GUÍA METODOLÓGICA PARA LA DEPURACIÓN DE DATOS DEL SISTEMA DE INFORMACIÓN Y REGISTRO CINEMATOGRÁFICO – SIREC, DEL MINISTERIO DE CULTURA

Elaborado por:

DIANA ROCIO GÓMEZ QUINTERO

LEIBY YAZMIN LÓPEZ MURILLO

UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA – UNAD -

ESCUELA DE CIENCIAS ADMINISTRATIVAS, CONTABLES, ECONÓMICAS Y DE NEGOCIOS – ECACEN

ESPECIALIZACIÓN EN GESTIÓN DE PROYECTOS

BOGOTÁ

2018

GUÍA METODOLÓGICA PARA LA DEPURACIÓN DE DATOS DEL SISTEMA DE INFORMACIÓN Y REGISTRO CINEMATOGRÁFICO – SIREC, DEL MINISTERIO DE CULTURA

Elaborado por:

DIANA ROCIO GÓMEZ QUINTERO

LEIBY YAZMIN LÓPEZ MURILLO

Tesis de grado para optar al título de Especialistas en Gestión de Proyectos, bajo la modalidad de monografía

Director

HÉCTOR HERRERA RAMÍREZ

UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA – UNAD

ZONA CENTRO BOGOTÁ - CUNDINAMARCA

ESCUELA DE CIENCIAS ADMINISTRATIVAS, CONTABLES, ECONÓMICAS Y DE NEGOCIOS – ECACEN

BOGOTÁ

2018

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

_____

_____

_____

Firma del presidente del Jurado

_____

Firma del Jurado

_____

Firma del Jurado

Bogotá, 19 de mayo de 2018

Dedicatoria

Dedicamos este trabajo a nuestras familias en especial a nuestros hijos, pues queremos ser ejemplo para ellos y demostrarles que todos nuestros sueños y objetivos en la vida se pueden lograr a pesar de las dificultades.

Agradecimientos

Agradecemos primeramente a Dios por permitimos dar un paso más en esta etapa de nuestras vidas, por su amor y protección.

Agradecemos a nuestras familias por su amor, acompañamiento y apoyo incondicional en todas las áreas de nuestras vidas.

Este logro alcanzado es con esfuerzo y orgullo, también es un paso más para continuar con nuestra área intelectual y laboral.

Resumen

Mediante este trabajo, se da a conocer los pasos previos para llevar a cabo una buena ejecución y monitoreo de la información, al momento de ingresar los datos en el Sistema de Información y Registro cinematográfico – SIREC. Teniendo en cuenta las validaciones en los campos y estructuras de las bases de datos, con formatos estándar, estructurados e interoperables que faciliten el acceso y reutilización de la información.

Algunos de los datos almacenados en el sistema de información, contienen errores debido a malas prácticas en el ingreso de la información lo cual conlleva a tomar decisiones erróneas, pérdida de tiempo, dinero y credibilidad. Esta situación ha capturado la atención de los investigadores, llevando al desarrollo de múltiples técnicas para detectar y corregir los problemas en los datos. Para lograr buenos resultados en procesos de limpieza de datos, la elección de la técnica es fundamental, pero no se conoce de alguna metodología que detalle la forma de realizar dicha selección de técnicas. Es por esto por lo que esta tesis construye una guía metodológica que oriente al analista de los datos del Sistema de Información y Registro Cinematográfico - SIREC hacia una selección de las técnicas adecuadas para aplicar a un conjunto de datos particular de un dominio específico que presentan tres problemas detectados en los registros de información que son: duplicados, valores atípicos incorrectos y valores faltantes. Para la construcción de la guía, se caracterizaron varias técnicas para cada uno de los tres problemas de datos bajo estudio, examinando su eficacia ante diferentes casos o situaciones problemáticas propuestas. Para realizar comparativos y validar la guía, se utilizaron tanto datos de prueba como reales pertenecientes a las bases de datos de agentes del sector cinematográfico colombiano,

inscritos en el SIREC.

Palabras claves: Calidad de datos, datos nulos, duplicidad, procesamiento de datos, heterogeneidad de datos.

Abstract

Through this work, the previous steps to carry out a good execution and monitoring of the information are made known, when entering the data in the Film Information and Registration System - SIREC. Considering the validations in the fields and structures of the databases, with standard, structured and interoperable formats that facilitate access and reuse of information.

Some of the data stored in the information system contain errors due to bad practices in the entry of information which leads to wrong decisions, loss of time, money and credibility. This situation has captured the attention of researchers, leading to the development of multiple techniques to detect and correct problems in the data. To achieve good results in data cleansing processes, the choice of technique is fundamental, but there is no known methodology to detail the way to perform this selection of techniques. Therefore, this thesis builds a methodological guide to guide the data analyst of the Information System and Cinematographic Registry - SIREC towards a selection of the appropriate techniques to apply to a data set of a specific domain that present three problems detected in the information registers that are: duplicates, incorrect outliers and missing values. For the construction of the guide, several techniques were characterized for each of the three data problems under study, examining their effectiveness in

different cases or problematic situations proposed. To make comparisons and validate the guide, both test and real data belonging to the databases of agents of the Colombian cinematographic sector, registered in SIREC, were used.

Keywords: Data quality, null data, duplicity, data processing, data heterogeneity.

Tabla de Contenido

1. Introducción…… 1

2. Definición del Problema…… 2

2.1. Planteamiento del Problema…… 2

2.2. Justificación…….. 3

2.3. Objetivo General…… 4

2.4. Objetivos Específicos…….. 4

2.5. Alcances y Delimitaciones…… 4

2.5.1. Alcances…… 4

2.5.2. Delimitaciones……... 5

3. Marco de Referencia…… 5

3.2. Antecedentes…… 5

4. Marco Teórico……...6

4.1. ¿Qué es Calidad de los Datos?..... 6

4.2. ¿Qué Dificulta Controlar la Calidad de Datos?..... 6

4.2.1. Procesos Que Afectan la Calidad de Datos…… 7

4.2.2. Descripción de las Fallas Presentadas en los Procesos Internos…….. 8

4.3. Dimensión de la Calidad de Datos…… 9

4.3.1. Área de Investigación en Calidad de Datos…… 9

4.4. ¿Qué es la Información?..... 10

4.5. Base de Datos y Calidad de la Información……. 10

4.5. Problemas en las Bases de Datos…….. 11

4.5.1. Tipos de Errores…….. 12

4.5.2. Valores de Atributo de una Tupla…….. 12

5. Marco Conceptual…… 13

6. Marco Legal…… 13

6.1. Ley 814 de 2003…… 13

6.1.2. Ley 1581 de 2012……. 14

7. Metodología…… 14

7.1. Fase de Análisis……. 15

7.1.1. Diagnóstico de las Bases de Datos……...15

7.1.2. Cantidad de Errores Encontrados…….. 16

7.1.3. Problemas que Enfrenta la Limpieza de Datos…….. 17

7.2. Errores Que se Presentan en el Sistema de Información…….. 19

7.2.1. Errores Asociados a la Captura de la Información……. 19

7.3. Controles de Error……. 20

7.3.1. Captura (On line)……. 22

7.3.2. Captura (Off Line)……. 23

7.3.3. Comparación Visual……... 24

7.3.4. Previsión de los Valores Desconocidos (Missing)…….. 24

7.3.5. Pre-validaciones, In- validaciones y Post-validaciones…….. 25

7.4. Errores Encontrados en los Formularios de Captura…….. 25

8. Errores en el Sistema por Dimensión…… 30

8.1. Dimensiones Exactitud y Unicidad…….. 31

8.2. Dimensiones Relacionadas con el Tiempo…… 36

8.2.1. Dimensión Completitud……. 40

9. Fase Hacer…….. 45

9.1. Proceso de Prevención de Errores……. 45

9.1.2. Reglas de Validación en el Formulario Captura de la Información de los Agentes….47

9.2. Validaciones del Formulario de Captura…….. 50

10. Proceso de Estandarización de la Información…… 53

10.1. Estandarización del Campo Dirección……. 53

10.2. Estandarización del Campo Número Telefónico…….. 57

10.3. Estandarización del Campo Celular…… 59

10.4. Estandarización de los Códigos de las Ciudades y Departamentos…… 60

11. Proceso de Unicidad de la Información…… 62

12. Proceso de Completitud de la Información…… 63

13. Proceso de Corrección de Información por Inexactitud Semántica…….. 64

14. Proceso de Corrección para Registro Inexistente……. 66

15. Fase Verificar…….. 67

15.1. Plan de Monitoreo…….. 68

16. Conclusiones…… 69

17. Bibliografía…… 70

Lista de Tablas

Tabla 1. Diccionario de Datos fuente: SIREC 16
Tabla 2. Resumen de errores. Fuente Base de datos SIREC 17
Tabla 3. Base de Datos – SIREC. Agentes, Análisis de Direcciones en Hoja de Cálculo 27
Tabla 4. Agentes, errores en el campo teléfono. Fuente SIREC 28
Tabla 5. Tabla de Agentes, errores en el campo celular. 29
Tabla 6. Tabla de Agentes, errores en el campo E-mail. 30
Tabla 7. Errores en el SIREC por Dimensión 31
Tabla 8. Consultas Agentes con Números Telefónicos en Hoja de Cálculo 36
Tabla 9. Base de Datos – SIREC. Agentes, Análisis de Direcciones en Hoja de Cálculo 27
Tabla 10. Reglas de Validaciones en Formulario de Captura 48
Tabla 11. Elementos Mínimos en un Registro de Dirección. 53
Tabla 12. Tabla de Abreviaturas para las Vías. 54
Tabla 13. Ejemplo de Registro de Números Telefónicos en el SIREC 57
Tabla 14. Valores que se presentan en el campo celular 58
Tabla 15. Códigos de los Municipios de acuerdo con el estándar DIVIPOLA 60

Lista de Figuras

Figura 1..Criterio de Calidad de Datos y Descripción.....7

Figura 2. Gestión de Calidad de Datos..... …….. ……11

Figura 3. Módulo Agentes – Sistema de Información – SIREC 15

Figura 5. Modelo relacional Base Agente, llave primaria AGE_ ID……22

Figura 6. Formulario de Datos de Contacto de un Agente……..26

Figura 7. Base de datos Agentes, Análisis de direcciones……26

Figura 8. Consulta SQL para Identificar los Números Telefónicos 35

Figura 9. . Registro Duplicado donde las Cédulas son Diferentes 38

Figura 10. Formulario Director 39

Figura 11.. Información de Divipola.....42

Figura 12.. Consulta SQL.....44

Figura 13 .Errores de Valores Nulos.....44

Figura 14. Formulario de datos de contacto de un Agente.....49

Figura 15. Validación para Fecha Registro y Fecha Actualización 50

Figura 16. Validación por nombre 50

Figura 17. Validación persona Jurídica ……...51

Figura 18. Formulario de Registro Campo Actual Dirección de un Agente …… 55

Figura 19. Campos de la Dirección ……..56

Figura 20.Caso de Contradicción en un Registro de Agente.....61

Figura 21. Opción Eliminar en el Formulario de un Agente Cinematográfico.....65

1. Introducción

El Sistema de Información y Registro Cinematográfico fue creado por la Ley de Cine 814 de 2003. con el fin de consolidar información del sector cinematográfico colombiano y para realizar el adecuado seguimiento del recaudo de la cuota para el desarrollo cinematográfico. Debido a su relevancia y su creciente papel en la toma de decisiones en la política cinematográfica de Colombia, es cada vez más importante, que la información almacenada esté libre de errores y así evitar tomar decisiones erróneas. Idealmente, los datos almacenados no deberían contener errores, pero es innegable que son una realidad que merecen atención. El presente trabajo ofrece a los analistas de los datos una guía metodológica que permite identificar los errores presentes en la base de datos de los agentes cinematográficos (Exhibidores, productores, distribuidores, directores, personal técnico y artístico) registrados en el sistema y las técnicas de corrección de los errores desde el proceso de captura.

La base de datos de agentes presenta tres tipos de errores: valores duplicados, valores faltantes (nulos) y valores atípicos, algunos de ellos incorrectos o variaciones tipográficas. Para cada tipo de error se sugiere una técnica para detectarlo y corregirlo. La calidad de la limpieza lograda sobre los datos depende de la técnica aplicada y la naturaleza de los datos específicos sobre los que se está trabajando. Contar con una guía metodológica que oriente la selección de la técnica a aplicar en un caso particular, es un apoyo significativo para las personas que deben realizar tareas de depuración a los datos. En esta tesis se construyó una metodológica que se estableció en tres fases: Planificar, Hacer y Verificar la cual permitirá un proceso de mejora continua en cada fase terminada.

2. Definición del Problema

2.1. Planteamiento del Problema

La Dirección de Cinematografía del Ministerio de Cultura tiene a cargo el funcionamiento de un sistema llamado SIREC - Sistema de Información y Registro Cinematográfico SIREC, el cual fue creado por la Ley 814 de 2003 cuyo propósito consiste en consolidar información, apoyar procesos de seguimiento y toma de decisiones de las políticas públicas del sector cinematográfico.

El sistema cuenta con 20 módulos y 12 funcionarios operando el sistema, de los cuales 7 registran información al sistema. El SIREC cuenta con 10.000 registros de agentes del sector cinematográfico colombiano que incluyen exhibidores, productores, distribuidores, directores, personal técnico y artístico, asociaciones etc.… La particularidad de este sistema es que un agente del sector puede estar registrado en diferentes roles o cargos, por ejemplo: un productor puede estar también registrado como director y en personal técnico o artístico, por otra parte, existen diferentes entradas de información al sistema de acuerdo con el rol.

El gran volumen de datos, las múltiples entradas de información, las malas prácticas en el tratamiento de la información y las pocas reglas de validación de los campos de los formularios de captura han generado inconsistencias en los datos, tales como valores duplicados, nulos y atípicos. Lo anterior conlleva a graves consecuencias como pérdida de tiempo por reprocesos de los datos, afectando la calidad y veracidad de la información estadística. Por lo tanto, se requiere crear una metodología que permita en tres fases (Analizar, Hacer y Verificar) realizar la corrección de la información inconsistente y mitigar las causas que generan errores desde la captura hasta su difusión.

¿En qué medida el diseño de los formularios de captura de la información y las múltiples entradas de información impactan la calidad de la información procesada en el SIREC?

2.2.Justificación

Los datos procesados adquieren un valor agregado, lo que conlleva a que su contenido sea fiable y corresponda a la realidad. Este valor agregado hace referencia a la calidad, la cual genera un nivel de confianza sobre la información arrojada, permitiendo que se publiquen datos exactos, medibles, comprobables que ayuden a una mejor toma de decisiones y colaborando en el cumplimiento de los objetivos del Ministerio de Cultura y las entidades que hacen uso de ella.

El Sistema de Información y Registro Cinematográfico – SIREC tiene la misión de ofrecer información estratégica para el diseño de la política cinematográfica nacional. En este se procesa información referente a espectadores y recaudo de taquilla en salas de cine nacional, contiene datos referentes a inversiones y donaciones a películas colombianas, guarda información de los largos y cortometrajes reconocidos como producto nacional, es una herramienta que permite gestionar el seguimiento del recaudo de la cuota para el desarrollo cinematográfico entre otros.

Dada la importancia de la información manejada es necesario que la información almacenada cumpla con los atributos de exactitud, relevancia, accesibilidad, transparencia, coherencia, oportunidad y puntualidad, de tal manera que se garantice la calidad de la información procesada y difundida oficialmente en el país, con el fin de mitigar los riesgos asociados a la perdida de información e inexactitud.

Lo anterior exige establecer una metodología que permita la validación y corrección de los datos que presentan errores de inconsistencias, como ausencia, duplicidad y valores atípicos.

2.3.Objetivo General

Diseñar una guía metodológica que permita realizar la depuración de la información de los datos de contacto de los agentes del sector cinematográfico registrados en el Sistema de Información y Registro Cinematográfico – SIREC, del Ministerio de Cultura.

2.4.Objetivos Específicos

1. Determinar qué errores presentan los datos almacenados en el SIREC y los factores asociados a los errores.
2. Establecer las técnicas de detección y corrección de errores a utilizar en cada caso.
3. Definir reglas de validaciones, que pueden ser aplicadas al momento de la captura de información en el sistema.

2.5.Alcances y Delimitaciones

2.5.1. Alcances.

Los datos analizados hacen parte de la base de datos de los agentes cinematográficos: productores, directores, exhibidores, distribuidores, personal técnico y artístico, los cuales contienen los siguientes campos: tipo de Id, identificación, nombres, apellidos, dirección, teléfono, celular, E-mail y página web,

Con la implementación de una metodología para la detección y corrección de los errores de duplicidad, datos faltantes o nulos y valores atípicos en el sistema de información SIREC se pretende eliminar los errores en la información, mediante procesos de unificación de los datos para el caso de duplicidad, imputación de datos para datos faltantes y corrección de información inconsistente por valores atípicos o tipográficos con el fin de mejorar la calidad y estructura de los datos del sistema mitigando así los riesgos de error desde la captura de la

información.

Con este trabajo se busca la optimización del diseño de los formularios de captura y la reducción de tiempo y costos operacionales, contribuyendo a la consolidación de la información estadística precisa, generando una cultura de confianza en la información recibida y difundida a nivel nacional desde el SIREC.

2.5.2. Delimitaciones.

La metodología será convalidada con base a los datos arrojados por el Sistema de Información y Registro Cinematográfico SIREC y soportados con la documentación existente (expedientes físicos de Proyectos y Productos); los datos identificados como duplicados no se eliminarán, se unificarán.

Para la consulta y detección de errores solo se utilizará consulta en SQL, y se exportarán los datos a hoja de cálculo. No se utilizará ninguna herramienta de análisis de datos por licenciamientos y políticas de seguridad de la entidad.

3. Marco de Referencia

3.2.Antecedentes

Múltiples trabajos se han realizado con respecto al tema de calidad de datos, los trabajos de investigación incluyendo aquellos de tipo metodológico, no se ocupan lo suficiente de la selección de las técnicas para la depuración de datos, en un caso particular. El trabajo de Oliveira, P., Rodríguez, F., Henriques, P., y Galhardas, H. (2005), plantea como detectar la anomalía de los datos sin indicar cual técnica usar para la detección y/o corrección. Tierstein, Leslie, aunque presenta una metodología que intenta cubrir todo el proceso de depuración de datos, se enfoca principalmente hacia el manejo de los datos históricos, no examina las técnicas existentes para la depuración, no se ocupa de recomendar una técnica y no tiene en

cuenta la naturaleza de los datos. Rosenthal, A., Wood, D., y Hughes, E. (2001), están orientados al enriquecimiento de los sistemas de bases de datos con metadatos, sin examinar ni recomendar técnicas de depuración.

