MINISTERIO DE AGRICULTURA Y DESARROLLO RURAL

OFICINA DE TECNOLOGIA DE LA INFORMACIÓN Y LAS COMUNICACIONES

OTIC

METODOLOGÍA PARA MEDIR LA CALIDAD DE LOS COMPONENTES DE INFORMACIÓN

Bogotá D.C., Noviembre 2016

HISTORIAL DE VERSIONES

Versión	Fecha	Descripción	Autor	Cargo
1.0	22/08/2016	Versión inicial.	Jaime Rodriguez	Gestor de información
1.0	14/09/2017	Versión inicial.	Jaime Rodriguez	Gestor de información

TABLA DE CONTENIDO

1. INTRODUCCIÓN 4
2. OBJETIVO 5
3. MARCO CONCEPTUAL 5
3.1. Datos 5
3.2. Información 5
3.3. Conocimiento y sabiduría 5
3.4. Ciclo de vida de los datos 6
3.5. Calidad de datos 7
3.6. Raíces de los problemas de calidad de datos 8
3.7. Características de Calidad de Datos 8
3.8. Evaluación de la calidad de los datos 10
4. ARQUITECTURA CONCEPTUAL DEL PROCESO DE CALIDAD DE DATOS .. 11
5. ANÁLISIS Y EVALUACIÓN DE LA CALIDAD DE LOS DATOS 13
5.1 Seleccionar los grupos de información 14
5.2 Identificar las fuentes de verificación de los datos 15
5.3 Diligenciar la ficha técnica de la evaluación cuantitativa 17
5.4 Seleccionar y documentar los tipos de reglas de validación a evaluar 19
5.5 Cuantificar y documentar los tipos de defectos en los datos 20
6. CONCLUSIONES 22
7. BIBLIOGRAFÍA 23
8. ANEXO I 24

1. INTRODUCCIÓN

Los avances tecnológicos de los últimos años en materia de comunicaciones han permitido el desarrollo de sistemas de información de gran presencia que brindan acceso a grandes volúmenes de información. La necesidad de acceder en forma uniforme a la información disponible de múltiples fuentes de datos, ya sean internas al MADR o accesibles a través de Internet, es cada vez más fuerte y generalizada. Dichos requisitos de información son generalmente resueltos implementando complejos procesos de manipulación de datos sobre fuentes de datos heterogéneas. A medida que aumenta la cantidad de datos producidos por las entidades Adscritas y Vinculadas como también de los operadores del MADR, los usuarios se interesan más y más en la calidad de los resultados. Debido a la heterogeneidad de las fuentes de datos resulta difícil evaluar la calidad de los datos para brindar a los usuarios respuestas uniformes y de alta calidad.

La calidad depende de la calidad interna de las fuentes (la coherencia, la completitud, la frescura, etc.), de la confianza sobre quién produce los datos de esas fuentes, y también de la forma de producir la información devuelta al usuario. En un contexto en donde la información es producida por algoritmos sofisticados de agregación, la evaluación de la calidad requiere un conocimiento fino del proceso de producción. Además, la heterogeneidad de las fuentes de datos (por ejemplo diferentes formatos o semántica de los datos) agrega complejidad a la evaluación.

En este documento se presenta una herramienta que permite realizar el análisis y evaluación de la calidad de los datos de los registros administrativos del MADR, se conforma de las siguientes secciones: Una sobre el marco conceptual, seguido la Arquitectura conceptual del proceso de calidad de datos, luego se describe la herramienta para el análisis y evaluación de la calidad de los datos, y por último se presentan las conclusiones y recomendaciones.

2. OBJETIVO

Elaborar un esquema general para implementar el plan de calidad de datos, en cuento al análisis y evaluación de la calidad del dato.

3. MARCO CONCEPTUAL

3.1. Datos

Según Larry English los datos son:

- Representaciones de las cosas o entidades en el mundo real.
- Representaciones de las características o hechos (atributos) de las entidades.
- El material bruto y básico del cual se deriva la información para tomas de decisiones y acciones inteligentes.
- Junto con los datos que describan y contextualicen (metadata) los datos se produce información.

3.2. Información

- La información son datos en contexto, datos usables o útiles, datos con significado que pueden ser interpretados y comprendidos.
- La calidad de la información depende de:
    - la calidad de la semántica de los datos
    - los valores correctos
    - la presentación o formato compresible para los usuarios
- Información = f(Datos + Definición + Presentación)

3.3. Conocimiento y sabiduría

- La información en contexto, comprendida y aplicada por la gente se convierte en conocimiento
    - **Conocimiento = f(Gente + Información + Significado)**
- El conocimiento es un valor agregado a la información a través de la experiencia y aplicación de la información en un área específica.

- Cuando se combina conocimiento correcto, experiencia e intuición comprendida es posible tomar decisiones y actuar adecuadamente ante situaciones específicas. Esta combinación es lo que se llama sabiduría, el conocimiento empoderado para actuar:
    - **Sabiduría = f(Gente + conocimiento + acción)**

Figura 1. Datos, Información, Conocimiento y Sabiduría

3.4. Ciclo de vida de los datos

1. **Adquisición de los datos**
    - Definir la vista: modelamiento lógico de los datos
    - Implementar la vista: diseño e implementación física de los datos físicos
    - Definir canales de captura
    - Asegurar la calidad de los datos en los canales de captura
    - Obtener los datos: poblar la base de datos
    - Actualizar registros: almacenamiento y mantenimiento de los datos, copias de seguridad de los datos, archivar los datos

