FORTALECER LA DOCUMENTACIÓN DE LOS DATOS, GENERADOS Y ALMACENADOS EN EL SISTEMA DE INFORMACIÓN DEL PROGRAMA NACIONAL DE MONITOREO DE LA “RED DE VIGILANCIA PARA LA CONSERVACIÓN Y PROTECCIÓN DE LA CALIDAD DE LAS AGUAS MARINAS Y COSTERAS DE COLOMBIA” REDCAM

JUAN SEBASTIAN GARZON ARIZA

UNIVERSIDAD SANTO TOMÁS
FACULTAD DE INGENIERÍA AMBIENTAL
ÁREA DE INVESTIGACIÓN CIENTÍFICA MARINA
BOGOTÁ
2016

FORTALECER LA DOCUMENTACIÓN DE LOS DATOS, GENERADOS Y ALMACENADOS EN EL SISTEMA DE INFORMACIÓN DEL PROGRAMA NACIONAL DE MONITOREO DE LA “RED DE VIGILANCIA PARA LA CONSERVACIÓN Y PROTECCIÓN DE LA CALIDAD DE LAS AGUAS MARINAS Y COSTERAS DE COLOMBIA” REDCAM

JUAN SEBASTIAN GARZON ARIZA
Autor

Documento final de pasantía para optar al título de Ingeniero Ambiental

Director(a): Ángela María Jaramillo Londoño, Bióloga marina, PhD
Docente Universidad Santo Tomás (USTA).

UNIVERSIDAD SANTO TOMÁS
FACULTAD DE INGENIERÍA AMBIENTAL
BOGOTÁ
2016

TABLA DE CONTENIDO

LISTADO DE GRÁFICOS

LISTADO DE TABLAS

RESUMEN

La base de datos del sistema de información del programa nacional de monitoreo de la “red de vigilancia para la conservación y protección de la calidad de las aguas marinas y costeras de Colombia” - REDCAM, genera información actualizada sobre el estado de la calidad de agua marino-costera y además funciona como instrumento de gestión y apoyo para atender y soportar metas, programas y políticas nacionales e internacionales de carácter ambiental [1]; por lo tanto, esta se encuentra constantemente en crecimiento y su administración se torna compleja.

Por este motivo, el personal encargado del manejo de la base de datos, puede estar omitiendo u ocultando información relevante, ya que se identificó que existen falencias en la documentación de sus datos y adicionalmente existe la necesidad de verificar y depurar algunos registros que pueden presentar errores técnicos o de digitación.

Para dar solución a estas problemáticas, por medio de apropiación de conocimiento, se realizó la documentación de los muestreos realizados en el año 2015 por los nodos integrantes de la REDCAM, por medio de la construcción de metadatos, teniendo en cuenta estándares nacionales e internacionales. Por otro lado, se realizó un ejercicio práctico de depuración de datos, identificando una metodología por medio del análisis multicriterio AHP, que permitiera dar solución a las necesidades identificadas por los administradores de la base de datos.

El desarrollo de estas actividades, permitió evidenciar que tan completa se encuentra la información que se genera, a partir del trabajo en campo y el análisis de laboratorio para las muestras de agua y sedimentos, ayudando a los administradores del sistema, a solicitar los datos a los diferentes nodos integrantes del sistema y complementar los registros. Así mismo, la aplicación de una metodología que permita identificar y limpiar información sospechosa, contribuyó al cumplimiento en tareas de verificación y depuración de la base de datos, ya que esta brinda soporte para la investigación, protección y manejo de los recursos marino-costeros de Colombia.

1. INTRODUCCIÓN

La Red de Vigilancia para la Conservación y Protección de la Calidad de las Aguas Marinas y Costeras de Colombia – REDCAM cuenta actualmente con 15 años de operación y 16 nodos participantes, siendo el INVEMAR, el nodo central de la REDCAM. Este sistema de información además de generar información actualizada sobre el estado de la calidad del agua marina y costera, funciona como instrumento de gestión y apoyo para atender y soportar las metas, programas y políticas nacionales e internacionales de carácter ambiental [1].

Es por esto que el INVEMAR con el fin de unificar y estandarizar la información generada por los nodos participantes del sistema de información, desarrolló el Manual de Funcionamiento del Sistema de Información [2] y el Manual de Técnicas Analíticas para la Determinación de Parámetros Fisicoquímicos y Contaminantes Marinos [3]; sin embargo la magnitud de la base de datos ha aumentado considerablemente y por ende su administración.

Actualmente, han transcurrido 16 años y la base de datos estructurada y alimentada constantemente, muestra aún vacíos en la documentación de la información porque no cuenta con metadatos que ilustren sobre las condiciones y características en las que se dio el muestreo, generando esto, dificultades para los usuarios que consultan y utilizan la base de datos, ya que en la ausencia de esta información suplementaria, se puede omitir u ocultar información relevante que describa el contexto y características empleadas en el muestreo que se realizó en una zona geográfica determinada.

Por lo tanto, el trabajo busca fortalecer la documentación de los datos, generados y almacenados en el sistema de información del programa nacional de monitoreo de la “red de vigilancia para la conservación y protección de la calidad de las aguas marinas y costeras de Colombia” - REDCAM, el cual será alcanzado por medio de la construcción de metadatos y la aplicación de una metodología que permita hacer limpieza de datos.

Por consiguiente, en el desarrollo del trabajo se encuentra un análisis de información secundaria correspondiente a los dos objetivos planteados, el cual funcionará como soporte teórico para abordar la problemática identificada. El desarrollo de la pasantía se encuentra vinculado directamente con la metodología propuesta, así que por medio de tres capítulos se presentarán los avances de cada actividad asociada a los objetivos y otras tareas culminadas en la pasantía.

Finalmente, con la metodología desarrollada, se concluye que la base de datos del sistema de información de la REDCAM, necesita un tratamiento que le permita organizar y consolidar la información que contiene, por medio de herramientas que le brinden obtener índices de calidad positivos. Por lo tanto, se recomienda que se diseñe o adopte modelos y guías de normalización de datos, que permitan estandarizar la información que contiene el programa nacional de monitoreo.

2. OBJETIVOS

2.1. Objetivo general

Fortalecer la documentación de los datos, generados y almacenados en el sistema de información del programa nacional de monitoreo de la “red de vigilancia para la conservación y protección de la calidad de las aguas marinas y costeras de Colombia” REDCAM

2.2. Objetivos específicos

- Construir los metadatos de los monitoreos realizados por el programa nacional REDCAM en el año 2015
- Aplicar una metodología existente para hacer limpieza de datos que se pueda aplicar en al menos una de las variables de la REDCAM

3. MARCO REFERENCIAL

3.1. ANTECEDENTES

El manejo de datos a través del tiempo se ha convertido en una tarea obligatoria para la acumulación y manejo de las bases de datos de diferentes instituciones que alimentan constantemente sus centros de documentación.

Algunos ejemplos del manejo de datos en Latinoamérica son los ejecutados por el Instituto Oceanográfico de la Armada – INOCAR, de Ecuador, en donde se formularon proyectos de recuperación y manejo de bases de datos oceanográficos como la “Creación de un sistema de arqueología de datos del INOCAR (servidores, equipos de comunicación y software), [4], adicionalmente en el año 2015 se encuentra el proyecto de “Estructuración de la base de datos así como la recopilación de la información hidrológica y geológica existente”, desarrollado por el INOCAR [5], [6].

En Colombia se comenzó a trabajar en metodologías de recuperación y manejo de datos con el apoyo de la Dirección General Marítima de Colombia – DIMAR y el Centro Control Contaminación del Pacifico, actual Centro de Investigaciones Oceanograficas e Hidrograficas del Pacifico - CCCP, por medio del proyecto “Metodología Archivística para la Recuperación de Información Oceanográfica del Pacifico Colombiano” en donde se realizó una propuesta que permita recuperar y manejar información oceanográfica por parte de investigadores para diferentes fines académicos y profesionales [7]. Uno de los proyectos que aportó al desarrollo de este trabajo fue realizado por el INOCAR, el cual fue denominado “Arqueología de datos oceanográficos: Contribución a la preservación de investigación ecuatoriana oceanográfica” teniendo como resultado la recuperación y conservación de los datos oceanográficos de los últimos 30 años de investigación en Ecuador [7].

Adicionalmente, se continúa con la capacitación de personal calificado para el manejo de datos, realizando cursos de profundización en Bélgica, México, Venezuela y EEUU. En Colombia se han realizado cursos como el “Curso Colombiano preparatorio en manejo de datos e información oceanográfico Odincarsa” en donde organizaron el IODE y la CCO, en el año 2003 [8]. En el año 2006 se formuló un proyecto para fortalecer la central de datos oceanográficos de DIMAR “RETROCEAN” en donde se generó la recuperación y control de datos oceanográficos, propiedad del Centro de Investigaciones Oceanográficas e Hidrográficas - CIOH [8], [9].

Referente a la construcción y utilización de metadatos, se han desarrollado tres cursos introductorios al perfil de metadatos marinos ISO 19115, en donde se profundizaba en la adopción de la norma por DIMAR para la creación de metadatos marinos. Estos cursos se han realizado en Tumaco y Cartagena – Colombia, el ultimo realizado en el año 2008 [8].

3.2. MARCO CONCEPTUAL

A continuación, será presentada la información básica que sirvió para dar sustento teórico al diseño y documentación de información por medio de metadatos, como también la revisión de información secundaria que permitiera la identificación y calificación de diferentes metodologías para que finalmente una fuera elegida, con el fin de dar cumplimiento a los objetivos planteados en este documento.

3.2.1. Dato

Los datos son considerados representaciones simbólicas de características y atributos que describen momentos, sucesos y entidades. Se tienen en cuenta símbolos numéricos, alfabéticos, imágenes y entre otros, que se pueden considerar resultado de mediciones y observaciones que toman importancia cuando son asociadas a un contexto determinado.[10]

3.2.2. Metadato

El metadato es considerado un dato sobre el dato desde 1995 por [11], [12], fue retomado y actualizado por [13], Expresando que el concepto debe ser visto desde una perspectiva mucho más amplia en la cual se deben integrar los diferentes campos profesionales para la generación de recursos digitales de importancia; Esta definición también es citada por [14].

También es definido dentro del concepto de Objeto de Aprendizaje como una estructura de información externa que facilita su almacenamiento, identificación y recuperación de información [15], [12].

Algunas de las ventajas del uso de metadatos son[15], [16] [8]:

- La construcción de metadatos incrementa el acceso a la información, es decir que la información suministrada tendrá una ruta más simple para su localización.
- Se puede tener un mayor control legal acerca del uso de la información.
- Aumenta la precisión en las búsquedas en la web.
- Evita pérdidas de tiempo dinero y trabajo [17].
- Adaptabilidad, para poder ajustarse a las necesidades de los usuarios.

Complementando estas definiciones, [16] cita que “los metadatos son datos secundarios como pueden ser el autor, el título, las palabras clave, el resumen, la fecha, u otros que describen los datos primarios o recursos de información, es decir, se emplean para suministrar información sobre datos producidos, ellos describen el contenido y otras características de los datos primarios para posibilitar a una persona o máquina, ubicar y entender los datos”.

Por lo tanto, para la construcción de metadatos, se debe tener presente el análisis de la norma internacional ISO19115, la cual proporciona una guía para establecer terminología, definiciones y procedimientos de aplicación para los metadatos geo-espaciales [18], y asegurar que los datos sean utilizados de forma correcta, teniendo en cuenta las observaciones y limitaciones que fueron característicos de la aplicación geográfica [19].

A partir de la norma ISO, se han diseñado diferentes tipos de metadatos como el CSDGM (Content Standard for Digital Geospatial Metadata) y el Dublin Core Metada (documentación de fondos bibliográficos como libros, artículos); teniendo en común, la introducción de atributos específicos de carácter obligatorio, que permiten vincular información detallada de acuerdo a la aplicación o modelo que se esté documentando [10].

Esta norma internacional define una serie de atributos básicos para la construcción de metadatos geográficos que permitirá el descubrimiento, acceso, transferencia y utilización de datos [18]. Estos atributos se encuentran en la tabla 8 de este documento.

