IDEAM - Instituto de Hidrología, Meteorología y Estudios Ambientales

Plan de calidad de Componentes de Información

TABLA DE CONTENIDO

LISTADO DE GRAFICAS

1. OBJETIVO

El presente plan de calidad de los componentes de información tiene como objetivo establecer los lineamientos que permitan dar la óptima solución a los problemas causantes de la mala calidad de los componentes de información con que se toman las decisiones tanto a nivel misional como a nivel administrativo del IDEAM. Entre los objetivos que pretende cumplir el presente plan de calidad de datos se enuncian los siguientes:

1.1. El objetivo principal de los proyectos de calidad de componentes de información es ayudar al negocio a interactuar con el cliente o interesado de manera eficiente, haciendo que la experiencia del interesado resulte lo más grata posible.

1.2. Convencer a toda la institución de las bondades de implementar un proceso de calidad de componentes de información. En este sentido, un patrocinador (sponsor) fuerte (alta dirección) y convencido de los resultados, generará la confianza y el tiempo para lograrlo.

1.3. Definir procesos que permitan apoyar la mejora de la calidad de los componentes de información, a partir del Gobierno de Datos

1.4. Definir un método que permita la limpieza y estandarización de componente de información para el mantenimiento, integridad y aseguramiento de la calidad de los mismos.

1.5. Mantener altos estándares de calidad de los componentes de información dentro del IDEAM.

1.6. Dotar al IDEAM de métodos preventivos y correctivos eficientes en la gestión de la calidad de los componentes de información, para el tratamiento de los errores, duplicidad e inconsistencias que estos puedan presentar, proveyendo control y los medios de depuración para minimizar el impacto que pueda traer en los procesos de gestión.

1.7. El plan de calidad de componentes de información pretende cubrir los procesos para el tratamiento y gestión en el diagnóstico de calidad de datos, información y demás, definir principios para orientar en la estrategia de calidad de componentes de información, limpieza y normalización de dichos componentes, asistir en proyectos de migración de datos, definir el control de duplicados, establecer oportunidades de mejora y la adecuada validación de los componentes de información.

Finalmente, el presente documento pretende enfocarse en las dimensiones más importantes de los componentes de información como son la exactitud, la integridad, la consistencia y la coherencia; es conveniente señalar que éstas deben ser definidas teniendo en cuenta las características propias de cada dependencia.

En esta lógica se basa este plan de calidad de componentes de información, el cual trata de cómo identificar y eliminar las causas de familias completas de errores para evitar errores futuros (más que para detectarlos y corregirlos) y de implantar a futuro una infraestructura de gestión que permita desarrollar adecuadamente este objetivo. Un elemento importante considerado en estos sistemas es el diagnóstico. En ese sentido también se darán pautas para el diagnóstico de la calidad de los componentes de información que, entre otros aspectos, inciden en la búsqueda de las causas que provocan

los problemas de calidad.

2. ALCANCE

Este plan tiene como alcance aplicar procesos de control de calidad específicamente a los datos, información, servicios de información y flujos de información del IDEAM, enfocados en las dimensiones más importantes de dichos componentes de información como son la exactitud, la integridad, la consistencia, la coherencia, la eficiencia, velocidad de proceso, usabilidad y gestión, además del valor agregado al servicio para la satisfacción total de los interesados.

3. NORMATIVIDAD

- Política de Gobierno Digital expedida por el Decreto 1008 del 14 de junio de 2018 del Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC), a través de la Dirección de Gobierno Digital, cuyo objetivo será incentivar el uso y aprovechamiento de las TIC para consolidar un Estado y ciudadanos competitivos, proactivos e innovadores, que generen valor público en un entorno de confianza digital.

- Estrategia de Gobierno en Línea, que se plasma en el Decreto Único Reglamentario del Sector de Tecnologías de la Información y las Comunicaciones 1078 de 2015, comprende cuatro grandes propósitos: lograr que los ciudadanos cuenten con servicios en línea de muy **alta calidad**, impulsar el empoderamiento y la colaboración de los ciudadanos con el Gobierno, encontrar diferentes formas para que la gestión en las entidades públicas sea óptima gracias al uso estratégico de la tecnología y garantizar la seguridad y la privacidad de la información.

- Lineamientos LI.ES.01, LI.INF.02, LI.INF.13 y LI.INF.10 La entidad hace monitoreo a la calidad y uso de los datos.

- Lineamientos LI.GO.04 y LI.GO.05, La entidad cuenta con un esquema de gobierno de TI que contemple políticas, procesos, recursos, gestión del talento y proveedores, compras, calidad, instancias de decisión, estructura organizacional e indicadores de la operación de TI.

- Lineamiento LI.INF.13. La entidad aplica los mecanismos adecuados de aseguramiento, control, inspección y mejoramiento de la calidad de los componentes de información.

- LI.SIS.20 y LI.SIS.21 La entidad aplica los mecanismos adecuados de aseguramiento, control, inspección y mejoramiento de la calidad de los sistemas de información.

- Ley de Protección de Datos Personales Ley 1581 de 2012, reconoce y protege el derecho que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos que sean susceptibles de tratamiento por entidades de naturaleza pública o privada.

- Ley 1712 de 2014, por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones.

- Decreto 235 de 2010, por el cual se regula el intercambio de información entre entidades para el cumplimiento de funciones públicas.

- Decreto 2280 de 2010, por el cual se modifica el artículo 3° del Decreto 235 de 2010.
- Decreto 1377 de 2013, por el cual se reglamenta parcialmente la Ley 1581 de 2012.
- Decreto 886 de 2014, por el cual se reglamenta el artículo 25 de la Ley 1581 de 2012, relativo al Registro Nacional de Bases de Datos.
- Decreto 103 de 2015, títulos I, II, III, IV, por el cual se reglamenta parcialmente la Ley 1712 de 2014 y se dictan otras disposiciones.
- Decreto 1081 de 2015, capitulo 4, por medio del cual se expide el Decreto Reglamentario Único del Sector Presidencia de la República.
- Sentencia C-748 de 2011, control constitucional al Proyecto de Ley Estatutaria No.184 de 2010 Senado; 046 de 2010 Cámara, “por la cual se dictan disposiciones generales para la protección de datos personales”.

4. DEFINICIONES

- **Algoritmo**: Es un conjunto de instrucciones o reglas definidas, ordenadas y finitas que permite realizar una actividad mediante pasos sucesivos.
- **Análisis**: Consiste en identificar los componentes de un todo, separarlos y examinarlos para lograr acceder a sus principios más elementales.
- **Aseguramiento de Calidad – QA**

  En el ámbito del desarrollo de software, la sigla QA significa **Quality Assurance**, o aseguramiento de la calidad. Se trata de un conjunto de actividades de evaluación de las distintas etapas del proceso de desarrollo para garantizar que el producto final sea de calidad. El concepto de calidad se presta a múltiples interpretaciones, pero siempre implica que el software satisfaga las necesidades del cliente.

  Más allá de las diferencias, un buen plan de QA no puede desconocer la importancia de los estándares. Con esto nos referimos a reglas escritas y no ambiguas sobre los objetivos del producto, las metodologías de diseño y a seguir y convenciones necesarias para guiar la tarea de los programadores (estilos de codificación, estructuras de datos, etc.).

  El plan de QA atraviesa el proceso de desarrollo desde el nacimiento de la idea hasta la implementación del software. En las primeras etapas, verifica que los objetivos estén bien planteados y los requerimientos sean precisos. En las fases de diseño y codificación, vigila el cumplimiento de los estándares fijados. Finalmente, revisa que el software en funcionamiento respete los requerimientos pedidos y que la entrega al cliente se haga en las condiciones adecuadas.

- **Automatizar**: Proceso por el cual se ejecutar de manera programada y repetitivamente diversas actividades sin la intervención del ser humano.
- **Bodegas de datos**: Repositorio centralizado de datos, que almacena por lo general información histórica, y sirve como base para una solución de Inteligencia de Negocios.

- **Calidad de datos**: Se refiere al mecanismo de verificación de los datos, y su correcta validez, para que estén aptos en diversos procesos de la empresa.

  Es el componente del dominio de información asociado con procesos de ajuste y depuración de datos masivos, y definición, medición y mejora continua de los indicadores de calidad del dato.

- **Componente de información**

  Componente de información es un término agrupador utilizado, en el MRAE para la gestión de TI, para referirse al conjunto de los datos, la información, los servicios de información y los flujos de información bajo un único nombre.

  A continuación, se describen los componentes de información:

- **Datos**: Es una representación simbólica de una característica particular de un elemento o situación, que pertenece a un modelo de una realidad. Tiene un tipo (numérico, cadena de caracteres o lógico) que determina el conjunto de valores que el dato puede tomar. En el contexto informático, los datos se almacenan, procesan y transmiten usando medios electrónicos, constituyendo los elementos primarios de los sistemas de información.

- Los datos son números, letras o símbolos que describen objetos, condiciones o situaciones. Son el conjunto básico de hechos referentes a una persona, cosa o transacción de interés para distintos objetivos, entre los cuales se encuentra la toma de decisiones. Ejemplo de datos: Cédula, nombre, dirección, nombre de un trámite, los cuales tiene un tipo, por ejemplo, cédula es de tipo numérico, nombre es de tipo carácter.