Como trabajos previos y referencia para el desarrollo de la monografía, se relacionan los siguientes:

1. Tesis de Maestría: Guía metodológica para la selección de técnicas de depuración de datos. Autor: Iván Amón Uribe (2012).
2. Tesis de pregrado Ingeniería Informática UPB: Caracterización de técnicas para detección de valores faltantes. Autor: Carolina Muñoz (2010).

4. Marco Teórico

4.1.¿Qué es Calidad de los Datos?

Los datos representan objetos del mundo real. Dichas representaciones resultan ser aplicables en contextos de diferentes y variadas características. Por otro lado, los datos pueden ser almacenados o sometidos a algún proceso o transformación, siendo siempre de suma importancia para garantizar el éxito las entidades.

Se puede decir, que la definición de la calidad de los datos está relacionada estrechamente con la exactitud, completitud, consistencia y actualidad de los datos.

4.2.¿Qué Dificulta Controlar la Calidad de Datos?

Los cambios continuos y las rápidas implementaciones de sistemas de información; los métodos, técnicas y herramientas para controlar la calidad no se han desarrollado al mismo ritmo que la construcción e implementación de los sistemas de información. Hay acciones desfavorables que conllevan a la mala calidad de los datos, tales como datos duplicados e integración. Los datos, son de calidad, si son adecuados para lo que se necesitan y depende

el uso de estos; cuando se le da un uso diferente al previsto en el diseño original degradan la calidad de la base de datos.

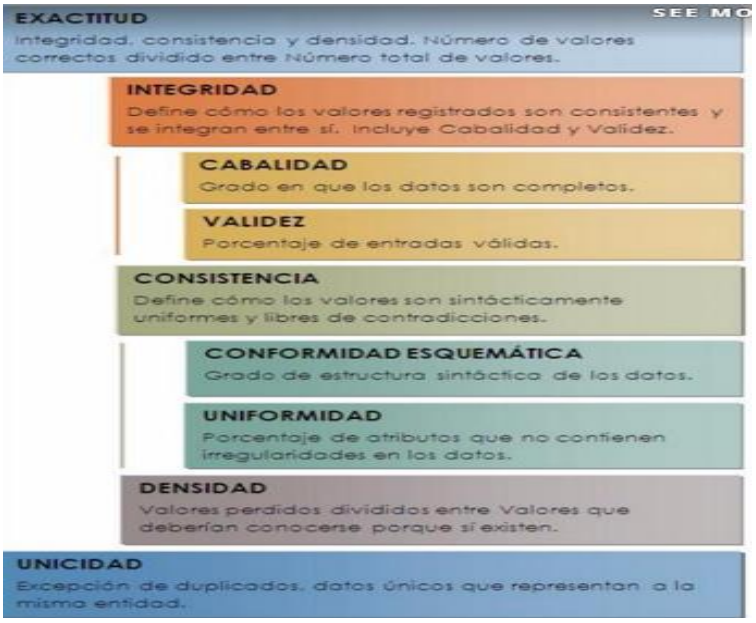


Figura 1. Criterio de Calidad de Datos. Modificado de: Ahmed & Aziz, (2010); Muller y Freytag,(2003).

4.2.1. Procesos Que Afectan la Calidad de Datos.

Procesos externos:

a. Migración de datos
b. Consolidación de sistemas
c. Entrada manual de datos (errores de digitación)
d. Interfaces en tiempo real
e. Campos vacíos
f. Valores desactualizados
g. Valor fuera del rango permitido
h. Errores de ortografía
i. Valores sin sentido

Deterioro Natural:

j. Nuevo Uso de los Datos
k. Cambios no registrados
l. Actualización de los sistemas
m. Automatización de procesos

4.2.2. Descripción de las Fallas Presentadas en los Procesos Internos.

Migración de los datos de un sistema, antiguo, a un nuevo sistema. El proceso requiere establecer la correspondencia entre la estructura original y la nueva estructura, por mencionar algunos:

a. Metadata incompleta
b. Condiciones específicas incorporadas en el código (programa fuente)
c. Valores faltantes o nulos
d. Las reglas de negocios del sistema nuevo son diferentes al sistema antiguo
e. Duplicidad de datos.

Consolidaciones: Constantemente los datos de la fuente se trasladan a una base de datos que ya contiene información, lo cual genera toda clase de conflictos de datos, tales como duplicados, es una de las principales causas de problemas de calidad de datos.

Entrada Manual: al ingresar los datos manualmente por interfaces o formularios los principales errores presentados son captura errada de los datos, formularios con fallas en el diseño, provocando el ingreso de datos errados, valores por defecto, falta de instrucciones adecuadas (Metadata).

Interfaz en tiempo real: los sistemas intercambian datos con interfaces en tiempo real, esto permite tener la información sincronizada, pero no brinda el tiempo para verificar que los datos sean los correctos.

4.3.Dimensión de la Calidad de Datos

Cada dimensión refleja un aspecto distinto de la calidad de los datos. Las mismas pueden estar referidas a la extensión de los datos (su valor) o a la intensión (su esquema). De esta manera se puede diferenciar entre calidad en los datos y calidad en los esquemas. El enfoque del presente documento es hacia la calidad de los datos.

Los conceptos como exactitud, actualidad y completitud. Todas estas características de los datos, se denominan dimensiones de la calidad de datos.

Se define factor de calidad como un aspecto particular de una dimensión. Una dimensión puede ser vista como un agrupamiento de factores de calidad que tienen el mismo propósito.

4.3.1. Área de Investigación en Calidad de Datos.

Logar la calidad de datos, es una tarea compleja debido a su importancia, naturaleza, y los diferentes tipos de datos y sistemas de información que pueden estar implicados.

Dentro del área de calidad de datos, se incluye los siguientes puntos:

a) Dimensiones: las mediciones sobre el nivel de calidad de los datos se aplican a las dimensiones de interés.

b) Metodologías: Proveen guías de acción.

c) Modelos: Representan las dimensiones u otros aspectos de calidad de datos.

d) Técnicas: Proveen soluciones a problemas de calidad de datos.

e) Herramientas: son necesarias para que las metodologías y técnicas que puedan ser implementadas de forma efectiva.

4.4. ¿Qué es la Información?

La información es un conjunto de datos acerca de algún suceso, hecho, fenómeno o situación, que organizados en un contexto determinado tienen su significado, cuyo propósito puede ser el de reducir la incertidumbre o incrementar el conocimiento acerca de algo.

Los datos se pueden convertir en información añadiéndoles valor (Martínez, 2012). La conversión de la información no es más que la transformación de un esquema de representación de los elementos de información a otro, incluyendo los siguientes elementos:

1. Datos ingresados o almacenados en un sistema de información.
2. Reportes Generados.
3. Parámetros de configuración del sistema.
4. Formularios manuales o automatizados.

4.5. Base de Datos y Calidad de la Información

Para que las bases de datos sean útiles y puedan ser procesadas, analizadas e interpretadas de manera eficaz y eficiente, requieren que los datos que las conforman sean los correctos.

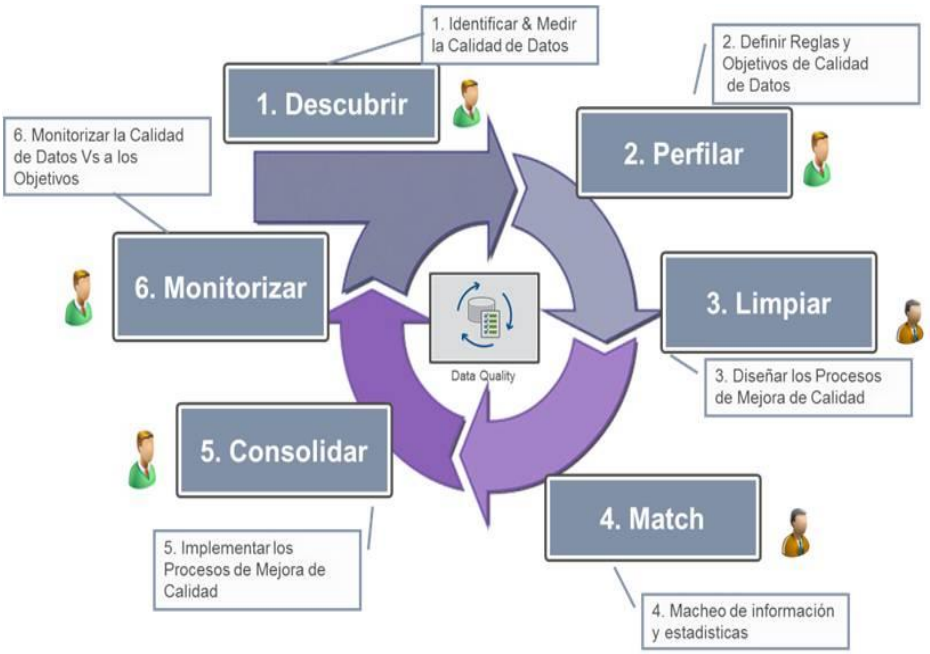


Figura 2 Gestión de la Calidad de Datos

No solo tener datos precisos o exactos se habla de datos de calidad, los cuales cumplen unos criterios que le dan un valor agregado a la información. Entre estos están la exactitud, integridad, cabalidad, validez, consistencia, uniformidad, densidad y unicidad. (Ahmed &Aziz, 2010); (Müller & Freitag, 2003). Pipino, Lee y Wanga (2002). Agregan que la calidad de los datos es un concepto multidimensional que abarca accesibilidad, cantidad apropiada de información, credibilidad, datos completos, representación consistente, facilidad de manipulación, libertad de error, interpretabilidad, objetividad, relevancia, seguridad, datos oportunos, facilidad de comprensión. Cuando los datos no cumplen con criterios de calidad se consideran "sucios" debido a que cualquier información inválida, inconsistente, incompleta o incorrecta, atenta contra la integridad de cualquier tipo de base de datos (Kitlas, 2012).

4.5. Problemas en las Bases de Datos

Existen diferentes tipos de errores en las bases de datos y su clasificación difiere mucho de una a otra ya que varía de un autor a otro, sin embargo, es posible identificar ciertos tipos comunes en las investigaciones de este tema. A continuación, se mostrará una clasificación.

4.5.1. Tipos de Errores.

Estos errores se dan en una tabla de la base de datos:

Errores en un sólo campo:

- Campos vacíos
- Valores desactualizados
- Valor no perteneciente al conjunto
- Errores de ortografía
- Valor no perteneciente al contexto
- Valores sin sentido
- Valores imprecisos

4.5.2. Valores de Atributo de una Tupla

- Tupla semi-vacía
- Inconsistencia entre valores de atributos

Múltiples fuentes de datos:

Estos tipos de errores se encuentran en múltiples bases de datos cuando éstas son analizadas como un todo.

- Inconsistencias de sintaxis
- Inconsistencia en la representación
- Existencia de sinónimos
- Existencia de homónimos
- Redundancia sobre una entidad
- Inconsistencia sobre una entidad

5. Marco Conceptual

Atributos: En bases de datos son el conjunto de propiedades de una entidad, esto está representado por las columnas de una tabla.

Valores Atípicos (Outliers): son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Los datos atípicos son ocasionados por:

- Errores de procedimiento
- Acontecimientos extraordinarios
- Valores extremos

Valores Faltantes: Los datos atípicos distorsionan los resultados de los análisis, y por esta razón hay que identificarlos y tratarlos de manera adecuada, generalmente excluyéndolos del análisis.

Datos Nulos (Null Data): Normalmente, los valores NULL indican que los datos son desconocidos, no aplicables o que se agregarán posteriormente.

Datos Erróneos: datos que son válidos, pero no se ajustan a la entidad real.

6. Marco Legal

6.1. Ley 814 de 2003

Capítulo 1, artículo 4. Competencias.

Mantener, para efectos del adecuado seguimiento y control a la cuota para el desarrollo cinematográfico y ejecución de los recursos del Fondo para el Desarrollo Cinematográfico y para el cumplimiento de las políticas públicas a su cargo, un sistema de información y

registro cinematográfico, que se denominará SIREC sobre agentes o sectores participantes de la actividad cinematográfica en Colombia, y, en general, de comercialización de obras en los diferentes medios o soportes, niveles de asistencia a las salas de exhibición.

6.1.2. Ley 1581 de 2012

La ley de protección de datos personales es una ley que complementa la regulación vigente para la protección del derecho fundamental que tienen todas las personas naturales a autorizar la información personal que es almacenada en bases de datos o archivos, así como su posterior actualización y rectificación. Esta ley se aplica a las bases de datos o archivos que contengan datos personales de personas naturales.

7. Metodología

Para realizar la detección de errores del Sistema de Información y Registro Cinematográfico SIREC, se define una metodología ágil y estructurada, que consta de tres fases que permiten realizar todo el proceso detección, depuración y verificación de los errores del sistema en las tablas relacionadas con los datos de contacto de los agentes registrados en el sistema.

Las Fases que se definen en la metodología son las siguientes:

Fase de Análisis: en esta fase se identificará los tipos de errores que se presentan en el sistema y los factores asociados que causan dichos errores.

Fase de Hacer: en esta fase se especifican todos los procesos que se deben realizar en el sistema para corregir los errores presentados y se proponen los mecanismos para la prevención de los errores.

Fase de Verificación; en esta fase de muestra la manera de realizar pruebas de verificación de los procesos de corrección de la información que se hicieron en la etapa de hacer.

7.1.Fase de Análisis

7.1.1. Diagnóstico de las Bases de Datos.

El Sistema de Información y Registro Cinematográfico SIREC, está desarrollado en .NET, actualmente cuanta con 20 Módulos en los cuales se tiene la opción de registrar información, y generar reportes, las acciones establecidas para cada módulo son: crear nuevo registro, actualizar, eliminar, consultar.

El módulo donde se ingresa directamente la información se llama Agentes, ver figura 3, a continuación, en la Tabla 1 y se describe los campos requeridos para el registro de un agente en el SIREC.

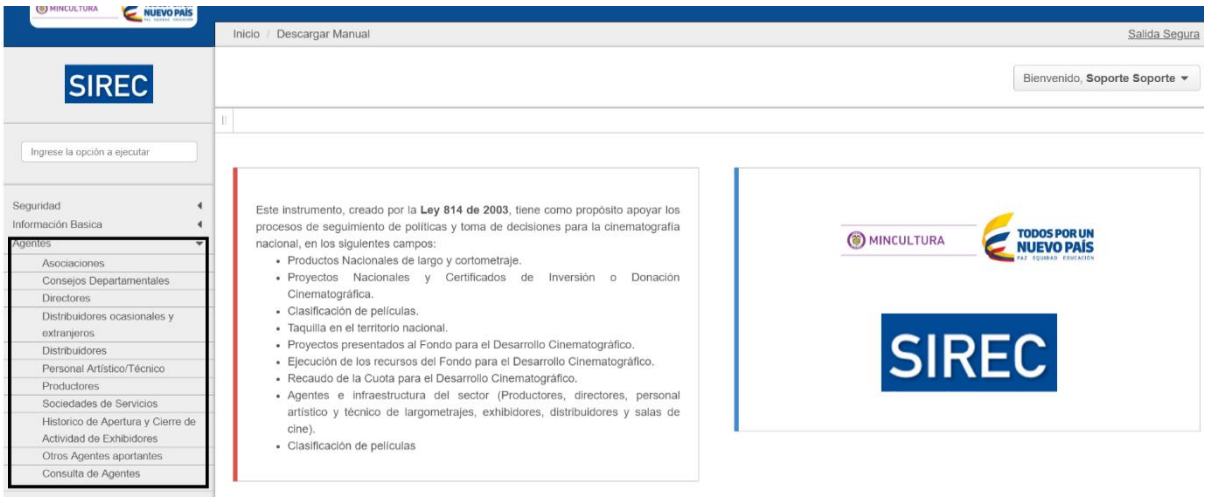


Figura 3 Módulo Agentes – Sistema de Información – SIREC.

DICCIONARIO DE DATOS DE LA BASE DE DATOS AGENTES

CAMPO	DESCRIPCIÓN DEL CAMPO	TIPO DE DATO
SOR_ID	Es el número de identificación de la solicitud, este número es automático generado por el sistema.	Varchar

SOR_FECHA_ REGISTRO	Fecha en la cual se realiza la solicitud del registro del agente	Datetime
SOR_FECHA _ACTUALIZACION	Fecha en la cual se realiza la solicitud del registro del agente	Datetime
AGE_TIPO_PERSONA	Tipo de Persona, Natural o Jurídica	Char
AGE_ID	Número de identificación del solicitante, para este proceso es un agente persona Natural o Jurídica	Varchar
AGE_NOMBRES	Nombre del agente	Varchar
AGE_APELLIDOS	Apellidos del agente	Varchar
AGE_TIPO_ID_REP_LEGAL	Tipo de Identificación del agente (NIT, CC, o, CE)	Varchar
AGE_ID_REP_LEGAL	Identificación del representante legal para persona Jurídica	Varchar
AGE_NOMBRES_REP_LEGAL	Nombre del representante legal para persona Jurídica	Varchar
AGE_APELLIDOS_REP_LEGAL	Apellidos del representante Legal para persona Jurídica	Varchar
AGE_DIRECCION	Dirección del agente	Varchar
AGE_TELEFONO	Teléfono del agente	Varchar
AGE_CELULAR	Celular del agente	Varchar
AGE_EMAIL	Email del Solicitante productor	Varchar
PAIS_ID	Identificación del país	Varchar
ZON_ID	Identificador de Ciudad o Zona	Int

Tabla 1. Diccionario de Datos fuente: SIREC

.

Actualmente la base de datos de agentes del SIREC contiene 10.000 registros de personas vinculadas al sector cinematográfico en Colombia, este número de registros será el conjunto por evaluar y determinar el tipo de errores contenidos en los campos pertenecientes a cada registro.

7.1.2. Cantidad de Errores Encontrados.

A continuación, se describen numéricamente el número de errores encontrados en cada campo del registro de un agente cinematográfico:

	ERRORES ENCONTRADOS		
	Inconsistencias	**Contradicciones**	**Nulos**
SOR_ID	0	0	0
SOR_FECHA_ REGISTRO	100	0	0
SOR_FECHA _ACTUALIZACION	85	0	0
AGE_TIPO_PERSONA	2036	560	85
AGE_ID	596	698	80
AGE_NOMBRES	159	91	75
AGE_APELLIDOS	190	37	55
AGE_TIPO_ID_REP_LEGAL	20	10	0

AGE_ID_REP_LEGAL	225	29	5
AGE_NOMBRES_REP_LEGAL	131	0	36
AGE_APELLIDOS_REP_LEGAL	789	5	52
AGE_DIRECCION	3895	910	236
AGE_TELEFONO	2051	91	2063
AGE_CELULAR	1563	59	1825
AGE_EMAIL	520	933	1000
PAGINA_WEB	398	123	5020
PAIS_ID	0	0	0
ZON_ID	200	0	0
TOTAL	12958	3546	10532

Tabla 2. Resumen de errores. Fuente Base de datos SIREC.