Tabla 1. Atributos para la construcción de metadatos geo-espaciales, según la ISO 19115.

ATRIBUTO	CARÁCTER
Título del conjunto de datos	Obligatorio
Fecha de referencia	Obligatorio
Idioma del conjunto de datos	Obligatorio
Categoría del tema	Obligatorio
Resumen descriptivo	Obligatorio
Fecha de los metadatos	Obligatorio
Parte responsable del conjunto de datos	Optativo
Resolución espacial del conjunto de datos	Optativo
Formato de distribución	Optativo
Extensión vertical y temporal	Optativo
Tipo de representación espacial	Optativo
Sistema de referencia	Optativo
Linaje	Optativo
Recurso en línea	Optativo
Identificador del fichero de metadatos	Optativo
Norma de metadatos	Optativo
Versión de la norma de metadatos	Optativo
Localización geográfica de los datos	Condicionales
Conjunto de caracteres del conjunto de datos	Condicionales
Idioma de los metadatos	Condicionales
Conjunto de caracteres de los metadatos	Condicionales
Punto de contacto para los metadatos	Condicionales

Se debe resaltar que los atributos que se incluyan en el desarrollo del metadato, serán dependientes del entorno e información que se desea documentar, incrementando la accesibilidad, la preservación del dato original y el control de versiones posteriores [12].

Por otro lado, la NTC 4611 [19], referencia el desarrollo del metadato mínimo debe responder a preguntas como:

- ¿Qué producto es?
- ¿Cuál es la zona geográfica que enmarca el producto?
- ¿Quién me puede dar información del producto?
- ¿Cuál es la fecha del producto?

Por lo tanto esta norma técnica Colombiana adopta la norma internacional, obteniendo los mismos fines. Sin embargo, se realiza la clasificación de metadatos geográficos mínimos y detallados, dejando abierta la posibilidad de utilizar la estructura para documentar diferentes aplicaciones y entornos [19].

El metadato, mínimo debe contener los siguientes atributos [19]:

- Información de la citación (Obligatorio).
- Descripción (Obligatorio).
- Dominio Espacial (Obligatorio).
    - Extensión geográfica (Obligatorio).
    - Límites geográficos (Obligatorio).
- Descriptores (Obligatorio).
- Calidad de los datos (Condicional).
- Distribución (Condicional).
- Citación (Obligatorio).
- Contacto (Obligatorio).

3.2.3. Arqueología de datos

El termino de Arqueología de datos no es común dentro de los términos informáticos, sin embargo Según [20], “el concepto de ‘arqueología de datos’ se usa para describir el proceso de búsqueda, restauración, evaluación, corrección e interpretación de conjuntos de datos históricos”.

Otra definición que se conoce, expresa que el término de arqueología se refiere a un “proceso de identificar, restaurar, evaluar, corregir, recuperar e interpretar archivos históricos de datos oceanográficos, a fin de que no se pierdan para la comunidad científica”. Esta definición es citada por metodología archivística para

la recuperación de información oceanográfica del pacífico colombiano, la cual fue tomada de [21].

Esta información fue recopilada por [7], en donde se documentan diferentes estudios realizados en la historia acerca de arqueología de datos.

3.2.4. Data Mining

El data mining ha tenido diferentes definiciones complejas pero la más cercana al tratamiento de los datos que se busca es la propuesta por [22] citando que es "la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión", pero no solamente lo deja en definición si no que plantea diferentes ejemplos para su uso.

Data mining según [23] “se refiere al análisis de datos y herramientas computacionales (software) en la búsqueda de características, reglas y regularidades en un gran conjunto de datos” y [24] dice que “es la extracción de patrones o información interesante (no trivial, implícita, previamente desconocida y potencialmente útil) de grandes bases de datos”.

3.2.5. Data Cleansing

Según [25], data Cleansing se define como el proceso de eliminación de los errores y las inconsistencias en los datos y la resolución del objeto identificado.

El proceso de data Cleansing posee una relación directa con la adquisición de datos y al cómo se puede mejorar la calidad de estos, por lo tanto [25], [26] define los siguientes pasos como un proceso de depuración de datos.

- Definir y determinar el tipo de error.
- Buscar e identificar casos de error.
- Corregir los errores al descubierto.

3.2.6. Data Ladder

DataLadder es una compañía que ofrece su software de calidad de datos a los usuarios para optimizar sus negocios y obtener el mayor provecho a sus bases de datos mediante las herramientas que proporciona la entidad. Este software utiliza procesos de lógica difusa, tecnología semántica y herramientas de enriquecimiento para calidad de datos. [27].

Es por esto que DataLadder nace como una solución a las necesidades que presentan las empresas que alimentan y almacenan bases de datos, ya que estas implementan herramientas que no les genera una respuesta rápida y efectiva, subestimando el alcance de sus problemas. Por el otro lado DataMatch se posiciona como una plataforma sólida asegurando mejoras en la productividad, optimización de tiempo labora, satisfacción del usuario y logrando que los datos inconsistentes sean utilizables en tiempos de respuesta cortos [28] [29].

Una de las funciones por las cuales DataLadder es una de las herramientas escogidas para la limpieza de datos es, la eliminación de duplicados. Esta función permite identificar y depurar información igual o sospechosamente similar que se genera por la combinación de bases de datos y errores en la digitación de información, convirtiéndolo en una tarea compleja y demorada para el usuario [29].

A continuación, serán descritos algunos de los pasos utilizados por la herramienta de DataLadder para la eliminación de duplicados.

- **Combinación de Bases de datos**

Inicialmente, la herramienta identifica que tipos de fuentes y archivos se encuentran en la base de datos, por lo que DataMach brinda opciones para importar, combinar y exportar la información en un formato único y compatible en un archivo; Esta función la realiza mediante el reconocimiento de campos similares, comparando diferentes fuentes entre sí. [29]

- **Duplicado**

Para la identificación y eliminación de información de información, la herramienta utiliza las siguientes prácticas que facilitaran los procesos:

  - Utilización de lógica difusa para definir los valores límite.
  - Codificación IBM para su coincidencia.
  - Limpieza y normalización de datos. [29]

- **Supervivencia**

Uno de los problemas que se presenta cuando se realiza eliminación de duplicados de información es la decisión que se debe tomar con respecto a los datos que fueron identificados anteriormente. Por esta razón DataMach brinda una herramienta que permite decidir si, crear, eliminar o combinar información en una única celda con el fin de limpiar o complementar los atributos de la base de datos [29].

Aplicaciones

DataLadder ha desarrollado diferentes proyectos que funcionan como ejemplo para identificar los alcances que tiene la herramienta. Algunos de estos proyectos son:

- TRACKING RECORDS ACROSS DATABASES: Por medio de la herramienta se combinaron dos bases de datos con información complementaria acerca de condiciones médicas y registros actuales de la universidad de Virginia. Además la universidad logro identificar cambios en los registros y realizar seguimiento a los expedientes médicos en tiempo real, ya que incorporó la limpieza de datos dentro de su sistema [30].

- FOR DATA DEDUPLICATION NEEDS, SIZE MATTERS: La empresa Buckle, deseaba tener un sistema para solucionar sus problemas en la base de datos. Encontrando en DataMatch la solución para desduplicar la información de sus compradores [31].

- DATA MATCHING TOOLS PROVIDE ADDED VALUE FOR TECHNOLOGY CONSULTANTS: EDP CONSULTING INC es una empresa dedicada a brindar software que permita solucionar problemas a diferentes compañías, sin embargo uno de sus clientes solicitaba trasladar su información a un nuevo sistema, por lo que la empresa tuvo que recurrir a DataMatch para iniciar con la limpieza y depuración de la base de datos, con el fin de extraer toda la información con un formato compatible [32].

- MASTERING AND MAINTAINING DATA QUALITY: AMEC es una empresa dedicada a la industria de la ingeniería ambiental, especializada en consultoría, ingeniería y servicios de gestión de proyectos ambientales e infraestructura. Esta empresa necesitaba migrar su información financiera y de recursos personales para una superficie que soportara una mayor magnitud de información, por lo que encontró en las herramientas de DataMatch la solución para la deduplicación de información y limpieza de sus datos [33].

- FUZZY MATCHING TOOL GIVES PROVIDES UNEXPECTED PROFIT CENTER: QRP es una empresa dedicada a las impresiones comerciales y posee bases de datos con bastante información correspondiente a un mismo usuario, por lo que necesitaba herramientas para combinar y eliminar toda la información innecesaria. DataMatch por medio de sus herramientas, brindó a QRP la posibilidad de combinar y unificar sus clientes con información de contacto actualizada y verificada [34].

- FUZZY MATCHING MAKES DEDUPLICATION A BREEZE FOR POWER EQUIPMENT RETAILER: Arlington es una empresa dedicada a suministrar piezas y mantenimiento de energía, por lo que su servicio fue creciendo y su base de datos de clientes también. Con el fin de limpiar toda la información recolectada en los años de funcionamiento, se utilizó las herramientas de DataMatch para eliminar información duplicada y generar una base de datos funcional para las ventas y servicios que ofrece la compañía [35].

3.2.7. Outliers

Existen diferentes definiciones para los valores atípicos presentes en bases de datos y otros campos en donde son utilizados, sin embargo, las definiciones siempre se encuentran orientadas a que son observaciones inconsistentes, discordantes, contaminantes, que difieren del conjunto de datos al que pertenecen [36].

Sin embargo, en esta oportunidad se usará la definición utilizada por [37], citando que la identificación de valores atípicos es el "problema de encontrar patrones de datos que no se ajustan al comportamiento esperado" .[36], asumiendo que en este proceso podemos encontrar situaciones en las cuales se tengan sesgos para su registro [38].

Por esta razón, actualmente existen diferentes métodos utilizados en diversos campos de acción y su elección, dependerá inicialmente del estado de normalidad de los datos que se encuentren en análisis; cabe resaltar que por medio de investigación, en la actualidad existen diferentes tipos de valor atípico que será característico de la situación que se esté evaluando [36] [39]. A continuación se describirán tres métodos para identificar valores atípicos.

- ***Prueba de Grubb's***

La prueba de Grubb's y también llamado como el método ESD (extreme studentized deviate) [36] tiene como objetivo, identificar valores atípicos en un conjunto de datos, que fue distribuidamente analizado anteriormente y presenta un comportamiento normal, siendo este el principal requisito para el uso de esta prueba.

Inicialmente, el procedimiento para la identificación los valores atípicos o también conocidos como "outliers", se realiza para los datos extremos de la ordenación numérica hasta que no se encuentren más errores, sin embargo la reiteración de este procedimiento puede bajar la probabilidad de encontrar errores por problemas de enmascaramiento [39].

Según la guía metodológica para la elección de una técnica de depuración [36], basado en información analizada del libro [40]**,** el procedimiento de la prueba de Grubb's es el siguiente.

1. Ordenar los datos ascendentemente ($X_1 < X_2 < X_3 < \ldots < X_n$).
2. Preguntar y definir si X_1 o X_n son sospechosamente valores atípicos.
3. Calcular el promedio (*P*) y desviación estándar (*S*).
4. Se debe calcular Z según el valor sospechoso que fue elegido anteriormente.

$$\text{Si } X_1 \text{ es sospechoso } \rightarrow Z = \frac{P - X_1}{S}$$

$$\text{Si } X_n \text{ es sospechoso } \rightarrow Z = \frac{X_n - P}{S}$$

5. Se debe elegir el nivel de confianza considerado para la prueba y calcular T, este valor debe ser comparado con la tabla estandarizada por la prueba. Si este valor es mayor al valor que arroja la tabla, se considera un valor atípico.

Sin embargo otros autores proponen y utilizan una ecuación de comparación directa para identificar valores atípicos por medio del método [38] [39]. La ecuación es la siguiente:

$$Z > \frac{N-1}{\sqrt{N}} - \sqrt{\frac{T^2\left(\frac{\alpha}{2N}\right), N-2}{N-2+T^2\left(\frac{\alpha}{2N}\right), N-2}}, Siendo\ \alpha\ el\ nivel\ de\ significancia$$

Los expertos aseguran que la prueba es fácil de usar pero se deben considerar los requisitos y condiciones, dentro de ellas se debe tener en cuenta que el tamaño de la serie de datos no debe ser grande y que su distribución debe ser normal. [36].