- **Información**: Es un conjunto de datos organizados y procesados que tienen un significado, relevancia, propósito y contexto (conjunto de circunstancias materiales o abstractas, que se producen alrededor de un hecho, o evento dado, que están fiablemente comprobadas), que resulta útil para que los usuarios finales, ejecuten de manera apropiada su proceso de negocio y pueda tomar decisiones. La información puede servir como evidencia de las actuaciones de las entidades. Un documento, un listado de contratistas o funcionarios, la satisfacción de usuarios frente a un servicio, indicadores del entorno se consideran ejemplos de información y deben ser gestionados como tal.

- **Servicios de Información**: Es la integración de actividades que satisfacen necesidades de información de uno o más grupos de interés. Los servicios de información son las diferentes formas de brindar acceso a la información. Un servicio de información se describe a través de un contrato funcional (qué recibe como entrada y qué produce como salida) y un conjunto de acuerdos de servicio que se deben cumplir. Por ejemplo, la Unidad de la Atención y Reparación Integral a las Victimas provee un servicio web de intercambio de información sobre víctimas del conflicto armado en Colombia, entre otros.

- **Flujos de Información**: Corresponde a la descripción explícita de la interacción entre proveedores de información y consumidores de información, con un patrón repetible de

invocación definido por parte de la entidad. Puede incorporar servicios de información, datos e información. Cada información tiene asociado un flujo.

- **Consolidación**: Mecanismo por el cual se centralizan diferentes componentes (generalmente tecnológicos) en un solo lugar, con el fin de garantizar una administración más sencilla.

- **Control de calidad**

  Es el conjunto de actividades destinadas a evaluar el trabajo para el desarrollo de un producto.

  Control de calidad = medición de la calidad de un producto

  Las tareas de aseguramiento de la calidad están interesadas en el proceso de desarrollo del producto, mientras que **testing** y el control de calidad están interesados en el desarrollo del producto en sí mismo.

- **Duplicidad**: Ejercicio de tener múltiples copias de la misma información en diferentes sistemas de datos. Es un proceso normal de las empresas, pero que dificultan los análisis de los datos.

- **Estandarización**: Proceso enfocado en la eliminación de datos erróneos, duplicados o que representen sinónimos dentro de la información.

- **Etl**: (Etract, Transform, Load) Siglas referentes a los procesos de Inteligencia de Negocios que permiten la extracción de datos de cualquier fuente de datos, diversas transformaciones dependiendo las reglas de negocio y posteriormente el cargue de esta información en diversos orígenes de datos.

- **Framework**: Es un conjunto estandarizado de conceptos, prácticas y criterios particular que sirve como referencia, para enfrentar y resolver nuevos problemas de índole similar, y en este caso particular, mediante el uso de tecnologías.

- **Iso**: Es el organismo encargado de promover el desarrollo de normas internacionales de fabricación de productos y servicios, comercio y comunicación para todas las ramas industriales a excepción de la eléctrica y la electrónica, cuyo estándar es IEEE.

- **Limpieza de datos**: Es el proceso de descubrir, corregir o eliminar datos erróneos de una base de datos.

- **Monitoreo de datos**: Proceso encargado de hacer seguimiento a los datos corregidos, datos correctos y a las sugerencias de corrección que la herramienta pueda generar sobre los datos analizados.

- **Normalizar**: es un proceso que clasifica relaciones, objetos, formas de relación y otros tantos elementos en grupos, en base a las características que cada uno posee. Si se identifican ciertas reglas, se aplica una categoría, si se definen otras reglas, se aplicará otra categoría

- **Origen de datos**: Proporciona acceso a repositorios de contenido externo, permitiendo importar contenido en el portal mediante el uso de rastreador y envió de documentos, cada fuente de datos está configurada para tener acceso a un repositorio de documentos.

- **Perfilamiento**: Etapa de revisión de datos existentes, y las estadísticas e información acerca de los mismos.

- **Plan de calidad**

  Según la norma ISO 9000 2015 “Fundamentos y vocabulario”, un plan de calidad es una especificación de los procedimientos y recursos asociados a aplicar, cuándo deben aplicarse y quién tiene que aplicarlos a un objeto específico. Un plan de calidad es información documentada que especifica qué procedimientos de trabajo y recursos se encuentran asociados y se deben aplicar en el proceso, quien son las personas que deben aplicarlos y cuándo tienen que aplicarse a un proyecto, producto, proceso o contrato específico. Los planes de calidad proporcionan una forma de relacionar los requisitos específicos del proceso, producto, proyecto o contrato con los métodos y prácticas de trabajo que apoyan la realización del producto o servicio ofrecido.

  El equipo de calidad realizará un conjunto de actividades que servirán para:

- **Reducir, eliminar y prevenir**.

  Las deficiencias de calidad de los productos a obtener.

- **Alcanzar una razonable confianza.**

  En que las prestaciones y servicios esperados por el cliente o el usuario queden satisfechas.

- **Testing / Testeo / Pruebas**

  Es el proceso de ejecución de un sistema con la intención de encontrar defectos, incluida la planificación de las pruebas previa a la ejecución de los casos de prueba. En la mayoría de los casos.

  Testing = control de calidad

- **Testeo unitario**: se prueba que cada módulo funcione bien por separado.

- **Test de integración**: los módulos probados independientemente durante el testeo unitario se acoplan y se prueban en conjunto.

- **Test funcional**: se prueba que el software ofrezca las funciones solicitadas.

- **Test de aceptación**: el usuario verifica que el producto satisfaga sus expectativas.

- **Tic:** Las TIC (Tecnologías de la Información y la Comunicación) agrupan los elementos y las técnicas usadas en el tratamiento y la transmisión de la información, principalmente la informática, Internet y las telecomunicaciones.

- **Prueba de stress:** se prueba la resistencia de la aplicación enviándole una cantidad de peticiones excesiva, buscando que colapse.
- **Diferencias entre garantía de calidad (QA) y control de calidad (QC)**

QA (Quality Assurance)	vs	QC (Quality Control)
Se diseñan y definen todos los parámetros de aceptación de un paquete de Software		Se controla el comportamiento del producto final
Es un sistema de **PREVENCIÓN** de fallos que predice casi todo sobre la seguridad, funcionamiento, normas de calidad y legalidad de un paquete de SOFTWARE y genera medidas correctivas para controlar y evitar que los productos o servicios defectuosos lleguen a la fase de producción		Es un sistema de **CORRECCIÓN** de fallos e introducción de mejoras
El departamento QA trabaja junto a desarrollo, ingenieros, managers y el cliente		El departamento de QC, trabaja junto a QA
El departamento QA está presente desde la fase del diseño del producto		El departamento de QC entra en acción cuando el producto está finalizado.
El QA está orientado al proceso		QC está orientado al producto
QA asegura que todos los desarrolladores siguen el mismos estándar de calidad, dentro de una gran corporación		QC asegura que el funcionamiento del producto, es el esperado
QA se diseña y ejecuta antes de tener un producto finalizado		QC se ejecuta durante la puesta en pre-producción
Se ejecutan las pruebas funcionales, unitarias, de integración y las pruebas de regresión, en otras palabras, se efectúa todo lo que es el "WHITE BOX TESTING" (pruebas a nivel de código fuente)		utan las pruebas funcionales, pruebas ¡a, de estrés, de rendimiento y de dad, etc., en otras palabras, se efectúa que es el "BLACK BOX TESTING" as a nivel de servicio)

QC, únicamente necesitan saber cómo funciona o cómo debería funcionar el producto, y estar orientado hacia el cliente.

- **Base de Datos:** Conjunto organizado de datos personales que sea objeto de Tratamiento.
- **Dato personal:** Cuando hablamos de datos personales nos referimos a toda aquella información asociada a una persona y que permite su identificación. Por ejemplo, su documento de identidad, el lugar de nacimiento, estado civil, edad, lugar de residencia, trayectoria académica, laboral, o profesional. Existe también información más sensible como su estado de salud, sus características físicas, ideología política, vida sexual, entre otros aspectos.
- **Integridad de Datos**

  La integridad de un dato alude a ese atributo o cualidad que es inherente a la

información cuando se considera exacta, completa, homogénea, sólida y coherente con la intención de los creadores de los datos que la conforman.

Esta cualidad, que va ligada al propio dato y no al lugar donde se almacena, al contrario de lo que sostienen creencias bastante extendidas; se obtiene cuando se impide eficazmente que el contenido de una base de datos, de un proceso o de un sistema se vea, accidental o intencionalmente:

Modificado, en base a su propio contenido o con ayuda de la inserción de nuevo.

Destruido total o parcialmente.

5. PLAN DE CALIDAD DE LOS COMPONENTES DE INFORMACIÓN

En primera instancia estableceremos una serie de conceptos técnicos importantes complementarios a los que se han definido en la sección anterior, y que se deben tener en cuenta en el presente plan de calidad de componentes de información a saber:

5.1. PALABRAS CLAVES

5.1.1. Big Data

Big Data se refiere a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

Aunque el tamaño utilizado para determinar si un conjunto de datos determinado se considera Big Data no está firmemente definido y sigue cambiando con el tiempo, la mayoría de los analistas y profesionales actualmente se refieren a conjuntos de datos que van desde 30-50 Terabytes a varios Petabytes.