2. **Procesamiento de los datos**
    - Depurar los datos
    - Consolidar e integrar los datos
    - Generar valor agregado a partir de los datos

3. **Uso de los datos**
    - Definir la subvista: diseñar la consulta
    - Recuperar los datos: procesar la consulta
    - Manipular los datos: ordenar, agregar, reformatear y analizar
    - Presentar resultados: diseñar reporte, la forma de presentación de los datos

3.5. Calidad de datos

- Calidad de alguna entidad, objeto o cualquier cosa hace referencia al nivel de satisfacción o cumplimiento consistente con los requerimientos, necesidades o expectativas de los usuarios.
- La calidad de los datos hace referencia al cumplimiento consistente (completo) de los requerimientos o necesidades de los consumidores de los datos. En ese sentido, la calidad de los datos es relativa al uso potencial de los datos.
- Según Jurán, los datos son de alta calidad si ellos son conformes a su uso previsto en operaciones, tomas de decisiones y planeación.
- En otras palabras, calidad de datos es el estado de completitud, validez, consistencia, oportunidad y exactitud que hace que los datos sean apropiados para un uso específico o permitan satisfacer un propósito dado.
    - Completitud
    - Validez
    - Consistencia
    - Oportunidad (temporalidad)
    - Exactitud

3.6. Raíces de los problemas de calidad de datos

- **Errores en la captura**
  - Errores de digitación
  - Procesos de entrada de datos: formularios sin lista, aceptan nulos, campos confusos
  - Errores deliberados, a propósito
  - Errores de sistema

- **Sistemas de información**
  - Modificaciones no documentadas
  - Manuales de usuario incompletos
  - Funcionalidad no conocida
  - Funcionalidad no adecuada
  - Bajo mantenimiento en los diccionarios de datos y repositorios

- **Problemas en procedimientos y políticas**
  - Seguridad, privacidad y reglas de uso
  - Inventariar activos de información
  - Disponibilidad de los datos
  - Arquitectura de los datos
  - Planeación
  - El rol de la calidad

3.7. Características de Calidad de Datos

1. **Calidad intrínseca**
   - **Confiabilidad**: La calidad de la información y su fuente evocan credibilidad basada sobre la información misma o la historia o reputación de la fuente.
   - **Exactitud**: Según Larry English mide el grado en el cual un dato correctamente representa los atributos de un objeto o evento del mundo real; los datos son certificados, libres de

error. Ejemplo: la edad o el sexo de una persona, la temperatura. Número de productores afiliados a una asociación X en un mes específico.

- **Objetividad**: Los datos son objetivos, libres de sesgo.
- **Validez**: Mide el grado en el cual los datos corresponden o cumplen con las reglas de negocio. Ejemplo: el área de terreno del predio no puede ser menor al número de hectáreas sembradas.
- **Correctitud**: Mide el grado en el cual los datos cumplen con estándares aprobados o convencionales. Ejemplo: la localización geográfica de un predio es codificada según la división política administrativa del Dane (Divipola); las agencias áreas, los tipos de naves y los aeropuertos se codifican según estándares establecidos por la IATA.
- **Precisión**: Mide el nivel correcto de granularidad de los datos. Es el número de dígitos significantes de un dato que representa una observación. Ejemplo: la temperatura, el tiempo.

2. **Calidad contextual**
    - **Completitud**: Mide el grado en el cual todos los datos requeridos son conocidos; es decir, la completitud hace referencia a la profundidad y el alcance de la información contenida de los datos. Ejemplo: la razón social, el NIT, la localización geográfica, la dirección, el teléfono, entre otros datos, son datos requeridos para identificar unívocamente a una asociación de productores. Si existen todos estos datos para una asociación específica, se cumple con la completitud.
    - **Cantidad apropiada de datos**: Volumen suficiente de datos para su interpretación.

3. **Calidad representacional**
    - **Interpretabilidad**: Los datos se pueden interpretar.
    - **Fácil de comprender**: Los datos son claros, legibles.
    - **Consistencia representacional**: La representación o formato de los datos es siempre la misma en todos los casos, y no existe redundancia en los datos. Ejemplo: el formato de fechas (ddmmyyyy/yyyymmdd/mmddyyyy).
    - **Consistencia semántica**. Los datos deben ser claros, no ambiguos y consistentes. Ejemplo: la variable sexo permitir valores 1, M, 0, F.

- **Consistencia estructural**. Los tipos de entidad y atributos deben tener la misma estructura básica y formato. Ejemplo: los datos básicos de las asociación de productores siguen una misma estructura y formato.
- **Representación concisa**: Datos bien presentados, concisos, bien representados, bien organizados, bien formateados y estéticos.

4. **Calidad de accesibilidad**
    - **Accesibilidad**: Habilidad para acceder los datos cuando son requeridos. Ejemplo: los datos se pueden acceder desde la página Web y por solicitud.
    - **Seguridad de acceso**:
        - **Reserva estadística**. El acceso a los datos debe ser restringido dependiendo de la audiencia o consumidores de los datos. Ejemplo: Los datos puntuales sobre el registro de productores son de reserva; solo pueden ser accedidos por los funcionarios del MADR.
        - **Custodia de los datos**. Ejemplo: la custodia es responsabilidad del equipo del área misional del MADR.
        - **Uso de tecnologías seguras**: protocolos de seguridad, autenticación, VLANs servicios Web, dispositivos móviles.
        - Administración del acceso a las bases de datos.
        - Procesos de entrada de datos que validen, homologuen y estandaricen los datos

3.8. Evaluación de la calidad de los datos

La evaluación de la calidad de los datos, se define como una valoración científica y estadística de un conjunto de datos para determinar si éstos son adecuados para ser utilizados, de acuerdo a un propósito específico, en cuanto a su calidad, cantidad y tipo.