- ***Prueba de Dixon***

La prueba de Dixon fue creada para la determinar e identificar si un valor ubicado en los extremos es un valor atípico o también llamado "outlier". Para la aplicación de esta prueba se debe tener en cuenta como requisito que la distribución de la serie de datos debe presentar normalidad y que la cantidad de datos debe estar en un rango entre 3 – 24 [36].

Según la guía metodológica para la elección de una técnica de depuración [36], basado en información analizada del libro [40]**,** el procedimiento de la prueba de Dixon es el siguiente.

1. Ordenar los datos ascendentemente ($X_1 < X_2 < X_3 < \ldots < X_n$).
2. Preguntar y definir si X_1 o X_n son sospechosamente valores atípicos.
3. Determinar la relación a de acuerdo al número de datos., utilizando la siguiente tabla (tomada de [36]).

Tabla 2. Relación de acuerdo al número de datos para el desarrollo de la prueba de Dixon.

Numero de datos	Relación (r) a calcular
3- 7	r_{10}
8 – 10	r_{11}
11 – 13	r_{21}
14 -24	r_{22}

4. Con la ayuda de la siguiente tabla (tomada de [36]), calcular el valor de Dixon.

Tabla 3. Ecuaciones solución para la prueba de Dixon.

r	Si X_n es sospechoso	Si X_1 es sospechoso
r_{10}	$\frac{X_n - X_{n-1}}{X_n - X_1}$	$\frac{X_2 - X_1}{X_n - X_1}$
r_{11}	$\frac{X_n - X_{n-1}}{X_n - X_2}$	$\frac{X_2 - X_1}{X_{n-1} - X_1}$
r_{21}	$\frac{X_n - X_{n-2}}{X_n - X_2}$	$\frac{X_3 - X_1}{X_{n-1} - X_1}$
r_{22}	$\frac{X_n - X_{n-2}}{X_n - X_3}$	$\frac{X_3 - X_1}{X_{n-2} - X_1}$

5. Se debe comparar el valor resultado obtenido de acuerdo al nivel de significancia en la tabla de valores atípicos de la prueba de Dixon; si este resultado es mayor que el valor de la tabla, se cataloga como valor atípico.

Para el uso de la metodología, se tiene que tener en cuenta los requisitos nombrados anteriormente y que pueden existir problemas de enmascaramiento. Si estos pasos e hipótesis son ejecutados correctamente, la metodología puede funcionar fácilmente para grupos de datos pequeños [36].

- ***Prueba de Tukey***

El diagrama de cajas o también llamado Boxplot, es un método que se considera como resumen de la información que se somete a un análisis, en el cual se puede identificar la variabilidad, tendencia, simetría y valores atípicos de un conjunto de datos [41].

El diagrama de cajas debe contener las siguientes medidas descriptivas de los datos:

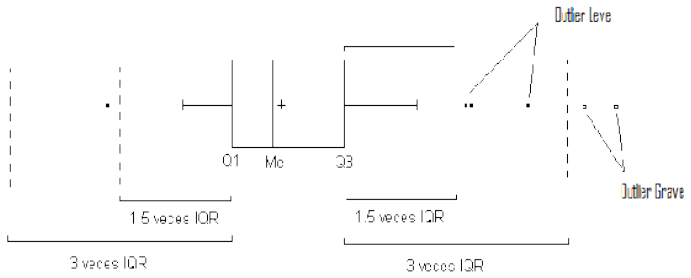
- Mediana.
- Primer cuartil (Q_1).
- Tercer cuartil (Q_3).
- Valor máximo.
- Valor mínimo

En donde el 50 % de los datos se verá representado en una caja o rectángulo, compuesto por Q_1, la mediana y Q_3; Los valores extremos serán considerados las cotas o bigotes [36] [42].

Para que la prueba pueda identificar los valores atípicos del conjunto de datos, se debe condicionar la longitud de los bigotes, esta operación se obtendrá multiplicando 1,5 o 3, la diferencia del tercer y primer cuartil. Los valores que se encuentren dentro del rango (1,5 y 3), serán considerados valores atípicos leves y los datos que se encuentren por encima de 3 veces el rango intercuartil, son considerados valores atípicos extremos [36], [42], [41].

Además del grafico de Boxplot, existen diagramas que pueden ayudar a identificar valores atípicos, entre ellos encontramos los de dispersión, sin embargo si se desea tener un resultado exacto, se deberá implementar un método numérico que sustente la determinación de valores atípicos u outliers [36].

Gráfico 1. Descripción del diagrama de cajas o Boxplot.



Tomado de la Guía metodológica para la selección de técnicas de depuración de datos. [36]

3.2.8. Quality Flag

Las banderas de calidad son una herramienta que permite limpiar los datos originales que no cumplen con estándares de calidad, seguridad y fiabilidad; generando interrogantes que deben ser ejecutados durante el tratamiento de la información primaria para su documentación final [43].

El uso de las banderas de calidad permitirá al administrador de los datos.

- Fusionar diferentes conjuntos de datos.
- Retener y almacenar información con la calidad original.
- Añadir y documentar información adicional acerca de la calidad pasada y actual de los datos.
- Aceptar o rechazar la información definitiva dependiendo de la finalidad a la cual será expuesta dicha información.
- Comparar su información original y final con otros registros nacionales e internacionales que cuenten con banderas de calidad, permitiendo añadir nuevas prácticas de calidad y resultados.

Según [44], se debe ter presente que no se deben generar banderas de calidad para los datos temporales y geográficos, así como la columna destinada a registrar las observaciones que se presentan en campo; esta medida es tomada porque “no se permite que esta información sea cuestionable o mala”.

Se debe aclarar que el procedimiento y uso de las banderas de calidad para todas las variables, cambiara de acuerdo al proyecto y objetivo de este. Por esta razón se debe identificar y analizar cuáles de las banderas existentes se acopla a las necesidades, ofreciendo al usuario un criterio de calidad para su uso. [43], [45]

Propuestas de banderas de calidad

- ***Esquema de dos niveles (IOC – UNESCO)***

El esquema consiste en asignar banderas de calidad a los datos para facilitar el intercambio y documentación de la información. Consiste en dos niveles y su uso dependerá al nivel de complejidad que necesite el usuario.

- Primer nivel

Tabla 4. Banderas de calidad de primer nivel según la IOC [43].

Valor	Nombre de bandera	Definición
1	Buena	Pruebas de control de calidad requeridas, aprobadas
2	No evaluada, no disponible o desconocido	Se utiliza cuando no hay pruebas de calidad o la información de calidad no está disponible
3	Cuestionable / Sospechosa	Cuando se presenta un error que no es crítico en métrica o prueba de análisis
4	Malo	Error critico definido por la prueba de calidad o asignado por el proveedor de los datos
9	Datos faltantes	Es el indicador de que el dato es faltante o no existente

- Segundo Nivel

Las banderas de calidad del nivel secundario se utilizan para complementar las del primer nivel, ya que son los resultados de las pruebas cuantitativas de calidad ejecutadas a los datos y todo el proceso histórico del procesamiento de la información. Se pueden utilizar a manera de ejemplo los picos máximos de datos, verificación de desviación y distribución, valores interpolados y valores corregidos.

Se debe aclarar que la aplicación de banderas de calidad de primer nivel, no obliga condicionalmente la aplicación de banderas de calidad de segundo nivel.

- ***ODV***

El modelo de banderas de calidad utilizado por el Ocean Data View (ODV), es el adoptado por el repositorio colombiano de datos oceanográficos, debido a su sencillez y amplio campo de implementación [44], [45].

Tabla 5. Banderas de calidad utilizadas por el Ocean Data View (ODV).

Código	Descripción
0	Bueno
1	Desconocido
4	Cuestionable
8	Malo

- ***QARTOD***

Este modelo fue tomado del manual y guía del Ocean Data Standards [43], lo cual es una propuesta generada en el año 2010, pero no es definitiva.

Tabla 6. Banderas de calidad de QARTOD

Código	Descripción
0	Calidad no evaluada
1	Malo
2	Sospechoso – cuestionable
3	Bueno
9	Dato faltante

Finalmente, no existe un estándar internacional que se deba seguir, sin embargo, el uso de las guías estandarizadas por organismos internacionales, permitirá obtener beneficio en la investigación, utilizando un conjunto de datos determinado.

3.2.9. R Project

R como software y herramienta es un lenguaje y un entorno computacional, que integra gráficos estadísticos; sin embargo, R dentro de su panel de ofrecimientos, tiene variedad de formas de presentar la información que se esté trabajando, de acuerdo a la complejidad requiera el usuario. Esto es permitido porque R maneja un lenguaje de programación compatible con otras plataformas existentes [46], [47]**.**

Ya que R es un entorno y un conjunto integrado de servicios de software, dedicado a la manipulación y análisis de información, se identificaron algunas herramientas que son aprovechables en su plataforma [46]. Estas herramientas son:

- Pruebas estadísticas clásicas.
- Análisis de series temporales.
- Clasificación y agrupación de información.
- Modelamiento de datos.
- Ingreso y salida rápida de información.
- Operación de vectores y matrices.
- Conjunto de operaciones para la manipulación de vectores y matrices.

Sin embargo R es una plataforma abierta y flexible para extender su catálogo, permitiendo añadir funcionalidad con la adición de nuevas funciones, métodos de análisis estadístico y múltiples aplicaciones [46], [48].

Es por esto que, usualmente los usuarios del software consideran que es un sistema netamente estadístico pero los creadores prefieren catalogarlo como un entorno en el cual se pueden agregar paquetes que complementan los procesos estadísticos. Cabe resaltar que el software y los paquetes adicionales son de libre descarga y pueden ser utilizados en cualquier plataforma o sistema operativo [46], [48].

Para lograr un buen aprendizaje y una aplicación efectiva del software, este cuenta con manuales y guías de introducción, desarrollo y aprovechamiento de todas las funciones y complementos que pueden ser ejecutados por R [48].

Aunque para algunos de investigadores y expertos, R agrupa características que emiten madurez, manejabilidad, recursos y seguridad, que se convierten en ventajas para ser el software elegido para el desarrollo de sus proyectos [49], otros investigadores consideran que el manejo es complejo [48]; concluyendo que existen muchas técnicas y métodos que abarcan todos los software estadísticos para el manejo de datos e información, pero se debe estudiar y trabajar profundamente en las capacidades que puede brindar una determinada metodología [47]**.**

Las principales técnicas de minería de datos incluyen la clasificación y predicción, la agrupación, detección de valores, reglas de asociación, análisis de secuencias, análisis de series de tiempo y minería de texto, y también algunas nuevas técnicas como el análisis de redes sociales y análisis de los sentimientos. [50]

Como se mencionó anteriormente, R puede complementarse con muchos paquetes que le ayudan a expandir su campo de acción. Es por esto que en el año 2012, por medio del libro “R and Data Mining”, se realizó un análisis y agrupación de información con la cual se pueden realizar procedimientos de limpieza de datos [50].

En este libro se mencionan las principales técnicas de minería, depuración y limpieza de datos [50]. Algunas de estas técnicas son:

- Clasificación y predicción.
- Detección de valores atípicos.
- Reglas de asociación.
- Análisis de secuencias.
- Análisis de series de tiempo.
- Minería de texto (se encuentran artículos en donde se desarrollan procesos estadísticos, lingüísticos e informáticos con fines de análisis, recuperación y depuración de información). [51], [52]
- Análisis en redes y sensibilidad.

La minería, depuración y limpieza de datos utilizando R como plataforma de apoyo y análisis, es un tema que se viene aplicando actualmente, sin embargo requieren conocimientos mínimos y previos para conseguir y optimizar los resultados. Es por esto que se realizan cursos de capacitación y práctica, con el fin de apropiar el conocimiento y aplicarlo en diferentes campos de acción.

3.2.10. Análisis multicriterio (AHP)

El método AHP es considerado una metodología que permite apoyar en la valoración cuantitativa a los tomadores de decisiones, generando escenarios con diferentes alternativas u opciones, sustentados en operaciones matemáticas. El resultado obtenido, tiene la posibilidad de ser satisfactoria para el usuario pero también puede ser insatisfactoria y equivocada [53],[54].