7.1.3. Problemas que Enfrenta la Limpieza de Datos.

Tanto la limpieza como la transformación de datos se encuentran relacionadas a resolver la misma problemática, ya que es necesario realizar transformaciones a nivel de la estructura, representación o contenido de los datos para lograr efectivamente su limpieza.

Los problemas que enfrenta la limpieza de datos se pueden clasificar como sigue:

- Problemas provenientes de una sola fuente de información.

La calidad de los datos depende en gran medida de las restricciones de integridad y el esquema en el cual se encuentran inmersos. Por ejemplo, las bases de datos tienen menor probabilidad de poseer errores e inconsistencias en los datos, a diferencia de los archivos de texto plano en los cuales no existe ningún tipo de reglas ni restricciones con respecto a los datos ni sus valores.

Se distinguen además problemas a nivel del esquema o a nivel de instancia. Estas últimas son las que conciernen a la calidad de los datos, y son ocasionados por ejemplo por errores de tipeo.

- Problemas provenientes de varias fuentes de información.

Cuando se integran varias fuentes de información, los problemas existentes para una

sola fuente se incrementan drásticamente. En este caso, se distinguen dos tipos de problemas a nivel del esquema:

- Conflictos de nombres.

Cuando se utiliza el mismo nombre para representar distintos objetos, o cuando distintos nombres representan el mismo objeto.

- Conflictos estructurales

Cuando el mismo objeto se representa de distinta manera en fuentes de información distintas.

A nivel de instancia, los conflictos que pueden suceder son:

Diferentes interpretaciones del mismo valor (por ejemplo, el campo fecha, en algunos formularios relacionan controles de calendario como: mes/ día/año, en algunos otros relacionan día/mes/año, dando confusiones para determinar la fecha real de registro o actualización de los datos en el sistema.

La estandarización de las fechas de registro de un agente y la fecha de actualización de la información de un agente, deben estandarizarse en el formulario y en los reportes asociados como formato día/mes/año con el fin de no generar errores de inconsistencias al momento de consultar la información de los agentes en un periodo determinado.

- Diferentes niveles de agregación
- Diferentes puntos en el tiempo.

Sin duda, una de las mayores problemáticas de la limpieza de datos es la identificación de datos que representan el mismo objeto del mundo real. Sin embargo, al momento de realizar esta tarea es necesario considerar que a pesar de que existe información redundante, en muchas ocasiones los datos que representan el mismo objeto podrían complementarse (por ejemplo, obtener la dirección y el teléfono a partir del registro de una persona, y su E-mail o página web a partir de otro registro de la misma persona).

7.2.Errores Que se Presentan en el Sistema de Información

¿Qué es un error?

El error es la inexactitud entre el valor de una medida y el valor real del objeto, está media de inexactitud puede darse por diversos motivos.

7.2.1. Errores Asociados a la Captura de la Información.

En el Sistema de Información y Registro Cinematográfico los principales factores asociados a los errores de la información están relacionados directamente en la captura de la información, a continuación, se mencionan los tipos de errores:

7.2.1.1.*Errores de Observación.*

Cuando el valor capturado es diferente al valor real de la variable por errores en el tratamiento de información

7.2.1.2.*Errores en el Formulario de Recolección.*

Se generan cuando el formato de registro presenta deficiencias en el diseño o no es claro su diligenciamiento, también cuando no se verifica la completitud del diligenciamiento del formulario y se presentan datos faltantes.

7.2.1.3.*Errores en la Digitación.*

Se cometen fallas en el ingreso de la información, debido a que el digitador no tuvo una capacitación apropiada o hubo falta de supervisión y control.

7.2.1.4.*Errores en el Procesamiento.*

Pueden ocurrir problemas en los procesos de captura de la información (pérdida de

información luego de una migración o integración de datos).

En esta etapa se realiza el análisis y valoración de los procesos relacionados a la recolección, procesamiento, depuración y divulgación de la información ingresada al SIREC, se determinan los objetivos estratégicos para mitigar los riesgos de errores en la información.

7.3. Controles de Error

La siguiente figura 4 presenta cada una de las fases de gestión de los datos y las diversas propuestas para reducir el error.

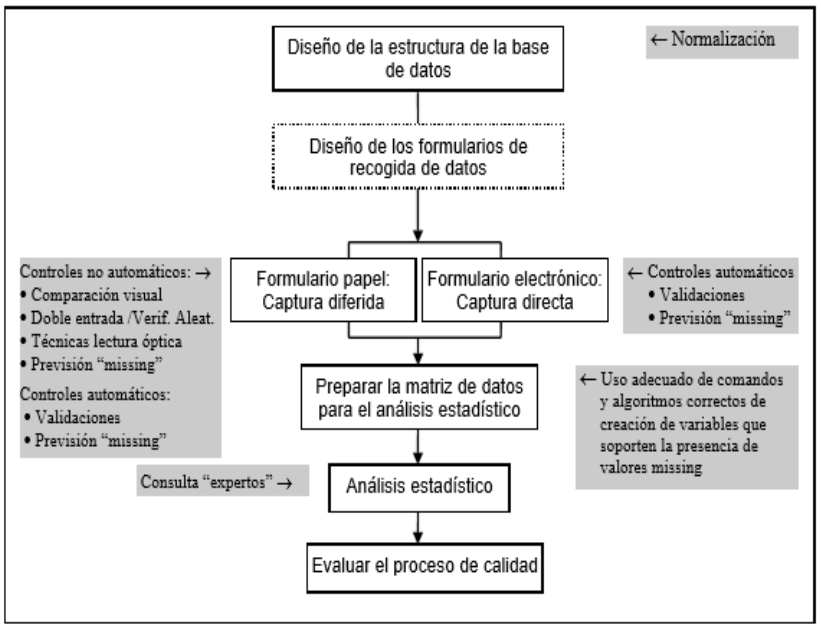


Figura 4: Fases y Controles Básicos de Calidad en la Gestión de Datos (modificada de Granero 1999).

Atendiendo la primera fase, es indispensable identificar el diseño de la estructura de la base de datos y el gestor de la base de datos, para el caso del SIREC, el sistema tiene un gestor de base de datos SQL con un modelo de datos relacional, como es habitual, este diseño conlleva a un proceso de normalización, el cual consiste en la aplicación iterativa de un conjunto de reglas conocidas.

En esta etapa los objetivos que se persiguen son:

1. Evitar redundancias, duplicidades y valores nulos.
2. Facilitar cambios que en el futuro se tengan que realizar en la estructura de las tablas.
3. Reducir el impacto de los cambios de la base de datos en las aplicaciones que accedan a los datos.

El proceso de normalización mejora la calidad de la gestión de datos. Por ejemplo, la aplicación de la primera forma normal garantiza que los registros están unívocamente identificados, evitando tanto registros duplicados como carentes de identificador, y que no existen grupos de repetición (campos de ocurrencia múltiple). Mediante la segunda y tercera formas normales se elimina la información redundante que persiste tras la aplicación de la primera forma normal, asegurando que sea independiente entre sí. Finalmente, la cuarta y quinta formas normales garantizan la actualización de los datos al mantener la integridad referencial y la recuperación de la información.

A continuación en la figura 5 se presenta el modelo relacional de un agente con el módulo de reconocimiento de producto nacional, la llave primaria con que se relaciona la base de agentes con todo el sistema es a través del campo AGE_ID, que corresponde al número de Id (identificación del agente), cabe mencionar que en el sistema una persona o agente puede tener diferentes roles, por ejemplo un agente puede ser productor, director etc. por lo cual su identificación es la llave primaria con la cual es reconocido en el SIREC.

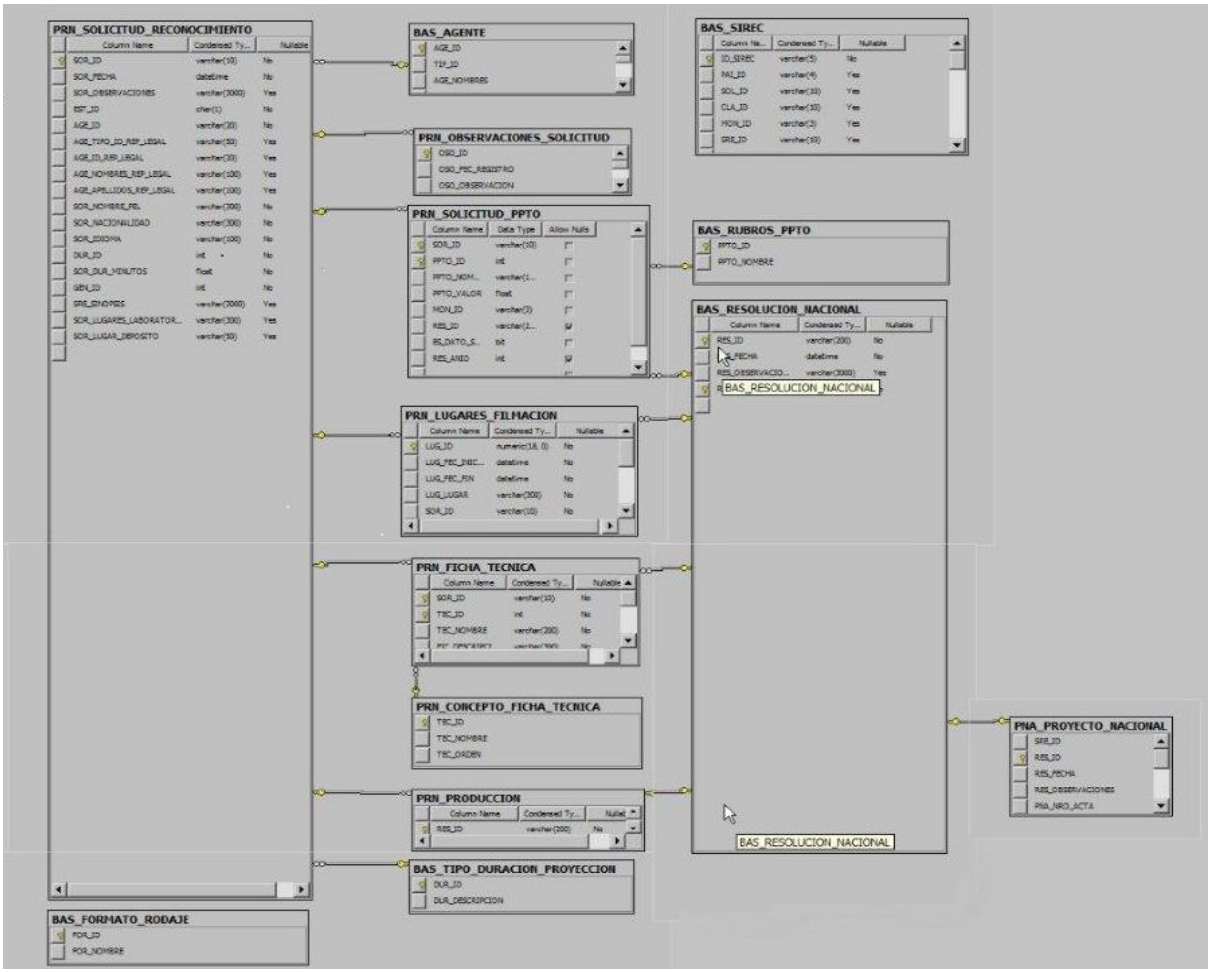


Figura 5. Modelo relacional Base Agente, llave primaria AGE_ ID. Fuente: SIREC.

Una vez definida la estructura de la base de datos se procede a la recogida y captura de los datos, la grabación puede realizarse de modo diferido (off line) o directo (on line).

En el modo (off line) los datos se encuentran contenidos en otros soportes, para el caso del SIREC, los registros administrativos y documentos se encuentran en papel y la tarea del operador del sistema o digitador consiste en migrarlos al sistema.

7.3.1. Captura (On line)

En la captura directa (on line) la grabación se realiza de forma simultánea al registro del dato. Las estrategias utilizadas para la mejorar la calidad de los datos durante su grabación deben garantizar que la información esté exenta de inconsistencias, que contengan el menor

número de valores desconocidos (missing) y que se ajuste a los momentos temporales de registro establecidos en el diseño.

La captura directa aventaja porque sólo es necesario introducir una vez la información, esta elimina el pre-registro de los datos y, por lo tanto, impide la falta de correspondencia entre el formato papel (tiende a ser eliminado) y el registro informático.

Sin embargo, y a pesar de sus ventajas, la captura directa plantea un problema, no asegura la consistencia de la información, y por lo tanto no garantiza por sí misma la calidad del registro. Para conseguir este objetivo se deben programar un conjunto de controles automáticos que incorporen todas las restricciones posibles y así asegurar los datos consistentes.

Es posible evaluar tres tipos de restricciones automáticas de datos (date, 1999): prevaricaciones, invalidaciones y post-validaciones. Estos controles garantizan que cada variable cumpla las condiciones de consistencia que afectan (Gassman, 1995) a la propia variable y otras contenidas en el mismo formulario.

7.3.2. Captura (Off Line)

En la captura diferida Off line es posible implementar controles de calidad no automáticos, entre los principales controles de calidad aplicables al Sistema de Información SIREC se encuentran: la comparación visual con verificación aleatoria y previsión de los valores desconocidos (Missing). El objetivo común de estos procedimientos es garantizar una correspondencia perfecta entra la información contenida en las fuentes originales y los datos grabados.

Es importante destacar que cuando se efectúa una captura diferida, el uso combinado de controles automáticos y no automáticos no equivale a la captura directa con protecciones.

Registrar la información de forma directa, programando los controles posibles, tiene ventaja pues existe la proximidad física al dato en el momento de la captura, permitiendo así que éste sea contrastado y si es preciso recuperado ante cualquier incidencia que pueda surgir.

7.3.3. Comparación Visual

La comparación visual puede realizarse a través de una comparación aleatoria que consiste en que una o más personas cotejen visualmente los datos grabados con los registrados en las fuentes originales. La mayor ventaja relacionada con esta técnica es que permite verificar la calidad de los datos fuentes: la que contiene los datos originales y la de los datos grabados.

7.3.4. Previsión de los Valores Desconocidos (Missing)

La existencia de los valores desconocidos es una constante en cualquier registro administrativo o documentación solicitada para un respectivo trámite que se requiera para el registro en el Sistema de Información – SIREC, ya sea por falta de cooperación de los solicitantes, por desconocimiento u olvido de los contenidos que se preguntan y por último falta de comprensión en los contenidos.

Se considera que la proporción de valores desconocidos constituye un indicador más de la calidad del proceso de recogida de los datos y de la información que se maneja, se analiza e interpreta.

Actualmente, se conceden gran importancia al uso de estrategias secuenciales que minimicen las consecuencias de la falta de respuesta (Taylor y Amir, 1994). Estas propuestas encuentran su aplicación desde la fase de recogida de datos. En caso de la captura diferida, el control relacionado con la falta de información comporta diseñar los formularios de captura de manera que incluyan la opción de valor desconocido para permitir posteriormente tratar esta “Falta de Información”. En el SIREC se tienen programada la opción sin asignar, no

aplica, en algunos campos cuando no se tiene la información de determinado registro.

7.3.5. Pre-validaciones, In- validaciones y Post-validaciones

Una pre-validación es una condición lógica asociada a un campo que se evalúa previamente a introducir ningún valor en él y que determina si se permite su edición o si se le asigna un valor automáticamente que puede ser “no aplicable”, desconocido o “deducible”.

Una in-validación es una protección que sólo permite grabar datos cuyo formato corresponda con el definido para el correspondiente campo. Está protección está implementada en todos los programas que contienen el concepto de “Formato de Campo”. Por Ejemplo, una in-validación no permite que se introduzca un valor cadena en un campo fecha, y tampoco un número con decimales si el formato del campo se ha definido para valores numéricos enteros.

Una post-validación es una protección que determina el rango de valores admisible en un campo y las condiciones que éstos deben cumplir respecto a los otros campos por ejemplo una post-validación no debe permitir grabar decimales para un campo que registra un número telefónico.

7.4.Errores Encontrados en los Formularios de Captura

Formulario datos de contacto de un agente.

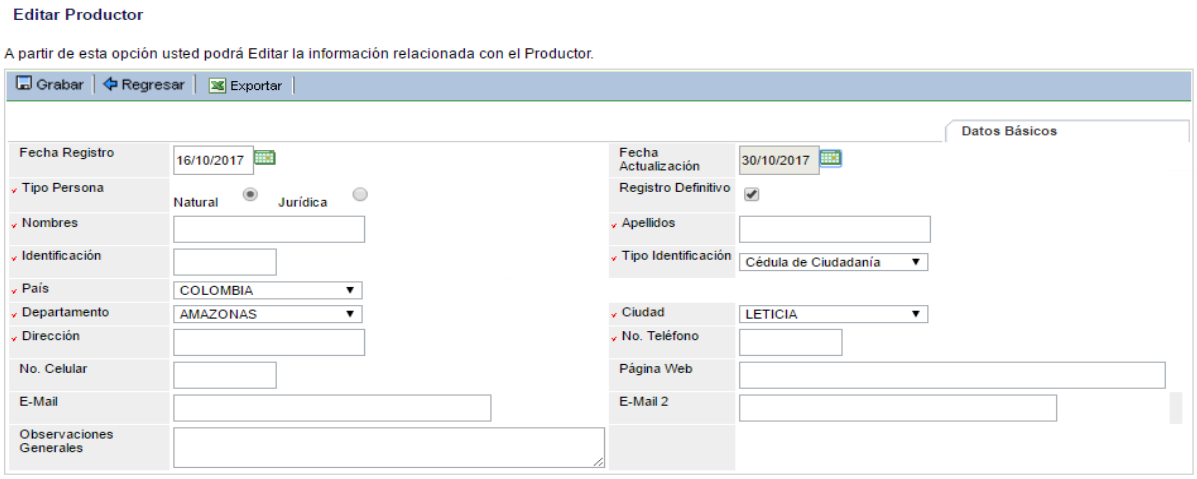


Figura 6. Formulario de Datos de Contacto de un Agente.. Fuente SIREC

a) Campo Dirección:

Las direcciones de los agentes que se encuentran registrados en el SIREC tienen diferentes nomenclaturas que inducen al error sintáctico como se muestra en la figura 7 y la tabla.3 .