La naturaleza compleja del Big Data se debe principalmente a la naturaleza no estructurada de gran parte de los datos generados por las tecnologías modernas, como los webs logs, la identificación por radiofrecuencia (RFID), los sensores incorporados en dispositivos, la maquinaria, los vehículos, las búsquedas en Internet, las redes sociales como Facebook, computadoras portátiles, teléfonos inteligentes y otros teléfonos móviles, dispositivos GPS y registros de centros de llamadas.

En la mayoría de los casos, con el fin de utilizar eficazmente el Big Data, debe combinarse con datos estructurados (normalmente de una base de datos relacional) de una aplicación comercial más convencional, como un ERP (Enterprise Resource Planning) o un CRM (Customer Relationship Management).

Las especiales características del Big Data hacen que su calidad de datos se enfrente a múltiples desafíos. Se trata de las conocidas como 5 Vs: Volumen, Velocidad, Variedad, Veracidad y Valor, que definen la problemática del Big Data. Estas 5 características del big data provocan que las empresas tengan problemas para extraer datos reales y de alta

calidad, de conjuntos de datos tan masivos, cambiantes y complicados.

Hasta la llegada del Big Data, mediante ETL podíamos cargar la información estructurada que teníamos almacenada en nuestro sistema ERP y CRM, por ejemplo. Pero ahora, podemos cargar información adicional que ya no se encuentra dentro de los dominios de la empresa: comentarios o likes en redes sociales, resultados de campañas de marketing, datos estadísticos de terceros, etc. Todos estos datos nos ofrecen información que nos ayuda a saber si nuestros productos o servicios están funcionando bien o por el contrario están teniendo problemas.

5.1.2. BODEGA DE DATOS

Un DataWareHouse es una colección de datos orientados a temas, integrados, no- volátiles y variante en el tiempo, organizado para soportar necesidades empresariales.

Orientado a temas: Los datos de esta base de datos permiten interpretar las diferentes entidades de la organización, como clientes, proveedores, productos, ventas, entre otros.

No volátil: La información perdura con el tiempo por el concepto de ser un repositorio histórico de información, que no se elimina ni se actualiza.

Variante en el tiempo: Según las operaciones de los sistemas transaccionales, una bodega de datos insertará todas las operaciones nuevas de la organización que esta considere, debe perdurar en el tiempo.

Una Bodega de Datos provee dos beneficios en una institución reales: Integración y acceso a datos. Con esto se identifica que con un sistema de Bodegas de Datos se puede integrar la información de los diferentes sistemas que la institución tenga para garantizar tener una única versión de los datos; esto de la mano de procesos de extracción, transformación, limpieza, calidad y carga de datos.

5.1.3. LIMPIEZA DE DATOS

Proceso que consiste en corregir datos que se encuentren mal digitados o que presenten problemas innecesarios de duplicidad dentro de una estructura transaccional de la información.

Existen cinco etapas que se deben tener en cuenta en una etapa de limpieza de datos:

- **Separación**: Toma cada palabra de un dato a estandarizar, y se estudia por separado. Por ejemplo, el dato “Auto Sur Av 42 – 31” se segmenta por cada palabra separadamente.
- **Estandarización**: Cambia palabras tales como Auto por Autopista, Av por Avenida, para garantizar que la información que se almacene sea estándar para toda la Base de Datos.
- **Verificación**: Garantiza que la información queda plenamente corregida bajo los estándares de calidad de datos definidos.
- **Búsqueda**: Búsqueda de homónimos en los registros existentes.
- **Agrupamiento**: Buscar información que cumplen criterios similares y que se puedan agrupar, por ejemplo, productos que pertenezcan al mismo proveedor.

5.1.4. EVALUACIÓN DE LA CALIDAD DE LA INFORMACIÓN

En un mundo en que la tecnología avanza a pasos agigantados, y en que la globalización y la sistematización continua hace parte de la vida humana, los datos se han convertido en el mayor insumo que cualquiera organización como el IDEAM puede tener, en contextos de administración, desarrollo y en procesos de toma de decisiones; siendo estos últimos, un factor importante para que las organizaciones cumplan sus objetivos eficaz y eficientemente. Por tanto, es inaceptable que una organización tenga datos con una pobre calidad, o con problemas de consistencia en la información que estén manejando. Lo importante es determinar una metodología que garantice estos estándares en los datos, y lograr un correcto funcionamiento de los sistemas en cuanto al manejo de la información que se dé.

De esta manera, se convierte en un factor importante hablar sobre la calidad de datos, y poner en práctica metodologías que permitan garantizar este proceso.

5.2. PLAN PARA LA CALIDAD DE LOS COMPONENTES DE INFORMACION

5.2.1. GESTIÓN DE CALIDAD DE LOS DATOS (DQM)

En el contexto de los proyectos y su gestión, los datos han tomado gran relevancia e importancia para la toma de decisiones, adaptar un enfoque basado en resultados, validar la Teoría de Cambio y evaluar el impacto de la intervención.

Desde el diseño de proyectos y las encuestas hasta la recolección (incluye estaciones convencionales, automáticas, radares, atención al ciudadano, información del Grupo de desarrollo y Talento Humano, etc), el manejo y análisis intervienen diversos factores que causan sesgos y errores en los datos, lo que crea la necesidad de un proceso en un marco de principios que asegure la calidad de estos. Dichos principios garantizan la integridad, claridad y transparencia de los indicadores, análisis y reportes, estos principios son:

- **Institucionalidad (Institutional):** cimentar una cultura que se enfoque en la calidad, haciendo hincapié en la objetividad y el profesionalismo.
- **Relevancia (Relevance):** apuntar a que los datos que se obtienen son los necesarios para los clientes de la información y que se hace mediante un uso eficiente de los recursos.
- **Oportunidad (Timeliness):** generar la información a través del tiempo de una forma oportuna, planeando la frecuencia de obtención de los datos y de actualización de los indicadores.
- **Precisión (Accuracy):** mantener el nivel de precisión y validez en la descripción del fenómeno o de la realidad.
- **Consistencia (Consistency):** asegurar que los datos e indicadores son comparables transversalmente y a través del tiempo. Se evita la duplicidad e inconsistencia.
- **Interpretabilidad (Interpretability):** Revelar como se interpreta las medidas estadísticas,

mantenerlas simples y entendibles para los usuarios de la información, adicionando toda información necesaria que ayude a entender las metodologías usadas y el nivel de confianza de los resultados.

- **Accesibilidad (Accessibility):** Hacer conocer la existencia de los datos y hacer accesible toda información que provea un mejor contexto a las partes interesadas.

Figura No. 1 - Marco de Principios de la Calidad de datos.
Fuente: LAC Documents

La Gestión de calidad de los componentes de información es un proceso transversal al diseño y la implementación de las herramientas que los colectan, como las encuestas, y a toda actividad que los manipulan o usan una vez se encuentran almacenados. El proceso busca cumplir los principios previamente mencionados por medio de buenas prácticas, actividades y herramientas, que pueden ir variando y actualizándose a medida que cambien los proyectos o mejoren las técnicas y tecnologías usadas.

La Gestión de calidad de los Componentes de Información se descompone en dos procesos complementarios:

- **Aseguramiento de Calidad (Quality Assurance)**: el objetivo es prevenir, reducir o limitar la aparición de errores desde el principio.

- **Control de Calidad (Quality Control)**: asegura que la calidad esté por encima de un nivel establecido, comparándolo con un estándar y tomando medidas correctivas si no se alcanza.

Por lo tanto, el Aseguramiento de Calidad anticipa problemas mientras que el control de calidad

responde a los problemas observados.

5.2.2. FUENTE DE LOS ERRORES

Ninguna medición puede ser ejecutada de manera absolutamente exacta. Inevitablemente el resultado de la medición contiene un error cuya magnitud es menor, mientras más exactos el método de una medición y el equipo de medición.

5.2.2.1. Fuentes de los errores según la metodología:

- **Error básico**: es el error en condiciones normales de empleo, es decir, el error básico ocurre mientras el objeto de estudio se encuentra en las condiciones esperadas.
- **Error adicional**: es el error por desviación de las condiciones de trabajo de los valores normales. Por ejemplo, una encuesta que es diseñada para un público objetivo y se le hace a alguien que está por fuera del perfil definido.

5.2.2.2. Fuentes de los errores según las personas implicadas en el proceso:

- **Imputables al investigador:**

- Mala definición del marco muestral y deficiente representatividad de la muestra por mal esquema de muestreo.
- Deficiente selección, formación y control de los entrevistadores.
- Deficiente diseño del cuestionario.

- **Imputables al entrevistador en levantamiento de los requerimientos:**

- Deficiencias en la formulación de las preguntas.
- Deficiencias en el control y en el registro de las preguntas.
- Mal seguimiento de las instrucciones.
- Fraude o falsificación de cuestionarios.
- Influencia del entrevistador sobre el entrevistado.

- **Imputables al entrevistado en levantamiento de los requerimientos:**

- Falta de repuesta por temor al no anonimato o debido a preguntas complejas.
- Deficiencias en las respuestas por falta de comprensión.
- Falta de sinceridad.
- Respuestas sesgadas o forzadas por complacencia con el sentido del cuestionario.