Por esta razón, algunos autores consideran que el método se caracteriza por ser flexible, ya que su contribución se encuentra operativamente, tácticamente y estratégicamente [54], [55]. Sin embargo, algunas ventajas atribuidas al uso de este método dentro de los proyectos son:

- Organización jerárquica del problema [56].
- Permite gestionar adecuadamente la información necesaria para la ejecución de las tareas [56].
- Asignación de pesos del decisor y otras partes, utilizando el método de Saaty [56].
- Comparaciones globales de las alternativas [56].
- No se requiere información cuantitativa para el desarrollo del método [56].
- Detecta y acepta o rechaza el método por error [56].
- Análisis por separado de todas las alternativas [54].
- Optimización de la eficiencia, eficacia y efectividad de los sistemas analizados [55].
- Adaptabilidad a todo tipo de sector [57].

Sin embargo, como desventaja para la ejecución del método, solo se pueden comparar entre 2 y 7 alternativas; esta escala es de acuerdo al número de Miller, pero existen estrategias que facilitan realizar conversiones para que el método sea aplicable [54].

A continuación será descrito el procedimiento para el desarrollo del método [53],[54].

1. Construcción de jerarquía de atributos, que deberá contener mínimo:
   i. Objetivo general, meta o propósito.
   ii. Criterios de evaluación.
   iii. Todas las posibles alternativas identificadas.
2. Construir y desarrollar la matriz de comparación por pares (MCPA). Consiste en evaluar por parejas las alternativas identificadas anteriormente, utilizando la escala de Saaty [56].

Tabla 7. Valores según la escala de Saaty.

Valor	Descripción
1	Igual importancia para dos criterios
3	Débil importancia de uno sobre el otro
5	Importancia esencial o fuerte de un criterio sobre otro
7	Importancia demostrada de un criterio sobre otro
9	Importancia absoluta de un criterio sobre otro

Se debe tener en cuenta que es posible asignar valores intermedios y en la evaluación de la alternativa contraria, se debe aplicar el valor reciproco de la evaluación inicial.

3. Desarrollar la matriz normalizada (MCN). Se obtiene dividiendo cada valor de la matriz (MCP), por la sumatoria de la columna a la cual pertenezca.
4. Generar el vector de prioridad para cada criterio, determinando el promedio de cada fila de la MCN correspondiente.
5. Determinar el coeficiente de consistencia (RC). Si este valor es menor al 10%, se considera aceptable, pero si es mayor, se deben reconsiderar los juicios y valores consignados en la MCP. [54]

Nota: Para determinar el coeficiente de consistencia se debe seguir el siguiente procedimiento.

i. Generar una matriz (NMax) que resulta de la sumatoria ponderada del producto de la primera celda de la MCP y el primer valor de la vector prioridad. Se debe seguir el procedimiento de multiplicación de matrices.
ii. Determinar el índice de consistencia, siendo (α), la sumatoria de los valores de la NMax y (n) el número de alternativas o criterios.

$$IC = \frac{\alpha - \mathrm{n}}{n - 1}$$

iii. Identificar el índice aleatorio (IR), de acuerdo a la siguiente tabla [53].

Tabla 8. Valores para determinar el índice aleatorio (IR).

(n) Alternativas	Índice Aleatorio (IR)
3	0.58
4	0.90
5	1.12
6	1.24
7	1.32
8	1.41

iv. Determinar el coeficiente de consistencia (CR).

$$CR = \frac{IC}{IR}$$

6. Elaborar una matriz prioridad (MP), la cual contendrá las alternativas por fila y los criterios por columna.
7. Construir la matriz de comparación por pares de criterios (MCPC)

8. Elaborar un vector prioridad global, multiplicando el vector de prioridad de la MCPC por la MP de las alternativas.

El valor que tenga mayor valor, será la recomendada a partir de ejecutar el método AHP, sin embargo se pueden realizar análisis por criterio, permitiendo evaluar todas las alternativas propuestas.

3.3. MARCO CONTEXTUAL

El Instituto de Investigaciones Marinas y Costeras “José Benito Vives de Andréis” – INVEMAR, “realiza investigación básica y aplicada de los ecosistemas marinos de interés nacional con el fin de proporcionar el conocimiento científico necesario para la formulación de políticas, la toma de decisiones y la elaboración de planes y proyectos dirigidos al manejo sostenible de los recursos, la recuperación del medio ambiente marino y cotero y al mejoramiento de la calidad de vida.” [58]

Para cumplir con sus funciones el instituto cuenta con cuatro programas de investigación y dos coordinaciones.

- Programa BEM - Biodiversidad y Ecosistemas Marinos.

Las líneas de investigación de éste programa son:

1. Línea de inventarios, taxonomía y biología de especies – ITE.
2. Línea de organización y dinámica de ecosistemas – ODI.
3. Línea de biología y estrategias de conservación – BEC.

- Programa CAM - Calidad Ambiental Marina.

Las líneas de investigación de éste programa son:

1. Línea de evaluación y seguimiento de la calidad ambiental – ESC.
2. Línea de prevención y protección de los ecosistemas marinos y costeros – PEM.
3. Línea de rehabilitación de ecosistemas marinos y costeros – RAE.

- Programa GEO - Geociencias Marinas y costeras.

Las líneas de investigación de éste programa son:

1. Línea de geología marina y costera - GMC.
2. Línea de oceanografía y clima - OCC.
3. Laboratorio e instrumentación marina - LABIMA.

- Programa VAR - Valoración y aprovechamiento de Recursos Marinos y costeros.

Las líneas de investigación de éste programa son:

1. Línea de valoración económica – VAE.
2. Línea de uso y producción sostenible – UPS.
3. Línea de bioprospección Marina – BIM.

- GEZ - Coordinación de Investigación e información para la Gestión Marina y Costera. Esta cuenta con:

1. Línea de análisis de información para la planificación.
2. Línea de cambio global y política marina – CGP.
3. Laboratorio de servicios de información – LABSIS.
4. Dependencia de comunicación científica – CMC.

- CSC - Coordinación de Servicios Científicos.

Dentro de las funciones de INVEMAR está realizar estudios "relacionados con la fijación de parámetros sobre emisiones contaminantes, vertimientos y demás factores de deterioro ambiental que puedan afectar el medio ambiente marino, costero e insular o sus recursos naturales renovables" [59], también cumple múltiples funciones alrededor de la información marítima y oceanográfica desde su obtención hasta su divulgación, además del manejo, seguimiento y aprovechamiento de los recursos y ecosistemas marinos para su protección y conservación.

Incluye procesos de investigación con apoyo y en coordinación con otras entidades como la Comisión Colombiana del Océano – CCO en el desarrollo de actividades de interés marino, La Dirección Marítima Nacional – DIMAR como proveedor de datos e información oceanográfica para la realización de estudios de interés científico marino, el Instituto de Hidrología, Meteorología y Estadios Ambientales – IDEAM en relación al manejo de información como el seguimiento de variables fisicoquímicas y ambientales para estudiar el comportamiento de medio y sus procesos, el Instituto de Investigación de Recursos Biológicos– "Alexander Von Humboldt" con todo lo relacionado al inventario de flora y fauna marino y la generación de estudios, el Ministerio de Medio Ambiente – Ministerio de Ambiente brindando información que permita evaluar el impacto ambiental de proyectos, el desarrollo de actividades, la definición de indicadores, modelos predictivos, entre otros, que intervengan en el medio marino costero y los recursos; además de otras instituciones que tienen como fin el fortalecer las políticas de interés nacional en relación a los recursos y ecosistemas marinos y coteros para el desarrollo sostenible del país.

4. DESARROLLO DE LA PASANTÍA

En el desarrollo de la pasantía fueron realizadas diferentes tareas, las cuales estuvieron encaminadas al manejo y estandarización de información en bases de datos, apoyo en la organización y ejecución del curso- taller del programa CAM – INVEMAR, análisis de información relacionada a fuentes de contaminación marina de los municipios costeros del departamento de Sucre y trabajo de campo para dar apoyo a proyectos que se ejecutaron por investigadores del instituto.

A continuación, será presentado un desarrollo de la metodología planteada para dar cumplimiento a los objetivos de este documento.

4.1. Capítulo 1. Construcción de metadatos

Para el cumplimiento del primer objetivo, se inició con la revisión de información secundaria, que permitiera contextualizar y orientar las demás actividades propuestas. Por lo tanto, a partir de la revisión realizada, el acompañamiento de la investigadora del programa CAM y el personal de apoyo del LABSIS, se realizó un documento en formato xls que contiene los atributos básicos que se deben desarrollar en el metadato para los muestreos realizados por la REDCAM en el año 2015.

Después del diseño del formato, se inició con la búsqueda y compilación de información primaria de los muestreos realizados en el año 2015, utilizando los planes de muestreo para cada departamento y los formatos establecidos por el instituto para el diligenciamiento de la información analizada en campo y en los laboratorios participantes. Se debe resaltar que los departamentos de Bolívar, Valle del Cauca, San Andrés y Antioquia, realizan análisis de variables en los laboratorios acreditados por sus corporaciones, por lo tanto se debe recopilar información de los reportes primarios de estos muestreos.

Dentro de los documentos utilizados para esta tarea, se encuentra el FT-MEE-002, correspondiente a las variables y métodos del LABCAM 2015; el registro MEE, correspondiente al control de entrega de datos externos; los planes de muestreo para los departamentos de Córdoba, Sucre, Magdalena, Chocó, Nariño, Cauca y Atlántico; la oferta de servicios de laboratorio de agua de CORPOURABA; la consulta realizada a la base de datos el 16 de Junio de 2016 y corroborado con los formatos FT-MEE-002 para cada departamento.

Adicionalmente, se tuvo soporte y acompañamiento por parte de la Unidad de Laboratorios de Calidad Ambiental Marina – LABCAM, para corroborar la información digitalizada y la solución de dudas acerca de los datos que se encontraban en estos registros.

Finalmente, utilizando la información existente de los planes de muestreo y el formato único de registro de datos del instituto, se realizaron los metadatos para los departamentos mencionados a continuación. Un ejemplo de ellos se encuentra en la Tabla 9. Metadato realizado para el muestreo realizado en el segundo semestre del año 2015 para el departamento del Magdalena.Tabla 9 del documento.

- San Andrés
- La Guajira
- Magdalena
- Atlántico
- Bolívar
- Sucre
- Córdoba
- Antioquia
- Choco
- Cauca
- Valle del Cauca
- Nariño

4.2. Capítulo 2. Aplicación de una metodología de limpieza de datos

La primera actividad de la metodología propuesta se basa en la identificación de errores y vacíos en una de las variables que se registran en la base de datos. Es por esto que la tarea se inició con la convocatoria a los investigadores del programa CAM y el personal de apoyo del LABSIS , que estuvieran vinculados al suministro y análisis de información.

A partir de la identificación de las necesidades, se definieron los criterios teniendo en cuenta la revisión de información secundaria realizada, alusiva a elección de metodologías y software para la solución de necesidades particulares; la definición de estos criterios se encuentra desarrollada en el numeral 5.2.2 Definición de criterios del capítulo de resultados. Posteriormente se realizó la búsqueda y análisis de información de las metodologías, que permitiera calificar los criterios y seguir con la elección de una de ellas, utilizando el análisis multicriterio AHP.

Se continuó con un ejercicio práctico en el cual se establecieron las banderas de calidad para la concentración, método analítico y límite de detección, ya que, según las necesidades identificadas, estos atributos son los que presentan los problemas en la base de datos identificaron errores.

Finalmente, se utilizaron las banderas de calidad en un conjunto de datos, con el fin de brindar una opción para la limpieza de información de la base de datos del sistema de información REDCAM.