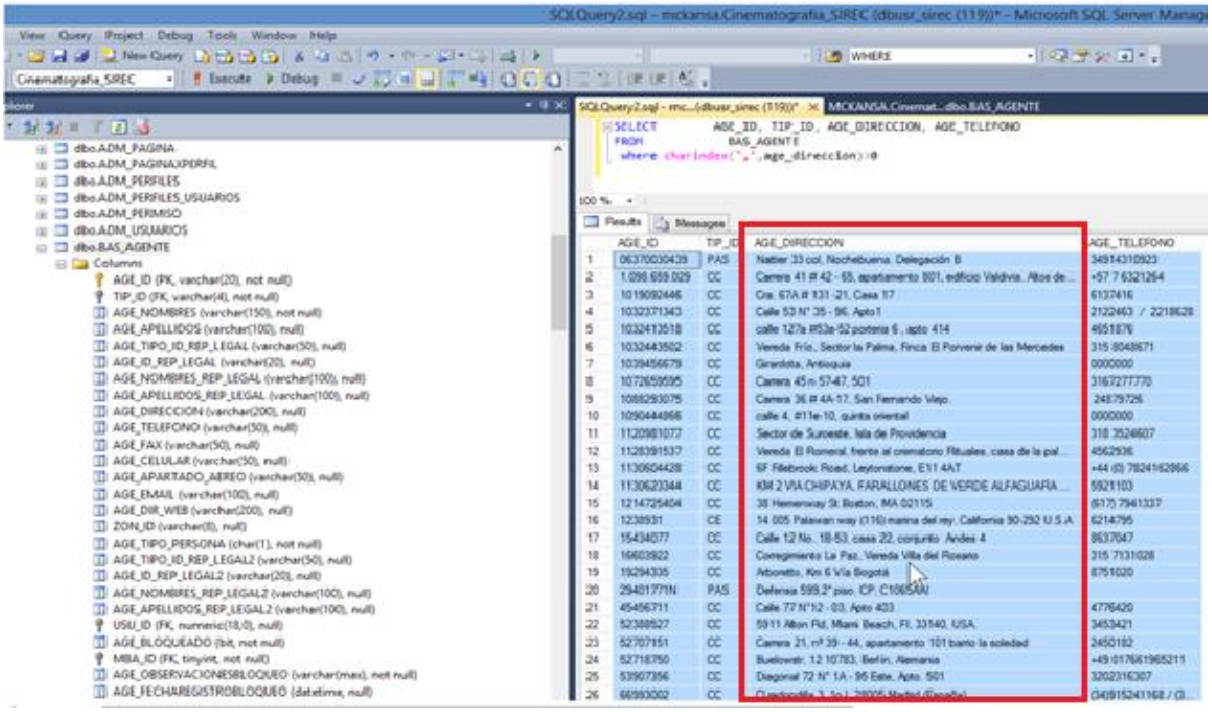


Figura 7. Base de datos Agentes, Análisis de direcciones. Fuente SIREC.

G12

	A	B	C	D
1	AGE_ID	TIP_ID	AGE_DIRECCION	AGE_TELEFONO
2	6370030439	PAS	Nattier 33 col, Nochebuena. Delegación B	34914310923
3	1.098.659.029	CC	Carrera 41 # 42 - 65, apartamento 801, edificio Valdivia. Altos de c	+57 7 6321264
4	1019092446	CC	Cra. 67A # 131 -21, Casa 17	6137416
5	1032371343	CC	Calle 53 N° 35 - 96, Apto1	2122463 / 2218628
6	1032413518	CC	calle 127a #53a-52 porteria 6 , apto 414	4651876
7	1032443502	CC	Vereda Frío, Sector la Palma, Finca El Porvenir de las Mercedes	315 8048671
8	1039456679	CC	Girardota, Antioquia	0
9	1072659595	CC	Carrera 45 n 57-47, 501	3167277770
10	1088293075	CC	Carrera 36 # 4A-17, San Fernando Viejo.	24879726
11	1090444866	CC	calle 4, #11e-10, quinta oriental	0
12	1120981077	CC	Sector de Suroeste, Isla de Providencia	318 3524607
13	1128391537	CC	Vereda El Romeral, frente al crematorio Rituales, casa de la palm	4562936
14	1130604428	CC	6F Fillebrook Road, Leytonstone, E11 4AT	+44 (0) 7824162866
15	1130623344	CC	KM 2 VIA CHIPAYA, FARALLONES DE VERDE ALFAGUARA CASA 13	5921103
16	1214725404	CC	38 Hemenway St Boston, MA 02115	(617) 7941337
17	1238931	CE	14 005 Palawan way (116) marina del rey, California 90-292 U.S.A	6214795
18	15434077	CC	Calle 12 No. 1B-53, casa 22, conjunto Andes 4	8637047
19	16603922	CC	Corregimiento La Paz, Vereda Villa del Rosario	315 7131028
20	19294335	CC	Arboretto, Km 6 Vía Bogotá	8751020
21	29401771N	PAS	Defensa 599,2° piso. CP: C1065AAI	
22	45456711	CC	Calle 77 N°12 - 03, Apto 403	4776420
23	52388527	CC	5911 Alton Rd, Miami Beach, FI, 33140, USA.	3453421
24	52707151	CC	Carrera 21, nº 39 - 44, apartamento 101 barrio la soledad	2450182
25	52718750	CC	Buelowstr. 12 10783, Berlín, Alemania	+49 017661965211
26	53907356	CC	Diagonal 72 N° 1A - 95 Este, Apto. 501	3202316307
27	66993002	CC	Cl redondilla 3, 1o I, 28005 Madrid (España)	(34)915241168 / (34)0787006806
28	71267956	CC	Km 5 Mz F, Lote 71	311 8437088

Tabla 3. Base de Datos – SIREC. Agentes, Análisis de Direcciones en Hoja de Cálculo.

Para reducir los errores es necesario que se creen pre-validaciones en el campo de captura Dirección, en el cual permita ingresar la información estandarizada por ejemplo la calle por Cl.

b) Campo Teléfono

Se encontraron en el campo teléfono errores de inconsistencia de la información, en el campo télefono se encuentran hasta dos números de télefono, hay números que corresponden a télefonos moviles, existen letras en el número teléfonico, hay campos vacios como se muestra en la Tabla 4.

TIP_ID	AGE_NOMBRES	AGE_APELLIDOS	AGE_DIRECCION	AGE_TELEFONO	AGE_CELULAR
CC	Agustin	Godoy Ronderos	Teodoro Garcia 2349	54155647 3766	54155647 3766
CC	Agustín	Hoyos Arango	Cl 45 A No 39 B - 02 (189)	3026461	3013358546
NIT	Ahumada Domínguez Inversiones S.A.S		Cl 95 No 13-55 (403)	5201433	310 7999182
CC	Aida	Morales	Tv 25 No 53 B - 60 (102)	300 5332635 / 313 4304691	300 5332635 / 313 4304691
CC	Aida Cristina	Bossa Bustamante	Cr 13 A No 78 - 31	2127534	3002099535
NIT	Aion Fundación para la Acción Creativa		Cl 184 No 20- 61 APTO 202	9210065	3112268291
NIT	AJNA SAS		Calle 134A N°53-82 T 9 Of. 301	5784652	3138725674
NIT	Akapana Producciones RL. Amparo Alfonso		CL 25C No 3 - 92 APTO 101	2436181	3167540320
CC	Akemi	Nakamura Gómez	Cr 45 No 9 Sur - 45	314 7910866	
NIT	Akira Cine Ltda		CR 19C No 86A - 43 APTO 203	4767967	3112361012
NIT	AkiraStudio S.A.S		Cr 96 B No. 17 B - 40	3102758950	3102758950
CC	Alba Lucía	Martínez Rueda	Cl 51 No 24 -06 (301)	301 2536668	301 2536668
CC	Alba Mery	Salazar Jimenez	Carrera 51 No. 73-32	2631444	3137480194
CC	ALBA SOFÍA RODRÍGUEZ ORTIZ				
CC	Albero Enrique	Serrano Vergel	Cl 102 No 16-26 (201)	6919546	3156020638
CC	Alberto	Arango Robledo	Cr 19 B No 85-72	310 5536969	310 5536969
CC	Alberto	Campuzano Sanchez	CL 2B No 66 - 56 APTO 401A	N/A	3155085410
CC	Alberto	Castellanos Cordoba	CL 12F No 2 - 91 APTO 14	3178757167	3178757167
CC	Alberto	Gómez Peña	Cl 116 No. 22 - 45 406	6376828	3002338224
CC	Alberto	Moncayo Cordoba	CL 35 No 21 - 24 Apto 101	8129226	3003245088
CC	Alberto	Rodríguez	Cr 28 No 86 - 32	311 5619409	311 5619409
CC	Alberto	Rodriguez Gómez	Cl 5 No. 80 - 109 (3 / 309)	310 3882237	310 3882237
CC	Alberto	Rodríguez Rodríguez	Cr 50 B No 64 A-18 (202)	311 2834566	311 2834566
CC	Alberto	Sanchéz Cristo	Cl 94A No 13-59 (303)	6237773	315 7970817
CC	Alberto Ignacio	Cortes Ruíz	Dg 53 No. 24 - 30	6254789	316 7178624 / 315 2051932 /3167178624
CC	Alberto Iván	Lemus Cuadro	Cr 52 No. 61 - 26	310 3876013	300 4026870
CC	Alberto Mario de Jesús	Villa Mendoza	Cl 166 No 9 - 70 (203)	320 8427259	320 8427259
CC	Alberto Rafael	Loaiza Sánchez	Cr 77 No 19 - 35 (1 / 1004)	7345260	
NIT	Alcaldia Municipal de la Mesa		Cl 8 Cr 21 esquina	8472221	
CC	Aldemar	Solano Peña	CL 4 No 16 - 47	3207376520	3207376520
NIT	ALEBRIJE ENTERTAINMENT LLC		No aplica	0	0
CC	Aleethia	Stendal	CR 2A No 66 - 52 APTO 122D	6094019	

Tabla 4. Agentes, errores en el campo teléfono. Fuente SIREC.

Para reducir los errores se debe crear pre-validaciones en el campo de captura teléfono, en el cual solo se permita ingresar un número telefónico de 7 dígitos, en la fase de hacer se presenta las soluciones de corrección de la información y las validaciones para el campo celular.

c) Campo Celular

Se encontraron en el campo celular errores de inconsistencia de la información, en este campo se encuentran hasta dos números moviles, hay números que corresponden a télefonos fijos, existen letras en el número teléfonico, hay campos vacios como se muestra en la siguiente tabla.

TIP_ID	AGE_NOMBRES	AGE_APELLIDOS	AGE_DIRECCION	AGE_TELEFONO	AGE_CELULAR	AGE_EMAIL
CC	Agustin	Godoy Ronderos	Teodoro Garcia 2349	54155647 3766	54155647 3766	agustinmgodoy@gmail.com / agustinjgody@hotmail.com
CC	Alejandro	Gómez				
CC	Alejandro	Matallana Beltrán	Cra 58 No 134-57 Etapa 4 Int. 2 Apto 501	2269867	3192782561	alematallana@gmail.com
CC	Alejandro Arturo	Buchelli Gomez	CL 12 No 29B - 78 APTO 501Z	4859927	3178708827	alejandro.bucheli@gmail.com
CC	Alejandro David	Landes Echavarria	301 OCEAN DR APTO 405 MIAMI BEACH FL 33139	N/A	N/A	alejandro.landes@gmail.com
CC	Alejandro Evaristo	Quintero Montero	Cl 73 No 11 - 52 (204)	6063744	315 3600725	alejandroquinterom@gmail.com
PAS	Alejandro Gabriel	Lázaro Alonso	Calle Nebulosas 2, portal C 1eroA. 28045	34 911437761		
NIT	Alejandro Simbaqueva Manrique		tranversal 15 b # 46 - 16 ofc 707 chapinero	7559496	3138642464	piragnastudios@gmail.com
CC	Alessandro	Angulo Brandestini	Cr 5 No 26 - 39	2866789	310 2124120	alessandro@laberinto.tv
CC	Alex Andrés	López Guevara	Cr 18 No 2 - 62	8211546	3105905315	kinephilos.alex@gmail.com
CC	Alex Díaz	Rubio Reyes	CR 2 No 43 - 23 PIEDRA PINADA BAJA	8 2 666350	301 201 84 06	alexrubio@creandes.org
CC	Alexander	González Tascón	Cr 26 No 15 - 56	3774534 / 3264282	317 4024297	alexandergonzalezt@gmail.com
CC	Alexander	Ramírez	CR 2A No 66 - 52 APTO 112B	3491842	311 2103684	jalexramirezo@gmail.com
CC	Alexander	Sterling Reyes	CR 83 No 43 - 70	3321428	3148973849	17centimetros@gmail.com
CC	Alexander	Valencia Martínez	Cr 5ª No 59ª-44	3488000	+ 57 321 3821578	
CC	Alexander	Vargas Cardona	Cl 35 No 21 - 43 (201)	300 7812324	300 7812324	thelittlehousofrecords@gmail.com
NIT	ALEXANDER KATZOWICK		No aplica	0	0	No aplica
CC	Alexandra	Cardona Restrepo	Av Cr 7 No 119 - 14 (519)	7532645-7591665	315 6013536	alexandracardonarestrepo@gmail.com
CC	Alexandra	Castañeda Giraldo	Calle 13 c 72 93	4847360	3108544682	sazacast@yahoo.com
CC	Alexandra	Garzón Garibello	CR 16 No 183 - 43 4 APTO 1101	7577311	312 5024189	alegaribello@gmail.com
CC	Alexandra	Molina	Cl 9 A No 1 E - 31	8210393		alemon2605@gmail.com
CC	Alexandra	Sarmiento	Cr 2 No 7 A - 15	6379502	6379502	totodan1@yahoo.es
CC	Alexis David	Durán Ventura	Km 11 vía la calera	3107728068	3107728068	alexis.duran@imaginariafilms.com

Tabla 5. Tabla de Agentes, errores en el campo celular. Fuente: SIREC.

Para reducir los errores se debe crear pre-validaciones en el campo de captura E-mail, en el cual solo se permita ingresar un solo E-mail, en la fase de hacer se presenta las soluciones de corrección de la información y las validaciones para el campo E-Mail.

d) Campo E-mail

Se encontraron en el campo email, errores de inconsistencia de la información, en este campo se encuentran hasta dos números E-mail, hay números que corresponden a télefonos fijos, existen letras en el número teléfonico, hay campos vacios como se muestra en la siguiente Tabla.

A	B	C	H	I	J	K
TIP_ID	AGE_NOMBRES	AGE_APELLIDOS	AGE_DIRECCION	AGE_TELEFONO	AGE_CELULAR	AGE_EMAIL
CC	Agustin	Godoy Ronderos	Teodoro Garcia 2349	54155647 3766	54155647 3766	agustinmgodoy@gmail.com / agustinjgody@hotmail.com
CC	Alejandro	Gómez				
CC	Alejandro	Matallana Beltrán	Cra 58 No 134-57 Etapa 4 Int. 2 Apto 501	2269867	3192782561	alematallana.com
CC	Alejandro Arturo	Buchelli Gomez	CL 12 No 29B - 78 APTO 501Z	4859927	3178708827	alejandro.bucheli@gmail.com
CC	Alejandro David	Landes Echavarria	301 OCEAN DR APTO 405 MIAMI BEACH FL 33139	N/A	N/A	alejandro.landes@gmail.com
CC	Alejandro Evaristo	Quintero Montero	Cl 73 No 11 - 52 (204)	6063744	315 3600725	alejandroquinterom@gmail.com
PAS	Alejandro Gabriel	Lázaro Alonso	Calle Nebulosas 2, portal C 1eroA. 28045	34 911437761		
NIT	Alejandro Simbaqueva Manrique		tranversal 15 b # 46 - 16 ofc 707 chapinero	7559496	3138642464	piragnastudios@gmail.com
CC	Alessandro	Angulo Brandestini	Cr 5 No 26 - 39	2866789	310 2124120	alessandro@laberinto.tv
CC	Alex Andrés	López Guevara	Cr 18 No 2 - 62	8211546	3105905315	kinephilos.alex@gmail.com
CC	Alex Díaz	Rubio Reyes	CR 2 No 43 - 23 PIEDRA PINADA BAJA	8 2 666350	301 201 84 06	alexrubio@creandes.org
CC	Alexander	González Tascón	Cr 26 No 15 - 56	3774534 / 3264282	317 4024297	alexandergonzalezt@gmail.com
CC	Alexander	Ramírez	CR 2A No 66 - 52 APTO 112B	3491842	311 2103684	jalexramirezo@gmail.com
CC	Alexander	Sterling Reyes	CR 83 No 43 - 70	3321428	3148973849	17centimetros@gmail.com
CC	Alexander	Valencia Martínez	Cr 5ª No 59ª-44	3488000	+ 57 321 3821578	
CC	Alexander	Vargas Cardona	Cl 35 No 21 - 43 (201)	300 7812324	300 7812324	thelittlehousofrecords@gmail.com
NIT	ALEXANDER KATZOWICK		No aplica	0	0	No aplica
CC	Alexandra	Cardona Restrepo	Av Cr 7 No 119 - 14 (519)	7532645-7591665	315 6013536	alexandracardonarestrepo@gmail.com

Tabla 6. Tabla de Agentes, errores en el campo E-mail. Fuente: SIREC.

8. Errores en el Sistema por Dimensión.

Los conceptos dimensionales como exactitud, completitud y actualidad son características de los datos, se denominan dimensiones de la calidad de los datos. Cada dimensión refleja un aspecto distinto de la calidad de los datos. Las mismas pueden estar referidas a la extensión de los datos (su valor), o a la intensión (su esquema).

De esta manera podemos distinguir entre calidad en los datos y calidad en los esquemas. El objetivo del presente proyecto es en la calidad inherente a los datos. Se define factor de calidad como un aspecto particular de una dimensión. En este sentido, una dimensión puede ser vista como un agrupamiento de factores de calidad que tienen el mismo propósito. Es claro que la mala calidad en los datos puede provocar varios problemas, así como también la mala calidad de un esquema (por ejemplo, un esquema de una base de datos relacional sin normalizar) podría provocar problemas mayores, tales como redundancias. Ambos tipos de dimensiones, tanto las referidas a los datos como a los esquemas, proveen una visión

cualitativa de la calidad, mientras que las medidas cuantitativas se representan mediante las métricas. Una métrica es un instrumento que define la forma de medir un factor de calidad. Un mismo factor de calidad puede medirse con diferentes métricas. Por otro lado, definimos método de medición como un proceso que implementa una métrica. A su vez, una misma métrica puede ser medida por diferentes métodos. Existen varias dimensiones que reflejan distintos aspectos de los datos, al considerar que los datos pretenden representar todo tipo de características de la realidad, desde espaciales y temporales, hasta sociales. A continuación, se describen algunas dimensiones de la calidad de datos y sus errores asociados. Dependiendo del contexto en el cual nos situemos elegiremos mejorar aquellas dimensiones que consideramos de mayor valor para la calidad de nuestros datos, y no tomar en cuenta las que no afectan de manera significativa.

A continuación, se presenta la siguiente tabla 7 con los tipos de errores que se presentan en el SIREC por dimensión.

Dimensión	Factor	Tipo de error
Exactitud	Exactitud sintáctica	Valor fuera de rango
		Estandarización
	Exactitud semántica	Registro inexistente
		Defecto mal registrado
		Valor fuera de referencial
Completitud	Densidad	Valor nulo
		Clasificación de defecto
Consistencia	Integridad intra-relación	Reglas de integridad intra-relación
		Valor único
	Integridad referencial	Referencia inválida
Unicidad	Duplicación	Registro duplicado
	Contradicción	Registro contradictorio

Tabla 7. Errores en el SIREC por Dimensión.

