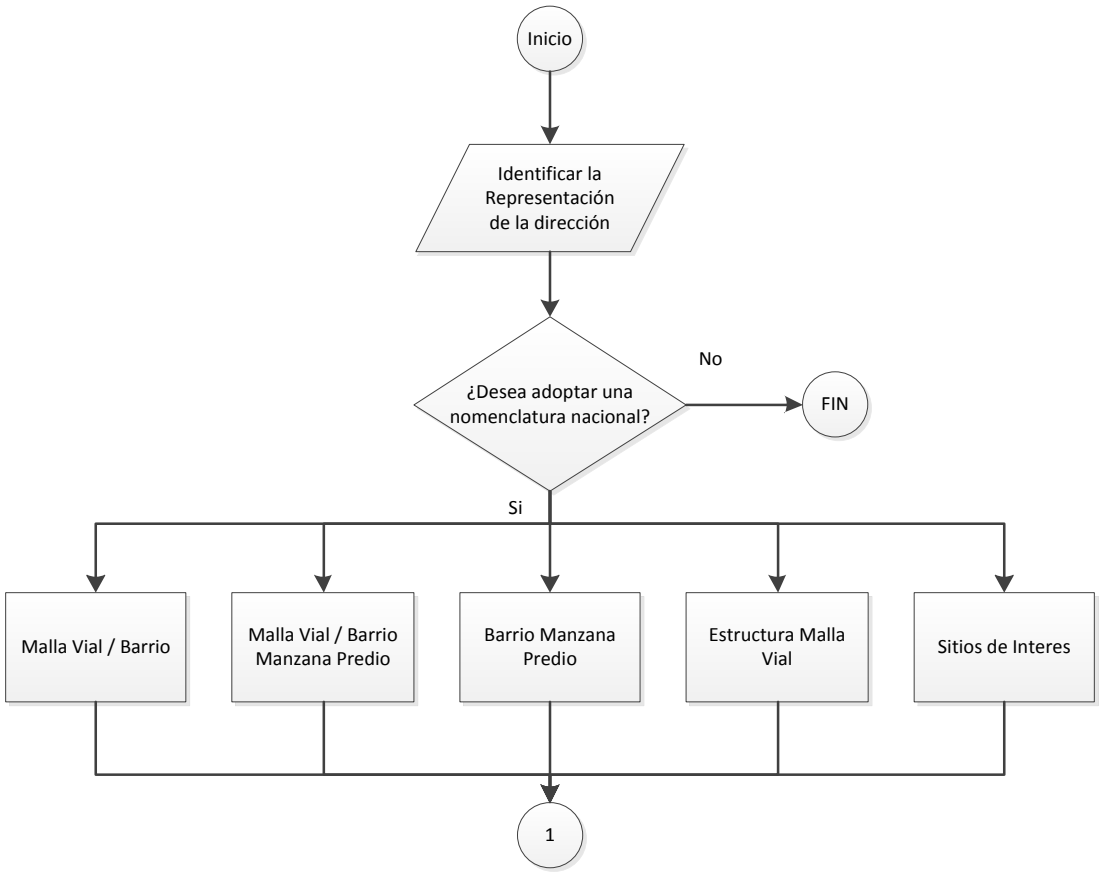
Sin embargo, es usual que muchas empresas no tengan definido un proceso de estandarización en el registro de sus direcciones residenciales. El ingreso de los datos se realiza como un texto libre sin tener ningún orden, dando lugar a diferentes tipos de errores tipográficos, sintácticos, ortográficos, etc. En otras ocasiones, no tienen el cuidado de definir algunas reglas de validación para que los datos registrados sean correctos. Por tal motivo, es fundamental que las empresas organicen sus direcciones residenciales con calidad de datos y de una manera estandarizada para que logren generar valor y tomar decisiones más acertadas.

De acuerdo con lo anterior, resulta necesario considerar lo desarrollado en el ítem 6.1 donde es mencionada la gran cantidad de accidentes geográficos que han generado diferencias en el sentido y orientación de las vías principales (Calles y Carreras) y las vías alternas (Avenidas, Transversales, Circulares y Pasajes) en las principales ciudades colombianas. Estos accidentes, a su vez han ocasionado algunas diferencias en el momento de estandarizar la nomenclatura urbana a nivel nacional.

Por otro lado, el proceso de estandarización de direcciones residenciales está inmerso dentro de un conjunto de pasos que incluyen la segmentación, la aplicación de técnicas de calidad de datos para limpiar el texto, corrección de errores, validación y búsqueda referencial de las direcciones residenciales, tal como fue mencionado en el ítem 6.2 del presente informe.