5. RESULTADOS OBTENIDOS

5.1. Capítulo 1. Construcción de metadatos

5.1.1. Proponer estructura de metadato

Con el acompañamiento de la administradora de la base de datos y personal del LABSIS, se elaboró el formato (Tabla 9) con el cual se elaboraron los metadatos. Este formato contiene los siguientes atributos:

- Título: Deberá contener la información necesaria y concreta, que oriente al usuario acerca de la información que incluye el metadato.
- Fecha de publicación del metadato: Esta fecha indicará, cuando el autor realizo el metadato.
- Cítese como: Se debe realizar una propuesta de citación del metadato; se recomienda incluir las personas que realizaron el muestreo.
- Cobertura temporal: Debe incluir las fechas en las cuales el muestreo fue realizado.
- Cobertura geográfica: Debe indicar la zona en la cual fue realizado el muestreo; para programa REDCAM, no es necesario documentar las coordenadas de las estaciones, ya que estas ya se encuentran almacenadas.
- Resumen: Para los muestreos realizados en el marco del programa REDCAM, este deberá contener.
    - Clasificación por sustrato agua y sedimento si fue realizado.
    - Numero indicativo.
    - Variable medida.
    - Unidades de medida.
    - Método de análisis en laboratorio.
    - Límite de detección correspondiente a cada método.
- Observaciones: Deberá contener cualquier inconveniente operativo, técnico y logístico que se haya presentado en campo, lo cual dificultó la toma del dato.
- Estaciones de monitoreo. Contendrá el código de las estaciones que fueron incluidas en el muestreo documentado.
- Palabras claves.
- Personas de contacto: Se debe diligenciar la información de contacto del responsable de la salida, el responsable del programa y el autor del metadato.

5.1.2. Construcción de metadatos

Se realizaron 22 metadatos correspondientes a los muestreos realizados en el año 2015 para los departamentos costeros de Colombia. A continuación se mostrará la estructura del metadato realizado para el segundo muestreo del departamento del Magdalena en el año 2015 (Tabla 9).

Adicionalmente, con la revisión y compilación de información, se realizó un diagnóstico (ejemplo, Tabla 10) en el cual se identificó que no se cuenta con la totalidad de información necesaria para la documentación completa de los muestreos, por esta razón se debe realizar la búsqueda y compilación de información por parte de los nodos integrantes del programa y se recomienda iniciar con la organización y desarrollo de las etapas de arqueología de datos.

Tabla 9. Metadato realizado para el muestreo realizado en el segundo semestre del año 2015 para el departamento del Magdalena.

Titulo	Conjunto de datos resultado del muestreo realizado en el Departamento A en el segundo semestre de 2015
Fecha Publicación Metadato	11/07/2016
Cítese como:	Juan Garzon., 2016, Conjunto de datos resultado del muestreo realizado en el Departamento A en el segundo semestre de 2015, INVEMAR, Santa Marta.
Cobertura temporal	21 y 22 de Septiembre de 2015
Cobertura geográfica	Departamento Costero de Colombia
Resumen	Se recolectaron y analizaron muestras de aguas marinas y costeras para n número de estaciones, tres (3) de ellas en sedimentos. Las muestras fueron procesadas y analizadas en el laboratorio de calidad de aguas. Los parámetros y métodos de análisis utilizados fueron los siguientes:
	Campos del documento [número consecutivo. código variable - unidad de medida - método de análisis - límite de detección] Se realiza una descripción detallada de los métodos de laboratorio para cada tipo de sustrato que se vaya a contemplar SUSTRATO - AGUA 1.Variable A - Unidad de medida (b) - método de análisis 1 - límite de detección SUSTRATO SEDIMENTOS 1.Variable B - Unidad de medida (b) - método de análisis 2 - límite de detección

Observaciones	Por problemas de orden público, no se realizó el muestreo en la estación 1.		
Estaciones de Monitoreo	[Código estación 1] [Código estación 2]...[Código estación n] La distribución espacial de los datos puede consultarse en siam.invemar.org.co >> herramientas >> Proyectos y estaciones.		
Palabras Claves	REDCAM, departamento A, datos calidad de aguas y sedimentos		
Personas de Contacto	**Responsable salida**	**Nombre**	Yoselin Nieto
		Cargo	Investigadora Científica
		Teléfono (s)	5-4328600
		E-mail	yoselin.nieto@invemar.org.co
		Dirección	Calle 25 No. 2-55, Playa Salguero, Santa Marta D.T.C.H., Colombia
	Responsable REDCAM	**Nombre**	Lizbeth Janet Vivas
		Cargo	Jefe Línea PEM
		Teléfono (s)	5-4328600
		E-mail	janet.vivas@invemar.org.co
		Dirección	Calle 25 No. 2-55, Playa Salguero, Santa Marta D.T.C.H., Colombia
Autor		**Nombre**	Juan Sebastián Garzón Ariza
		Cargo	Pasante línea PEM
		Teléfono (s)	5-4328600
		E-mail	est.juan.garzon@invemar.org.co
		Dirección	Calle 25 No. 2-55, Playa Salguero, Santa Marta D.T.C.H., Colombia

Tabla 10. Ejemplo de diagnóstico de vacíos de información para la documentación de los muestreos realizados en el año 2015.

Departamento	Muestreo	Información faltante
Departamento A	1M	• Información de investigador que realizó el muestreo. • Consultar variables FQ y Mb (métodos y LD).
	2M	• No hay muestreo

5.2. Capítulo 2. Aplicación de una metodología

5.2.1. Identificación de necesidades

Por medio de la socialización por parte de la administradora de la base de datos del sistema REDCAM al estudiante, se identificaron las siguientes necesidades de

depuración en el conjunto de información para la variable de Coliformes Termotolerantes.

- Verificar la descripción realizada en el atributo de observaciones, contrarrestándolo con los formatos que contienen la información en “bruto”
- Rectificar los valores consignados como inferiores a los límites de detección de la técnica analítica.
- Revisar que todos los métodos analíticos estén codificados y registrados en la base de datos.
- Información de concentración sospechosa por ser valores atípicos de una variable determinada o por sobrepasar los límites de detección de las técnicas analíticas.

5.2.2. Definición de criterios

A continuación se realizará la descripción de los criterios que serán base para la evaluación cuantitativa, buscando identificar ventajas y falencias de las metodologías (Tabla 11).

- **Costo**

En el diseño de un proyecto es de vital importancia realizar un estimado del presupuesto que se va a destinar a este, convirtiéndose en el primer obstáculo [60]. Es por esto que los costos para la implementación de una metodología, varían dependiendo de las necesidades que se presenten [61].

Para la evaluación de costos de las metodologías de limpieza de datos, se tendrá en cuenta, el valor de la adquisición, mantenimiento y soporte técnico. El valor será asignado en una escala de 0 a 3, siendo 3 el menor costo y 0 el mayor.

- **Usabilidad**

La usabilidad se refiere a la capacidad de la metodología para que sea agradable para usuario, es decir que se va a evaluar que tan fácil será su aprendizaje, que tan agradable puede ser el manejo de su plataforma y que nivel de conformidad puede obtener el usuario con su uso.

Estos sub criterios serán evaluados de la siguiente manera

- Aprendizaje: El valor estará en una escala entre 2 y 0, siendo 2 fácil y 0 un aprendizaje difícil.
- Plataforma: El valor estará en una escala entre 2 y 0, siendo 2 fácil y 0 un entendimiento difícil.
- Conformidad de uso: El valor estará en una escala entre 2 y 0, siendo 2 buena y 0 una mala conformidad.

- **Funcionalidad**

Se evaluará la capacidad de ejecutar diferentes funciones, es por esto que las necesidades identificadas, se agruparán como subcriterios para asignarle un valor cuantitativo. Por lo tanto la importancia de la funcionalidad se evidencia en la evaluación de estos subcriterios porque se obtiene el margen de aplicación de cada metodología [62].

Se asignara un valor de (1) a cada subcriterio que pueda ser ejecutado por la metodología. Las necesidades fueron agrupadas en los siguientes subcriterios de funcionalidad:

- Identificación de vacíos e información sospechosa
- Comparación de columnas de información
- Depuración de información
- Corrección de información sospechosa
- Condicionar el ingreso de información futura

En la tabla 11, se presentan los valores asignados para los criterios definidos anteriormente; estos valores serán agrupados para posteriormente realizar la evaluación cuantitativa de las metodologías preseleccionadas.

Tabla 11. Valores asignados para la evaluación de los criterios.

	Criterios de elección			Valor
Valores de Ponderación	**Costo**	$ 0		3
		$ 1 - $ 1.000.000		2
		$ 1.000.001 - $ 100.000.000		1
		Más de $ 100.000.000		0
	Usabilidad	aprendizaje	Fácil	2
			Moderado	1
			Difícil	0
		Plataforma	Fácil	2
			Moderado	1
			Difícil	0
		Conformidad de uso	buena	2
			regular	1
			mala	0
	Funcionalidad	Identificación de vacíos e información sospechosa		1
		Comparación de columnas de información		1
		Depuración de información falsa		1
		Corrección de información sospechosa		1
		Condicionar el ingreso de información futura		1

5.2.3. Elección de la metodología

- **Análisis de información secundaria**

A partir de los de la definición de criterios de valuación, se inició la búsqueda de metodologías que permitieran realizar la limpieza de datos, para reducir la mayor cantidad de necesidades que se presentan en los registros de la base de datos, identificando un gran número de estas, pero se eligieron cuatro que tuvieran diferentes grados de dificultad y aplicación, con el fin de que pudieran ser comparadas a partir de los criterios establecidos.

A continuación, será presentada la descripción y experiencia obtenida al abordaje de las metodologías seleccionadas.

- Data Ladder

Este es un software líder en el mercado de la administración de bases de datos, siendo actualmente el soporte para la generación, manejo y representación de la información, generada por sectores y empresas que deben almacenar y soportar grandes cantidades de datos en tiempos cortos y largos.

Para obtener información de su plataforma, se tuvo contacto con empleados de la empresa vía correo electrónico. En esta comunicación se divulgaron las necesidades identificadas por los investigadores del instituto y se recibió la oferta ideal para la solución de algunos problemas. Sin embargo la empresa ofrece la posibilidad de utilizar una versión de prueba de su software, ya que el paquete de venta tiene un costo alto de adquisición.

A partir de la instalación de la versión de prueba del software y la revisión de casos de estudio publicados por la entidad, se analizó que para el manejo de las herramientas y procesos con los que cuenta la metodología, se necesita capacitación y acompañamiento continuo por parte de personal idóneo en el manejo del software; a pesar de que su plataforma es amigable, se necesitan conocimientos previos en manejo de bases de datos, soporte profesional y técnico continuamente. Por otro lado, los resultados obtenidos y divulgados por medio de los casos de estudio, siempre es positivo y recomiendan la vinculación de DataLadder dentro de sus organizaciones.

Cabe resaltar que las aplicaciones que fueron encontradas de la utilización de Data Ladder, se orientan hacia procesos de mercadeo, organización de información personal y financiero. Sin embargo no se descarta que la herramienta pueda vincular tareas en diferentes campos profesionales.

- R Project

Como lo definen los creadores de la herramienta, R Project es un lenguaje y entorno computacional [46], que ofrece muchas funciones aplicables a diferentes campos de acción. Sin embargo su manejo esta netamente codificado en lenguaje de programación, que no para todos los usuarios es de fácil aprendizaje y entendimiento.

Actualmente la limpieza, minería, y depuración de información se está implementando, con la utilización de R Project como base matemática y estadística, dando soporte para el análisis de bases de datos o series de información.

Para el desarrollo de estos campos de acción, se construyó y divulgo el libro “R and Data Mining” [50], en el cual se muestran todas las funciones y utilidades que se le pueden ejecutar al software. Sin embargo su contenido es denso para su aprendizaje y desarrollo pruebas con sets de datos, por lo que el autor del manual, ofrece cursos de capacitación dentro de su plataforma porque considera que se debe tener un conocimiento previo en lenguajes de programación y estadístico, para lograr resultados óptimos del contenido ejecutable del software.

Para explorar la plataforma de R Project y poder evaluar la funcionalidad y usabilidad, se descargó el software ya que es gratuito para cualquier sistema operativo; obteniendo una primera impresión negativa, ya que su presentación no tiene guías de procedimientos, sin embargo, el soporte técnico virtual de la herramienta, suministra manuales guía y listas de funciones que pueden ser ejecutables.