8.1.Dimensiones Exactitud y Unicidad

La exactitud se puede definir como la cercanía que existe entre un valor del mundo real, y su representación. De acuerdo con el enfoque teórico, la exactitud se define como una correcta y precisa asociación entre los estados del sistema de información y los objetos del

mundo real. Existen tres factores de exactitud: exactitud semántica, exactitud sintáctica y precisión. La exactitud sintáctica se refiere a la cercanía entre un valor v y los elementos de un dominio A. Esto es, si v corresponde a algún valor válido de A (sin importar si ese valor corresponde a uno del mundo real). Para poder medir la exactitud sintáctica se puede utilizar la comparación de funciones, métrica que mide la distancia entre un valor P y los valores en el dominio. Otras alternativas posibles son la utilización de diccionarios que representen fielmente el dominio, o el chequeo de los datos contras reglas sintácticas. La exactitud semántica se refiere a la cercanía que existe entre un valor P y un valor real P‟. Esta dimensión se mide fundamentalmente con valores booleanos (indicando si es un valor correcto o no), para lo cual es necesario conocer cuáles son los valores reales por considerar. En este caso, interesa medir que tan bien se encuentran representados los estados del mundo real. Una de las métricas utilizadas es la comparación de los datos con referenciales considerados válidos. La precisión, por otra parte, se refiere al nivel de detalle de los datos. El enfoque hasta ahora ha sido en la exactitud a nivel de valores, o sea, del valor de una celda (o campo) de una tupla.

Para aclarar los conceptos se plantea un ejemplo sencillo. En la base de datos agentes se almacena el nombre y el número telefónico de determinadas personas. Para el dato "Teléfono" se especifica que su valor estará en el rango exacto de 7 dígitos, que se usa para los números de teléfonos fijos en Colombia. Además, se sabe que existe una persona llamada Diana López, con número de teléfono 7179351. Se consideran entonces los siguientes casos: Si existe un registro para una persona donde el campo teléfono tiene el valor 717625, entonces se trata de un error sintáctico (valor fuera del rango de 7 dígitos). Si existe un registro para Diana donde el campo teléfono tiene el valor 7175193, entonces se trata de un error semántico, ya que es sabido que Diana no tiene el número telefónico 7175193, sino que tiene el número 7179351 (en este caso no hay error sintáctico, pues el número corresponde a

7 dígitos, es un valor válido para el campo teléfono). Cuando ocurren errores de digitación ambos tipos de exactitud coinciden. Al modificar su valor, se logrará exactitud sintáctica, ya que el valor escrito correctamente se corresponderá con alguno del dominio, y semántica, ya que existirá un valor real asociado al valor escrito correctamente. Una forma de chequear la exactitud semántica es comparar diferentes fuentes de datos, y encontrar a partir de estas el valor correcto deseado. Esto también requiere de la resolución del problema de identificación de objetos, el cual consiste en identificar si dos tuplas representan el mismo objeto en el mundo real. En el caso en que la exactitud sea considerada en un conjunto de valores, es necesario considerar también la duplicación. Dicha problemática ocurre cuando un objeto del mundo real se encuentra presente más de una vez (más de una tupla representa exactamente el mismo objeto).

Sin embargo, podrían existir también tuplas que representan el mismo objeto del mundo real, pero con diferentes claves. Este aspecto es considerado por la dimensión de Unicidad. Es importante destacar aquí que existen diferentes situaciones que pueden llevar a la duplicación de datos:

a. Cuando la misma entidad se identifica de diferentes formas
b. Cuando ocurren errores en la clave primaria de una entidad
c. Cuando la misma entidad se repite con diferentes claves

Distinguimos dos factores de la dimensión Unicidad:

Duplicación: la misma entidad aparece repetida de manera exacta.

Contradicción: la misma entidad aparece repetida con contradicciones**.**

Para la dimensión Exactitud, se identifican los tipos de errores que se mencionan a continuación:

a. Dentro del factor exactitud sintáctica:

- Valor fuera de rango.
- Estandarización.

b. Dentro del factor exactitud semántica:

- Registro inexistente.
- Defecto mal registrado.
- Valores fuera de referencial.

A continuación, se describe cada uno de estos tipos de errores de manera detallada.

a. Factor Exactitud Sintáctica

- Valor Fuera de rango

Resulta de interés que los valores de los tiempos y líneas de código se encuentren dentro de un rango determinado, el cual debe ser previamente definido. Un valor fuera de rango en las líneas de código no permitiría identificar, por ejemplo, cuáles son los registros de defectos que corresponden al mismo defecto de la realidad. Es importante considerar, sin embargo, que si un valor cae fuera del rango determinado no significa necesariamente que dicho valor sea erróneo. El hecho de que existan valores "anómalos" pero correctos es difícil poder distinguir de los errores reales.

Error con valor fuera de rango que se presenta en el sistema. Campo **AGE_CELULAR**
Este campo corresponde al número de celular de un agente productor, en este campo se presentó inconsistencias debido a que el campo aparecían números con espacio por ejemplo 321 356987 y hasta dos celulares en el mismo campo.

- Estandarización

Para el procesamiento y entendimiento de la información del sistema se hace necesario crear una estandarización para el ingreso de los datos para el campo dirección debido a que se presenta en este campo errores de inconsistencias semánticas a causa del mal ingreso de la información en la fuente primaria y al momento de digitarla en el sistema.

Para visualizar este error consiste en verificar mediante la ejecución de consultas SQL si

los valores corresponden al estándar definido y si se encuentran en el rango establecido.

En la figura 8 y tabla 8 se puede observar que no existe una estandarizan para las direcciones.

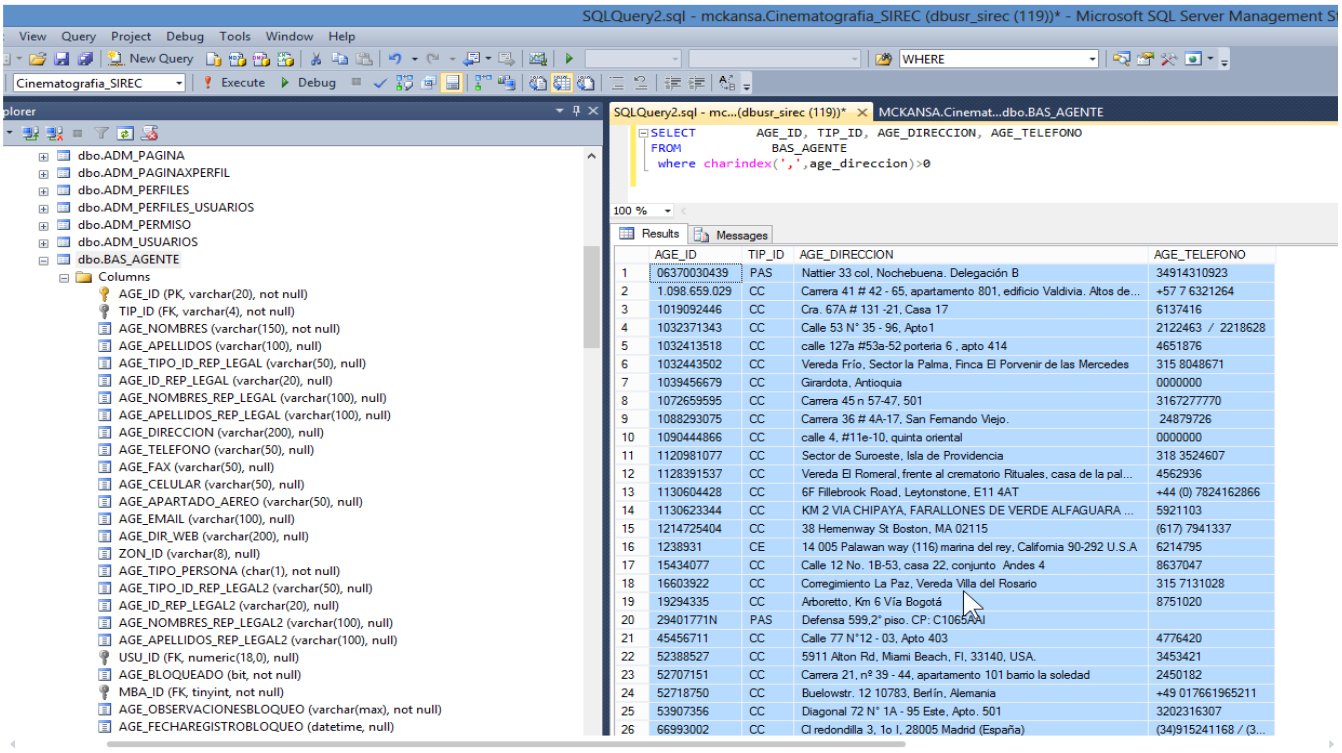


Figura 8. Consulta SQL para Identificar los Números Telefónicos. Fuente: SIREC.

G12

	A	B	C	D
1	AGE_ID	TIP_ID	AGE_DIRECCION	AGE_TELEFONO
2	6370030439	PAS	Nattier 33 col, Nochebuena. Delegación B	34914310923
3	1.098.659.029	CC	Carrera 41 # 42 - 65, apartamento 801, edificio Valdivia. Altos de c	+57 7 6321264
4	1019092446	CC	Cra. 67A # 131 -21, Casa 17	6137416
5	1032371343	CC	Calle 53 N° 35 - 96, Apto1	2122463 / 2218628
6	1032413518	CC	calle 127a #53a-52 porteria 6 , apto 414	4651876
7	1032443502	CC	Vereda Frío, Sector la Palma, Finca El Porvenir de las Mercedes	315 8048671
8	1039456679	CC	Girardota, Antioquia	0
9	1072659595	CC	Carrera 45 n 57-47, 501	3167277770
10	1088293075	CC	Carrera 36 # 4A-17, San Fernando Viejo.	24879726
11	1090444866	CC	calle 4, #11e-10, quinta oriental	0
12	1120981077	CC	Sector de Suroeste, Isla de Providencia	318 3524607
13	1128391537	CC	Vereda El Romeral, frente al crematorio Rituales, casa de la palm	4562936
14	1130604428	CC	6F Fillebrook Road, Leytonstone, E11 4AT	+44 (0) 7824162866
15	1130623344	CC	KM 2 VIA CHIPAYA, FARALLONES DE VERDE ALFAGUARA CASA 13	5921103
16	1214725404	CC	38 Hemenway St Boston, MA 02115	(617) 7941337
17	1238931	CE	14 005 Palawan way (116) marina del rey, California 90-292 U.S.A	6214795
18	15434077	CC	Calle 12 No. 1B-53, casa 22, conjunto Andes 4	8637047
19	16603922	CC	Corregimiento La Paz, Vereda Villa del Rosario	315 7131028
20	19294335	CC	Arboretto, Km 6 Vía Bogotá	8751020
21	29401771N	PAS	Defensa 599,2° piso. CP: C1065AAI	
22	45456711	CC	Calle 77 N°12 - 03, Apto 403	4776420
23	52388527	CC	5911 Alton Rd, Miami Beach, FI, 33140, USA.	3453421
24	52707151	CC	Carrera 21, nº 39 - 44, apartamento 101 barrio la soledad	2450182
25	52718750	CC	Buelowstr. 12 10783, Berlín, Alemania	+49 017661965211
26	53907356	CC	Diagonal 72 N° 1A - 95 Este, Apto. 501	3202316307
27	66993002	CC	Cl redondilla 3, 1o I, 28005 Madrid (España)	(34)915241168 / (34)0787006806
28	71267956	CC	Km 5 Mz F, Lote 71	311 8437088

Tabla 8. Consultas Agentes con Números Telefónicos en Hoja de Cálculo. Fuente: SIREC.

8.2.Dimensiones Relacionadas con el Tiempo

Los cambios y actualizaciones de los datos son conceptos muy importantes para tener en cuenta. Es posible afirmar que en determinados contextos un dato no actualizado es de mala calidad y puede llegar a ocasionar problemas de exactitud.

Como ejemplo, un registro de un productor de cine, fue registrado en el sistema, cuando nacionalizó su obra nacional, sin embargo los datos de registro como los son: el correo electrónico, el número de celular, teléfono, dirección; son datos que pueden estar actualizándose, cada vez que llega una nueva solicitud de reconocimiento como obra nacional (nacionalización), si en el caso, no se registra la fecha de actualización de los datos ingresados en el sistema, se creería que la información no está actualizada, y por ende podría llegar a confusiones de datos desactualizados.

Se describen las siguientes dimensiones relacionadas con el tiempo:

- Actualidad: trata sobre la actualización de los datos y su vigencia. Esta dimensión puede ser medida de acuerdo con la información de “última actualización".
- Volatilidad: se refiere a la frecuencia con que los datos cambian en el tiempo. Una medida para esta dimensión es la cantidad de tiempo que los datos permanecen siendo válidos.

- Consistencia

Esta dimensión hace referencia al cumplimiento de las reglas semánticas que son definidas sobre los datos.

De acuerdo con el enfoque teórico, la inconsistencia de los datos se hace presente cuando existe más de un estado del sistema de información asociado al mismo objeto de la realidad. Una situación que ocasionó la generación de inconsistencias en los datos incorporados al SIREC fue ocasionada por una migración externa realizada en el año 2005, en la cual se subió al sistema diferentes bases de datos de agentes que existía en el motor de bases de datos Access que se manejaba en la época.

Un ejemplo es el siguiente caso del señor Andrés López Forero, que se encuentra registrado dos veces en el módulo de personal técnico y artístico, pero con diferente cédula en una aparece el número 795544999 y en el otro 79554999, en este caso se está presentando una contradicción de la información en el número de cédula como se muestra en la siguiente figura 9.

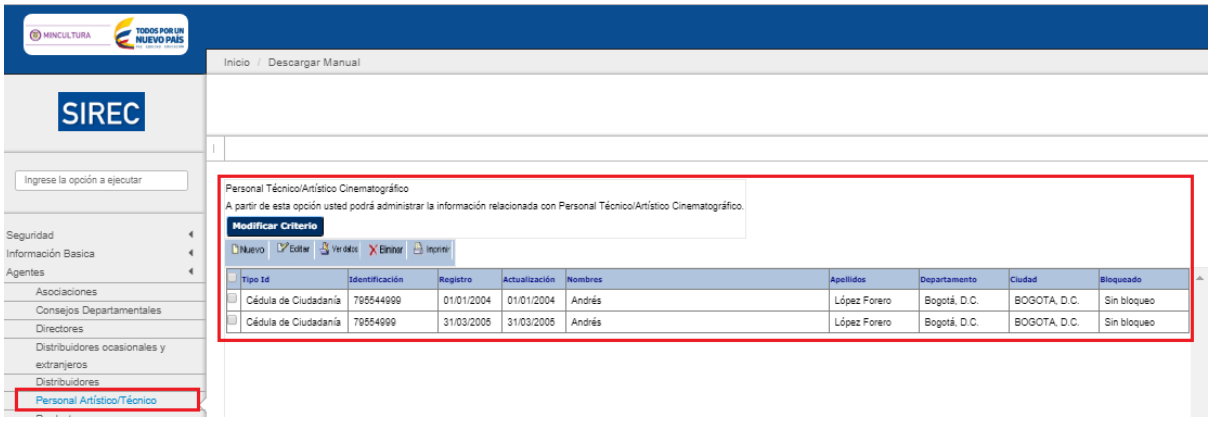


Figura 9. Registro Duplicado donde las Cédulas son Diferentes. Fuente: SIREC.

b. Factor Exactitud Semántica

- Registro Inexistente

En este caso se identifican aquellos registros (tuplas) que no corresponden a ningún objeto de la realidad. Esto es, registros que se encuentran almacenados en la base de datos, pero que se asocian a un objeto que en la realidad no existe. Los registros inexistentes no deberían formar parte de la base en cuestión ya que no reflejan la realidad, además de que su consideración a la hora de analizar los datos afectaría el resultado obtenido. Es por este motivo que interesa identificarlos.

Se analizan dos casos:

- Registro de Defectos

La causa conocida de este error corresponde a las migraciones realizadas en el año 2005, y también por el ingreso de información de agentes participantes, a través del módulo convocatorias, periodo 2004 -2008.

Se subcontrató personal para el ingreso de dicha información, no se tuvo supervisión, ni se validó la información registrada.

A continuación, se pude observar uno de los errores presentados, como lo son: campo celular, (ningún celular inicia con un valor (697674753), error en campo teléfono (se

encuentra vacío), se validó campos de ciudad y departamento, se contrasta con la información física, se observa que existe asociada el registro a una ciudad diferente a la relacionada con el registro. Ciudad Cali, Dpto. Valle del Cauca (según expediente físico). Ver Figura 10.

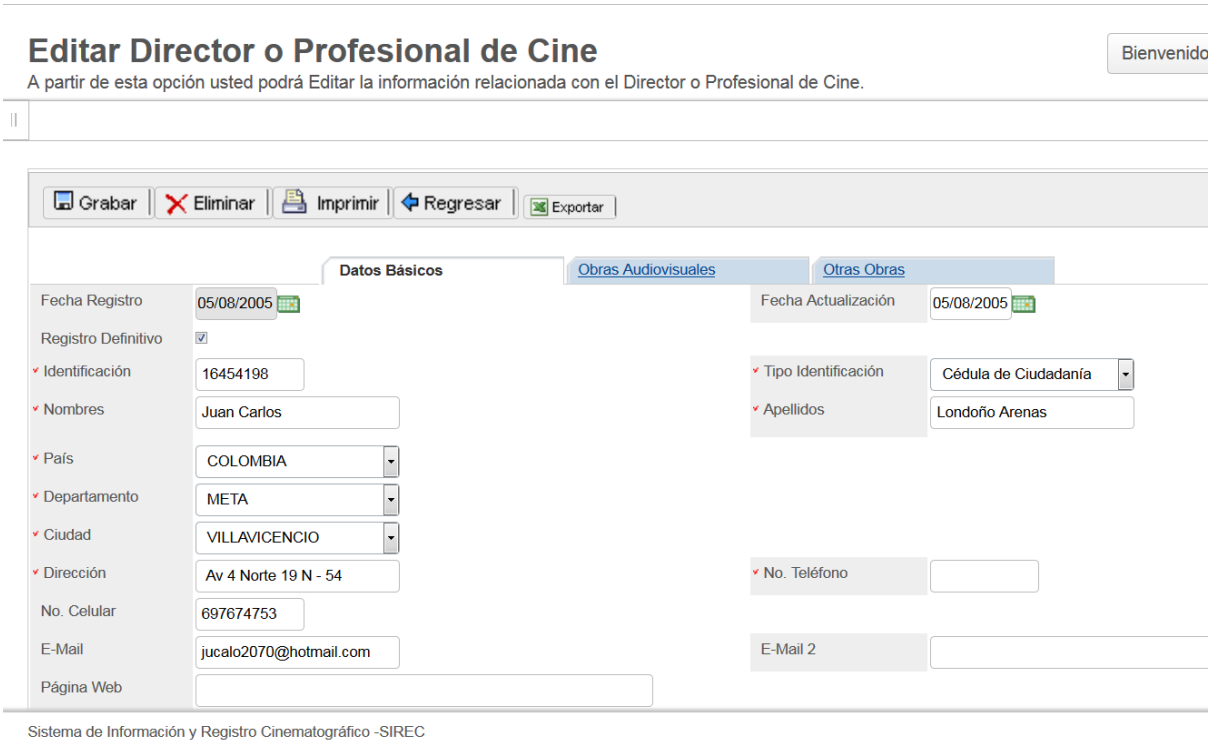


Figura 10. Formulario Director. Fuente: SIREC.