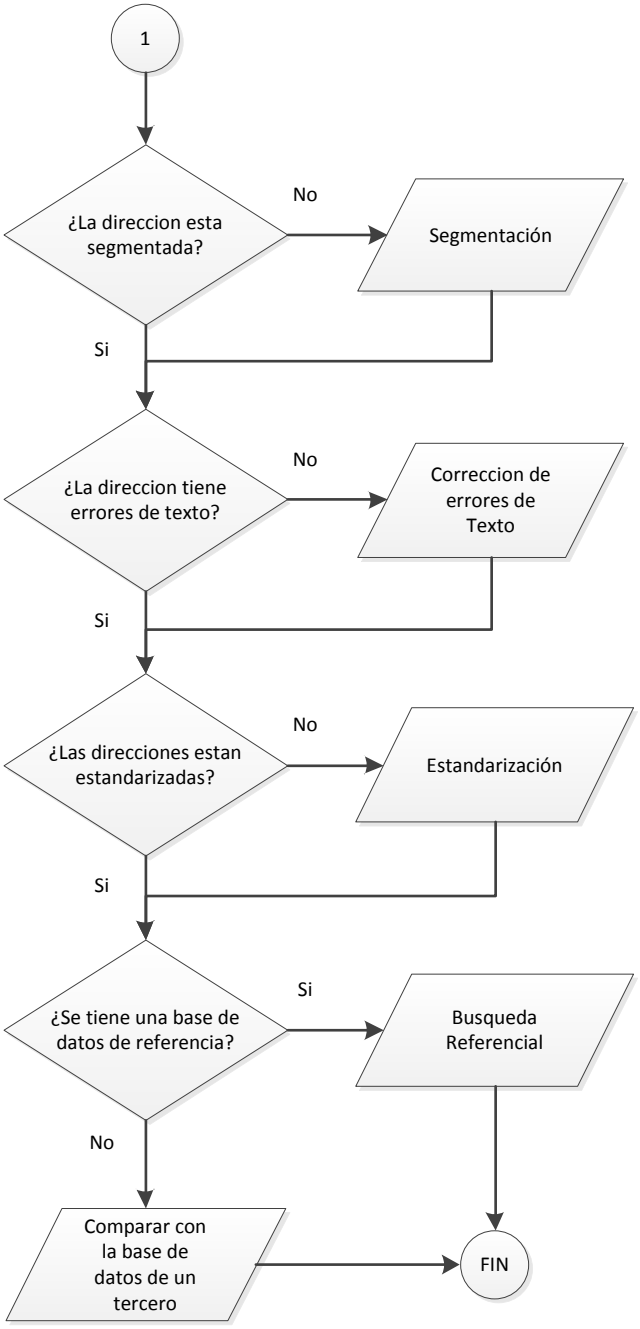
Partiendo de estas consideraciones preliminares, surge la Guía Metodológica para orientar a las empresas colombianas en el manejo de errores en el registro de direcciones residenciales para que puedan tener direcciones estandarizadas y con calidad de datos (Figuras 26 y 27).

Figura 26. Propuesta de Guía Metodológica - Identificar la representación de una dirección



Fuente: Elaboración propia

Figura 27. Propuesta de Guía Metodológica - Manejo y depuración de errores



Fuente: Elaboración propia

A continuación se describen cada una de las Fases que componen la Guía y en el caso de estudio presentado en la sección 6.4, puede verse aplicado a datos reales de direcciones residenciales de la ciudad de Medellín.

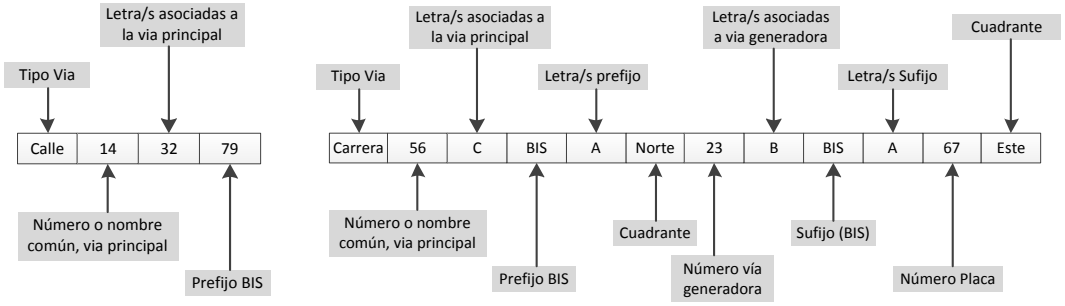
Identificación de la Representación:

El primer paso en la Guía es identificar la representación de las direcciones almacenadas en las fuentes de datos de la empresa. Este paso es fundamental porque a partir de él es definido el tipo de nomenclatura urbana que se toma como referencia en el proceso de estandarización de las direcciones residenciales. Partiendo de la sección 6.3.1 Estructuras residenciales en Colombia, la representación de las direcciones pueden ser del tipo: Malla Vial, Barrio-Manzana-Predio, Malla Vial/Barrio-Manzana-Predio, Malla Vial/Barrio y Sitios de interés (Figura 26).

Segmentación:

En esta segunda fase, las direcciones son descompuestas en unidades atómicas para analizarse por separado. Teniendo en cuenta la estructura de Malla Vial, en ésta pueden existir direcciones que puedan ocupar hasta 13 campos y en otros casos existirán direcciones que solamente ocupen 4 campos. Por ejemplo, la dirección: Carrera 56C BIS A Norte # 23B BIS A – 67 ESTE tiene 12 campos y la dirección: Calle 14 # 32 - 79 tiene 4 campos (Figura 28).

Figura 28. Ejemplos de direcciones en la estructura de Malla Vial



Fuente: Elaboración propia

Para el proceso de descomposición de direcciones, es necesario utilizar un método que permita separar diferentes bloques dentro de una cadena de texto. Varios autores coinciden en que la utilización de Modelos Ocultos de Markov (HMM) permite realizar una segmentación con muy buenos resultados (Viola & Narasimhan, 2005) y (Borkar et al., 2001). Este método no se detalla por salirse del alcance de este trabajo.