A partir de la revisión de información secundaria, se evidencio que los resultados generados a partir de las funciones que posee el software son positivos, brindando un amplio campo de herramientas que facilitan el análisis matemático y estadístico de información. Además en el libro [52], se evidenció que el software tiene la posibilidad de solucionar gran porcentaje de las necesidades identificadas por el instituto.

- Outliers

A partir de la revisión de información secundaria, se identificó que la identificación y eliminación de outliers o también llamados valores atípicos, funciona como estrategia de limpieza de datos de un conjunto de información. Para el desarrollo de esta metodología se puede utilizar cualquier software matemático o estadístico que integre funciones básicas, o se puede realizar de forma manual por el usuario.

Se encontraron tres pruebas existentes para identificación y eliminación de outliers, las cuales fueron documentadas en el marco referencial de este trabajo; sin embargo se realizó un análisis de acuerdo a las necesidades identificadas y se obtuvo que esta metodología posee un bajo porcentaje de cumplimiento de acuerdo a las problemáticas de la base de datos.

Por otro lado, estas pruebas son de fácil aprendizaje y entendimiento, lo cual facilita su implementación en conjuntos de datos y no requiere costo alguno para su implementación.

- Quality Flag

Según la revisión de información secundaria realizada, las banderas de calidad son una herramienta que permite identificar y optimizar la calidad de los datos. Para lograr este objetivo, las banderas de calidad pueden ser utilizadas en cualquier plataforma en la cual se manejen bases de datos, ya que se convierte en un atributo más de esta.

Para su implementación de esta metodología, se requieren conocimientos básicos en estadística y un manejo acorde de la plataforma en la cual se almacena la información, acompañado de idoneidad de la información que se maneja. Estos conocimientos previos, permitirán determinar cuáles serán las banderas y sus respectivos algoritmos por parte del usuario.

Siendo así, las banderas de calidad pueden brindar una solución a las necesidades identificadas por los administradores de la base de datos, dependiendo de la plataforma que sea utilizada para su codificación, se debe resaltar que de esta plataforma o software utilizado, dependerá el aprendizaje y su entendimiento final.

- **Calificación de metodologías**

A partir del análisis teórico-práctico realizado anteriormente de las metodologías seleccionadas y la definición de los criterios (numeral 5.2.2), el estudiante le asignó una calificación individual según los criterios establecidos (Tabla 12).

Tabla 12. Calificación de las metodologías según los criterios de evaluación

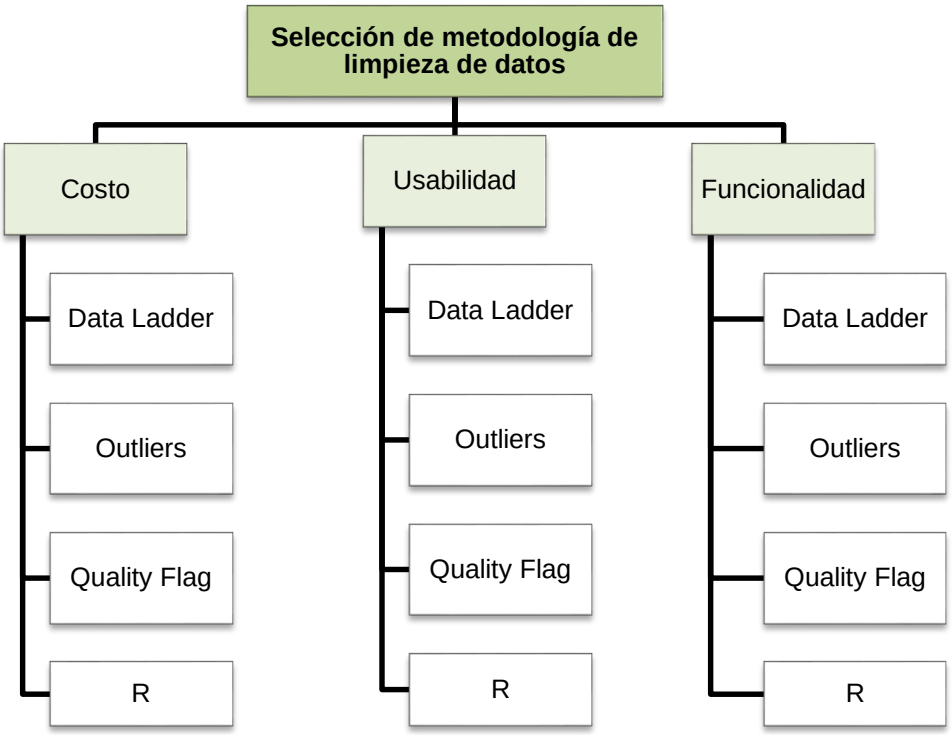
Criterios de elección			Metodología			
			Data Ladder	Outliers	Quality Flag	R
Costo	$ 0			x	x	x
	$ 1 - $ 1.000.000					
	$ 1.000.001 - $ 100.000.000					
	Más de $ 100.000.000		x			
Usabilidad	aprendizaje	Fácil		x		
		Moderado			x	
		Difícil	x			x
	Plataforma	Fácil		x		
		Moderado	x		x	
		Difícil				x
	Conformidad de uso	Buena	x		x	x
		Regular		x		
		Mala				
Funcionalidad	Identificación de vacíos e información sospechosa		x		x	x
	Comparación de columnas de información		x		x	x
	Depuración de información falsa		x	x	x	x
	Corrección de información sospechosa		x	x	x	x
	Condicionar el ingreso de información futura				x	

- **Evaluación de criterios**

A partir de la metodología existente para el análisis multicriterio AHP, se evaluaron los criterios y alternativas definidas anteriormente. A continuación se mostrará el desarrollo de los pasos establecidos para esta evaluación.

Teniendo en cuenta la organización de los criterios y alternativas, Se inició con la realización del árbol de jerarquías (Grafico 2), seguido de la evaluación de cada criterios por medio de la matriz de comparación por pares de las alternativas (MCPA), seguido de la matriz de comparación normalizada (MCN) (Tabla 13, 15, 17,19), para que finalmente se obtenga el vector de ponderación y corroborado por medio del índice y coeficiente de consistencia. (Tabla 14, 16, 18, 20)

Gráfico 2. Árbol de jerarquías para el desarrollo del análisis multicriterio



- **Costo**

Tabla 13. MCPA y MCN del criterio Costo.

MCPA (Costo)

	Data Ladder	Outliers	Quality Flag	R
Data Ladder	1	0,2	0,2	0,2
Outliers	5	1	1	1
Quality Flag	5	1	1	1
R	5	1	1	1

MCN (Costo)

	Data Ladder	Outliers	Quality Flag	R
Data Ladder	0,063	0,063	0,063	0,063
Outliers	0,313	0,313	0,313	0,313
Quality Flag	0,313	0,313	0,313	0,313
R	0,313	0,313	0,313	0,313

Tabla 14. Valor de ponderación final del criterio costo y los valores del índice de consistencia (IC) y coeficiente de correlación (CR).

Valor de Ponderación
6,250%
31,250%
31,250%
31,250%

Matriz NMax
0,250
1,250
1,250
1,250

N max	IC	CR
4,000	0,000	0,000

Según los pesos o valores asignados para las alternativas, se identificó que para una elección de la metodología según el costo, sería rechazada Data Ladder, ya que la adquisición de su plataforma tiene un costo elevado en comparación de las otras. El coeficiente de correlación se encuentra por debajo del 10%, indicando que no existe sesgo en la asignación de pesos para las alternativas.

- **Usabilidad**

Tabla 15. MCPA y MCN del criterio Usabilidad.

MCPA (Usabilidad)

	Data Ladder	Outliers	Quality Flag	R
Data Ladder	1,000	0,200	0,333	3,000
Outliers	5,000	1,000	3,000	7,000
Quality Flag	3,000	0,333	1,000	5,000
R	0,333	0,143	0,200	1,000

MCN (Usabilidad)

	Data Ladder	Outliers	Quality Flag	R
Data Ladder	0,107	0,119	0,074	0,188
Outliers	0,536	0,597	0,662	0,438
Quality Flag	0,321	0,199	0,221	0,313
R	0,036	0,085	0,044	0,063

Tabla 16. Valor de ponderación final del criterio Usabilidad y los valores del índice de consistencia (IC) y coeficiente de correlación (CR).

Valor de Ponderación	Matriz N Max
12,187%	0,492
55,789%	2,356
26,335%	1,099
5,689%	0,230

N max	IC	CR
4,177	0,059	0,065

Según los pesos o valores asignados para las alternativas, se identificó que para una elección según la usabilidad, la metodología que sería aplicada es la identificación de Outliers, ya que el aprendizaje y plataforma son fáciles de desarrollar, sin embargo, los métodos para realizar esta metodología presentan desventajas que no emiten una buena conformidad de su uso.

Por otro lado, la aplicación de R como metodología para la solución de las necesidades, es la que posee el menor valor de ponderación, ya que es un software con un aprendizaje y un manejo de su plataforma difícil, sin embargo, si se logra tener las capacidades técnicas para el manejo de este software, los resultados obtienen un buena conformidad de uso para el usuario.

Se debe tener en cuenta que el coeficiente consistencia se encuentra por debajo del 10%, indicando que no existe un sesgo en la asignación de pesos para las alternativas, según el criterio de usabilidad.

- **Funcionalidad**

Tabla 17. MCPA y MCN del criterio Funcionalidad.

MCPA (Funcionalidad)				
	Data Ladder	**Outliers**	**Quality Flag**	**R**
Data Ladder	1,000	5,000	0,333	1,000
Outliers	0,200	1,000	0,167	0,200
Quality Flag	3,000	6,000	1,000	3,000
R	1,000	5,000	0,333	1,000

MCN (Funcionalidad)				
	Data Ladder	**Outliers**	**Quality Flag**	**R**
Data Ladder	0,192	0,294	0,182	0,192
Outliers	0,038	0,059	0,091	0,038
Quality Flag	0,577	0,353	0,545	0,577
R	0,192	0,294	0,182	0,192

Tabla 18. Valor de ponderación final del criterio Funcionalidad y los valores del índice de consistencia (IC) y coeficiente de correlación (CR).

Valor de Ponderación	Matriz N Max
21,514%	0,885
5,666%	0,228
51,306%	2,144
21,514%	0,885

N max	IC	CR
4,141	0,047	0,052

Según los pesos o valores asignados para las alternativas, se identificó que para una elección según la funcionalidad, las banderas de calidad (Quality Flag) sería la metodología aplicada, ya que por medio de esta alternativa se puede solucionar la mayor cantidad de necesidades presentes en la base de datos. Se debe resaltar que la aplicación se puede realizar en una plantilla Excel, ya que la alternativa funciona como un atributo adicional a la base de datos.

Se debe tener en cuenta que el coeficiente consistencia se encuentra por debajo del 10%, indicando que no existe un sesgo en la asignación de pesos para las alternativas, según el criterio de funcionalidad.

- **Criterio**

Tabla 19. MCPA y MCN para la evaluación de los criterios.

MCPA (Criterio)			
	Costo	**Usabilidad**	**Funcionalidad**
Costo	1,000	0,333	0,200
Usabilidad	3,000	1,000	0,333
Funcionalidad	5,000	3,000	1,000

MCN (Criterio)			
	Costo	**Usabilidad**	**Funcionalidad**
Costo	0,111	0,077	0,130
Usabilidad	0,333	0,231	0,217
Funcionalidad	0,556	0,692	0,652

Tabla 20. Valor de ponderación final para la evaluación de los criterios y los valores del índice de consistencia (IC) y coeficiente de correlación (CR).

Valor de Ponderación	Matriz N Max
10,616%	0,320
26,050%	0,790
63,335%	1,946

N max	IC	CR
3,055	0,028	0,048

Por otro lado, la matriz de comparación por pares de los criterios determinó que la funcionalidad tiene un mayor peso para la elección de una de las alternativas evaluadas; el costo posee el menor valor de ponderación, indicando que tiene menor importancia para la elección de la metodología de limpieza de datos.

Los pesos asignados fueron corroborados por medio del coeficiente de consistencia, teniendo un valor menor al 10% e indicando que no existen sesgos en el proceso.