5.2.2.3. Fuentes de los errores según el origen:

- **Errores de tipo aleatorio**: son los errores que ocurre sin intención aparente, consecuencia de factores externos y que generalmente no son posibles de controlar por los diseñadores y encargados de las encuestas. Esto pueden ocurrir en cualquier momento del proceso estadístico de adquisición o tratamiento de los Componentes de Información, y las consecuencias pueden tener diferentes impactos.
- **Precisión**: es el grado de ausencia de error aleatorio. Es decir, cuanto mayor sea el tamaño muestral menor es el papel que juega el azar en nuestras estimaciones (el intervalo de confianza será menor y aumenta la precisión).
- **Errores sistemáticos**: son los errores que son consecuencia de algún factor interno y por lo general imputables a los involucrados en el proceso de diseño, captura y análisis de los Componentes de Información. Este tipo de error es más común que el aleatorio y puede tener un mayor impacto.
- **Validez**: un estudio tendrá validez si realmente mide lo que pretendemos medir libre de sesgos o errores sistemáticos.

5.2.2.4. Fuentes de los errores según la magnitud:

- **Error absoluto**: Es la cuantificación del error en términos de los Componentes de Información obtenidos o en términos de la misma respuesta, por ejemplo, número de datos en blanco o conteo de edades atípicas.
- **Error relativo**: Es el error que cuantifica la magnitud del error absoluto en cuanto al total de la muestra o el valor real.

5.2.2.5. Fuente del error desde el punto de vista del muestreo:

- **No muestrales**: Cuando las características de interés son medida de forma incorrecta.
- **Muestrales**: Cuándo ocurren diferencias entre la medida muestral y la poblacional.

5.2.2.6. Fuentes de los errores según el tipo de error cometido:

- Comisión: registro de Componentes de Información como por ejemplo los datos incorrectos o imprecisos.
- Omisión: no registro de los Componentes de Información como datos o metadata.

5.2.2. PROCESO DE CALIDAD DE LOS COMPONENTES DE INFORMACION

El proceso de calidad de los componentes de información (datos, información, servicios de información y flujos de información), es necesario para mantener los Componentes de Información, indicadores y reportes, relevantes, precisos y consistentes, dado que en el proceso de recolección y almacenamiento se puede generar Componentes de Información defectuosos que impacten la legitimidad de los análisis y los resultados y por ende la información final. Con Componentes de Información como por ejemplo datos limpios se pueden tomar mejores decisiones y asegurar que los resultados obtenidos se asemejan a la realidad del objeto de estudio.

Entre los objetivos del control de la calidad de los componentes de información se tiene:

- Entender y documentar la calidad y confiabilidad de los componentes de información.
- Descubrir en los componentes de información los problemas de calidad que deben ser resueltos.
- Asegurar la estandarización e integración de los componentes de información comunes en las diferentes operaciones estadísticas.
- Especificar las reglas de transformación y validación que deben aplicarse a los componentes de información, para asegurar el nivel de calidad que se requiere en una migración hacia el repositorio de información.

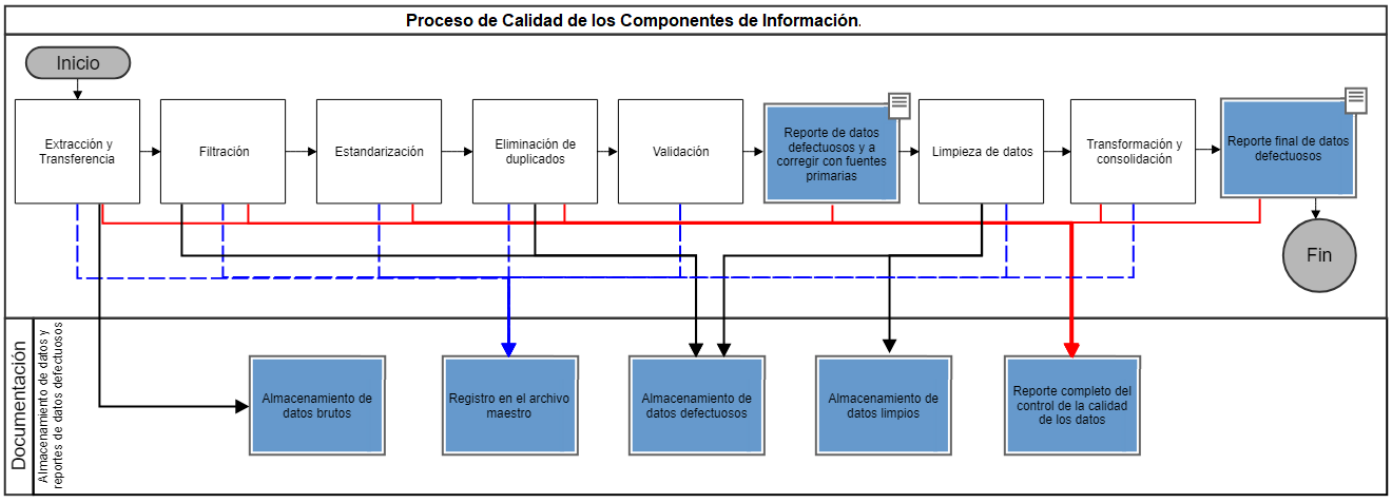

Figura No. 2 - Marco de Principios de la Calidad de Componentes de Información.
Fuente: LAC Documents

A continuación, se especifican cada uno de los pasos del Plan de Calidad de componentes de información:

5.2.2.1. PASOS DEL PLAN DE CALIDAD PARA LOS COMPONENTES DE INFORMACIÓN.

5.2.2.1.1. Extracción y Transferencia

Este proceso consiste en extraer los Componentes de Información tales como datos de los servidores receptores de los datos transmitidos por los radares, las estaciones hidrometeorológicas, ambientales y la nube (cuando el IDEAM implemente esta tecnología TI) o software donde han llegado después de su recolección, agregar unas variables relevantes y almacenarlo en un repositorio de bases de Componentes de Información como datos. Cada vez que los Componentes de Información como por ejemplo datos son almacenados en la estructura de datos, el software implementado para la gestión y procesamiento con seguridad asigna unos códigos de estaciones, sensores, etc., y actualiza unas variables importantes para perfilar los Componentes de Información (datos, información, etc), se generan Componentes de Información para las variables como fecha y hora de la recepción de Componentes de Información y de las demás variables hidrometeorológicas

y ambientales definidas por las áreas misionales del IDEAM, caso similar sucede con los Componentes de Información provenientes de las áreas de apoyo y estratégico.

Este proceso tiene como propósito:

- Tener una copia de seguridad (backup) de los Componentes de Información.
- Registrar la extracción.
- Agregar la información necesaria para que sea fácil de identificar y verificar el proceso en cualquier momento que sea necesario.
- En esta parte del proceso se excluyen las pruebas (test) para probar los Componentes de Información.
- Es muy importante almacenar estos Componentes de Información crudos como vienen, en bruto, sin ningún tipo de modificación de tal manera que se pueda diferenciar entre el archivo que se quiere mantener en crudo o bruto y aquel que va a ser objeto de limpieza o que lo ha sido.
- incluir un backup de la metadata para asegurar la replicabilidad de la estructura de los Componentes de Información.

Documentar:

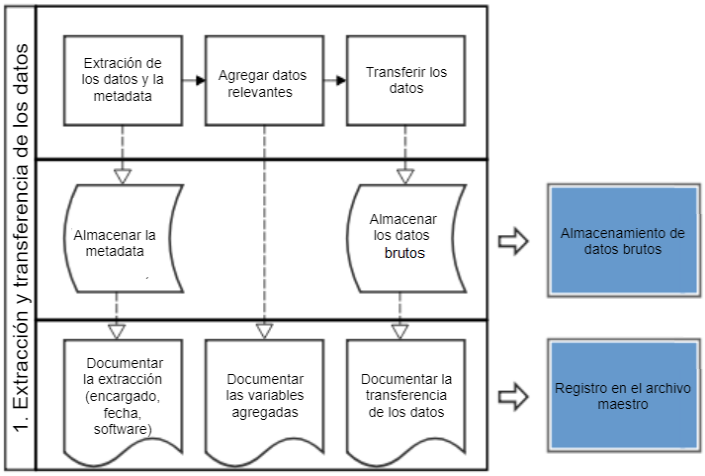

Figura No. 3 – Extracción y transferencia de los Componentes de Información.
Fuente: LAC Documents

5.2.2.1.1. 1. Extracción de los Componentes de Información y la metadata

Los Componentes de Información se deben extraer para la agregación periódica de tiempo (diaria, semanal, mensual, otros que se disponga) en el backup en lugar de hacer un reemplazo completo del backup, porque los Componentes de Información pueden sufrir modificación en las plataformas de donde vienen o se procesan, por ejemplo:

- Un objeto que se actualice cada vez que llegan los Componentes de Información, solo va a mostrar los últimos que han llegado, lo que igual puede pasar durante el mismo día o la

misma hora, pero siendo esto más fácil de identificar y de hacer el debido proceso en la determinada extracción de la hora, día, mes en que ocurrió.

- Los Componentes de Información defectuosos se corrigen y los registros duplicados se eliminan (opcionalmente, aunque no es obligatorio, almacenar los registros duplicados eliminados en una tabla de datos por aparte) para no afectar la interpretación, precisión y relevancia de los reportes generados.

- Durante el ciclo de vida del proyecto se pueden hacer muchas modificaciones en el sistema que pueden afectar los Componentes de Información, como las migraciones o modificación al modelo de objetos.