4. ARQUITECTURA CONCEPTUAL DEL PROCESO DE CALIDAD DE DATOS

La arquitectura a emplear en una solución de calidad de datos está integrada por varios componentes y a su vez, cada uno de esos componentes está especializado en ofrecer soluciones concretas y eficientes. En este sentido, presentamos siete procesos que deben ser aplicados en orden lógico a saber:

1. **Proceso de Extracción.** Este proceso tiene el propósito de **seleccionar los datos requeridos** de la fuente de datos.
2. **Proceso de Transferencia.** Se **transfieren los datos desde la fuente de datos** hasta el área de trabajo en la plataforma del repositorio destino.
3. **Proceso de Filtración y Estandarización.** La filtración busca **descartar aquellos datos que no se requieren** en el repositorio destino. La estandarización se refiere a descomponer los datos a unidades básicas, redefinir los formatos y valores de los datos.

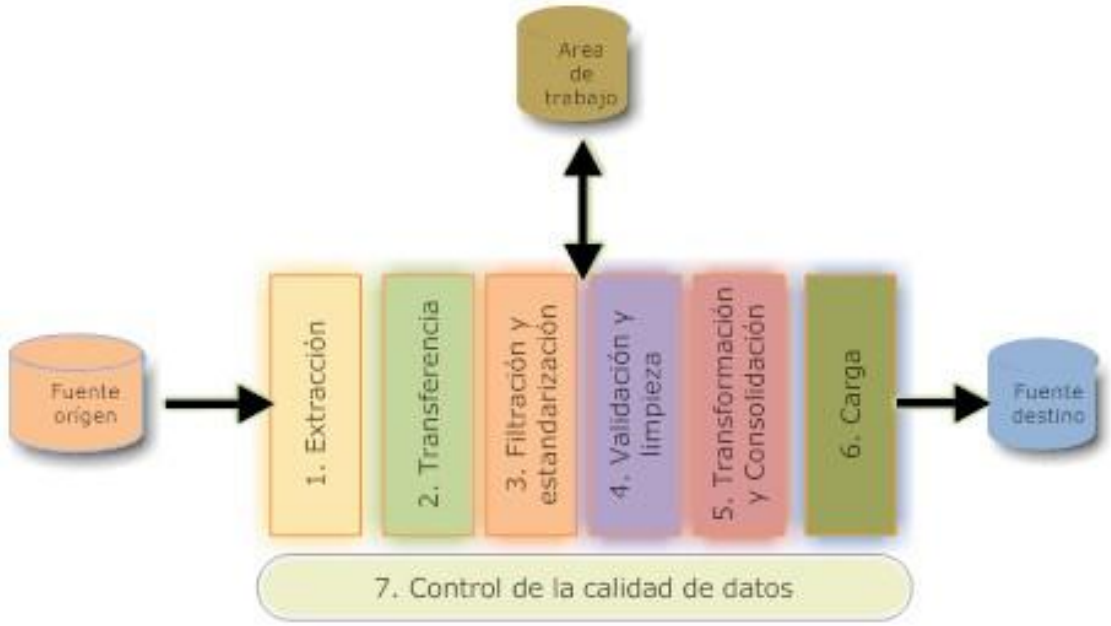

Figura 3. Arquitectura Conceptual del Proceso de Calidad de Datos

4. **Proceso de Validación y Limpieza.** Permite hacer el descubrimiento, corrección o eliminación de datos erróneos de una base de datos. Este proceso admite identificar datos

incompletos, incorrectos, inexactos, no pertinentes, etc. y luego substituir, modificar o eliminar estos datos sucios. Después de la limpieza, la base de datos podrá ser compatible con otras bases de datos similares en el sistema.

5. **Proceso de Transformación y Consolidación**. Consiste en la **aplicación de una serie de funciones o reglas de negocio sobre los datos extraídos para convertirlos en datos que, a continuación, serán consolidados en la nueva fuente.**
6. **Proceso de Carga.** Este proceso final se enfoca en **cargar los datos a las estructuras o tablas** del repositorio destino.
7. **Proceso de Control de Calidad.** Este proceso permite auditar, controlar y monitorear la ejecución de los demás procesos. Es un proceso transversal a todos los demás pasos, permitiendo asegurar la calidad del proceso y los datos que finalmente se cargan en el repositorio destino.

5. ANÁLISIS Y EVALUACIÓN DE LA CALIDAD DE LOS DATOS

Para estas actividades suponemos que, mediante las primeras reuniones o visitas de inspección a las dependencias misionales o entidad productoras de información ya sean adscritas, vinculadas u operadores, ya se ha identificado el inventario de las fuentes de datos; las bases de datos, las tablas de referencia y los archivos internos utilizados en la producción de las operaciones o registros administrativos que deben ser analizadas y evaluadas. Para todos ellos debe existir la documentación de la estructura y definición de los datos.

Por su contenido, los archivos fuentes de una operación o registros administrativos son de dos categorías:

Los archivos de eventos: características o propiedades de la unidad de observación, sus valores están determinados por una ocurrencia de las características de la unidad de observación investigada.

Los archivos o tablas de referencia: sus valores le dan contexto a las propiedades de la unidad observada.