Corrección de Errores en Texto:

En esta tercera fase, se realiza un proceso de análisis y limpieza de datos fundamentado en una selección de técnicas de depuración de datos (Amon, 2010). Además, es importante elaborar un perfilamiento de datos a través de herramientas como DQAnalyzer, con el propósito de revisar datos nulos, duplicados, frecuencia de aparición de las diferentes cadenas y demás estadísticas de calidad de datos (Naumann, 2013).

Para el manejo de errores ortográficos o tipográficos, pueden emplearse técnicas de record linkage, en especial el algoritmo de distancia de edición o "edit distance" mencionados por el autor (Ranzijin, 2013). La finalidad de la utilización de estas técnicas es lograr una buena preparación del texto para la siguiente fase.

Estandarización:

En la cuarta fase, las direcciones son corregidas y pasan a un estándar definido por cada empresa en particular. Cada elemento debe ser normalizado de acuerdo con alguna tabla de abreviaturas predefinida. La normalización puede variar y ajustarse a abreviaturas de dos o tres letras. Lo recomendable es que la representación sea de dos letras pensando en facilitar un posterior proceso de geo codificación. Por ejemplo, el Tipo de Vía: Calle va a cambiar por CL, Carrera por CR, Avenida por AV, Circular por CQ, Diagonal por DG y Transversal por TV.

Búsqueda Referencial:

En esta última fase, las direcciones que han sido estandarizadas son validadas realizando un proceso de comparación con una fuente de datos que tenga calidad de datos y esté debidamente estandarizada. Hay que aclarar que la Búsqueda Referencial debe realizarse en una base de datos o una herramienta que proporcione la misma estructura de representación seleccionada en la primera fase de la Guía Metodológica.

6.4 Caso de Estudio

A continuación, se presentan los resultados del cuarto objetivo específico: "*Realizar un caso de estudio de estandarización y calidad de datos para identificar posibles direcciones inválidas en la ciudad de Medellín".*

La guía metodológica es explicada a través de un caso de estudio de 210 registros de direcciones residenciales de la ciudad de Medellín. Los datos fueron ingresados en modo de texto libre en un formulario Web de registro de clientes de la empresa Compuredes Enlace Operativo.

Fase de Identificación de la Representación:

La representación del conjunto de direcciones es similar a la estructura de la nomenclatura Malla Vial, mencionada anteriormente en la sección de estructuras residenciales en Colombia (Camacho & Tellez, 2009). Para el desarrollo de las fases posteriores, fue adoptada ésta representación de Estructura de Malla Vial. La Tabla 9 tiene ejemplos de direcciones de tipo Carrera, Calle y Circular encontradas en los registros de clientes:

Tabla 9 Ejemplos de representación de Carrera, Calle y Circular

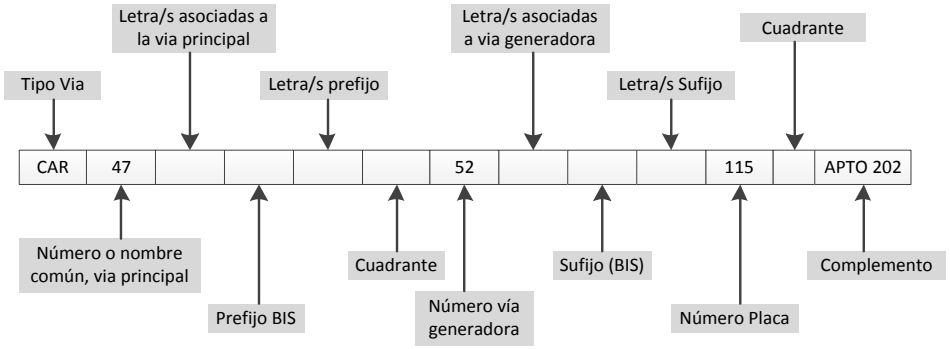
Carrera	Calle	Circular
CAR 47 52 115 APT 202	CALL 43 29-15	CIR 2DA # 73-32
CRR 35 # 4 A SUR 15	CALLE 30 A # 89 - 19	CIR. 74 76 C 66
KRA 50A 38 39	CLE 36 SUR # 27 B 04	CIRC 73 A # 38 35 AP 402
CARRERA 37 # 2 SUR 65 APTO 1202	CLLE 16 SUR # 43 A 49	CIRCULAR 73B 39-10 APTO 1102

Fuente propia

Fase de Segmentación:

En la segunda fase, el proceso de Segmentación fue realizado manualmente porque el conjunto de registros es pequeño, pero en una cantidad considerable de registros es recomendable apoyarse en herramientas como los Modelos Ocultos de Markov (HMM) para que realicen el cálculo computacional en el proceso de descomposición. Tomando los ejemplos de la Tabla 9, la dirección CAR 47 52 115 APT 202 es descompuesta de la siguiente manera: Tipo Vía: CAR – Número de vía principal: 24 – Número de vía generadora: 52 – Número de placa: 115 – Complemento: APTO 202 (Figura 29). Posteriormente, cada dirección es segmentada de acuerdo con el anterior ejemplo.

Figura 29. Segmentación de una dirección



Fuente propia

Fase de Corrección de Errores en Texto:

En esta tercera fase, es realizado un proceso de análisis, limpieza de datos y perfilamiento de datos a través de la herramienta DQAnalyzer con el propósito de revisar datos nulos, duplicados, frecuencia de aparición de las diferentes cadenas y demás estadísticas de calidad de datos (Figura 30a y 30b) (Naumann, 2013).