- Debido a que, terminado el periodo de tener el total de los Componentes de Información, no es posible tener la totalidad de estos porque puede haber errores de transmisión en las estaciones o dispositivos encargados de transmitir los Componentes de Información hidrometeorológicas y ambientales u otro tipo de siniestro que impida recolectar el dato, es prudente estar verificando esto.

A continuación, se sugiere una manera de realizar la limpieza de Componentes de Información, quizá esta no sea idéntica al contexto real de cómo se tratan los Componentes de Información en el IDEAM, pero pretender brindar una idea de una estrategia de cómo se podría llegar a implementar un proceso de limpieza similar:

Una organización que se encarga de realizar encuestas, ha culminado el mes de trabajo de encuestas, pero no es posible tener la totalidad de los Componentes de Información porque existen encuestas marcadas como completas, pero no sincronizadas, para ello verifican esto continuamente, por lo que se recomienda hacer el backup provisional de los dos meses inmediatamente anteriores y consolidar el mes anterior a los dos provisionales (backup -2 meses).

La organización decide hacer el backup provisional de los dos meses inmediatamente anteriores y consolidar el mes anterior a los dos provisionales (backup -2 meses). Suponga que se recientemente se cumplieron 3 meses de un proceso de encuestas, y se planea hacer un backup. Por ser el mes 3 se hace una revisión del mes 1, comparando la coherencia con el backup provisional que se hizo en el mes 2 (del mes 1), y si se encuentran nuevos registros se agregan al almacenamiento de Componentes de Información brutos y siguen el proceso de limpieza de Componentes de Información, para finalmente declararse como consolidados.

A los Componentes de Información provisionales del mes 2 se les hace el mismo proceso anterior, con la diferencia de que no se declaran consolidados, sino que siguen siendo provisionales; y los del mes 3 se declaran inmediatamente provisionales (porque no hay con que hacerles comparación).

Una vez cumplidos los 4 meses, el mes 1 se encuentra consolidado por lo que no hay que hacerle revisión; a los Componentes de Información del mes 2 se le hace comparación con el backup obtenido en el mes 4 y se consolida; al mes 3 se le hace una comparación que continua en provisional (hasta el próximo mes); y a los nuevos Componentes de Información obtenidos en el mes 4 se almacenan para compararlos con los backups que se harán en el mes 5 y 6 para declararlos como consolidados.

Esto quiere decir que se espera que el proceso de limpieza de Componentes de Información (de determinado mes) dure entre 1 y 3 meses, los beneficios de trabajar de esta forma son:

- Hay una verificación exhaustiva de los registros que son sincronizados tiempo después de finalizado el mes.
- Asegura que todos los registros pasen por el proceso de limpieza de Componentes de Información.
- Se hace backup de los registros en bruto reduciendo la posibilidad de que sean modificados.
- Hay un seguimiento en el tiempo de los registros a los que se le han hecho limpieza de Componentes de Información y a los que no.
- La metadata permite entender, evaluar y replicar la estructura de los Componentes de Información y los Componentes de Información mismos por lo que es importante hacer un almacenamiento de forma agregada, es decir que en este si reemplazaría por completo el archivo anterior, registrando los cambios incrementales que se le hace, por ejemplo, la adición de nuevos campos. En el caso en el que la metadata sufre cambios drásticos que no permite que sea comparable o que no encajen se debe generar un nuevo backup para los Componentes de Información y para la metadata.
- El registro del archivo maestro se conformará de las siguientes variables como se describe a continuación:

  Registro en el archivo maestro

  - Proyecto
  - Herramienta
  - Encuesta
  - Versión de la encuesta
  - Responsable de la extracción
  - Fecha y hora de la extracción
  - Salida; forma en la que se extrajo (por reportes, un AppExchange en caso de servicio en la nube u otros)
  - Nombre del archivo de la metadata
  - Cambios en la metadata

Finalmente, muchas soluciones que se dan en los proyectos y para el almacenamiento de Componentes de Información, incluye el uso de diferentes softwares, herramientas o lenguajes de programación, por lo que se hace necesario relacionar toda la documentación del uso de las herramientas tecnológicas, por ejemplo, el uso de Scripts, triggers y workflows deben ser especificados en la documentación de la extracción de los Componentes de Información.

5.2.2.1.1. 1. 2 Agregar Componentes de Información relevantes

En el pecoso de extracción y transferencia se debe almacenar todos los Componentes de Información disponibles, sin importar si son defectuosos o no, estos Componentes de Información incluyen los generados por las estaciones y radares al momento de trasmitirlos, el software utilizado para recepcionarlos y almacenarlos.

Registro en el archivo maestro sugerido

- Entrada.
- Nombre del reporte o archivo (. miss, otros).
- Periodo de inicio de recolección de los Componentes de Información.

- Variables agregadas en el proceso de extracción, en caso de que se creen.
- Documentación del software y el script que se usó para agregar las variables en caso de que se dé.
- Nombre del archivo bruto de los Componentes de Información.
- Ubicación del archivo bruto.
-

5.2.2.1.1. 1. 3 Exportación de la data (Data Export).

Almacenamiento de Componentes de Información brutos

El almacenamiento de los Componentes de Información crudos o en bruto, es el repositorio único de Componentes de Información el cual se está implementando en base de Componentes de Información no relacional Cassandra, que contiene todos los Componentes de Información generados en las estaciones automáticas, convencionales y radares sin haber pasado por algún tipo de manipulación, estos Componentes de Información también sirven como respaldo (backup) en caso de pérdida de la información durante el proceso de limpieza de Componentes de Información porque las propiedades del proyecto pueden cambiar o se puede cometer errores.

5.2.2.1.2. Filtración

El proceso de filtración consiste en eliminar aquellos Componentes de Información que no deben ser incluidos porque violan ciertas condiciones. Está filtración obedece a propiedades del proyecto más que a los Componentes de Información mismos, por ejemplo, si no nos interesan algunos registros, o si se establece que debe haber un tiempo mínimo de diferencia entre cada registro de Componentes de Información transmitido y hay algunos que violan está condición, deben ser excluidos del archivo que sigue en el proceso de limpieza.

Tenga en cuenta para el proceso de filtración las siguientes recomendaciones:

- Los Componentes de Información que son excluidos no se deben eliminar, sino que se deben almacenar como Componentes de Información defectuosos para que puedan ser objeto de validación.
- Tener en cuenta la exclusión de otros registros lógicamente relacionados (pueden estar en otros archivos).
- En esta parte se incluye aquellos registros de Componentes de Información que se consideren como no válidos y se deben excluir.

Documentar:

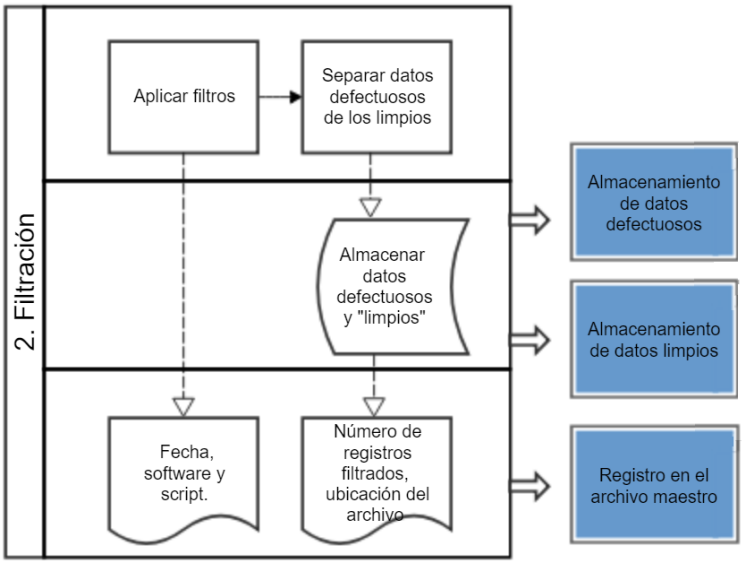

Figura No. 4 – Proceso de Filtración.
Fuente: LAC Documents

El registro en el archivo maestro es el que se sugiere:

Registro en el archivo maestro

- Fecha de filtración
- Software usado y script
- Reglas de filtración
- Número de registros totales
- Número de registros filtrados
- Nombre del archivo de Componentes de Información defectuosos
- Ubicación de los Componentes de Información defectuosos

Almacenamiento de Componentes de Información defectuosos

La idea general es separar el archivo de Componentes de Información crudos o brutos en dos archivos diferentes, uno con todos los registros excluidos porque no cumplen con algún criterio del proyecto y los duplicados (Ver proceso 4. Eliminar Duplicados); y otro con todos los Componentes de Información limpios, que es usado como base para los análisis y actualizar los indicadores. Es importante distinguir en el archivo de Componentes de Información defectuosos el criterio de exclusión que se aplicó a cada uno de los registros. Este documento se almacena solamente en una ruta específica para ello.

Almacenamiento de Componentes de Información limpios

El almacenamiento de los registros que no violan criterios propios del proyecto, se hace en el repositorio de Componentes de Información limpios para que sigan en el proceso de limpieza, y sigan el ciclo de vida de los Componentes de Información. El almacenamiento de los Componentes de

Información limpios se hace en un repositorio destinado para ello, esto ante la posibilidad de tener registro que se encuentren el proceso de validación y que no hayan sido limpiados completamente.