De acuerdo al método o sistema de recolección, los datos de una operación o registros administrativos pueden provenir de diversas fuentes, entre las cuales consideramos las siguientes posibilidades:

- Censos
- Muestras
- Registros administrativos
- Observaciones sobre el terreno

Una estrategia de análisis y evaluación de la calidad de los datos puede tener uno o más de los siguientes objetivos:

- Entender y documentar la calidad y confiabilidad de los datos.
- Descubrir en los datos los problemas de calidad que deben ser resueltos durante los procesos de preparación y carga hacia el repositorio de información.

- Asegurar la armonización, estandarización e integración de los datos comunes en los diferentes registros administrativos.
- Especificar las reglas de transformación y validación que deben aplicarse a los datos, para asegurar el nivel de calidad que se requiere en una migración hacia el repositorio de información básica.

Los pasos a seguir durante el análisis y la evaluación de calidad de los datos, son:

- Seleccionar los grupos de información y los archivos a evaluar.
- Identificar las fuentes de verificación de los datos.
- Diligenciar la ficha técnica de la evaluación cuantitativa.
- Seleccionar y documentar los tipos de reglas de validación a evaluar.
- Cuantificar y documentar los tipos de defectos en los datos.

A continuación se describe cada uno de los pasos que se deben llevar a cabo durante el análisis y evaluación de la calidad de los datos. Cada paso se debe documentar en los formatos diseñados para tal fin.

5.1 Seleccionar los grupos de información

Dado que los archivos fuentes de un registro administrativo pueden ser numerosos, el objetivo de este paso consiste en identificar aquellos conjuntos de datos en donde una mala calidad cause un impacto desfavorable y significativo en los productos a difundir. El grupo de información puede incluir varios atributos de uno o más archivos, tomando siempre el origen de los datos, incluyendo los comunes, para determinar la efectividad del proceso de recolección o captura. Un grupo de datos puede corresponder a un producto, medida o indicador estadístico crítico o a una agrupación lógica de datos centrados alrededor de una clase temática de la unidad de observación.

Cada grupo debe ser calificado cualitativamente por la importancia relativa que causan los datos incorrectos en los indicadores o productos que generan, con “A =Impacto Alto”, “M =Impacto Medio”, o “B =Impacto Bajo”. Debe anotarse que un impacto alto requiere de una corrección real

de los datos defectuosos; un impacto medio aceptaría una imputación del valor por un método estadístico de estimación; mientras que un impacto bajo se podría resolver con una asignación del valor por defecto.

Se deben seleccionar, por lo menos, aquellos grupos de información que tienen un impacto alto o medio en el (los) producto(s) que generan. Un grupo de datos manejable para evaluación puede consistir de uno a diez archivos, con sus correspondientes campos identificados (de 3 a 30), dependiendo del registro administrativo por evaluar.

5.2 Identificar las fuentes de verificación de los datos

Además de los datos externos, se deben identificar las fuentes autorizadas de información, las cuales son utilizadas para verificar y corregir, por métodos humanos o electrónicos, los datos errados o sospechosos. Estas fuentes autorizadas son de dos clases:

- **Primarias**.- Están constituidas por el objeto mismo de investigación, o por observación física de los eventos.
- **Sustitutas**.- Las conforman documentos de soporte originales, o registros administrativos, que reflejan auténticamente la fuente.

Para cada categoría de Entidad se deben determinar las fuentes primarias y las sustitutas, dado que dentro de una categoría las fuentes de verificación son similares. Estas categorías son seis:

- Persona.
- Empresa.
- Objeto físico.
- Concepto.
- Localización.
- Evento.

Ejemplos de fuentes de verificación por clase de datos:

Clase de Datos	Fuente Primaria	Fuente Sustituta
***Persona Natural*:** (Toda persona con sus roles, por ej., Empleado, Contratista, etc.)	Las mismas personas	Registros administrativos, Formatos diligenciados por las personas, etc.
***Empresa*:** (Institución, Industria, Comerciante, etc.)	Depto. de Relaciones Industriales de la Empresa.	Registros administrativos, por ej., Archivos de la Cámara de Comercio, etc.
***Objeto Físico*:** (Equipo, Edificio, Terreno, etc.)	Observación o Muestra de los objetos físicos: por ej., materiales de inventario, equipos, etc.	Manuales técnicos, especificaciones técnicas de equipos y materiales, etc.
***Concepto*:** (Producto, Proceso, Especificación, Norma, etc.)	Gerente de producto, de Operaciones o Técnico.	Documentos oficiales, especificaciones, fórmulas, etc.
***Localización*:** (País, Departamento, Dirección, Ciudad, Mapa geográfico, etc.)	Observación directa de la Localización o Dirección.	Archivos de directorios telefónicos, Oficina de correos, Registros o Sistemas de información geográficos, etc.
***Evento*:** (Nacimiento, Factura, Reclamo, Pago, Orden de servicio, etc.)	Observación del evento.	Documentación escrita del evento, registros administrativos de soporte, facturas, etc.

Para propósitos de exactitud, es importante diferenciar la confirmación de valores entre las fuentes sustitutas y las primarias, teniendo en cuenta que aquellas podrían ser menos exactas, aunque también menos costosas.

En cuanto a los registros administrativos, o archivos secundarios que se identifiquen y seleccionen para completar o verificar los datos, antes de utilizarlos es necesario conocer:

- La definición y significado de los datos.
- La fuente y metodología de recolección.
- La fecha de recogida y su frecuencia de actualización.
- El nivel de calidad de la información; su grado de cobertura, confiabilidad y margen de error.
- Si se usaron técnicas de estimación para completar los datos, qué porcentaje ha sido estimado y con qué método.