Figura 30. (a) Frecuencia de valores en el Campo Tipo Vía (b) Frecuencia de valores corregidos y normalizados

Frequency Analysis

Range: none

Value	Count	%
CARRERA	37	17.37%
CALLE	28	13.15%
CLLE	21	9.86%
CRR	20	9.39%
CAR	18	8.45%
CLE	12	5.63%
KRA	12	5.63%
DIAG	8	3.76%
CALL	7	3.29%
CIR	6	2.82%

(a)

Extremes

First Values	Frequency
AV	4
AVENIDA	1
CALL	7
CALLE	28
CAR	18

Last Values	Frequency
TRANSV	2
TRANS.	1
TRANS	6
TRAN	1
KRA	12

(b)

Fuente: Elaboración propia

Por ejemplo, en la identificación del campo de Tipo Vía, la representación de los elementos: Avenida, Calle, Carrera, Circular, Diagonal y Transversal fueron encontrados con errores tipográficos (Tabla 10a). Posterior a la identificación, se normaliza el conjunto de datos dejando una representación única para cada elemento (Tabla 10b).

Tabla 10. (a) Campo Tipo Vía con errores tipográficos (b) Campos corregidos y normalizados

(a)	(b)
Avenida: AV, AVENIDA	**Avenida:** AVENIDA
Calle: CALL, CALLE, CLE, CLLE	**Calle:** CALLE
Carrera: CAR, CAR., CRR, KRA, CARRERA	**Carrera:** CARRERA
Circular: CIR, CIR., CIRC	**Circular:** CIRCULAR
Diagonal: DIA, DIAG, DIAG., DIG, DIAGONAL	**Diagonal:** DIAGONAL
Transversal: TRAN, TRANS, TRANS., TRANSV	**Transversal:** TRANSVERSAL

Fuente: Elaboración propia

Fase de Estandarización:

En la cuarta fase, los campos Tipo Vía, Cuadrante, Prefijo BIS, Sufijo BIS y Complemento son abreviados a un valor normalizado, utilizando algunas de las abreviaturas propuestas por el MEN (Camacho & Tellez, 2009). Surgieron algunas excepciones en las abreviaturas para Carrera (CR) y Apartamento (APTO) ya que en la propuesta del MEN está como (KR) y (AP) respectivamente (Tabla 11).

Tabla 11. (a) Campo Tipo Vía con elementos normalizados (b) Campos estandarizados

(a)	(b)
Avenida: AVENIDA	**Avenida:** AV
Calle: CALLE	**Calle:** CL
Carrera: CARRERA	**Carrera:** CR
Circular: CIRCULAR	**Circular:** CQ
Diagonal: DIAGONAL	**Diagonal:** DG
Transversal: TRANSVERSAL	**Transversal:** TV

Fuente: Elaboración propia

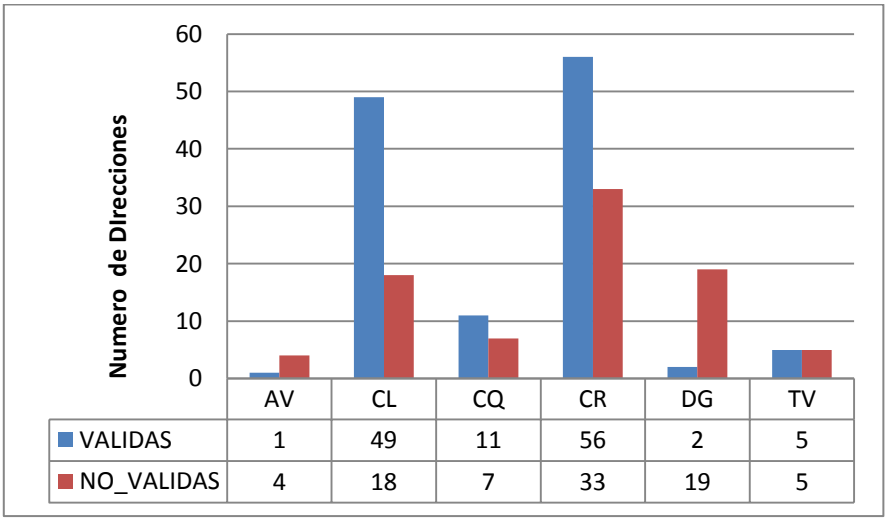
Algunas empresas pueden tener una tabla de normalización que difiera de los estándares propuestos por el MEN. En las facturas de servicios públicos como las de EPM (Empresas Públicas de Medellín) o las de Tigo-UNE para el departamento de Antioquia, en el campo Complemento es utilizada la abreviatura (INT) para referirse a un Apartamento en vez de un Interior. Por ejemplo, la dirección Calle 34C # 88B - 55 Apto 124 está representada como: CL 34 C 88 B 55 INT 124

Fase de Búsqueda Referencial:

En la última fase, cabe recordar que las direcciones residenciales tienen la estructura de representación de Malla Vial y van a ser comparadas utilizando la herramienta AddressDoctor que permite la búsqueda de direcciones con este tipo de representación.