5.2.2.1.3. Estandarización

La estandarización consiste en la codificación de las variables (hidrometeorológicas y ambientales) y opcionalmente sus resultados o Componentes de Información almacenados en ellas, no solo para un mejor tratamiento, sino también para asegurarse de que tengan el mismo formato, por ejemplo, fecha en MM/DD/AAAA.

Algunos beneficios de la codificación estandarizada es el mejoramiento de la comunicación entre los equipos interdisciplinarios de análisis de los Componentes de Información, como son las áreas temáticas y en su momento las áreas de apoyo y estratégicas, entre los equipos de desarrollo de software, lo que permite reducir los errores en los Componentes de Información al momento de intercambiarlos, de procesarlos, de usarlos para la toma de decisiones, para la programación y mejora la calidad del software. Lo anterior repercute en la competitividad de las instituciones (Wang, Wang, Li, Li, & Du, 2010) y en la productividad de sus trabajadores porque se mejora la comunicación, y la decisión en los análisis de resultados y Componentes de Información, teniendo esto impacto en positivo en la competitividad, eficiencia y productividad de las entidades. Lo anterior se toma de las recomendaciones que hacen Deitel & Deitel (2004) y Humphrey (2009) para el establecimiento de las reglas de calidad para la codificación estandarizada.

Entre las recomendaciones que da el autor mencionado para la estandarización de las variables se tiene:

- Otorgue a los identificadores para valores variables un nombre significativo (Deitel & Deitel, 2004, pág. 29), es decir, que con el puro nombre se pueda o ayude a deducir a qué se refiere o cuál es su funcionalidad. Evite abreviaciones o variables de una sola letra (Humphrey, 2009, p. 51).

- La primera letra de la palabra debe ser una letra en mayúscula, las demás en minúscula. Cuando dos palabras describan mejor a una variable, la primera letra de la segunda palabra deberá ser mayúscula y el resto en minúscula (Deitel & Deitel, 2004, pág. 29).

- Si el nombre de la variable requiere un número, escríbalo contiguo a las letras. Dele preferencia a usar el número al final del nombre.

- la que se han extraído las recomendaciones para los errores comunes de programación, buenas prácticas de programación, tips para prevenir errores, tips de rendimiento, tips de portabilidad y observaciones de ingeniería de software.

- Un segundo objetivo es concientizar sobre la importancia de contar con documentos que fomenten la codificación estandarizada desde que el recurso humano está en formación.

Lo anterior significa que se pueda hacer agrupación de resultados y cambios de formatos.

Documentar:

Una vez hecha la estandarización sobre los archivos limpios, se guardan sobre escribiéndose y se

registran los cambios en el archivo maestro.

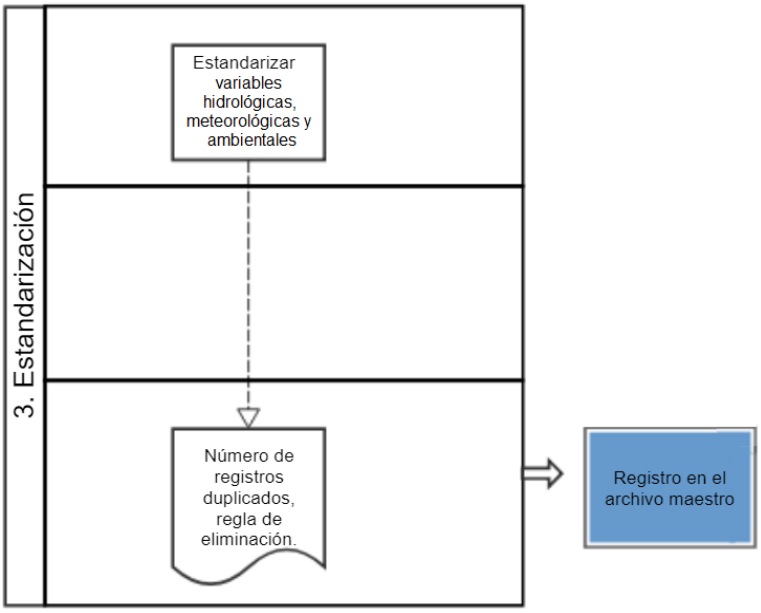

Figura No. 5 – Proceso de Estandarización.
Fuente: LAC Documents

Registro en el archivo maestro

- Variables codificadas, número de variables codificadas.
- Reglas de estandarización.

5.2.2.1.4. Eliminar duplicados

El proceso de limpieza de Componentes de Información inicia eliminando los registros duplicados, es necesario excluirlos para que no afecten la relevancia de los reportes.

Seguir las siguientes recomendaciones:

- Estos Componentes de Información se deben almacenar en el archivo de Componentes de Información defectuosos para que puedan ser fuente de validación.

- Las reglas de eliminar los registros duplicados pueden cambiar de acuerdo al proyecto, pero se debe adoptar un proceso que estandarice la forma en que se eliminen los duplicados.

Documentar:

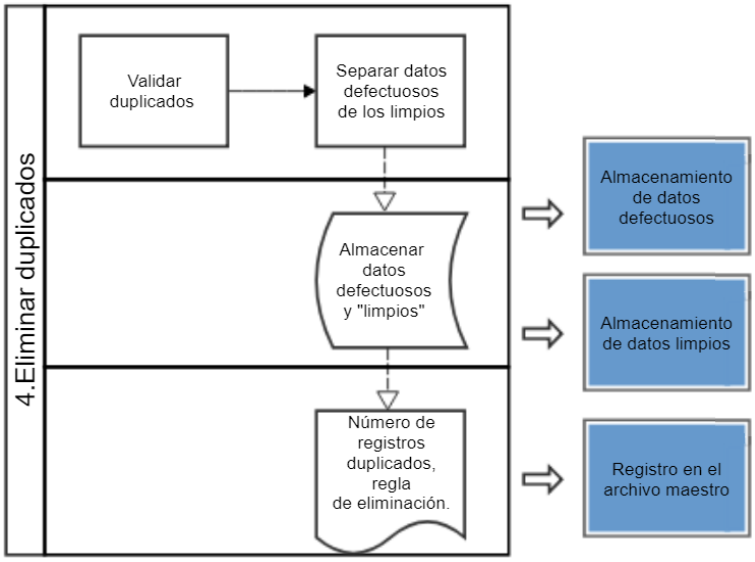

Figura No. 6 – Proceso Eliminación de duplicados.
Fuente: LAC Documents

Reglas de decisión para eliminar duplicados

La forma en que se generan los Componentes de Información en el proyecto determina las reglas que se pueden adoptar para eliminar los duplicados, lo importante es mantener consistentes las reglas para no perder la relevancia de los registros que siguen en el proceso de limpieza. A continuación, se presenta un proceso sugerido para eliminar duplicados:

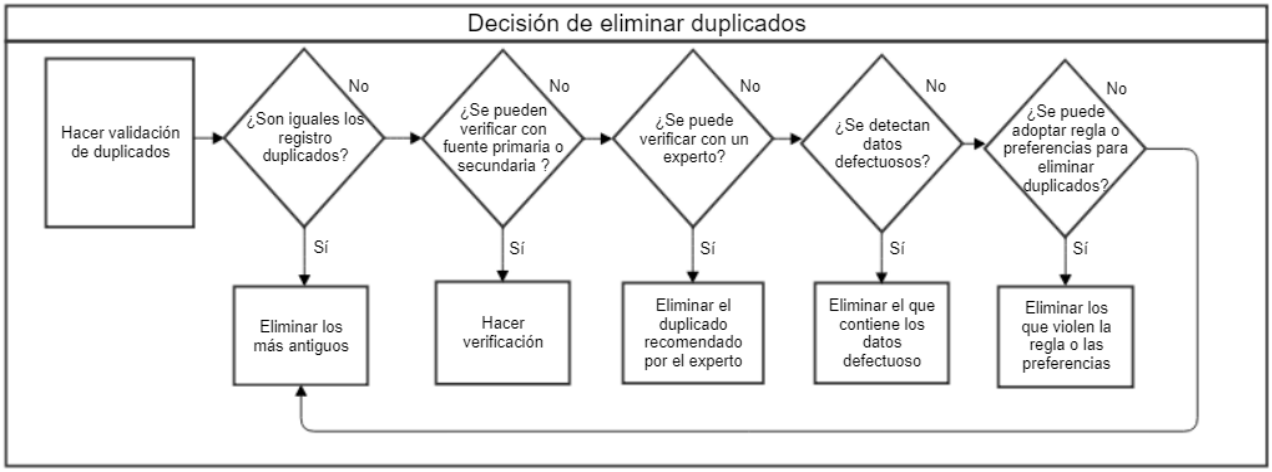

Figura No. 7 – Proceso Sugerido Eliminación de duplicados.
Fuente: LAC Documents

Registro en el archivo maestro

- Registros duplicados, Número de registros duplicados.
- Regla de validación de los duplicados.
-

Almacenamiento de Componentes de Información defectuosos

Los Componentes de Información eliminados deben almacenarse en el archivo de Componentes de Información defectuosos con su respectiva descripción del tipo de regla de decisión que se tomó para eliminarla como duplicado.

Almacenamiento de Componentes de Información limpios

Al archivo que se le hizo la filtración y la estandarización, ahora se le eliminaron los registros duplicados, y sigue en el proceso de limpieza.