5.3 Diligenciar la ficha técnica de la evaluación cuantitativa

En este paso se deben identificar los tipos y las características de evaluación de calidad de los datos.

La ficha técnica debe indicar la clase y forma como se realizaron las pruebas y las clases de pruebas que se realizaron. Es similar al reporte de auditoria de los estados financieros.

El tipo de evaluación define si ella se va a efectuar en forma manual o electrónica; sobre la población total o sobre una muestra seleccionada estadísticamente, en cuyo caso se deben documentar los parámetros de selección del tamaño, el nivel de confianza y el porcentaje de error.

En una evaluación electrónica los datos en medios magnéticos se procesan en un computador para analizar el cumplimiento de las reglas de validación. Sin embargo, cuando el volumen de los datos es pequeño, dichas reglas de validación se pueden comprobar manualmente. En una evaluación física se verifica la calidad de los datos de manera que sean correctos y estén acordes con la realidad que representan, utilizando las fuentes primarias y sustitutas.

Las características de evaluación son los aspectos de calidad que interesan y son los más relevantes en la elaboración de las estadísticas. Para identificarlas en este paso se debe tener en cuenta los métodos cuantitativos de obtención de sus medidas. A continuación, se presentan las características de calidad más importantes que deben considerarse, y el método de obtención de sus medidas:

- **Consistencia de la definición**.- Es la concordancia del contenido del dato1 con su definición. Su medida se obtiene por la existencia o no de la concordancia.
- **Cubrimiento de valores**.- Es la característica de tener todos los datos con los valores requeridos. Su medida es el porcentaje de registros que tienen un valor NOTNULL (valores no faltantes) para un dato específico. El complemento es el porcentaje de valores faltantes. Para el cálculo del porcentaje de faltantes no se debe tener en cuenta aquellos valores que no aplican al sujeto investigado.
- **Cumplimiento de las reglas de validación**.- Ver documento anexo: “Reglas de Validación”: Los datos deben cumplir los diferentes tipos de reglas, precisadas en dicho anexo. Su medida

es el grado de cumplimiento de las éstas y se expresa como el porcentaje de registros y de campos con valores que cumplen con las mismas para un dato específico. A manera de ejemplo, las clases de pruebas de validación incluyen: valores de dominios, rangos, llaves primarias únicas sin duplicados, integridad referencial, reglas de dependencia, consistencia de formatos, etc. Es importante anotar que un valor de campo, por ser válido, no necesariamente significa que sea exacto o correcto.

- **Exactitud con la fuente sustituta**.- Significa que el dato es el mismo que contiene una fuente sustituta autorizada, generalmente el documento de origen o una forma electrónica externa no alterada. Su medida es una evaluación del porcentaje de registros y de campos cuyos valores para un dato específico son exactos, comparados con los valores de la fuente original autorizada. Se debe anotar que esta forma de evaluación no es perfecta, ya que las personas que llenan los documentos de soporte cometen errores, lo mismo que las formas electrónicas pueden introducir errores de varios tipos.
- **Exactitud con la fuente primaria**.- El dato refleja exactamente la realidad del objeto o evento que describe. Es el más alto grado de exactitud de información posible, ya que se está evaluando el dato contra su valor, observado físicamente. Para la evaluación con la realidad se requiere confirmar el valor del dato con la medida del objeto o la observación del evento. Su medida es el porcentaje de registros y de campos cuyos valores para un dato determinado han sido confirmados como correctos con sus valores actuales. La escogencia entre la evaluación con una fuente sustituta o con la realidad, depende del impacto negativo que causan los datos incorrectos en el producto estadístico que afectan. Los costos involucrados son más altos en la evaluación con la realidad, pero se pueden disminuir con la toma de muestras estadísticas. Un ejemplo de evaluación con la realidad es la observación en el terreno de las características de una vivienda de interés social rural construida.
- **Ocurrencias Duplicadas**.- Es un caso especial de validación de integridad y representa el grado de correlación unívoca (uno a uno) entre el registro y el objeto o evento que representa. Su medida es el porcentaje de registros que son duplicados de otros registros dentro en una colección de datos. Por registros duplicados no sólo queremos decir que tengan valores idénticos de identificación, sino también que los registros sean representaciones duplicadas del mismo objeto o evento del mundo real.

- **Equivalencia de datos redundantes**.- Es el grado con el cual los datos de una colección o base de datos son semánticamente equivalentes a los datos del mismo objeto o evento en otra colección o base de datos. Semánticamente equivalentes significa que los valores son iguales en su concepto; es decir, significan lo mismo en ambos lugares, aun cuando tengan distinto valor. Por ej., si el sexo de P. Mesa es M por masculino en uno y en el otro lugar es 1 por masculino para el mismo P. Mesa, hay equivalencia. La medida de equivalencia es el porcentaje de campos en los registros de una colección que son semánticamente equivalentes a sus respectivos campos en la otra colección de datos.
- Sobre los datos comunes o redundantes en archivos se debe determinar la fuente de datos más confiable, ya que posteriormente se deberán consolidar.
- **Las principales guías para seleccionar las fuentes más confiables, son:**
    - Los datos tienden a ser más confiables, entre más crítico sea el proceso para la entidad productora de la información-, que, como fuente, los crea y actualiza. Por ej., el salario de una persona tomado del departamento de personal de la empresa donde trabaja, es más confiable que el registrado en una solicitud de apertura de cuenta bancaria.
    - Los datos actualizados frecuentemente son, en general, más exactos que los datos viejos.
    - -Los datos creados y actualizados por personas responsables del proceso son más confiables que los actualizados por terceros, cuando no tienen incentivos para capturarlos correctamente. Si la labor de un funcionario es medida por la calidad de la información que produce, ésta tenderá a ser más confiable.
    - -Los datos históricos, como “fecha de iniciación del servicio” deben ser extractados de la ocurrencia más antigua del registro.