Dentro de los 210 registros existen en total 5 direcciones con la abreviatura de Avenida (AV), 67 direcciones con la abreviatura de Calle (CL), 18 direcciones con la abreviatura de Circular (CQ), 89 direcciones con la abreviatura de Carrera (CR), 21 direcciones con la abreviatura de Diagonal (DG) y 10 direcciones con la abreviatura de Transversal (TV). Todos los registros fueron ingresados en la herramienta AddressDoctor, tal como fue explicado en la sección 6.2, encontrando que 124 direcciones son válidas y 86 no válidas. Los resultados son clasificados por las abreviaturas del campo Tipo Vía de cada registro y son presentados en la Figura 31.

Figura 31. Validación de direcciones con AddressDoctor



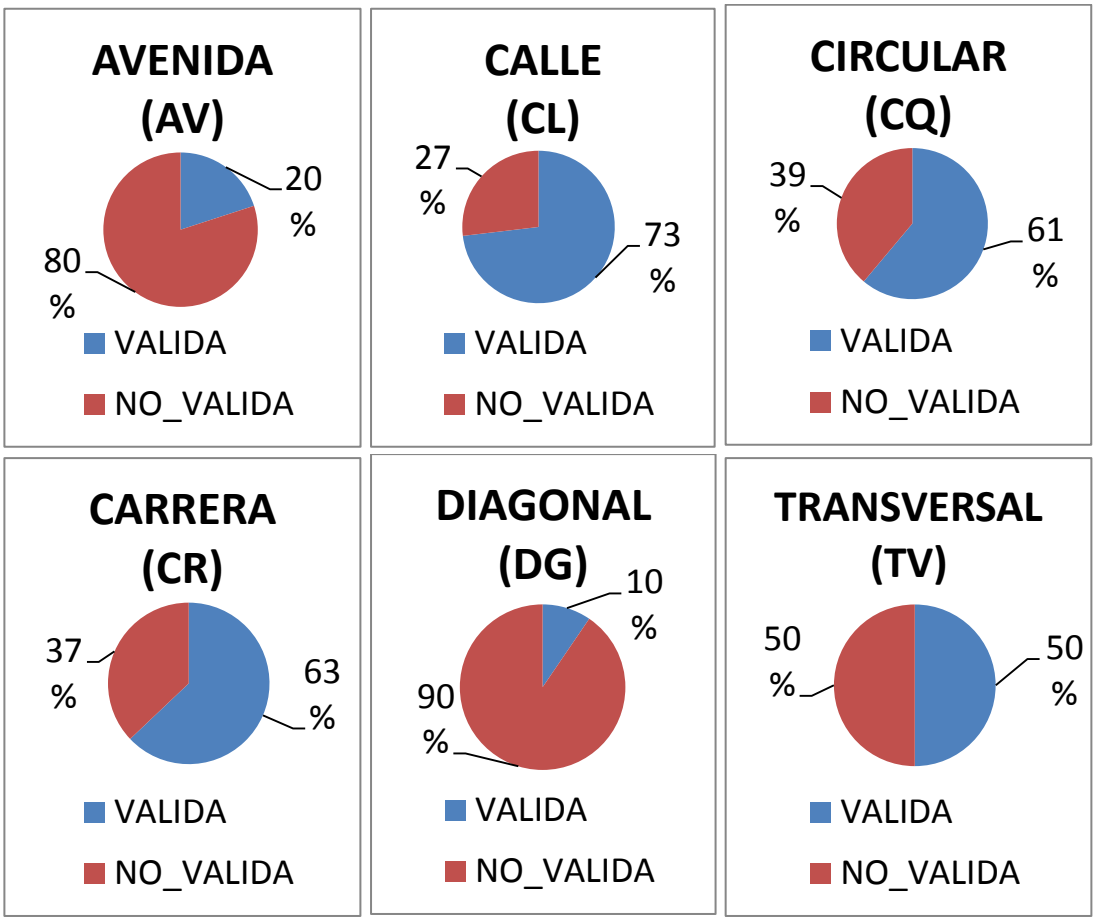
	AV	CL	CQ	CR	DG	TV
VALIDAS	1	49	11	56	2	5
NO_VALIDAS	4	18	7	33	19	5

Fuente: Elaboración propia

Partiendo del hecho de que todas las direcciones residenciales fueron validadas con los usuarios en el momento de ingresarlas en el formulario Web, se encuentra que la herramienta AddressDoctor tuvo un 59% de aciertos (124 direcciones) y un 41% de desaciertos (86 direcciones) en el momento de validar las direcciones en su base de datos.

Además, es posible observar que AddressDoctor tuvo mejores resultados validando direcciones que tienen las abreviaturas CL, CR y CQ. No tuvo resultados óptimos validando direcciones con abreviaturas AV y DG. Por otro lado, obtuvo un 50% en las abreviaturas TV (Figura 32).

Figura 32 Porcentaje de acierto con AddressDoctor



Fuente: Elaboración propia

Este ejercicio presenta la identificación de la representación de la dirección residencial, su posterior segmentación de forma manual, corrección de errores en texto, estandarización y búsqueda referencial empleando la herramienta AddressDoctor para comparar las direcciones registradas.

De esta manera, fue abordada la propuesta de Guía Metodológica para el manejo de errores en la calidad de datos y en la estandarización de las direcciones residenciales colombianas. El trabajo parte de la identificación de las características de la nomenclatura nacional, continúa con una selección de herramientas software para la revisión de calidad de datos y finaliza con una validación de direcciones a través del desarrollo de un caso de estudio con direcciones reales de la ciudad de Medellín.