Medición

La métrica utilizada consiste entonces en la revisión manual de las tuplas involucradas en el presente tipo de error, con el fin de identificar si las mismas corresponden o no a objetos de la realidad.

Defecto Mal Registrado

Este tipo de error corresponde al caso de defectos que sí existen en la realidad, pero se "asocian" de manera incorrecta a un registro de defecto. Esto significa que los datos que se registran de un defecto no son válidos para ese defecto dado (a pesar de que sí son valores

sintácticamente correctos).

Un defecto mal registrado no permite, en general, lograr identificarlo dentro del código a partir de sus datos, ya que los mismos son incorrectos.

Cualquiera de los campos que se ingresan al registrar un defecto podría contener errores ocasionados por una equivocación por parte del verificador, ya sea por falta de comprensión, un error de concepto o distracción.

Con respecto a la clasificación del defecto, resulta sumamente complejo verificar si la misma se realizó de manera correcta, motivo por el cual no se analizará este error en el marco del presente proyecto.

Medición

Se realizará a nivel de celda, ya que involucra campos del registro de defecto.

No se identifica una forma automática de medir este tipo mediante una sentencia SQL.

La métrica utilizada consiste entonces en la revisión manual de las celdas involucradas en el presente tipo de error, con el fin de identificar si las mismas corresponden o no a datos válidos.

8.2.1. Dimensión Completitud

La completitud se puede definir como la medida en que los datos son de suficiente alcance y profundidad, esta dimensión se define como la capacidad del sistema de información de representar todos los estados significativos de una realidad dada. Existen dos factores de la completitud: cobertura y densidad. La cobertura se refiere a la porción de datos de la realidad que se encuentran contenidos en el sistema de información. La densidad se refiere a la cantidad de información contenida, y la faltante acerca de las entidades del sistema de información. Completitud de Datos Relacionales. La completitud en un modelo relacional puede caracterizarse por los siguientes aspectos: Valores nulos: el significado de los valores

nulos puede ser variado. Un valor nulo puede indicar que dicho valor no existe en el mundo real, que el valor existe en el mundo real pero no se conoce, o que no se sabe si el valor existe o no en el mundo real.

Para la dimensión Completitud, se identifican los tipos de errores que se mencionan a continuación.

Dentro del factor densidad:

- Valor nulo.
- Clasificación de defectos.

Dentro del factor cobertura, se identificó que la base de datos no se encuentra acorde a los códigos y nombres según la DIVIPOLA (Codificación de la División Político Administrativa de Colombia). Para solucionar esta inconsistencia, se debe ajustar internamente la base de datos, a los códigos y nombres de los municipios y departamentos que determina DIVIPOLA. Para descargar los códigos actuales, se debe hacer a través de la página del DANE. En el siguiente link: https://www.dane.gov.co/index.php/sistema-estadistico-nacional-sen/normas-y-estandares/nomenclaturas-y-clasificaciones/nomenclaturas/codificacion-de-la-division-politica-administrativa-de-colombia-divipola

Figura 11. Información de Divipola. Fuente: Página del Dane

c. Factor Densidad

- Valor Nulo

Interesa conocer qué información fue registrada y cuál fue omitida. Para aquella información que fue omitida, interesa conocer la causa de la omisión, y en caso de que sea posible, determinar el valor que debería tomar en lugar de nulo.

En el SIREC, muchos agentes tienen como dato nulo, el número de contacto fijo, debido que en la actualidad las personas referencian su teléfono móvil para ser contactados, se observa que existen muchos agentes entre el periodo 2004 - 2005 que no tienen relacionado

el campo del teléfono celular, esta información no fue suministrada, a través del formulario de recolección de datos.

Resulta necesario identificar en primera instancia cuáles son los campos que admiten nulos y cuáles no, según el esquema actual de la base de datos, se debe crear una validación que no permita grabar la información, si los campos teléfono y celular están “vacíos”.
Si no se cuenta con dicha información, se podría ingresar como dato siete o diez ceros en los campos mencionados.

Luego, se identifican aquellos campos que admiten nulos, pero deberían en la realidad contener algún valor distinto de vacío (el hecho de que admitan nulos es un error en el diseño de la base de datos). Este último caso es el que interesa medir. Se asume que el control sobre los campos declarados como no nulos se realiza correctamente por el SGBD. El motivo de omisión de los campos podría ser cualquiera de los siguientes:

1. El digitador no ingresa el valor (puede ser por omisión accidental o por no saber determinarlo).
2. Un error en el manejo de los datos (ya sea de la aplicación web, o de la base) que ocasiona que el valor ingresado por el digitador no se almacene correctamente.
3. Falta de completitud de la información en la fuente primaria debido a que en los formularios de registros de agentes no se diligenciaron completamente los datos, por lo tanto, no aparece información de algunos datos como teléfono, email y celular, una de las razones asociadas es que en el formulario de recogida de la información a estos campos no se les especificó como campos obligatorios.
4. El agente no tenga un número telefónico fijo, un número celular o un correo de contacto.
5. Porque los datos fueron cargados en un tiempo que no era tan usual utilizar correo electrónico o teléfono móvil.

Medición

La medición es a nivel de celda, ya que intervienen aquellos campos que deberían tener un valor distinto de nulo. Luego de identificar cuáles son los campos involucrados en este tipo de error, su forma de medición consiste en verificar mediante la ejecución de consultas SQL si los mismos contienen valores nulos.

A continuación, se presenta en las siguientes figuras 12 y 13, las cuales muestran una combinación de errores que se presentan de valores nulos, por falta de completitud de la información.

A través de la siguiente consulta de SQL se evidencias los agentes que contienen errores faltantes:

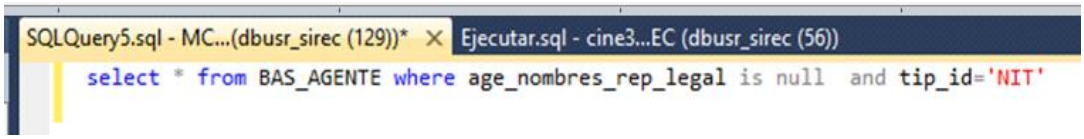
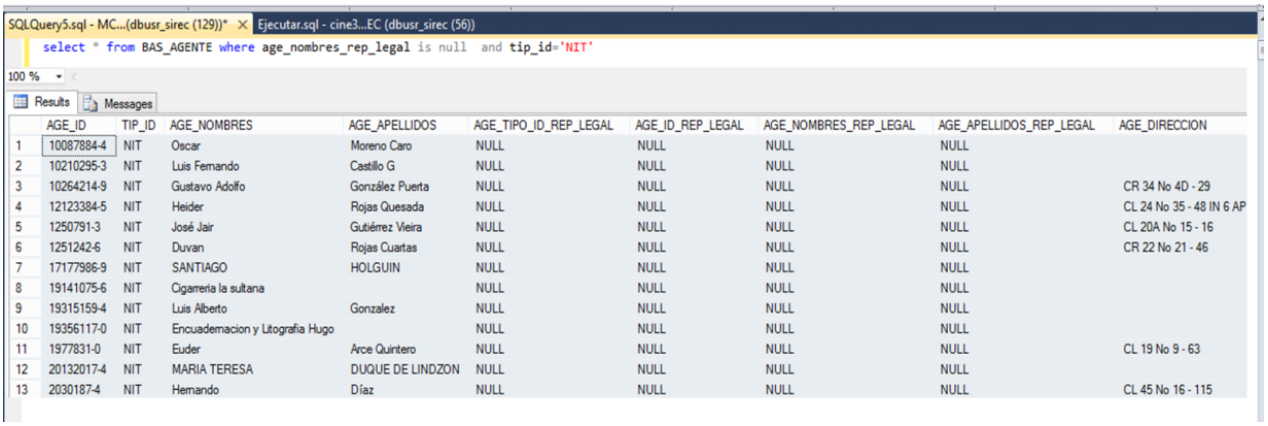


Figura 12. Consulta SQL. Fuente: SIREC.



	AGE_ID	TIP_ID	AGE_NOMBRES	AGE_APELLIDOS	AGE_TIPO_ID_REP_LEGAL	AGE_ID_REP_LEGAL	AGE_NOMBRES_REP_LEGAL	AGE_APELLIDOS_REP_LEGAL	AGE_DIRECCION
1	10087884-4	NIT	Oscar	Moreno Caro	NULL	NULL	NULL	NULL	
2	10210295-3	NIT	Luis Fernando	Castillo G	NULL	NULL	NULL	NULL	
3	10264214-9	NIT	Gustavo Adolfo	González Puerta	NULL	NULL	NULL	NULL	CR 34 No 4D - 29
4	12123384-5	NIT	Heider	Rojas Quesada	NULL	NULL	NULL	NULL	CL 24 No 35 - 48 IN 6 AP
5	1250791-3	NIT	José Jair	Gutiérrez Vieira	NULL	NULL	NULL	NULL	CL 20A No 15 - 16
6	1251242-6	NIT	Duvan	Rojas Cuartas	NULL	NULL	NULL	NULL	CR 22 No 21 - 46
7	17177986-9	NIT	SANTIAGO	HOLGUIN	NULL	NULL	NULL	NULL	
8	19141075-6	NIT	Cigarreria la sultana		NULL	NULL	NULL	NULL	
9	19315159-4	NIT	Luis Alberto	Gonzalez	NULL	NULL	NULL	NULL	
10	19356117-0	NIT	Encuadernacion y Litografia Hugo		NULL	NULL	NULL	NULL	
11	1977831-0	NIT	Euder	Arce Quintero	NULL	NULL	NULL	NULL	CL 19 No 9 - 63
12	20132017-4	NIT	MARIA TERESA	DUQUE DE LINDZON	NULL	NULL	NULL	NULL	
13	2030187-4	NIT	Hernando	Díaz	NULL	NULL	NULL	NULL	CL 45 No 16 - 115

Figura 13.Errores de ValoresNulos. Fuente SIREC.

9. Fase Hacer

En esta fase de Hacer, se debe implementar todas las posibles acciones que ayuden a corregir los datos inconsistentes y también a prevenir futuros errores en la captura.

En esta etapa se deben realizar las acciones necesarias para limpiar la base de datos e implementar controles en el programa automáticos en la estructura de la base de datos.

9.1.Proceso de Prevención de Errores

Consiste en evitar que ocurran inconsistencias en los datos a futuro. Para ello es necesario identificar primero cuáles son las causas de los errores y cómo lograr eliminarlas de manera permanente.

La localización y corrección de errores se lleva a cabo para datos cuya creación y actualización es poco frecuente. Sin embargo, la prevención de errores a través del manejo de procesos es utilizada en mayor medida cuando los datos son actualizados y creados de manera frecuente. Se incluyen controles a los procesos en los cuales los datos son creados y/o actualizados para evitar que sucedan inconsistencias.

Los *edits* también pueden ser utilizados para la prevención de errores y la mejora de procesos, evitando la ocurrencia de ciertas inconsistencias en la base.

Otra forma de prevención de errores consiste en identificar cuáles con las actividades manuales en las cuales suelen ocurrir la mayor cantidad de errores, y buscar su automatización.

Teniendo en cuenta que los errores detectados se presentan al momento de la captura de la información es necesario programar en el formulario de recogida validaciones que permitan minimizar los errores de dominio, sintácticos de la información de los agentes.

A continuación, se presenta en la siguiente Tabla 10. Las reglas de validación para los campos del formulario.

9.1.2. Reglas de Validación en el Formulario Captura de la Información de los Agentes.

Nombre del Campo	Descripción del Campo	Dominio o lista de valores	Tipo de datos	Longitud del campo	Otras reglas de validación
		[Liste los valores válidos, patrón o rango para el campo]	[Numérico, Cadena, Carácter, Entero, otro]	[Según el tipo de dato indicar la longitud del campo]	[Enumere claramente cada regla que se debe aplicar al campo o archivo]
AGE_TIPO_PERSONA	Tipo de persona del agente (Productor) (Natural o Jurídico)	CC, NIT	Carácter	1	169-RV400 REGLAS DE DOMINIO 1) el valor del campo debe tener solo los valores (CC, NIT) 156-RV320 MUTUAMENTE DEPENDIENTES 2) Si el valor del campo TIP_ID ='NIT' entonces la variable AGE_ID debe tener como mínimo 11 (once) caracteres
AGE_ID	Identificación del productor	>0	carácter	20	169-RV400 REGLAS DE DOMINIO 1) el valor del campo debe ser mayor a (0) cero.
AGE_NOMBRES	Nombres del Productor (Persona Natural o Jurídica)		carácter	150	169-RV400 REGLAS DE DOMINIO 1) el valor del campo no debe ser nulo
AGE_APELLIDOS	Apellidos del Productor (Persona Natural)		carácter	100	156-RV320 MUTUAMENTE DEPENDIENTES 1) si TIP_ID = 'CC' entonces el campo AGE_APELLIDOS no puede ser nulo
AGE_TIPO_ID_REP_LEGAL	Tipo de Identificación del representante legal (CC, o, CE)		carácter	50	156-RV320 MUTUAMENTE DEPENDIENTES

					1) si AGE_TIPO_ID_REP_LEGAL NO ES NULO entonces el campo AGE_NOMBRES_REP_LEGAL no puede ser nulo
AGE_ID_REP_LEGAL	ID del primer representante legal para el caso de productor persona Jurídica		carácter	20	156-RV320 MUTUAMENTE DEPENDIENTES 1) si AGE_TIPO_ID_REP_LEGAL NO ES NULO entonces el campo AGE_ID_REP_LEGAL no puede ser nulo
AGE_NOMBRES_REP_LEGAL	Nombres del principal representante legal para el caso de Productor persona Jurídica		carácter	100	156-RV320 MUTUAMENTE DEPENDIENTES 1) si AGE_TIPO_ID_REP_LEGAL NO ES NULO entonces el campo AGE_NOMBRES_REP_LEGAL no puede ser nulo
AGE_APELLIDOS_REP_LEGAL	Apellidos del principal representante legal para el caso de Productor persona Jurídica		carácter	100	156-RV320 MUTUAMENTE DEPENDIENTES 1) si AGE_TIPO_ID_REP_LEGAL NO ES NULO entonces el campo AGE_APELLIDOS_REP_LEGAL no puede ser nulo
AGE_CELULAR	Número de celular del Productor (Persona Natural o Jurídica)	>0	carácter	50	169-RV400 REGLAS DE DOMINIO 1) el valor del campo debe ser mayor a (0) cero. 2) El campo debe contener 10 dígitos únicamente 3) Si el agente no cuenta celular se debe digitar 10 ceros, 0000000000.
AGE_TELEFONO	Teléfono del agente productor		carácter	50	169-RV400 REGLAS DE DOMINIO 1) el valor del campo no debe ser nulo
AGE_DIRECCION	Dirección del agente Productor (Persona Natural o Jurídica)		carácter	200	169-RV400 REGLAS DE DOMINIO 1) el valor del campo no debe ser nulo

AGE_EMAIL	Email del Productor (Persona Natural o Jurídica)		carácter	100	169-RV400 REGLAS DE DOMINIO 1) el valor del campo no debe ser nulo
AGE_DIR_WEB	Dirección Web del Productor (Persona Natural o Jurídica)		carácter	200	169-RV400 REGLAS DE DOMINIO 1) el valor del campo debe contener el carácter especial @ y el punto.
PAIS_ID	ID del país del agente		carácter	4	
ZON_ID	Identificador de Ciudad - Zona	Códigos Divipola	numérico entero	6	122-RV150 REGLAS DE INTEGRIDAD REFERENCIAL. 1) El código del campo debe existir en la tabla CD_DIVIPOLA, en el campo COD_DPTO 155-RV310 DEPENDIENTE DEL ESTADO DE LA ENTIDAD 1) si el valor del campo AGE_DIRECCION no es nulo entonces el campo ZON_ID tampoco debe ser nulo
ZON_NOMBRE	Nombre de la ciudad o municipio	Nombres de la tabla Divipola	carácter	100	122-RV150 REGLAS DE INTEGRIDAD REFERENCIAL. 1) El código del campo debe existir en la tabla CD_DIVIPOLA, en el campo COD_DPTO 155-RV310 DEPENDIENTE DEL ESTADO DE LA ENTIDAD 1) si el valor del campo AGE_DIRECCION no es nulo entonces el campo ZON_ID tampoco debe ser nulo
REGISTRO_FECHA	Fecha de registro del agente		Fecha	8	169-RV400 REGLAS DE DOMINIO 1) La fecha debe tener formato fecha mm/dd/aaaa
ACTUALIZACION_FECHA	Fecha de actualización de los datos del agente		Fecha	8	169-RV400 REGLAS DE DOMINIO 1) La fecha debe tener formato fecha mm/dd/aaaa

Tabla 10. Reglas de Validaciones en Formulario de Captura. Fuente: Los Autores.

9.2.Validaciones del Formulario de Captura

Es indispensable una buena planeación de los campos requeridos para la captura de los datos los datos de los agentes cinematográficos, a continuación, se presenta en la figura 12, el formulario de captura de información.

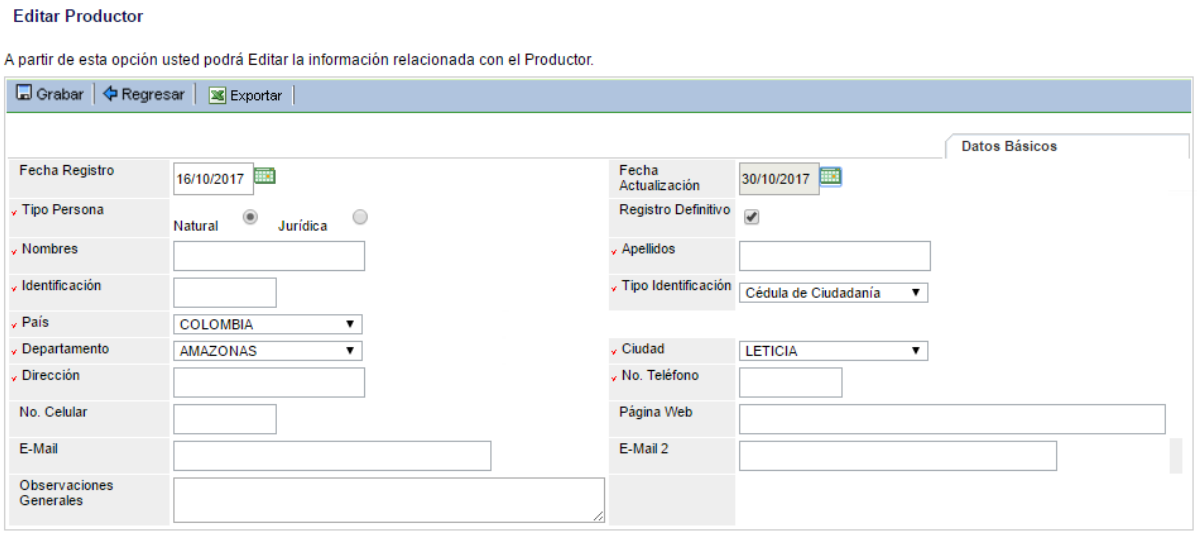


Figura 14. Formulario de datos de contacto de un Agente. Fuente: SIREC.