5.4 Seleccionar y documentar los tipos de reglas de validación a evaluar

En este paso se especifican y describen, en detalle, los diferentes tipos de reglas de validación que debe aplicarse a los datos para evaluar su nivel de calidad.

El analista de datos debe ayudarse de un modelo lógico para descubrir y especificar cada una de las reglas, de acuerdo con las guías documentadas en el anexo “Reglas de Validación”.

5.5 Cuantificar y documentar los tipos de defectos en los datos

En este paso se analizan los datos para descubrir y cuantificar sus defectos. Las medidas son el resultado cuantificado de la evaluación de las características y reglas seleccionadas anteriormente.

Durante la evaluación se extraen los datos fuentes, se observan y se analizan para confirmar sus definiciones y sus contenidos, o para descubrir los defectos o anomalías que deben cuantificarse durante el proceso de evaluación.

El descubrimiento de los datos sospechosos puede hacerse sobre el 100% de los registros para archivos pequeños, o utilizando muestreo estadístico para consultas manuales o automatizadas sobre los datos, como las de aceptación de atributos, descubrimiento de atributos, materialidad, etc., de manera que podamos hacer la inferencia de los resultados sobre el total de la población.

Al finalizar el análisis de calidad de los archivos, o bases de datos fuentes, se deben interpretar y presentar los resultados: clasificar los problemas o patrones de calidad en los datos, listar y analizar dos o tres ejemplos representativos de cada tipo y cuantificar los tipos de defectos en los datos, estimando su frecuencia.

Es importante anotar que los errores más frecuentes, no necesariamente significan los errores que mayores impactos negativos producen; por lo tanto, el impacto o el costo del error deben tenerse en cuenta al presentar los reportes.

En general, según se describió en las características de evaluación, se puede decir que los tipos de errores están clasificados en las siguientes categorías:

- 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.
- Incorrectos, cuando no concuerdan con la realidad.
- Duplicados, es decir, existen varias identificaciones del mismo sujeto.
- Discrepancia con datos redundantes.

Los reportes deben presentarse gráficamente, o en hojas de cálculo, incluyendo, en lo posible, diagramas de Pareto, en las siguientes formas:

- **Reportes de Resumen de medidas**.-Incluyen generalmente representaciones gráficas del resumen de los resultados. Resumen de errores por grupo de información, por registro, por campo.
- **Reportes Detallados de calidad**.-En este reporte se presentan unos resultados por tipos de error en el campo, y otros por tipos de error en el registro. Por ej., el campo dirección, puede tener errores en el formato de la calle, o la calle es incorrecta, o falta la ciudad, o está incompleta, etc.
- **Reportes de Discrepancia.-**Estos reportes comparan la equivalencia o consistencia de los datos cuando existen registros del mismo objeto o evento en múltiples bases de datos. Ellos ilustran la sincronización entre archivos redundantes y muestran sus problemas de inconsistencia.
- **Reportes de Excepción.-**Simplemente muestran el descubrimiento de los datos encontrados errados y sus valores correctos.

6. CONCLUSIONES

- El análisis y evaluación de la calidad de datos es un insumo fundamental para el proceso de migración e integración de datos hacia Productores 360 proveniente de los registros administrativos que producen las entidades adscritas y vinculadas como los operadores del MADR.
- La aplicación rigurosa del método permite garantizar la calidad de los datos que serán incorporados y difundidos a través de productores 360.
- El análisis y evaluación de la calidad de datos es un proceso de mejoramiento continuo que asegura la difusión de información confiable para la toma de decisiones y acciones en los sectores relacionados, como también un mecanismo de auditoría que se tiene con las entidades adscritas y vinculadas como los operadores del MADR.
- Se recomienda que todo registro administrativo debe previamente documentarse para poder planear el análisis y evaluación de la calidad de datos.

7. BIBLIOGRAFÍA

Nicolas Dib. Descripción de los Procesos de Calidad de Datos en el Repositorio de Información Básica. (Primera Parte). Revista de la información básica ib. Año 1 – No 2. Págs. 115-122. Diciembre de 2006.

Nicolas Dib. Evaluación de Calidad de los Datos Estadísticos. Revista de la información básica ib. Año 2 – No 3. Págs. 75-83. Junio de 2007.

English, Larry P., Improving Data Warehouse and Business Information Quality. New York: John Wiley and Sons, 1999.

English, Larry P., Information Quality Assessment, Data Cleansing and Transformation Processes. The Data Warehousing Institute: The fifth annual Implementation Conference, 2000.

Brackett, Michael H., Data Resource Quality: Turning Bad Habits into Good Practices. Englewood Cliffs, NJ: Addison Wesley Longman, 2000.

Olson, Jack E., Data Quality: The accuracy Dimension. San Francisco, Morgan Kaufmann, 2003.

Ross, Ronald G., The Business Rule Book: Classifying, Defining & Modeling Rules, Business Rule Solutions (1997).

Duncan, Karolyn and Wells, David, Rule Based Data Cleansing. The Journal of Data Warehousing, Fall, 1999.

Definición sobre el término Calidad de datos. Consultado en http://en.wikipedia.org/wiki/Data_quality. marzo 2007.