5.2.2.1.5. Validación

La validación de los Componentes de Información es el proceso que se encarga de encontrar los Componentes de Información defectuosos, y aquellos Componentes de Información que no cumplan con las reglas y/o criterios de validación serán los que pasaran al proceso de limpieza.

Antes de pasar a la validación de los Componentes de Información primero hay que definir las fuentes de verificación y el impacto de las variables, para tener claro que se va hacer con los Componentes de Información defectuosos.

Fuente de verificación de los Componentes de Información

La fuente de verificación de los Componentes de Información es aquella de donde vamos a poder obtener el valor correcto en caso de los errores y validar el valor en caso de los valores atípicos (outliers). existen 3 fuentes de verificación.

- **Fuente primaria**: Es el origen mismo de los Componentes de Información, es decir las estaciones y demás dispositivos que generan y trasmiten los Componentes de Información hidrometeorológicas y ambientales.

- **Fuente sustituta**: Son otras bases de Componentes de Información en las que se puede verificar los resultados, pueden ser rangos de valores internacionales o tendencias de valores en el tiempo a nivel nacional.

- **Fuentes alternas**: Cuando no se tiene ni la fuente primaria ni la fuente sustituta se puede recurrir a técnicas estadísticas, preguntas a expertos, a Componentes de Información de estaciones cercanas a la fuente primaria u otras metodologías de validación que permitan corregir los Componentes de Información.

Impacto de las variables

Las variables tienen un impacto diferente de acuerdo a su tipo, importancia y frecuencia, es decir un **outlier** en un dato meteorológico tiene un gran impacto en los procesos de análisis y tomas de

decisiones. Esta parte se debe definir desde el diseño del proyecto, para que desde el aseguramiento de Calidad (QA) se prevenga errores en las variables más importantes. El Impacto de las variables se clasifican de acuerdo a su importancia, frecuencia, riesgo de errores, sensibilidad del dato y tipo de variable:

Impacto alto: Se presenta en aquellas variables en las que se enfoca el proyecto y que finalmente se usarán para medir la efectividad y los resultados, también son aquellas que contienen una mayor probabilidad de presentar errores, ya sea por el tipo de variable o el tipo de pregunta.

Por ejemplo:

Si el aumento de la productividad va a ser el principal indicador, se debe elegir esta variable como de alto impacto.

Impacto medio: Se presenta en variables importantes pero que, por el tipo de Componentes de Información que se estudia, no implican un riesgo alto para el nivel de confianza del análisis, es decir que no tiene un efecto grande si se presenta un error en los Componentes de Información.

Impacto bajo: Se presenta en variables que no son tan relevantes para el proyecto, tienen una baja probabilidad de presentar errores o que por su tipo no es posible detectar el error, es decir no implica un gran riesgo para la integridad del análisis.

El propósito de clasificar las variables es definir una fuente de verificación y un tratamiento diferente a cada uno de ellas:

Impacto alto: Se verifican con la fuente primaria y al 100% de los Componentes de Información erróneos.

Impacto medio: Se verifican con la fuente primaria de acuerdo el tipo de variable y no se es tan riguroso (<100%), también se verifica con las fuentes sustitutas.

Impacto bajo: Se valida con las fuentes alternas y pasan a un tratamiento estadístico u otra metodología de validación.

Reglas de validación

Finalmente, después de definir las fuentes y el impacto se pasa a definir las reglas de validación, este es un proceso flexible porque a medida que se hace la validación pueden surgir otras reglas, pero lo importante es hacer una precisa documentación de las reglas utilizadas.

Las principales reglas de validación a utilizar son los rangos y límites, porque permite detectar fácilmente los Componentes de Información defectuosos, estos se determinarán en el momento en que se diseñe el proyecto, con la ayuda de expertos: áreas temáticas o misionales, áreas de apoyo y estratégicas.

Otras reglas que se pueden utilizar, que en muchos casos son necesarias para evitar tener esa clase de errores (QA), son:

- Reglas de Identidad
- Reglas de Integridad Referencial
- Reglas de Cardinalidad
- Reglas de Herencia
- Reglas de Relaciones Dependientes
- Dependientes del estado de la Entidad
- Mutuamente Dependientes
- Mutuamente Excluyentes
- Relaciones Recursivas
- Reglas de Dominio
- Reglas de Atributos Dependientes
- Derivadas
- Restringidas
- Por valor
- Por Relación

Validación

Luego de elegir las reglas de validación, pasamos a hacer el análisis de las variables de acuerdo a su impacto para detectar alguno de los siguientes problemas:

- Inconsistentes por definición.
- Faltantes, es decir, no existe el dato o es nulo o blanco, cuando debiera existir.
- Inválidos, cuando no cumplen alguna regla de validación; en este caso, se debe especificar la regla violada.
- Exactitud con la fuente sustituta
- Exactitud con la fuente primaria
- Incorrectos, cuando no concuerdan con la realidad.
- Discrepancia con Componentes de Información redundantes.

Documentar:

- En el proceso de validación solo se documenta en el archivo maestro dado que se preparan los Componentes de Información para ser reportados para su verificación y limpieza.

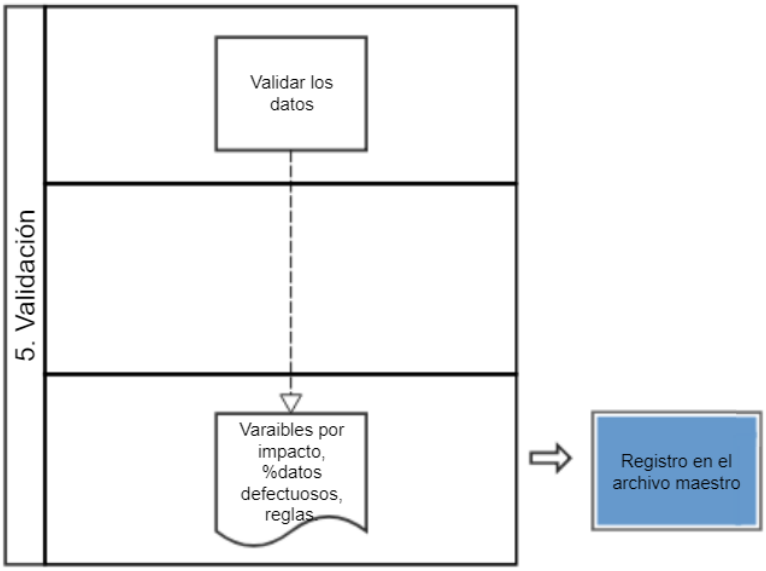

Figura No. 8 – Proceso de validación.
Fuente: LAC Documents

Registro en el archivo maestro

- Fuente de verificación de los Componentes de Información
- Número de variables por impacto
- Reglas de validación utilizadas
- Software y script usado en la validación
- Porcentaje de Componentes de Información defectuosos
- Porcentaje de Componentes de Información defectuosos por impacto

5.2.2.1.6. Reporte de calidad de los Componentes de Información, datos defectuosos y a corregir con fuentes primarias

Una vez hecha la validación se tienen los datos disponibles y se deben crear dos tipos de reportes:

- **Reporte de análisis de la calidad de los Componentes de Información**. El objetivo de este reporte es informar a los interesados (stakeholders) sobre la calidad de los Componentes de Información que se están obteniendo, mostrando reportes de medidas centrales, gráficas, cuadros, tipos de errores, variables con más errores, y finalmente el nivel de confianza de los Componentes de Información, para que se tomen decisiones que puedan ayudar a mejorar la calidad de los Componentes de Información.

- **Reporte de Componentes de Información defectuosos y a corregir con fuentes primarias**. El objetivo de este reporte es presentar de una forma comprensible la lista de Componentes de Información defectuosos para que sean corregidas en campo.

Este último reporte implica tres desafíos, los cuales se exponen como sugerencia en el proceso de

validación:

Desafío	Descripción del problema	Descripción de la solución
1. Elaboración eficiente y presentación oportuna	Debido a que el proceso de validación y la constante recolección de datos puede tomar mucho tiempo y hacer de este reporte una tarea ardua.	Ejecutar esta tarea mediante un software que contenga los rangos y límites, haga la validación y genere el reporte.
2. Formato adecuado	El reporte debe ser eficiente y facilitar la verificación en campo, por lo que el formato en que se presente puede llegar a influir en este proceso.	El formato del reporte por ejemplo debe ser agrupado por el recolector o tomador del dato en las estaciones automáticas o en estaciones ambientales, porestación y por variable. El reporte debe contener junto al dato por verificar: la variable, el código de la estación/radar/sensor y una breve descripción del problema.
3. Seguimiento a los datos verificados	Es indispensable utilizar una buena herramienta que permita hacerle seguimiento a los datos defectuosos que se han podido verificar en campo.	El seguimiento se debe hacer mediante un archivo (Excel o SmartSheet) replica del reporte y que permita aparte de agregar los nuevos casos que se van generando, distinguir la fecha en que se agregó a la lista por verificar en campo, la fecha en que se verificó y el valor correcto.