- **Ponderación Final**

Para la determinación de las ponderaciones finales para la elección de una metodología, se elaboró el árbol de criterios y alternativas (gráfico 3) para facilitar el comportamiento y análisis numérico.

Finalmente, en la (tabla 21) se encuentran los valores finales de las alternativas analizadas por el método de análisis multicriterio AHP, indicando que las banderas de calidad (Quality Flag) son las adecuadas para solucionar la mayor cantidad de necesidades presentes en la base de datos, teniendo en cuenta los criterios de costo, usabilidad y funcionalidad.

Gráfico 3. Árbol de criterios y alternativas con valores de ponderación individuales.

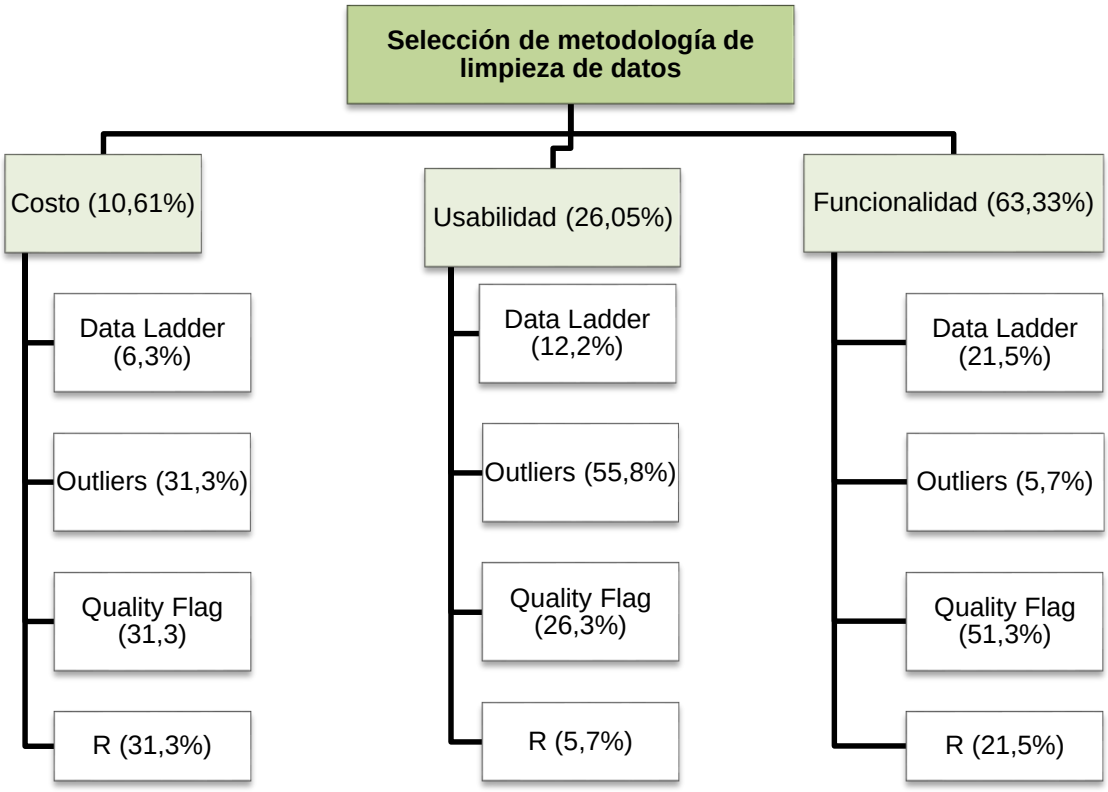


Tabla 21. Valores de ponderación final de las alternativas analizadas

Data Ladder	17,46%
Outliers	21,44%
Quality Flag	42,67%
R	18,43%

5.2.4. Ejercicio práctico

- **Definición de banderas de calidad**

EL ejercicio práctico fue realizado para el parámetro de Coliformes Termo-tolerantes (CTE), al cual se le establecieron banderas de calidad según el modelo ODV, para los atributos de concentración, límite de detección y método analítico, con el fin de identificar los errores que se presentan y poder realizar correcciones en la base de datos.

- **Concentración**

En este atributo resultado del análisis de la muestra, se evidenció que existen datos que pueden ser erróneos por los límites que presentan, es decir que se presentan datos muy altos y muy bajos que deben ser confirmados con la ayuda de expertos y la comparación de otros atributos, ya que durante el tiempo de funcionamiento del programa, el ingreso de información a la base de datos ha presentado diferentes formas de digitación.

Tabla 22. Bandera de calidad para la columna de Concentración.

Código	Descripción
1	Bueno: El dato se encuentra dentro del rango permisible conocido por los expertos
2	Desconocido: No se registra dato
4	Cuestionable: El dato puede ser sospechoso de ser errado por no encontrarse dentro del rango permisible
8	Malo: el dato es nulo por fallas en laboratorio

- **Método de análisis**

Este atributo de la base de datos es relacionado según los códigos establecidos para los métodos de análisis utilizados por los laboratorios, para el análisis de muestras recolectadas por los nodos integrantes del programa REDCAM. Por lo tanto, este atributo deberá añadir una bandera de calidad para confirmar que el código que se está digitando es el adecuado para la variable analizada.

Tabla 23. Bandera de calidad para la columna de Método de Análisis.

Código	Descripción
1	Bueno: El dato corresponde al método de análisis adecuado para determinar Coliformes Termo-tolerantes.
2	Desconocido: No se registra dato
4	Cuestionable: El dato puede ser sospechoso de ser errado
8	Malo: el dato es nulo por no corresponder al método de análisis correspondiente para determinar Coliformes termo-tolerantes

- **Límite de Detección**

El valor de este atributo permite la comparación de la información generada en el tiempo, ya que corrobora la concentración mínima que puede ser detectada por un método analítico determinado. Por lo tanto, indicar el límite de detección permitirá reportar y comparar datos con un grado de confianza.

Tabla 24. Bandera de calidad para la columna de Límite de Detección.

Código	Descripción
1	Bueno: El dato corresponde a la técnica analítica utilizada
2	Desconocido: No se registra dato
4	Cuestionable: El dato puede ser sospechoso de ser errado
8	Malo: LD (N/A)

- **Prueba de datos**

Para realizar la prueba de datos utilizando banderas de calidad, se utilizó procedimiento citado por Cecoldo [44], identificando los límites físicos con los cuales será agrupada la información.

- Concentración límite inferior (1,8); límite superior (1600); los valores que se encuentren fuera de este rango, son considerados cuestionables ya que puede ser información sospechosa de ser errada.
- Método de análisis: Para el determinar la concentración de CTE, se utiliza la fermentación en tubos múltiples como método analítico con código (10). La columna deberá tener este código o cualquier otro dependiendo al método utilizado por los nodos integrantes, para el caso del INVEMAR se utilizará esta codificación. Los valores sospechosos pueden estar marcados por errores de digitación y los valores que no correspondan al método determinado será un valor malo.
- Límite de detección: A partir del método utilizado para determinar la concentración de CTE, el límite de detección tendrá un único valor que determina la concentración mínima que se puede determinar. La fermentación por tubos múltiples es la técnica utilizada en el LABCAM y su único valor aceptado de límite de detección es 1,8.

A partir del establecimiento de los límites físicos, se realizó un ejercicio en un conjunto de datos en el cual se pudiera evidenciar los usos de las banderas de calidad (Tabla 25).

Tabla 25. Banderas de calidad en un conjunto de datos

Estación	Concentración	QF	Código Método	QF	Limite Detección	QF	Observación
Estación 1	1300	1	10.00	1	(en blanco)	2	(en blanco)
Estación 2	230	1	10.00	1	(en blanco)	2	(en blanco)
Estación 3	20	1	10.00	1	(en blanco)	2	(en blanco)
Estación 4	0	4	10.00	1	1.8	1	Menor a 1,8 NMP/100 mL
Estación 5	0	4	10.00	1	1.8	1	Menor a 1,8 NMP/100 mL
Estación 6	20	1	10.00	1	(en blanco)	2	(en blanco)

Estación 7	45	1	10.00	1	(en blanco)	2	(en blanco)
Estación 8	4	1	10.00	1	(en blanco)	2	(en blanco)
Estación 9	0	4	10.00	1	1.8	1	Menor a 1,8 NMP/100 mL
Estación 10	0	4	10.00	1	1.8	1	Menor a 1,8 NMP/100 mL
Estación 11	18	1	10.00	1	(en blanco)	2	(en blanco)
Estación 12	330	1	10.00	1	(en blanco)	2	(en blanco)
Estación 13	2400	4	10.00	1	(en blanco)	2	(en blanco)

Tabla 26. Resultado porcentual de las banderas de calidad en el conjunto de datos.

	QF Concentración	QF Método Analítico	QF límite de Detección
Bueno	61.5 %	100,0 %	30,8 %
Desconocido	0 %	0,0 %	69,2 %
Sospechoso	38.5 %	0,0 %	0,0 %
Malo	0,0 %	0,0 %	0,0 %

Según los resultados de la prueba de las banderas de calidad en el conjunto de datos (Tabla 26), se identificó que la columna de concentración, los datos sospechosos son identificados por encontrarse fuera del rango establecido técnicamente, sin embargo existen filas en las cuales el dato asignado es cero (0) porque se encontraba por debajo del límite de detección.

En la columna de límite de detección se encontró que un alto porcentaje no cuenta con el dato documentado, sin embargo en algunas filas, esta información puede ser diligenciada ya que el valor se encuentra en el atributo de observaciones.

5.3. Recomendaciones

- Se recomienda que se debe continuar con el proceso de documentación de la información generada en los muestreos realizados, en los años de funcionamiento del programa nacional de monitoreo “Red de Vigilancia para la Conservación y Protección de las Aguas Marinas y Costeras de Colombia” – REDCAM, con el fin de fortalecer la base de datos y llevarla a cumplir con los estándares nacionales e internacionales de calidad de datos.

- Se recomienda que previo a iniciar con la depuración de la base de datos, se diseñe o adopte una guía de normalización de datos para unificar los atributos que seas modificables, teniendo en cuenta que existen datos primarios que deben mantener su originalidad.; esto para que la limpieza de datos se realice bajo unos términos de referencia actuales y se condicione el ingreso de información futura.

- Se recomienda utilizar las herramientas estadísticas del entorno de programación R Project, que aporta una mejora de los datos, brindándole un soporte teórico que mejora los análisis realizados con esta información y fortalece la documentación en los metadatos, ya que es un atributo opcional que contribuye al crecimiento y fortalecimiento de la base de datos.

- Se recomienda iniciar con las etapas metodológicas de arqueología de datos con el fin de recuperar información perdida o faltante que se refunde en el tránsito de la información, desde la fuente hasta la base de datos.

6. CONCLUSIONES

- Las herramientas identificadas y desarrolladas en este documento, son instrumentos que pueden funcionar para cumplir con los objetivos de diferentes proyectos que se ejecutan en el programa de Calidad Ambiental Marina – CAM, ya que pueden ser adaptados y modificados para cumplir diferentes metas.

- Para la aplicación de banderas de calidad se deben conocer y estipular los límites físicos para poder establecer los rangos permisibles, por lo tanto, se deben contemplar herramientas que permitan realizar la búsqueda y recuperación de información con el fin de complementar los datos primarios para que sean almacenados en la base de datos.

- La implementación de banderas de calidad como metodología que permita realizar la identificación y limpieza de datos, permitirá reducir tiempo y esfuerzo por parte de los investigadores ya que la base de datos alcanzará y documentará información con mayores niveles de calidad.

7. BIBLIOGRAFIA

[1] O. Garcés *et al.*, “DIAGNÓSTICO Y EVALUA CIÓN DE LA CALIDAD D E LAS AGUAS MARINAS Y COSTERAS D EL CARIBE Y PACÍFICO COLOMBIANOS,” Santa Marta, 2016.

[2] B. Marín, J. Garay, A. Vélez, J. Arias, J. Bohórquez, and N. Calvano, *REDCAM - Manual de Funcionamiento Del Sistema de Informacion*. Santa Marta, 2002.