Definición de Calidad de datos. Consultado en http://it.csumb.edu/departments/data/glossary.html en marzo 2007.

- Arthur D. Chapman. Principles of data quality. 2005.

8. ANEXO I

Reglas de validación y consistencia

Las reglas de validación y consistencia son aquellas que prueban la integridad de los datos. Los modelos de datos, cuando existen, ayudan a localizar dichas reglas. Debemos recordar que un dato válido no necesariamente significa que sea correcto.

Durante los procesos de validación, se evalúa la integridad de los datos de cada archivo, mientras que en los procesos de limpieza se definen las acciones a tomar en caso de errores o inconsistencias en los mismos. Por ejemplo, efectuar imputaciones. Por lo tanto, se puede afirmar que las reglas de validación y limpieza se implementan en dos pasos:

a) Las reglas que prueban la integridad de los datos.

b) La especificación de las acciones a tomar cuando se encuentra una violación de integridad.

Las reglas de validación se pueden utilizar para evaluar la calidad de los datos y/o, para filtrarlos y/o, para corregirlos.

Recordemos que una estrategia de validación de los datos tiene, entre otros, uno o más de los siguientes objetivos:

- Entender y documentar la calidad y confiabilidad de los datos, es decir, su perfil.
- Descubrir en los datos los problemas de calidad que deben ser resueltos durante los procesos de validación y limpieza.
- Especificar las reglas de validación que deben aplicarse a los datos para asegurar el nivel de calidad que se requiere en una migración, una conversión o una carga de datos a un repositorio.

Los modelos lógicos de datos representados en un diagrama de entidad relación ERD-, permiten identificar un conjunto relativamente robusto de reglas de validación, analizando su estructura. Por lo tanto, es de suponerse que el profesional responsable de la evaluación, no sólo debe estar

familiarizado con el modelamiento de datos sino que, además, ha debido construir, previamente al análisis y determinación de las reglas a validar, el modelo lógico correspondiente a los datos de su interés.

El modelo de datos que se muestra en la figura No. 1, corresponde a una porción del repositorio de los datos del Registro de Usuarios Universitarios. La mayoría de los ejemplos utilizados para explicar los diferentes tipos de reglas de validación son tomados de dicho modelo.

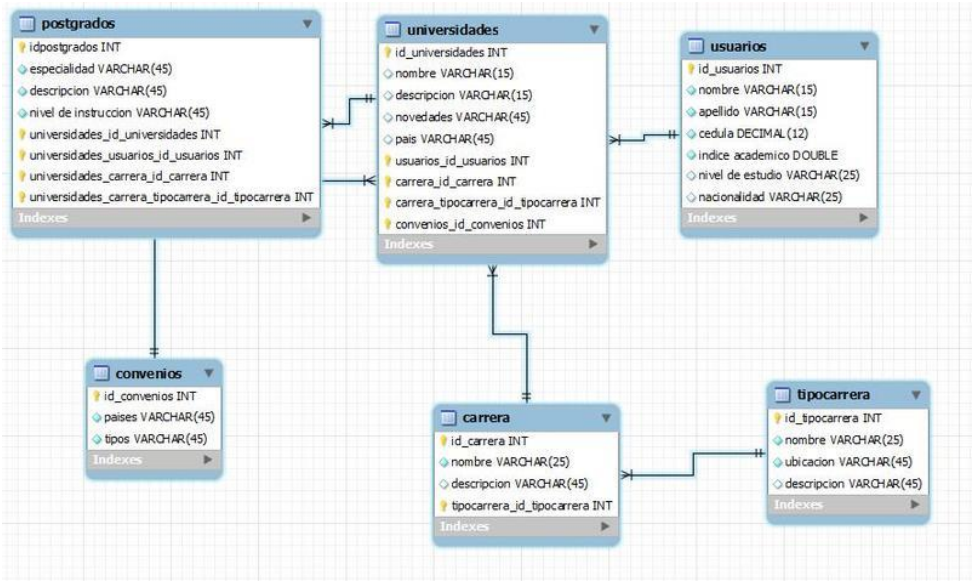

Figura No. 1.- Diagrama Entidad Relación del Registro de Usuarios Universitarios

Las reglas de validación de los datos, las clasificamos de la siguiente manera:

- 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
- Dependientes del estado de la Entidad
- Mutuamente Excluyentes
- Mutuamente Dependientes
- Derivadas
- Restringidas
- Por valor
- Por Relación

Los primeros cuatro tipos de reglas están “explícitas” en el modelo lógico de datos de entidad relación y pueden ser extraídas directamente del modelo cuando éste ya existe. Los tipos de reglas restantes están implícitas en el modelo de datos y requieren de algún análisis o investigación para descubrirlas o identificarlas específicamente. De todas maneras, ya sea el tipo de regla explícito o implícito, el método de usar el modelo de datos para documentarlas es un paso importante y necesario para evaluar su calidad.

Las reglas de validación se pueden describir en términos comunes, o en términos técnicos -usando la terminología de las bases de datos-, v.g., llave primaria, llave foránea, valores nulos, tablas, columnas, etc. Se deben usar los términos adecuados, según la audiencia a la que se dirijan.

A continuación, pasamos a describir cada uno de los tipos de reglas que, en su conjunto, representan una clasificación relativamente exhaustiva, para que sean utilizadas como referencia, en una etapa de evaluación en la calidad de los datos.

En los ejemplos que se presentan usaremos términos comunes para describir algunas de las reglas.