Validaciones requeridas para mitigar los errores de captura de los datos en el formulario de agentes:

a. **Fecha de Registro:** Se requiere control calendario de Microsoft, que permita visualizar la fecha con el formato dd/mm/aaaa.

b. **Fecha Actualización:** Se requiere control calendario de Microsoft, que permita visualizar la fecha con el formato dd/mm/aaaa

Validación: la fecha de actualización no puede ser inferior a la fecha de registro del agente, de lo contrario no permitir grabar, generar mensaje de error. Figura 13.

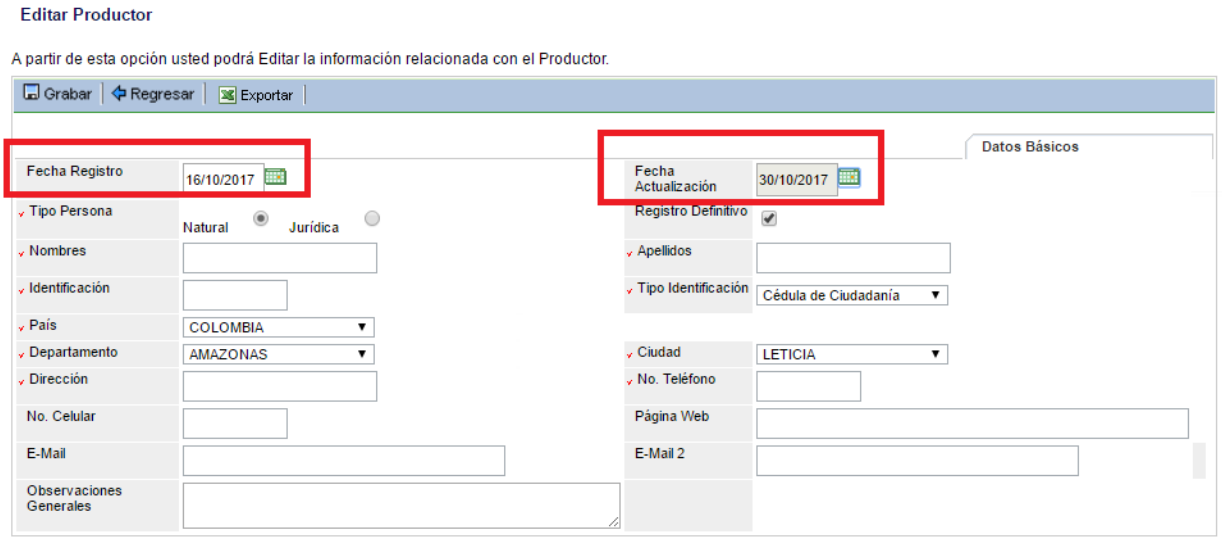


Figura 15. Validación para Fecha Registro y Fecha Actualización. Fuente: SIREC.

c. Tipo de Persona: Hace referencia a que un agente del sector cinematográfico puede ser una persona natural o una persona Jurídica.

Validaciones:

- Persona Natural: Se debe habilitar los campos Nombres y Apellidos, el campo Identificación, solo debe permitir ingresar un número máximo de 10 dígitos. **¡Error! No se encuentra el origen de la referencia.**14.

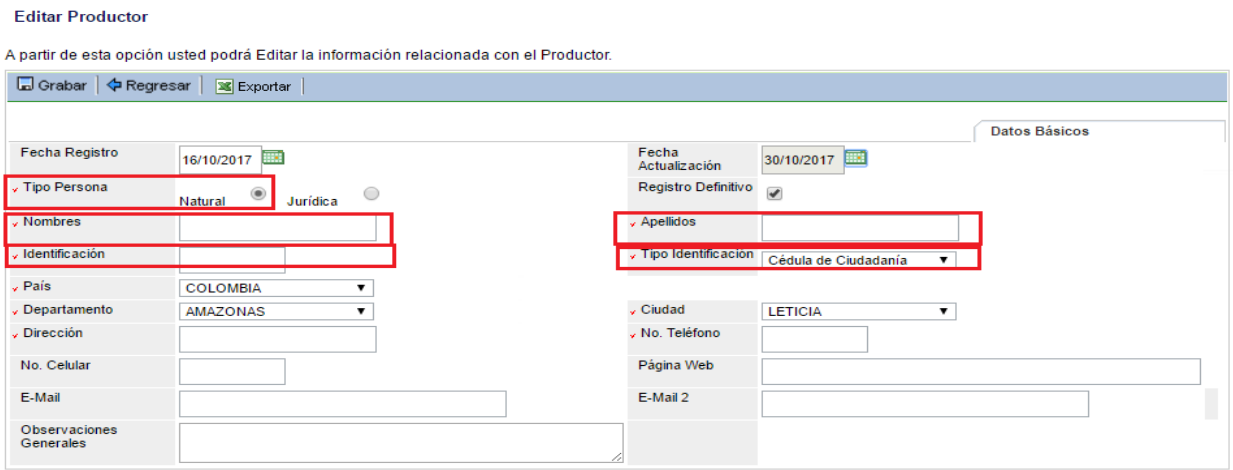


Figura 16. Validación por Nombre. Fuente: SIREC.

- Persona Jurídica: se debe habilitar los campos nombres, El tipo de Identificación de la persona jurídica debe quedar predeterminado a Nit, el campo identificación es de tipo carácter y solo puede permitir 10 caracteres, si el campo no contiene los 10 caracteres se debe generar un error y el sistema no debe permitir grabar. Figura 15.

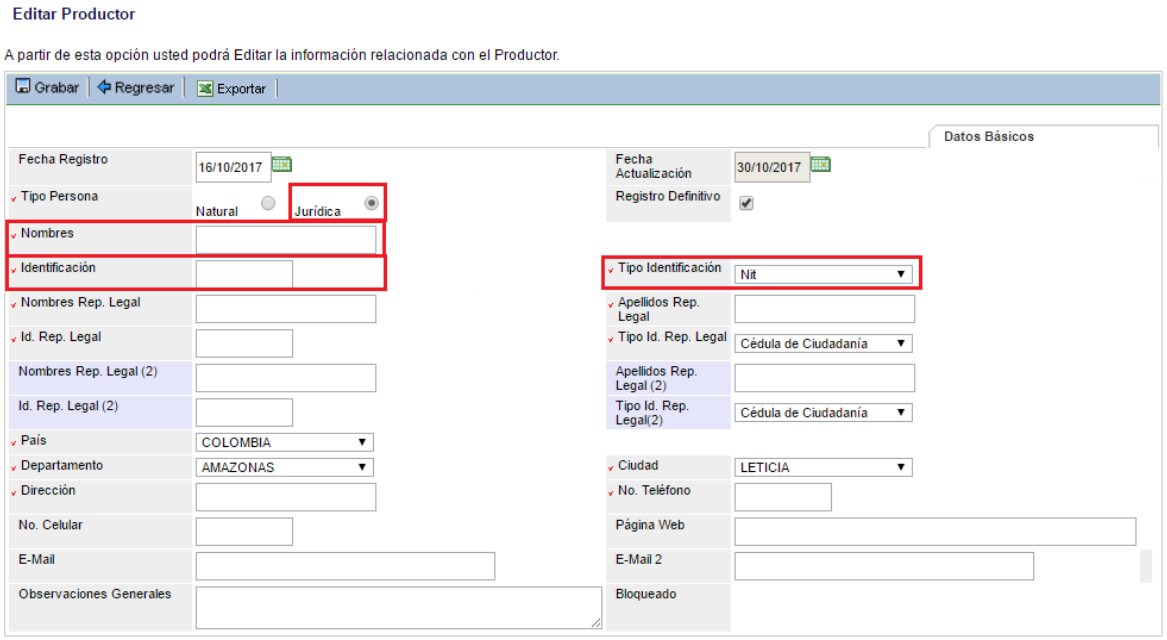


Figura 17. Validación persona Jurídica. Fuente: SREC.

d. Campo Teléfono: el campo no debe ser nulo, el sistema solo debe aceptar un número de 7 dígitos, en el caso que no se relacione un número telefónico se debe digitar en el campo 7 ceros (0000000).

e. Campo Celular: el campo no debe ser nulo, el sistema solo debe aceptar un número de 10 dígitos, en el caso que no se relacione un número celular se debe digitar en el campo 10 ceros (0000000000).

f. Campo Email: el campo es opcional, el sistema debe validar que haya un único carácter especial arroba (@) y mínimo un carácter punto (.).

10. Proceso de Estandarización de la Información

El proceso de estandarización es una manera de optimizar resultados y facilitar el proceso de búsqueda de la información, además de beneficiar el manejo, consulta y posterior análisis sobre la información.

En el SIREC, es necesario estandarizar las direcciones, teléfono y celular con el fin de almacenar la información actualizada, precisa y que corresponda a un objeto de la realidad.

10.1. Estandarización del Campo Dirección

Una dirección urbana puede ser escrita de muchas maneras, el problema más común es la forma como se escribe cada elemento que compone una dirección como por ejemplo el tipo de vía (Carrera, Cra, Kra, Cr, etc.). Otro problema que se presenta tiene que ver con la organización de estos elementos. Es necesario establecer algunos parámetros para la captura y almacenamiento de direcciones, y especificar una forma normalizada de escribir esta información. Un estándar se puede definir como el conjunto de reglas que proporcionan una forma de estructurar datos para facilitar su desempeño automático e interoperable.

Identificación de la información necesaria

Esta primera parte identifica la información necesaria que debe contener el registro de direcciones y la información de referencia (Malla vial - cartografía), la identificación, el contenido y la estructura de esta información es la base para la comparación, búsqueda y posterior localización de puntos sobre la cartografía. Como mínimo cada registro de dirección debe contener 6 elementos básicos: Tipo de vía, nombre o número de vía, prefijo o cuadrante, numero de vía generadora, prefijo o cuadrante de la vía generadora y un número de placa. Por ejemplo: Calle 11 Sur Número 23 A BIS – 50. Elemento Valor Tipo de vía. Calle Nombre o número de vía. 11 prefijo o cuadrante. SUR Número de vía generadora. 23 prefijo o

cuadrante de vía generadora. A BIS Número de placa 50

Elemento	Valor
Tipo de vía.	Calle
Nombre o número de vía.	11
Prefijo o cuadrante.	SUR
Número de vía generadora.	23
Prefijo o cuadrante de vía generadora.	A BIS
Número de placa	50

Tabla 11. Elementos Mínimos en un Registro de Dirección. Fuente: https://www.mineducacion.gov.co/1621/articles-193290_estandar_direcciones_urbanas.pdf

Adicionalmente a cada registro de dirección se debe especificar el departamento, el municipio-

Descripción General de los Elementos que Componen un Registro de Dirección Tipo de Vía.

Una vía puede ser clasificada de acuerdo con su orientación y diseño en:

Calle: Se codifica como CL. Vía pública con orientación predominante y sentido de crecimiento numérico de acuerdo con el modelo que cada ciudad le haya asignado.

Carrera: Se codifica como KR. Vía pública generalmente perpendicular a la calle con orientación predominante y sentido de crecimiento numérico de acuerdo con el modelo que cada ciudad le haya asignado.

Diagonal: Se codifica como DG. Vía pública que generalmente tiene el mismo sentido de la calle sin ser paralela a ésta, puede o no generar nomenclatura predial.

Transversal: Se codifica como TV. Vía pública que generalmente tiene el mismo sentido de la carrera sin ser paralela a ésta, puede o no generar nomenclatura predial, entre otros.

A continuación, se muestra las abreviaturas para tipo de vía.

Elemento	Abreviatura
Carrera	KR
Calle	CL
Transversal	TV
Avenida	AV
Diagonal	DG
Autopista	AU
Vía	VIA
Circular	CIR
Pasaje	PJ

Tabla 12. Tabla de Abreviaturas para las Vías. Fuente: https://www.mineducacion.gov.co/1621/articles-193290_estandar_direcciones_urbanas.pdf

Número nombre común de la vía principal: Valor numérico o nombre común que identifica la vía, en este caso la vía principal, por lo general las avenidas o vías principales tienen asociado un nombre común como, por ejemplo: "REGIONAL", "ORIENTAL", "MEDELLIN", "LAS PALMAS", etc. **Letra, letra-letra o letra-número-letra que acompaña la nomenclatura principal:** Campo alfanumérico, sirve para diferenciar las vías internas, generalmente siguen un orden lógico, ya sea alfabético o numérico o sea combinación de ambas.

Cuadrante (NORTE, SUR, ESTE, OESTE): Cuadrante geográfico en donde está ubicado el eje vial, indica el cuadrante al que pertenece en este caso la vía principal. Número de la vía generadora: Valor numérico con que se identifica el eje vía, en este caso la vía generadora.

Letra-letra o letra-número-letra que acompañan la vía generadora: Campo alfanumérico, sirve para diferenciar las vías internas, generalmente siguen un orden lógico, ya sea alfabético o con números consecutivos o sea combinación de ambos.

Número de la placa: Valor numérico, generalmente indica la distancia en metros desde la intersección entre la vía principal y la vía generadora hasta el acceso al predio. Corresponde con el segundo valor de la placa domiciliaria o predial (número que esta después del guion).

Cuadrante (NORTE, SUR, ESTE, OESTE): Campo que indica el cuadrante al que pertenece en este caso la vía generadora, (IGAC, 2009).

Procedimiento

Para el campo dirección se debe quitar la caja de texto actual que permite grabar la dirección, el cual es campo alfanumérico con un rango de 50 caracteres.

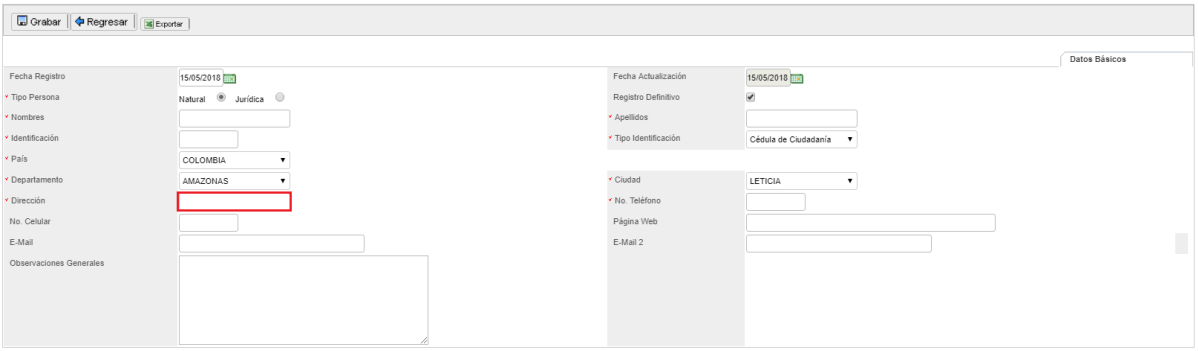


Figura 18. Formulario de Registro Campo Actual Dirección de un Agente. Fuente: SIREC.

El cambio se debe realizar por 4 listas desplegables que permitan seleccionar el tipo de vía, otro campo de texto donde se introduzca el primer número, seguido de una lista desplegable para seleccionar la letra, luego una caja del número, la letra, luego otra caja de texto para escribir el segundo número, para el caso de una segunda letra se crea una lista desplegable y por último la segunda caja de texto para digitar el tercer número, como se muestra en la siguiente figura 17.

Figura 19. Campos de la Dirección.

10.2. Estandarización del Campo Número Telefónico

Identificación de la información

Un número de teléfono es una secuencia de dígitos utilizada para identificar una línea telefónica dentro de una Red Telefónica Conmutada (RTC). El número contiene la información necesaria para identificar el punto final de la llamada. En Colombia los números telefónicos tienen una longitud fija de 7 dígitos.

Procedimiento

Para los registros de los números telefónicos registrados en el SIREC, se debe estandarizan los datos existentes, actualmente en la base de datos se encuentran muchos registros de números inconsistentes que presentan valores fuera de rango e información que no corresponde a la realidad como se muestra en la siguiente tabla 13.

	A	B	C	D
1	AGE_ID	TIP_ID	AGE_DIRECCION	AGE_TELEFONO
2	1019092446	CC	Cra. 67A # 131 -21, Casa 17	6137416
3	1032371343	CC	Calle 53 N° 35 - 96, Apto1	2122463 / 2218628
4	1032413518	CC	calle 127a #53a-52 porteria 6 , apto 414	4651876
5	1032443502	CC	Vereda Frío, Sector la Palma, Finca El Porvenir de las Mercedes	315 8048671
6	1039456679	CC	Girardota, Antioquia	0
7	1072659595	CC	Carrera 45 n 57-47, 501	3167277770
8	1088293075	CC	Carrera 36 # 4A-17, San Fernando Viejo.	24879726
9	1090444866	CC	calle 4, #11e-10, quinta oriental	0
10	1120981077	CC	Sector de Suroeste, Isla de Providencia	318 3524607
11	1128391537	CC	Vereda El Romeral, frente al crematorio Rituales, casa de la palm	4562936
12	1130604428	CC	6F Fillebrook Road, Leytonstone, E11 4AT	+44 (0) 7824162866
13	1130623344	CC	KM 2 VIA CHIPAYA, FARALLONES DE VERDE ALFAGUARA CASA 13	5921103
14	1214725404	CC	38 Hemenway St Boston, MA 02115	(617) 7941337
15	1238931	CE	14 005 Palawan way (116) marina del rey, California 90-292 U.S.A	6214795
16	15434077	CC	Calle 12 No. 1B-53, casa 22, conjunto Andes 4	8637047
17	16603922	CC	Corregimiento La Paz, Vereda Villa del Rosario	315 7131028
18	19294335	CC	Arboretto, Km 6 Vía Bogotá	8751020
19	29401771N	PAS	Defensa 599,2° piso. CP: C1065AAI	
20	45456711	CC	Calle 77 N°12 - 03, Apto 403	4776420
21	52388527	CC	5911 Alton Rd, Miami Beach, FI, 33140, USA.	3453421
22	52707151	CC	Carrera 21, nº 39 - 44, apartamento 101 barrio la soledad	2450182
23	52718750	CC	Buelowstr. 12 10783, Berlín, Alemania	+49 017661965211
24	53907356	CC	Diagonal 72 N° 1A - 95 Este, Apto. 501	3202316307
25	66993002	CC	Cl redondilla 3, 1o I, 28005 Madrid (España)	(34)915241168 / (34)0787006806
26	71267956	CC	Km 5 Mz F, Lote 71	311 8437088
27	71319667	CC	C/ Avinyó 34, 2o 3a 08002	34659801202
28	72201601	CC	Avenida de Roma 137, 4-1	+34 931402751
29	72357402	CC	VEREDA ROZO, FINCA LA GIOCONDA, CASA RITACUBA	0

Tabla 13. Ejemplo de Registro de Números Telefónicos en el SIREC.