7 CONCLUSIONES

Hay que destacar que la Nomenclatura Urbana de una población va mucho más allá de la señalización de Vías, Calles y Carreras, es una oportunidad para crear una base de datos de Información Urbana. Los datos recolectados van a utilizarse para localizar fácilmente una Dirección Residencial. Esta base de datos, que puede tomar la forma de un SIG (Sistema de Información Geográfica), representa la ventaja principal y más innovadora de la Nomenclatura, en especial en los países con un intenso crecimiento urbano como lo ha sido Colombia en las últimas décadas.

Ninguna de las herramientas estudiadas es lo suficientemente buena en la corrección de errores en texto, no soportan problemas como bloques en desorden, datos sobrantes, separadores no estándar y todos los esfuerzos previos por mejorar la calidad de los datos en la captura inicial influyen en el resultado final de la validación. Las funciones de limpieza de direcciones que soporten búsqueda referencial todavía no están lo suficientemente desarrolladas para las características de una dirección de Colombia, algunas de ellas pueden mejorar su calidad en algunos casos en áreas urbanas registradas, pero todavía no cubren la mayor parte del territorio.

El ingreso de las direcciones residenciales a los formularios de las páginas y/o aplicaciones empresariales debe tener reglas de validación. Esta precaución reduce el número de datos incorrectos optimizando la posterior limpieza y corrección de errores de texto en las fuentes de datos.

Sin importar la herramienta utilizada para validar direcciones, resulta esencial organizar el formato del texto y en lo posible estandarizar la dirección previamente. Si bien algunas herramientas corrigen algunos errores en texto y soportan ciertas variables en la escritura de los bloques de la dirección, éstas son limitadas y se sugiere hacer limpiezas de formato y texto previas con herramientas de limpieza de textos convencionales.

La Guía Metodológica surge como una herramienta de apoyo que permite elaborar un proceso de estandarización y calidad de datos al interior de las empresas. Es desarrollada a través de un conjunto de pasos necesarios para la depuración y limpieza de los datos almacenados en las fuentes de datos empresariales con el propósito de realizar un proceso de estandarización, teniendo en cuenta la estructura de las direcciones residenciales colombianas.

Es necesario que el país tenga bases de datos de direcciones y que estén abiertas al público para que las empresas puedan validar las direcciones almacenadas en sus fuentes de datos. Cuando la empresa no cuenta con una base de datos para realizar búsquedas referenciales, es recomendable que utilice las fuentes de datos externas. Actualmente este servicio lo realizan empresas privadas encargadas de levantar información en campo y que continuamente están actualizando y depurando las direcciones residenciales de las poblaciones.

El caso de estudio permite evidenciar que el acierto con AddresDoctor es relativamente bajo en términos de concordancia o *matching* de direcciones residenciales, obteniendo unos resultados muy variables en las diferentes abreviaturas CL, CR, CQ, AV, DG y TV. AddressDoctor puede emplearse como una herramienta alternativa para la búsqueda de direcciones residenciales, pero resulta recomendable comparar las direcciones de la organización con fuentes de datos actualizadas.

En el caso de estudio puede apreciarse que cada una de las fases propuestas contribuye al proceso de limpieza y estandarización de las direcciones residenciales. La calidad de datos aplicada a las direcciones residenciales permite realizar búsquedas más precisas en las fuentes de datos de direcciones y los resultados posteriores estarán reflejados en una reducción de la incertidumbre en la toma de decisiones de la organización.

8 TRABAJOS FUTUROS

Hacia futuro puede automatizarse el proceso de corrección de errores que resulta ser una de las partes más complejas en la estandarización de direcciones.

Se sugiere continuar explorando las funcionalidades de servicios de mapas y localización como Google Maps u OpenStreetMaps para calidad de datos de direcciones.

Además, el caso de estudio puede extenderse a otras empresas y regiones del país para que pueda aplicarse en otras ciudades que tengan problemas particulares en el proceso de calidad de datos y en la estandarización de sus direcciones residenciales.

9 REFERENCIAS

Amon, I. (2010). *Guía metodológica para la selección de técnicas de depuración de datos*. Universidad Nacional de Colombia. Retrieved from http://www.bdigital.unal.edu.co/2033/1/71644758.20101.pdf

Amon, I. (2014). Anomalias de los Datos: Entrevista sobre Calidad de Datos. Medellin: Universidad Pontificia Bolivariana.

Ballesteros, J. (2014). Anatomía de una Dirección: Entrevista sobre direcciones y Georeferenciaccion en Colombia. Medellin: Empresa Gisco S.A.S.

Batini, C., & Cappiello, C. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. doi:10.1145/1541880.1541883

Borkar, V., Deshmukh, K., & Sarawagi, S. (2001). Automatic segmentation of text into structured records. *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data - SIGMOD '01*, 175–186. doi:10.1145/375663.375682

Camacho, O. F., & Tellez, S. (2009). *Propuesta de Estándar de las Direcciones Urbanas para los Equipamientos del Ministerio de Educación*. Bogota.

Castillo, D. (2007). *Herramientas y Metodologias de Inteligencia Competitiva*. Universidad Pontificia Bolivariana.

Cely, C. A., & Ballesteros, J. (2010). *Anatomia del proceso de geocodificación sobre mallas viales en Colombia caso municipal de Sabana Larga*. Universidad de Antioquia.