[3] Instituto de Investigaciones Marinas y Costeras “Jose Benito Vives de Andreís” (INVEMAR), *MANUAL DE TÉCNICAS ANALÍTICAS PARA LA DETERMINACIÓN DE PARÁMETROS FISICOQUÍMICOS Y CONTAMINANTES MARINOS (AGUAS, SEDIMENTOS Y ORGANISMOS)*. Santa Marta, 2003.

[4] Instituto Oceanográfico de la Armada, “INFORME DE RENDICIÓN DE CUENTAS, ENERO-AGOSTO 2011,” Guayaquil, 2012.

[5] Instituto Oceanográfico de la Armada, “FORMULARIO DE INFORME DE RENDICION DE CUENTAS - 2015,” Guayaquil, 2016.

[6] Instituto Oceanográfico de la Armada, “INFORME DE RENDICIÓN DE CUENTAS 2015,” Guayaquil, 2016.

[7] J. L. Hernández J, R. V Ortíz-Martínez, and I. Suárez P, “Metodología archivística para la recuperación de información oceanográfica del Pacífico colombiano,” *Boletín Científico CCCP*, no. 14, pp. 123–150, 2007.

[8] Dimar, *Gestión de datos e información oceanográfica Colombiana*, vol. 6. San Andrés de Tumaco, 2008.

[9] Centro de Investigaciones Oceanográficas e Hidrográficas, “Bienvenidos al sitio web del CIOH.” [Online]. Available: http://www.cioh.org.co/dev/presentacion/lineas_inv.html. [Accessed: 24-Jun-2016].

[10] M. Estrada *et al.*, “Reflexiones sobre la gestión y custodia de datos oceanográficos en España. Recursos existentes y recomendaciones para el futuro,” España, 2010.

[11] P. Caplan, “You Call It Corn, We Call It Syntax-Independent Metadata for Document-Like Objects,” *Public Access-Computer Syst. Rev.*, vol. 6, no. 4, 1995.

[12] J. A. Senso and A. De La Rosa, “El concepto de metadato: algo más que descripción de recursos electrónicos,” *Ciência da Informação*, vol. 32, no. 2, pp. 95–106, 2003.

[13] M. Agudelo, “Los metadatos,” *Gestión Contenidos Educ. Virtual Calid.* Antioquia, Colombia, p. 5, 2009.

[14] J. Jiménez, “Renovación del metadato en Internet para la recuperación de la información,” *Biblios Rev. electrónica Bibl. …*, vol. 8, no. 1, pp. 1–9, 2001.

[15] E. Méndez, “Descripción de contenidos y documentación digital: introducción a los metadatos,” 2015, p. 1.

[16] I. Daudinot, “Organización y recuperación de información en Internet: teoría de los metadatos,” 2008. [Online]. Available: http://bvs.sld.cu/revistas/aci/vol14_5_06/aci06506.htm. [Accessed: 13-Oct-2016].
[17] R. Torréns and Z. Méndez, “Taller de manejo de datos y metadatos para las ciencias ecológicas,” 2005, p. 92.
[18] A. Sánchez, J. Nogueras, and D. Ballari, “Normas sobre metadatos (ISO19115, ISO19115-2, ISO19139, ISO 15836),” *Mapp.*, vol. 123, pp. 48–57, 2008.
[19] ICONTEC, “NTC 4611 - Información Geográfica. Metadato Geográfico,” 2011. [Online]. Available: http://e-normas.icontec.org.bdatos.usantotomas.edu.co:2048/icontec_enormas_mobile/visor/HTML5.asp. [Accessed: 10-Oct-2016].
[20] A. Rifkin and R. Khare, “Capturing the State of Distributed Systems with XML, by Rohit Khare and Adam Rifkin,” *archiving.html*, vol. 1, no. 44, 1997.
[21] Comisión Oceanográfiaca Intergubernamental, “Quinto Taller Regional para Estados Miembros del Caribe y America del Sur: GODAR-V (Proyecto Globar en Arqueología y Recuperación de Datos Oceanograáficos),” Cartagena de Indias - Colombia, 1996.
[22] L. Molina, “Data mining : torturant les dades fins que confessin,” Barcelona, 2002.
[23] R. B. Machado, “USO DE DATA MINING E SISTEMAS DE INFORMAÇÕES GEOGRÁFICAS NO APOIO A TOMADA DE DECISÕES,” UNIVERSIDADE FEDERAL DE SANTA CATARINA, 2005.
[24] L. Azaña and C. Ruz, “Qué es DataMining ?” pp. 1–23, 2007.
[25] J. Maletic and A. Marcus, “Data Cleansing: Beyond Integrity Analysis.,” *IQ2000*, pp. 1–10, 2000.
[26] H. Galhardas, D. Florescuand, E. Simon, and D. Shasha, “An Extensible Framework for Data Cleaning,” *16th Int. Conf. Data Eng.*, 2000.
[27] Dataladder, “Dataladder - The Leader in Data Cleansing Software,” 2016. [Online]. Available: http://dataladder.com/. [Accessed: 17-Sep-2016].
[28] Data Ladder, “COMPARISON STUDY,” *IN-HOUSE VS. BEST IN CLASS DATA MATCHING SOLUTIONS While*, 1995. [Online]. Available: http://dataladder.com/wp-content/uploads/2016/04/InHouse-Brochure-V3.pdf. [Accessed: 22-Sep-2016].
[29] Data Ladder, “Dataladder - Remove Duplicate - Dataladder,” 2016. [Online]. Available: http://dataladder.com/remove-duplicate/. [Accessed: 01-Jan-2016].
[30] Data Ladder, “CASE STUDY - TRACKING RECORDS ACROSS DATABASES,” 2016. [Online]. Available: http://dataladder.com/wp-content/uploads/2016/04/West-Virginia-University-Case-Study.pdf. [Accessed: 01-Jan-2016].
[31] Data Ladder, “CASE STUDY - Buckle Denim Retailer,” 2016. [Online]. Available: http://dataladder.com/wp-content/uploads/2016/04/Buckle-Denim-Retailer-Case-Study.pdf. [Accessed: 17-Sep-2016].
[32] Data Ladder, “CASE STUDY - EDP Consulting Group,” 2005. [Online].

Available: http://dataladder.com/wp-content/uploads/2016/04/EDP-Consulting-Group-Case-Study.pdf. [Accessed: 22-Sep-2016].
[33] Data Ladder, “CASE STUDY - AMEC Global,” 2016. [Online]. Available: http://dataladder.com/wp-content/uploads/2016/04/AMEC-Global-Env-Eng-Case-Study.pdf. [Accessed: 22-Sep-2016].
[34] Data Ladder, “CASE STUDY - Quick Reliable Printing FUZZY MATCHING TOOL GIVES PROVIDES UNEXPECTED PROFIT CENTER,” 2016. [Online]. Available: http://dataladder.com/wp-content/uploads/2016/04/Quick-Reliable-Printing-Case-Study.pdf. [Accessed: 22-Sep-2016].
[35] Data Ladder, “CASE STUDY - Arlington Power Equipment,” 2016. [Online]. Available: http://dataladder.com/wp-content/uploads/2016/04/Arlington-Power-Equipment-Case-Study.pdf. [Accessed: 22-Sep-2016].
[36] I. Amón, “GUÍA METODOLÓGICA PARA LA SELECCIÓN DE TÉCNICAS DE DEPURACIÓN DE DATOS,” Universidad Nacional de Colombia - Facultad de minas, Escuela de sistemas, 2010.
[37] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.
[38] STATGRAPHICS, “Identificación de Valores Atípicos,” 2007. [Online]. Available: http://www.statgraphics.net/wp-content/uploads/2011/12/tutoriales/Identificacion de Valores Atipicos.pdf. [Accessed: 24-Sep-2016].
[39] E. Dan and O. Ijeoma, “STATISTICAL ANALYSIS/METHODS OF DETECTING OUT LIERS IN A UNIVARIATE DATA IN A REGRESSION ANALYSIS MODEL,” *Int. J. Educ. Res.*, vol. 1, no. 5, p. 24, 2013.
[40] J. K. (John K. Taylor and C. Cihon, *Statistical techniques for data analysis.* Chapman & Hall/CRC, 2004.
[41] L. Castro, “Análisis de tendencia y homogeneidad de series climatológicas,” Universidad del Valle, 2010.
[42] C. Yrigoyen and I. L. Klein-Dpto de Economía Aplicada, “MÉTODOS GRÁFICOS DEL ANÁLISIS EXPLORATORIO DE DATOS ESPACIALES,” Madrid, 2009.
[43] Intergovernmental Oceanographic Commission of UNESCO, “Recomendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data,” *IOC Manuals Guid.*, vol. 3, no. 54, p. 13, 2013.
[44] R. V. Ortiz, “Normalización de conjuntos de datos oceanográficos y de meteorología marina,” 2010.
[45] O. J. S. Parra, F. J. Puente, and R. V. Ortiz, “Desarrollo de una herramienta computacional para constrastar la calidad de datos oceanográficos,” *Ingeniería*, vol. 13, no. 1, pp. 77–83, 2007.
[46] R Project, “What is R?,” 2010. [Online]. Available: https://www.r-project.org/about.html. [Accessed: 26-Sep-2016].
[47] R. D. C. Team, *Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos*, R Developm., vol. 1. 2000.
[48] E. Paradis and J. a Ahumada, “R para Principiantes,” *Evolution (N. Y).*, vol.

42, no. 75, p. 61, 2003.
[49] A. J. Arriaza Gómez, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, S. Pérez Plaza, and A. Sánchez Navas, “Estadística básica con R y R-Commander,” *Univ. Cádiz*, pp. 1–160, 2008.
[50] Y. Zhao, *R and Data Mining : Examples and Case Studies*, no. December 2012. Academic Press, Elsevier, 2013.
[51] I. Feinerer, K. Hornik, and D. Meyer, “Text Mining Infrastructure in R,” *J. Stat. Softw.*, vol. 25, no. 5, pp. 1–54, 2008.
[52] I. Feinerer, “An Introduction to Text Mining in R,” *R News*, vol. 8, no. 2, pp. 19–22, 2008.
[53] H. Roche and C. Vejo, “Métodos Cuantitativos Aplicados a la Administración. Analisis multicriterio en la toma de decisiones.,” 2005.
[54] S. Berumen and F. Redondo, “LA UTILIDAD DE LOS MÉTODOS DE DECISIÓN MULTICRITERIO (COMO EL AHP) EN UN ENTORNO DE COMPETITIVIDAD CRECIENTE,” *Cuad. Adm.*, vol. 20, no. 34, pp. 65–87, 2007.
[55] J. Moreno Jiménez, “EL PROCESO ANALÍTICO JERÁRQUICO (AHP).,” Zaragoza, 2002.
[56] E. Font, “GESTIÓN DE LA INFORMACIÓN EN LA UTILIZACIÓN DEL PROCESO ANALÍTICO JERÁRQUICO PARA LA TOMA DE DECISIONES DE NUEVOS PRODUCTOS.,” *An. Doc.*, vol. 3, pp. 55–66, 2000.
[57] S. Reyna, “Valoración ahp de los ecosistemas naturales de la comunidad valenciana,” 1995.
[58] INVEMAR, “Programas de investigación - INVEMAR,” 2016. [Online]. Available: http://www.invemar.org.co/web/guest/programas-de-investigacion. [Accessed: 14-Oct-2016].
[59] INVEMAR, “Funciones - INVEMAR,” 2016. [Online]. Available: http://www.invemar.org.co/web/guest/funciones. [Accessed: 14-Oct-2016].
[60] R. MORERA, “INSTRUMENTOS DE SELECCIÓN DE SOFTWARE PARA LA GESTIÓN DE ARCHIVOS,” *Bidulma*, vol. 14, pp. 301–333, 2000.
[61] L. Florez and F. Grisales, “FORMULACION DE CRITERIOS PARA LA SELECCION DE METODOLOGIAS DE DESARROLLO DE SOFTWARE,” PEREIRA, RISARALDA, 2014.
[62] M. F. Morales and R. C. Arley, “Automatización de unidades de información: Matriz técnica para la evaluación de software libre,” *Rev. Interam. Bibl.*, vol. 36, no. 3, pp. 207–219, 2013.