Figura No. 9 – Desafíos del reporte de Componentes de Información defectuosos.
Fuente: LAC Documents

5.2.2.1.7. Limpieza de Componentes de Información

Después de que los Componentes de Información defectuosos son detectados y verificados con la fuente primaria y/o sustitutas, se inicia la edición e imputación para que los Componentes de Información queden limpios, (usando el archivo de seguimiento de los Componentes de Información verificados). Sin embargo, no es posible verificar todos los Componentes de Información, siendo necesario usar fuentes alternas como la opinión de expertos o técnicas estadísticas para hacer la imputación.

- Hacer un proceso de imputación de Componentes de Información
- Técnicas de limpieza de Componentes de Información

Documentar:

Al final del proceso se debe reportar el cambio en la calidad de los Componentes de Información y las medidas tomadas debido al proceso de limpieza:

- A: Variable de impacto alto, no corregido
- B: Variable de impacto medio, no corregido
- C: Variable de impacto alto, corregido con fuente primaria
- D: Variable de impacto medio, corregido con fuente primaria
- T: Variable de impacto alto, corregido con fuente sustituta
- U: Variable de impacto medio, corregido con fuente sustituta
- X: Variable de impacto alto, corregido con fuente alterna

- Y: Variable de impacto medio, corregido con fuente alterna
- Z: Variable de impacto bajo, corregida con fuente alterna

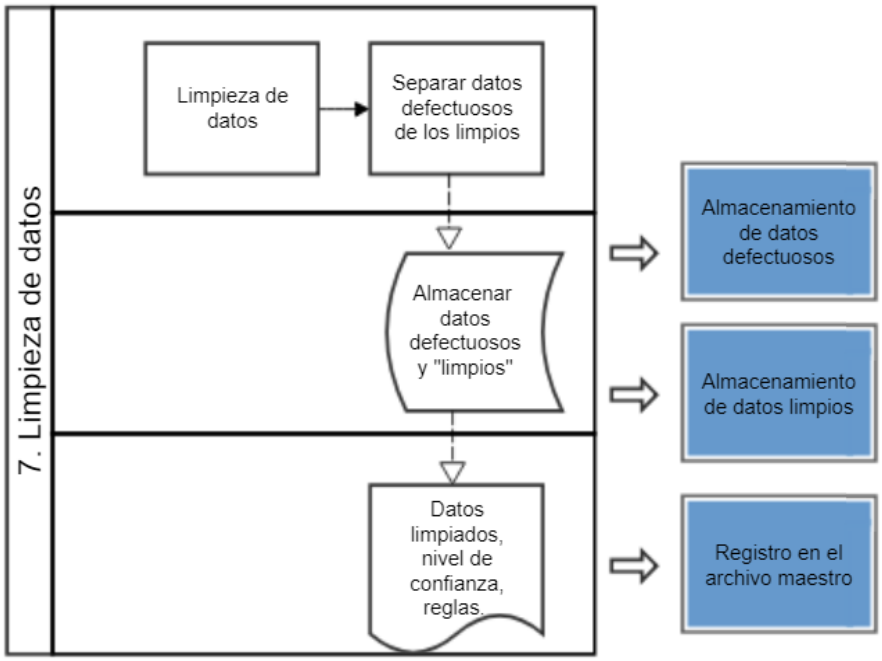

Figura No. 10 – Limpieza de Componentes de Información.
Fuente: LAC Documents

Registro en el archivo maestro

- % de datos corregidos
- % de datos corregidos por impacto
- % de datos inválidos o sospechosos
- % de datos inválidos o sospechosos por impacto
- Número de registros eliminados
- Número de medidas tomadas (A B C D T U X Y Z)
- Nuevo nivel de confianza de los datos

Almacenamiento de Componentes de Información defectuosos

Durante la validación y verificación de los Componentes de Información pueden surgir decisiones de excluirlos por la pobre calidad de los Componentes de Información, o porque se encuentran problemas de riesgo en la toma de decisiones y análisis definitivos.

Almacenamiento de Componentes de Información limpios

Al archivo que se le hizo la filtración, estandarización, y se le eliminaron los registros duplicados, ahora se le hizo limpieza de los Componentes de Información defectuosos, estando listo para el análisis de Componentes de Información.

5.2.2.1.8. Transformación y consolidación

Después de limpiar los Componentes de Información estos se preparan para su almacenamiento en el repositorio, por lo cual se pueden transformar y consolidar para agregar valor a las bases de Componentes de Información. La transformación consiste en hacer cambios a las variables, o derivar nuevas de las existentes. La consolidación consiste en unificar o agrupar estructuras de datos que contienen una relación, por ejemplo, varias estaciones que miden una misma variable hidrometeorológica o ambiental.

Reglas de transformación:

- **Copia simple de datos**: Los campos son copiados de un registro a otro, sin modificación, es decir, no hay transformación.
- **Conversión de dominio**: El dominio en el campo fuente es convertido al dominio en el campo de destino. La conversión de dominio se realiza para unificar el valor del campo, su código, la unidad de medida, su formato, el tipo de dato, su longitud, etc, entre aplicaciones.
- **Codificación o Clasificación de datos textuales**: Los datos textuales o en formato libre son analizados y se crea una clasificación de códigos para identificar categorías.
- **Filtro Vertical**: Cuando se descubre que un campo es usado con propósitos diferentes, se deben identificar cada uno de los usos y definirlos en el conjunto de valores respectivos del dominio. El atributo se abre en varios campos y se definen sus reglas de transformación.
- **Concatenación**: Dos o más valores de los datos atómicos son unidos en uno solo para dar significado a los campos. Los campos de nombres de estaciones y direcciones son ejemplos de este tipo de transformaciones.
- **Datos Derivados**
- **Datos Agregados**

Documentar:

Registro en el archivo maestro

Campos transformados

- Tipo de transformación
- ¿Hubo consolidación?
- Tipo de consolidación
- Fuente de los Componentes de Información

5.2.2.1.9. Reporte final de Componentes de Información defectuosos

Documentación

Al final de todo el proceso se espera obtener la siguiente documentación:

- Almacenamiento de Componentes de Información crudos o brutos

Los Componentes de Información en crudo que se obtienen de las estaciones y que no han sido modificados (no han sido objeto de validación, eliminación de duplicados o limpieza de Componentes de Información), el objetivo es mantener los Componentes de Información originales, dado que en ocasiones las reglas de validación pueden cambiar, o se pueden cometer errores en la eliminación de duplicados o en la limpieza de Componentes de Información.

- Registro en el archivo maestro

  El archivo maestro donde se hace el registro del proceso de limpieza de Componentes de Información, es decir que se puede encontrar el historial del control de la calidad que se ha hecho, además de las principales variables y medidas tomadas en cada uno de los sub procesos.

- Almacenamiento de Componentes de Información defectuosos

  Los Componentes de Información excluidos por filtración, eliminación de duplicados, validación o verificación en campo. El propósito de este archivo es poder acceder a los registros eliminados para validación de las partes interesadas. debe de incluir el motivo de exclusión de cada uno.

- Almacenamiento de Componentes de Información limpios

  Este archivo representa la versión más precisa y válida de los Componentes de Información que se puede obtener de acuerdo al uso eficiente de los recursos, quedando disponible para su uso oportuno y relevante.

- Reporte completo del control de la calidad de los Componentes de Información

  Este reporte busca ilustrar el perfil de los Componentes de Información y todo el proceso que han pasado para declares limpios y alcanzar el nivel de confianza actual.

6. BIBLIOGRAFIA

(1) La calidad de los datos: Su importancia para la gestión empresarial, Jobany José - Heredia Rico, junio 5 de 2009.

http://www.unilibrecali.edu.co/images2/revista-libre-empresa/pdf_articulos/volumen6/la_calidad_de_los_datos_su_importancia_para_la_gestion_empresarial_43_50.pdf

(2) Qué es un plan de QA, productora digital, software factory.

http://www.4rsoluciones.com/blog/que-es-un-plan-de-qa-2/

(3) Diferencias entre QA y QC, Novanotio.

http://www.novanotio.es/diferencias-entre-garantia-de-calidad-qa-y-control-de-calidad-qc/

(4) Normativa protección de datos personales, DNP.

https://colaboracion.dnp.gov.co/CDT/Programa%20Nacional%20del%20Servicio%20al%20Ciudadano/NORMATIVA%20PROTECCI%C3%93N%20DE%20DATOS%20PERSONALES.pdf

(5) Qué se entiende por integridad de datos, PowerData.

https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/348870/qu-se-entiende-por-integridad-de-los-datos

(6) Herramienta para el mantenimiento, integridad y aseguramiento de la calidad de datos, Luis Alberto Rivera Tavera, 2013.

http://polux.unipiloto.edu.co:8080/00000862.pdf

(7) Proceso del control de la calidad de los datos

https://applab.atlassian.net/wiki/spaces/LAC/pages/88342616/Proceso+del+control+de+la+calidad+de+los+datos+limpieza+de+datos

(8) Reglas de calidad para la codificación estandarizada, Dr. Edgar Danilo Domínguez Vera., dic 2015.

http://eprints.uanl.mx/9199/1/Reglas.pdf

7. HISTORIAL DE CAMBIOS

Versión	Fecha	Descripción
01	07/11/2018	Versión inicial.

ELABORÓ:	REVISÓ:	APROBÓ:
Eduardo Ramírez Acosta	Eduardo Ramírez Acosta	Leonardo Cárdenas Chitiva
Profesional Especializado	Profesional Especializado	Jefe oficina Informática
Líder Arquitectura Empresarial	Líder Arquitectura Empresarial	