Company, H. H. (2015). Global Locator for Address Geocoding. *Trillium Software*. Retrieved August 26, 2015, from http://www.trilliumsoftware.com/home/products/globallocator/index.aspx

Copano Ortiz, L. (2014). Gestión de direcciones y viarios: Dificultades para la generación e integración de un sistema georreferenciado. *Revista de Estudios Andaluces*, *31*(31), 54–84. doi:http://dx.doi.org/10.12795/rea.2014.i31.03

DIAN. (2014). Generador de Direcciones. *Direccion de Impuestos y Aduanas Nacionales*. Retrieved September 24, 2015, from https://muisca.dian.gov.co/WebRutMuisca/visor/formularios/f19/v4/direcciones/direcciones.jsp?

Fang, L., Yu, Z., & Zhao, X. (2010). The Design of a Unified Addressing Schema and the Matching Mode of China. *Igarss 2010*, *1*, 3987–3990.

Farvacque-Viitkovic, C., & Chavez, R. (2011). Sistema de numeración de las edificaciones. *Nomenclatura urbana y administracion de ciudades*. Retrieved September 5, 2015, from http://www.cca.org.mx/ps/funcionarios/cursos/nomvial/m4/ventanas/act4.html

Farvacque-Viitkovic, C., & Godin, L. (2005). *Nomenclatura y gestión urbana*. Washington, D.C.: Banco Internacional para la Recontrucción y el Desarrollo.

Friedman, T., & Judah, S. (2013). *Gartner Research: Magic Quadrant for Data Quality Tools*. Retrieved from http://www.existbi.com/wp-content/uploads/2014/03/Magic-Quadrant-for-Data-Quality-Tools.pdf

Garcia, H. (2010). Calles sin nombre. *Kirai - Un geek en Japon*. Retrieved from http://www.kirainet.com/calles-sin-nombre/

GitHub. (2015). OpenRefine. *OpenRefine*. Retrieved August 26, 2015, from https://github.com/OpenRefine/OpenRefine

Goldberg, D. W., Ballard, M., Boyd, J. H., Mullan, N., Garfield, C., Rosman, D., … Semmens, J. B. (2013). An evaluation framework for comparing geocoding systems. *International Journal of Health Geographics*, *12*(1), 50. doi:10.1186/1476-072X-12-50

IGAC. (2005). *Manual de Reconocimiento Predial IGAC*. Santa fe de Bogota.

Informatica. (2014). AddressDoctor Product Documentation. Germany: Informatica Corporation. Retrieved from https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/other/iad-560_user-guide.pdf

Jackson, K. T. (1985). *Crabgrass Frontier: The Suburbanization of the United States*. New York: Oxford University Press.

Jae-un, L. (2013). New address system takes full effect next year. *Korea.net*. Retrieved from http://www.korea.net/NewsFocus/Policies/view?articleId=115258

Jakobiak. (1992). Exemples commentés de veille technologique. In *De la vigilancia tecnológica a la inteligencia competitiva* (Les éditio., p. 15). Paris.

Kravets, N., & Hadden, W. C. (2007). The accuracy of address coding and the effects of coding errors. *Health & Place*, *13*(1), 293–8. doi:10.1016/j.healthplace.2005.08.006

Li, D., Wang, S., & Mei, Z. (2010). Approximate Address Matching. *2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 264–269. doi:10.1109/3PGCIC.2010.43

MinTIC. (2014). ¿A quienes beneficia el uso del Codigo Postal? *Codigo Postal Colombia*. Retrieved from http://www.mintic.gov.co/portal/604/articles-5529_archivo_pdf.pdf

Moore, S. (2007). “Dirty Data” is a Business Problem, Not an IT Problem, Says Gartner. *Gartner Research*. Retrieved from http://www.gartner.com/newsroom/id/501733

Naumann, F. (2013). An introduction to data Profiling. Potsdam: Hasso Plattner Institut.

Newman, K., & Haanen, A. Rural and urban addressing. , Pub. L. No. AS/NZS 4819 (2011). Australia / New Zeland: Standards Australia Limited/Standards New Zealand.

Nieto, J. (2015). IGAC entrega cartografía digital urbana y nomenclatura de la ciudad de Cúcuta. *IGAC*, pp. 1–2. Cucuta. Retrieved from http://www.igac.gov.co/wps/wcm/connect/c42bc18040bcb32a8be7cf4234eae643/IGAC+entrega+cartografía+digital+urbana+y+nomenclatura+de+la+ciudad+de+Cúcuta.pdf?MOD=AJPERES

Niglio, O. (2014). Kioto, la antigua capital del Japón y el modelo chino de la ciudad ideal. In F. de Arquitectura (Ed.), *Arquitectura y Urbanismo* (Vol. XXXV, pp. 92–96). Cuba: Instituto Superior Politécnico José Antonio Echeverría.

Olano, R., & Morales, A. (2006). Las Calles y Carreras de Medellin. In J. Osorio (Ed.), *Medellin en la memoria de Ricardo Olano* (2006th ed., p. 335). Medellin: Instituto Tecnologico Metropolitano.

Padrón Torres, L. (2007). Estudio de Herramientas para limpiar Direcciones Postales. Cuba: Empresa de Telecomunicaciones de Cuba S.A. Retrieved from http://www.monografias.com/trabajos39/limpieza-de-datos/limpieza-de-datos.shtml

Pérez Machado, R. (2008). Procesos de Geocodificación Urbana: Los Casos De São Paulo Y Barcelona. *Revista Catalana de Geografia*, *XIII*(33), 1–14.