Nombre	Reglas de Identidad
Código	RVC-01
Descripción	Cada ocurrencia de la entidad está unívocamente identificada. En términos técnicos, la llave primaria es única (no hay duplicados) y no puede ser nula (es NOT NULL).

Ejemplos	Dos municipios no pueden tener el mismo código de identificación y dicho código no puede tener valor nulo. Dos empleados no pueden tener el mismo código de identificación y dicho código no puede tener valor nulo.
Nro. de Reglas	Cada entidad tiene una regla de identidad
Proceso de análisis	Examinar cada entidad para determinar el atributo o combinación de atributos que la identifican unívocamente (llave primaria). Describir la(s) regla(s) como se muestra en los ejemplos.
Notas	1.- Observe que en la misma regla se especifican ambas condiciones, pero también es aceptable escribir una regla para cada condición. 2.- Debemos entender que un valor "nulo" representa ausencia total de un valor y no se debe confundir con el número o conjunto de caracteres que se use como "valor por defecto". 3.- Cuando la llave primaria es compuesta, se deben verificar cada uno de sus componentes antes de validar su identidad.

Nombre	**Reglas de Integridad Referencial**
Código	RVC-05
Descripción	Una regla de integridad referencial define la relación o asociación de existencia o identidad entre dos entidades. En términos técnicos, impone lo que se denomina las reglas de integridad, aplicadas a las tablas o archivos de referencia.
Ejemplos	Todas las personas registradas como productores deben pertenecer a un hogar; un hogar debe estar compuesto por al menos una persona. Una variable estadística pertenece a una unidad de observación; Una unidad de observación contiene variables estadísticas.
Nro. de Reglas	Cada relación entre dos entidades tiene dos reglas de integridad referencial. Aun cuando las relaciones uno-a-uno y uno-a-muchos sólo requieren de una regla para su implementación, las relaciones muchos-a-muchos requieren dos reglas para su implementación.
Proceso de análisis	Examinar cuál entidad puede estar relacionada o asociada con otra y describa su relación como se indica en los ejemplos.
Notas	1.- Las reglas de integridad referencial tienen que ver con la existencia de algo y no deben confundirse con las de cardinalidad, las cuales tienen que ver con la cantidad de ocurrencias en las que participa cada entidad. 2.- Las reglas de integridad referencial aplicadas a las tablas de referencia se implementan especificando métodos y procedimientos adecuados para la creación, actualización y eliminación de registros.

Nombre	**Reglas de Dominio**
Código	RVC-10
Descripción	Las reglas de dominio describen el conjunto de valores permitidos para cada atributo, ya sea por tipo de dato, formato o valores del dominio.

	Ellas verifican que los atributos tengan valores válidos y con significado.
Ejemplos	El sexo de una persona debe ser 1=Hombre o 2=Mujer. La edad de una persona debe ser un número entre 0 y 110. El municipio de nacimiento de una persona debe existir en la lista de municipios del departamento de nacimiento de dicha persona. La fecha de nacimiento de una persona debe tener formato "dd/mm/aaaa".
Nro. de Reglas	Cada atributo tiene por lo menos una regla que describe el conjunto de valores permitidos. La regla puede ser simple o compleja, combinando varias formas.
Proceso de análisis	Examinar cada atributo y determinar el conjunto de valores permitidos, expresado en una o varias de las siguientes formas: • Una lista de valores • Un rango de valores • Una restricción de valores • Designación de un formato o de un conjunto de caracteres permitidos • Un patrón de máscara (editado) • Una precisión dada en unidades • Formato libre (sin restricción)
Notas	Es importante anotar que nunca se debe dejar en blanco la regla de dominio del atributo. Cuando no tiene ninguna restricción, es preferible indicar que es de "texto en formato libre".

Nombre	**Reglas de Atributos Dependientes**
Código	RVC-15
Descripción	Este tipo de reglas describe las situaciones en donde el conjunto de valores permitidos para un atributo depende del valor de otro atributo que describe el estado de la entidad.
Ejemplos	Si una persona es menor de edad, no puede haber aprobado 16 años de estudio. El tipo de energía utilizado para cocinar debe ser numérico y puede tomar cualquiera de los siguientes valores 1, 2, 3, 4, 5, 6 ó, 9, a menos que el lugar donde cocinan sea = 6 (no cocinan).
Nro. de Reglas	Cada dependencia de estado origina una regla.
Proceso de análisis	Examinar cada entidad para determinar si existe un atributo que describa su estado. Si existe, analizar todos los otros atributos con las siguientes preguntas: Está el conjunto de valores del atributo restringido por el estado de la entidad ? Si lo está, qué estados permiten qué valores ? Qué estados permiten valores nulos ? Qué estados obligan valores nulos ?
Notas	1.- La mayoría de las veces, éste análisis se aplica a atributos dentro de la misma entidad, no entre entidades.

	2.- Los atributos dependientes del tiempo se pueden considerar un caso especial de éste tipo de reglas.

Como se puede observar, un modelo de datos puede proporcionar la descripción de un gran número de reglas de validación. El número de reglas inherentes, sin contar las de dependencia, puede ser aproximadamente estimado usando la siguiente expresión:

Número de reglas = (número de entidades) + (número de relaciones*3) + (número de atributos) + (número de sub-tipos*2) + (número de super-tipos*2)

Para la identificación de estas reglas, existe una variedad de técnicas entre las cuales la más utilizada es la construcción del modelo lógico de datos. El diseño de los cuestionarios, su documentación y las tablas de referencia entre otros, también ayudarán a verificar las reglas encontradas en el modelo.