Para el caso que hay números móviles en el campo teléfono, se debe pasar el dato a el campo celular y dejar el campo teléfono con 7 ceros 0000000.

Por otra parte, si existen dos números telefónicos en el mismo campo solo se debe dejar uno que corresponda al primer número telefónico. Se debe suprimir en la información del teléfono los signos +, (), los indicativos y las extensiones.

Para los campos telefónicos que tienen un número 0 que presentan valor nulo se debe agregar al campo 7 ceros, en futuras búsquedas de información se podrá seleccionar los agentes cuyos números telefónicos no registran con el fin de solicitar un proceso de actualización de la información.

10.3. Estandarización del Campo Celular

Identificación de la información

En Colombia un número celular tiene una secuencia de 10 dígitos, en el caso del campo celular actualmente se registran valores fuera de rango y que no corresponde a la realidad como por ejemplo números móviles de más de 10 dígitos existen espacios entre los números como se muestra en la siguiente tabla 14.

	A	B	C	H	I	J
1	TIP_ID	AGE_NOMBRES	AGE_APELLIDOS	AGE_DIRECCION	AGE_TELEFONO	AGE_CELULAR
2	CC	Agustin	Godoy Ronderos	Teodoro Garcia 2349	54155647 3766	54155647 3766
3	CC	Alejandro	Gómez			
4	CC	Alejandro	Matallana Beltrán	Cra 58 No 134-57 Etapa 4 Int. 2 Apto 501	2269867	3192782561
5	CC	Alejandro Arturo	Buchelli Gomez	CL 12 No 29B - 78 APTO 501Z	4859927	3178708827
6	CC	Alejandro David	Landes Echavarria	301 OCEAN DR APTO 405 MIAMI BEACH FL 33139	N/A	N/A
7	CC	Alejandro Evaristo	Quintero Montero	Cl 73 No 11 - 52 (204)	6063744	315 3600725
8	PAS	Alejandro Gabriel	Lázaro Alonso	Calle Nebulosas 2, portal C 1eroA. 28045	34 911437761	
9	NIT	Alejandro	Simbaqueva Manrique	tranversal 15 b # 46 - 16 ofc 707 chapinero	7559496	3138642464
10	CC	Alessandro	Angulo Brandestini	Cr 5 No 26 - 39	2866789	310 2124120
11	CC	Alex Andrés	López Guevara	Cr 18 No 2 - 62	8211546	3105905315
12	CC	Alex Díaz	Rubio Reyes	CR 2 No 43 - 23 PIEDRA PINADA BAJA	8 2 666350	301 201 84 06
13	CC	Alexander	González Tascón	Cr 26 No 15 - 56	3774534 / 3264282	317 4024297
14	CC	Alexander	Ramírez	CR 2A No 66 - 52 APTO 112B	3491842	311 2103684
15	CC	Alexander	Sterling Reyes	CR 83 No 43 - 70	3321428	3148973849
16	CC	Alexander	Valencia Martínez	Cr 5ª No 59ª-44	3488000	+ 57 321 3821578
17	CC	Alexander	Vargas Cardona	Cl 35 No 21 - 43 (201)	300 7812324	300 7812324
18	CC	Alexandra	Cardona Restrepo	Av Cr 7 No 119 - 14 (519)	7532645-7591665	315 6013536
19	CC	Alexandra	Castañeda Giraldo	Calle 13 c 72 93	4847360	3108544682

Tabla 14. Valores que se presentan en el campo celular – Base de Datos de Agentes

Procedimiento

Para los campos de celular que presentan más de 10 dígitos se debe corregir la inconsistencia buscando la fuente principal de la información, es decir los registros

administrativos que presentaron los agentes para la solicitud de registro con el fin de identificar el error y poder corregir la información por el valor real.

Para los números celulares que presentan espacio después de los primeros tres números que indican el operador móvil, se debe eliminar el espacio.

En los campos que se presentan letras se debe eliminar las letras y poner 10 ceros 0000000000 con el fin de facilitar futuras búsquedas de información para procesos de actualización de la información de agentes.

En el caso que un registro presenta campo nulo se debe poner 10 ceros 0000000000 atendiendo a las pre -validaciones del campo que indican que el dominio del campo debe ser mayor a 0 y contener solamente 10 dígitos.

10.4. Estandarización de los Códigos de las Ciudades y Departamentos

Actualmente la base de datos del SIREC debe actualizar la información de los códigos de los departamentos y municipios de acuerdo la codificación de la División político-administrativa de Colombia, Divipola, con el fin de los datos cumplan con los estándares nacionales,

El Departamento Administrativo Nacional de Estadística, DANE, dentro de sus funciones tiene la de diseñar y ejecutar las operaciones que garanticen la georreferenciación de la información estadística que requiera el país para la planeación y toma de decisiones por parte del Gobierno Nacional y las entidades territoriales identificadas como departamentos y municipios, para lo cual se responsabiliza de actualizar y divulgar la codificación de la División político-administrativa de Colombia, Divipola. La adecuada administración de esta herramienta facilita la asociación, integración y agregación de información temática demográfica, Es así, que se conformó la División político-administrativa de Colombia, Divipola, la cual es un estándar nacional que consolida en un inventario, la identificación y codificación de los departamentos, municipios y distritos (referidos a algunas ciudades). La Divipola tiene como objetivo esencial el dar a conocer a los ciudadanos, la

codificación de las entidades territoriales y centros poblados que conforman el territorio nacional.

Para el caso del SIREC, la tabla departamentos y ciudades en la base de datos debe asociarse de la siguiente manera como se muestra en la tabla XX con el fin de tener la información la codificación y nombres actualizada dos de los municipios colombianos.

DIVIPOLA- Codigos municipios

Código Departamento	Código Municipio	Nombre Departamento	Nombre Municipio
05	05002	Antioquia	Abejorral
54	54003	Norte de Santander	Ábrego
05	05004	Antioquia	Abriaquí
50	50006	Meta	Acacías
27	27006	Chocó	Acandí
41	41006	Huila	Acevedo
13	13006	Bolívar	Achí
41	41013	Huila	Agrado
20	20011	Cesar	Aguachica
68	68013	Santander	Aguada
17	17013	Caldas	Aguadas
85	85010	Casanare	Aguazul
41	41016	Huila	Aipe
25	25019	Cundinamarca	Albán
52	52019	Nariño	Albán
18	18029	Caquetá	Albania
44	44035	La Guajira	Albania
68	68020	Santander	Albania
76	76020	Valle del Cauca	Alcalá

Tabla 15. Códigos de los Municipios de acuerdo con el estándar DIVIPOLA. Fuente: https://www.dane.gov.co/index.php/sistema-estadistico-nacional-sen/normas-y-estandares/nomenclaturas-y-clasificaciones/nomenclaturas

El DANE trimestralmente realiza la actualización de los códigos Divipola por lo que se deberá consultar en la página del DANE en el siguiente Link: https://www.dane.gov.co/index.php/sistema-estadistico-nacional-sen

11. Proceso de Unicidad de la Información

En la información evaluada en el SIREC no se encontraron errores de duplicación cabe recordar que la duplicación consiste en que la misma entidad aparece repetida de manera exacta, gracias a que el modelo relacional de la base de datos es acorde y óptimo. Por otra parte, si se encontraron errores de contradicción en donde la misma entidad aparece repetida con contradicciones. Por ejemplo:

La señora Clare Feale Weiskopf aparece dos veces con número diferente de cédula, dirección, E-mail y teléfono diferente como se muestra en la siguiente figura XX.

Figura 20. Caso de Contradicción en un Registro de Agente. Fuente: SIREC.

Para corregir este caso no se debe eliminar ninguno de los registros, el proceso que se debe realizar en la base de datos es la unificación del registro, para ello es necesario consultar la fuente principal de donde se obtuvo el dato, ya sea por registros administrativos o por los aplicativos de trámite en línea, con el fin de identificar cual es la información correcta y actualizada, una vez se tiene identificada la información correcta, por un procedimiento de bases de datos se realiza la unificación y actualización de los datos dejando un solo registro.

Sin embargo, también es posible que el administrador del sistema identifique el registro correcto y edite la información en ese registro, posteriormente el registro incorrecto se puede eliminar.

12. Proceso de Completitud de la Información

La completitud en un modelo relacional puede caracterizarse por los siguientes aspectos: Valores nulos: el significado de los valores nulos puede ser variado. Un valor nulo puede indicar que dicho valor no existe en el mundo real, que el valor existe en el mundo real pero no se conoce, o que no se sabe si el valor existe o no en el mundo real.

En la información analizada en la base de datos de agentes se identificaron valores nulos para los campos teléfono, celular y E-mail.

Para el campo teléfono, se pudo identificar que los valores corresponden a registros en donde el agente no registra teléfono fijo, pero si número celular, por lo que hace parte de una realidad en donde las personas hacen mayor uso del teléfono móvil que del teléfono fijo, para este caso se hace necesario que, por reglas de validación del campo, se asignen 7 ceros y así en un futuro actualizar la información.

En el campo celular se identificó que existen valores nulos debido a que los registros corresponde a un periodo en donde no había un auge en general o acceso a los teléfonos móviles, los registros son del año 2004 a 2006, muchos agentes registrados no contienen esa información por lo que se puede deducir que el valor no existe para el mundo real, en este caso no se debe dejar el campo celular nulo sino que se debe digitar 10 ceros que corresponde a la secuencia única de un celular, con el fin de que en el futuro se pueda realizar búsquedas y actualización de la información de contacto de un agente cinematográfico.

Proceso de Actualización de la información de contacto de los agentes

Para complementar el proceso de completitud de los datos de contacto de los agentes es necesario que se cree una opción en el sistema que valide la fecha de registro y la fecha de actualización de la información de un agente con el fin de que desde el SIREC se le envíe al E-mail del agente, una notificación de solicitud de actualización de su información de contacto cada dos años, con el fin de contar con información actualizada, precisa y que corresponda a la realidad del objeto o entidad.

Para está opción el administrador del sistema cada mes deberá realizar una consulta de los registros que caducan por llevar más de dos años sin actualizar su información, y a través de la opción notificar generará una selección de los registros con el fin de enviar automáticamente la solicitud de actualización de los datos a sus correos electrónicos registrados.

13. Proceso de Corrección de Información por Inexactitud Semántica

Para corregir los datos por inexactitud semántica en la fase de limpieza de datos se deben realizar los siguientes pasos:

1. Análisis de datos: esta fase consiste en determinar los errores e inconsistencias que deberán eliminarse. Para ello se realiza una inspección manual o se utilizan programas de análisis de datos, para el caso del SIREC solo se debe usar Hoja de cálculo en Excel debido a que no se cuentan con licenciamientos de otros programas que permitan realizar el proceso.

Existen dos enfoques:

a. **Data profiling**: consiste en analizar los datos de la base de datos de agentes y a partir

de estos obtener propiedades que se cumplen en la misma. Se centra en el análisis de los atributos: su contenido, estructura, dependencias en una relación, solapamiento con atributos de otras relaciones, valores faltantes y duplicados.

b. Data mining: se ocupa de la identificación de patrones en conjuntos de datos (por ejemplo, definir una relación entre distintos atributos).

Ejemplos:

- Para errores de tipeo: ordenar los campos de manera tal que los valores con errores se sitúen cerca de los reales.
- Para valores faltantes: cantidad de nulos, presencia de valores por defecto pueden indicar también la falta de un valor.
- Variación en la representación de valores: comparar columnas iguales de tablas (fuentes) distintas.
- Duplicados: ordenar los valores por cantidad de ocurrencias.

2. Detección y Corrección de Errores

Utilizar el término error puede resultar demasiado amplio, teniendo en cuenta el concepto multifacético con el que se define la calidad de datos. Por lo tanto, se puede poner foco en:

- Detectar y corregir inconsistencias: básicamente se trata de detectar registros que no cumplan con determinadas reglas, y luego modificar los datos, por ejemplo, buscando los datos en la fuente original o a partir de la obtención de nueva información, para que cumplan con las reglas. Esta tarea incluye asegurar que la información se encuentra consistente (sin contradicciones) y libre de redundancias.

Una técnica para la localización de errores es la llamada *Data editing*, la cual consiste en la definición de reglas (*edits*) que deben ser respetadas por cierto conjunto de datos, para lograr de esta manera la detección de inconsistencias. Los *edits* representan condiciones de

error, por lo cual deben ser consistentes y no redundantes. Los datos de un registro deben ser ajustados de manera tal que cumplan con las reglas, pero minimizando las modificaciones a los datos.

A modo de ejemplo, se tiene la tabla que almacena los datos de contacto de un agente si es persona natural, tiene cedula de ciudadanía, luego, es posible definir una regla que especifique que, si tiene cedula de ciudadanía, entonces el campo Rep_Legal debe ser *false*. A partir de esta regla, se pueden identificar los registros que no la cumplan, y corregirlos. Existen varias formas de corregir los errores detectados:

- Refrescar la base de datos con nuevos datos.
- Utilizar los *edits* definidos de manera tal que cuando no se cumple una regla, se imputa un valor que haga que la misma sea verdadera.

14. Proceso de Corrección para Registro Inexistente

En este caso se identifican aquellos registros que no corresponden a ningún objeto de la realidad. Estos registros que se encuentran almacenados en la base de datos, pero que se asocian a un objeto que en la realidad no existe por lo cual se deben eliminar por la opción eliminar que se encuentra en el menú del formulario como se puede identificar en la siguiente figura 19.

Figura 21. Opción Eliminar en el Formulario de un Agente Cinematográfico

15. Fase Verificar

En esta etapa lo que se busca es asegurar que la información se encuentra consistente (sin contradicciones) y libre de redundancias, para ello se llevará a cabo unas revisiones técnicas, las cuales deben ser realizadas por los especialistas del área de calidad o en su defecto por un especialista designado por el administrador del sistema, el cual debe asumir, además, la responsabilidad de dejar constancia de los resultados de las revisiones en los registros habilitados (para este caso en el informe de hallazgos).

A continuación, se lista los siguientes pasos, con el fin de verificar cada una de las etapas relacionadas en la etapa aplicar la metodología propuesta.

1. Familiarizarse con el sistema, esto implica los formularios de recolección de datos e interfaces de captura de información en el sistema.
2. Identificar cuáles son las actividades manuales en las cuales suelen ocurrir la mayor cantidad de errores, y buscar su mejora.
3. Verificar que el sistema esté tomando las validaciones previamente programadas en el sistema.
4. Determinar los requerimientos, condiciones, recursos, tiempos y responsables del diseño, validaciones y pruebas.
5. Comprobar los campos con los cuales se realizarán las pruebas en términos del sistema.

   a. Categorización de errores

   b. Correcciones y ajustes.
6. Diseñar los casos de pruebas. Especificar las entradas, resultados, condiciones y procedimientos para la realización de las pruebas de funcionalidad.

7. Conducir las pruebas. Realización de las pruebas mediante el siguiente procedimiento:
    a. Ejecución de las acciones definidas para cada caso.
    b. Registro, en el formato de hallazgos las observaciones por parte del usuario para cada caso de prueba. Registro, en el formato de caso de prueba.
    c. Visto bueno, en el formato por parte del responsable de la prueba por cada caso.
8. Entrega de los resultados de las pruebas al supervisor del Plan de Pruebas.
9. Análisis de los resultados.
10. Reporte Informe final de pruebas.

15.1. Plan de Monitoreo

Desarrollar el Plan de Monitoreo de acuerdo con la guía metodológica establecida en el documento. El formato de reporte de hallazgos presenta la información y la guía para completar el formato del reporte de monitoreo el cual estará compuesto por los siguientes campos:

a. Fecha de registro de monitoreo
b. Número de errores presentados
c. Nombre del campo involucrado en el hallazgo.
d. Fecha en la cual se registró dicho hallazgo.
e. Descripción de la falla
f. Solución dada
g. Fecha de corrección en el sistema.

El especialista usará como soporte un checklist de verificación para tomar las acciones correctivas: El reporte tendrá los hallazgos identificados respecto a la implementación y operación del proyecto.

16. Conclusiones

La metodología propuesta es dinámica y fácilmente adaptable a los cambios y las mejoras por introducir en el SIREC, la aplicación del modelo (analizar, hacer, verificar) es basado en el concepto de mejora continua, la cual permitirá los siguientes resultados:

- La toma de decisiones sobre la seguridad de la información que arroja el sistema de información (auditorias e indicadores de evaluación).
- Se orienta a la mejora continua, a través de la gestión de acciones correctivas y preventivas.

El proceso de estandarización es una manera de optimizar resultados y facilitar el proceso de geocodificación, además de beneficiar el manejo, consulta y posterior análisis sobre la información.

17. Bibliografía

[1] [Online]. Disponible: http://www.bdigital.unal.edu.co/. [Consultada 2 de junio de 2017].

[2] [Online]. Disponible: http://dc-pubs.dbs.uni-es/Maletic2000DataCleansingBeyond.pdf. [Consultado 18 de junio de 2017].

[3]Online].Disponible:http://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf. [Consultado 15 de julio de 2017].

[4] [Online] Disponible: https://www.infrastructure-journal.net/abstracts.[Consultado 22 de agosto de 2017].

[5] Oliveira, P., Rodriguez, F., Henriques, P., y Galhardas, H. 2005. A Taxonomy of Data Quality Problems. En: Second International WorkShop on Data and Information Quality IQIS 2005 (Porto, Portugal).

[6] Tierstein, Leslie. A Methodology for Data Cleansing and Conversion, White paper W R Systems Ltd.

[7] Rosenthal, A., Wood,D, y Hughes, E. 2001. Methodology for Intelligence Database Data Quality.

[8] I. A. Uribe, "Guía Metodológica para la selección de técnicas de depuración de datos," Medellín, 2012.

[9] C. Muñoz, Caracterización de técnicas para detección de valores faltantes, Medellín: Tesis. UPB. Facultad de Ingeniería Informática, 2010.

[10][Online].Disponible:https://eva.fing.edu.uy/pluginfile.php/178384/mod_resource/content/1/3-Métricas%20de%20calidad.pdf [Consultado el 18 de febrero de 2018].