Planeacion, D. (2010). *Manual de Reconocimiento Predial de Antioquia*. Medellin. Retrieved from http://www.ceppia.com.co/Documentos-tematicos/TERRITORIAL/MANUAL-DE-RECONOCIMIENTO-PREDIAL.pdf

Portafolio. (2011, July 6). Cambio de direcciones afecta al 64,4% de empresarios. *Seccion de Negocios*, p. 2. Bogota. Retrieved from http://www.portafolio.co/negocios/cambio-direcciones-afecta-al-644-empresarios

Ranzijin, B. (2013). *A Geocoding Algorithm Based On A Comparative Study Of Address Matching Techniques*. Erasmus Universiteit Rotterdam.

Rosa, M. (2014, December 15). Se eliminarán transversales y diagonales en Cali para evitar enredos. *Periodico El Pais*, p. 2. Cali - Valle. Retrieved from http://www.elpais.com.co/elpais/cali/noticias/calenos-le-dicen-adios-direcciones-enredadas

Sanchez, J. (2002). *Empleo de herramientas de software que soportan Sistemas de Inteligencia Competitiva*. Universidad Carlos III de Madrid. Retrieved from http://disi.unal.edu.co/~jmsanchezt/documentos/tesina completo.pdf

Sánchez, J. M., & Tamayo, L. I. (2014). Vigilancia Tecnologica e Inteligencia Competitiva - Herramienta para la toma de decisiones. Medellin: UPB - Centro de Investigacion para el Desarrollo y la Innovacion.

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy and geographic bias of four methods of

geocoding in epidemiologic research. *Annals of Epidemiology*, *17*(6), 464–70. doi:10.1016/j.annepidem.2006.10.015

Servinformacion. (2015). SitiData Direcciones. *Soluciones integrales de Localización Inteligente*. Retrieved August 22, 2015, from http://www.servinformacion.com/?q=82/datos-con-calidad/sitidata-direcciones

SmartyStreets. (2015). Features Address Verification API. *The API to Address It All*. Retrieved July 24, 2015, from https://smartystreets.com/features

SQLPower. (2015). SQL Power DQguru - Data Quality. *Data Cleansing & Address Correction*. Retrieved August 26, 2015, from http://www.sqlpower.ca/page/dqguru

Svennerberg, G. (2010). *Beginning Google Maps API 3*. (M. Wade, Ed.)*Intel Whitepaper. Available …*. New York: Springer Verlag. Retrieved from www.apress.com.\nhttp://www.intel.cl/content/dam/www/public/us/en/documents/best-practices/mobile-app-development-framework.pdf

Tantner, A. (2009). Addressing the Houses: The Introduction of House Numbering in Europe. *Histoire & Mesure*. Viena. Retrieved from https://histoiremesure.revues.org/3942

Team, D. (2015). Datacleaner Reference Documentation. Retrieved from http://datacleaner.org/resources/docs/4.5/pdf/datacleaner-reference.pdf

Torres, G. (2015). Nueva nomenclatura de Cúcuta le dará orden a la ciudad y respetará los hitos y el arraigo cultural. *IGAC*, p. 4. Cucuta. Retrieved from http://www.igac.gov.co/wps/wcm/connect/e678bb8045d4ac238100bf543564456d/Nomenclatura+de+Cucuta.pdf?MOD=AJPERES

Vargas, J., & Horfan, D. (2013). Proceso de Geocodificación de direcciones en la ciudad de Medellín, una técnica determinística de Georreferenciación de direcciones. *USBMed*, *4*(1), 6–21.

Viola, P., & Narasimhan, M. (2005). Learning to extract information from semi-structured text using a discriminative context free grammar. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '05*, 330. doi:10.1145/1076034.1076091

Whitsel, E. A., Quibrera, P. M., Smith, R. L., Catellier, D. J., Liao, D., Henley, A. C., & Heiss, G. (2006). Accuracy of commercial geocoding: assessment and implications. *Epidemiologic Perspectives & Innovations : EP+I*, *3*(1), 8. doi:10.1186/1742-5573-3-8

Wilberg, B. V. (1993). Map Showing the New House Numbering System in the City of Chicago. In *New house numbering system in the city of Chicago* (p. 1). Chicago: Bureau of Maps & Plats. Retrieved from https://chicago.bibliocms.com/wp-content/uploads/sites/3/2015/01/MapNewOldHouseNumbersLG.jpg

Wu, P. Y., & Rathswohl, E. (2010). Address Matching : An Expert System and

Decision Support Application for GIS. *Information Systems Educators Conference ISECON Proceedings*, *27*, 1–6.

Xu, S., Flexner, S., & Carvalho, V. (2012). Geocoding Billions of Addresses: Toward a Spatial Record Linkage System with Big Data. In *Workshop on GIScience in the Big Data Age* (Internatio., Vol. 1, pp. 17 – 26).

Yang, D.-H., Bilaver, L. M., Hayes, O., & Goerge, R. (2004). Improving geocoding practices: evaluation of geocoding tools. *Journal of Medical Systems*, *28*(4), 361–70.

Zandbergen, P. A. (2008, May). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, *32*(3), 214–232. doi:10.1016/j.compenvurbsys.2007.11.006

Zandbergen, P. a. (2011). Influence of street reference data on geocoding quality. *Geocarto International*, *26*(1), 35–47. doi:10.1080/10106049.2010.537374