Leonardo Franco Calpa

Anderson Ramírez Hormiga

Bryan Camilo Grueso

**1) Identificación del problema**

**Se identifican de manera adecuada las necesidades que se presentan en el problema propuesto así como sus síntomas y condiciones por los que debe ser resuelto.  
Identificación de sus síntomas y necesidades:**-El director de Colciencias desea poder identificar los principales grupos de investigación del país.

-Generar y consolidar reportes gráficos de la información por regiones, ciudades, área de investigación y clasificación.

-Ubicar en un mapa los distintos grupos de investigación que existen en el país.

-Visualizar aleatoriamente en una región que grupos de investigación pertenecen a esa ciudad.

De acuerdo a la información planteada en el enunciado se establece una lista de requerimientos con el fin de trazar una ruta de planeación y desarrollo de acuerdo a las necesidades y la identificación de la situación problemática.

Definición del problema:

El director de Colciencias desea implementar un programa que teniendo en cuenta a los mejores grupos de investigación del país, pueda generar un informe acerca del estado de dichos grupos, mostrar gráficos para una posible clasificación y una mejor visualización de los mismos al igual que poder ubicarlos geográficamente en un mapa.

El programa debe cumplir los siguientes requerimientos:

|  |  |
| --- | --- |
| Nombre | RF1- Registro de datos |
| Resumen | El programa debe de estar en la capacidad de poder registrar los datos de la base de datos dada por Colciencias |
| Entradas | |
| -Información de los grupos de investigación. | |
| Resultados | |
| -Almacenar correctamente los datos ingresados en el programa. | |

|  |  |
| --- | --- |
| Nombre | RF2- Actualización de datos |
| Resumen | El programa debe poder actualizar los datos en cualquier momento sin ningún error |
| Entradas | |
| -Datos nuevos que van a ser actualizados en el sistema. | |
| Resultados | |
| -Datos actualizados correctamente y sin perjudicar al programa. | |

|  |  |
| --- | --- |
| Nombre | RF3- Visualización de los grupos de investigación |
| Resumen | El programa debe poder visualizar toda la información de los grupos de investigación |
| Entradas | |
| · Grupos de investigación   * Información de los grupos de investigación. * Cantidad de grupos de investigación. * Listado de los grupos de investigación. | |
| Resultados | |
| · Toda la información extraída de la base de datos visualizada en el programa. | |

|  |  |
| --- | --- |
| Nombre | RF4- Generación de reportes. |
| Resumen | El programa debe generar reportes de cada grupo de investigación y poder actualizarlos. |
| Entradas | |
| · Grupos de investigación   * Información de los grupos de investigación. | |
| Resultados | |
| · El correcto funcionamiento de la generación de reportes por cada grupo de investigación. | |

|  |  |
| --- | --- |
| Nombre | RF5- Visualización de la información por ciudades, regiones y áreas de investigación y clasificarlas. |
| Resumen | El programa debe de visualizar ciudades, zonas y áreas donde se encuentren los grupos de investigación y acceder a su información. |
| Entradas | |
| · Grupos de investigación   * Cantidad de grupos de investigación por áreas. * Información de los grupos de información de dichas áreas. | |
| Resultados | |
| · Visualización de los grupos de investigación por áreas y su respectiva clasificación | |

|  |  |
| --- | --- |
| Nombre | RF6-Informes de los artículos más frecuentados de cada grupo de investigación |
| Resumen | El programa podrá mostrar cuales son los artículos más frecuentados por cada grupo de investigación. |
| Entradas | |
| · Grupos de investigación.  - Artículos de cada grupo. | |
| Resultados | |
| · Informe con los artículos más frecuentados de los grupos de investigación. | |

**Requerimientos:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | Registrar | | Visualizar | | Actualizar | | Información de los grupos de investigación |

|  |  |
| --- | --- |
| Visualizar | Información consolidada y clasificada:  -Regiones  -Ciudades  -Zonas  -Grupos de investigación |

|  |  |
| --- | --- |
| Generar | Reportes Gráficos |

|  |  |
| --- | --- |
| Consultar | Conjunto de artículos más frecuentados |

**Definición del problema:**

Colciencias y la Asociación Colombiana para el Avance de la Ciencia -ACAC- requieren un software para el manejo, registro, visualización y actualización de grandes volúmenes de datos almacenados en forma de texto para generar reportes gráficos que consoliden y clasifiquen la información y permitan realizar consultas de los grupos de investigación para apoyar el desarrollo científico y tecnológico.

**2) Recopilación de información necesaria**

**Análisis exploratorio de datos.**

Dado un volumen considerable de datos en un archivo de texto resulta difícil tratar la información con los datos en ‘bruto’ y dependiendo del tamaño de la información se facilita realizar análisis con distintas herramientas que que realicen la clasificación y agrupamiento de datos para darle mayor valor a la información ahí almacenada usando técnicas de KDD (Knowledge Discovery in Databases) para la exploración de datos.

Por eso se encontró algunas estrategias que se proponen en un sitio informativo para el análisis de datos con el fin de convertirlos en información valiosa. Algunas son:

• La comprobación de los formatos de las distintas variables que tenemos en nuestro conjunto de datos.

• La visualización de las primeras y últimas filas, para saber que su importación en el entorno que estemos usando es el correcto.

• Realizamos una descripción básica a nivel estadístico: valores medios máximos, mínimos y desviaciones, lo cual nos da una foto inicial precisa del conjunto de datos sin necesidad de recurrir a inspeccionar visualmente todas las filas, algo imposible y muy subjetivo.

• Siempre realizamos gráficos exploratorios: diagramas de barras, histogramas de frecuencia o de densidad, que nos ayudan a ganar conocimiento sobre el conjunto de datos.

• Además, prestaremos especial interés en esta fase, a la presencia de valores nulos que nos indican que hay variables que no están informadas, y por tanto carecen de información.

* El proceso para tratar con grandes volúmenes de datos consiste principalmente en:
* Adquisición de datos
* Limpieza y ordenamiento
* Elección de la herramienta para tratar los datos de acuerdo a su volumen(bajo, medio, alto)
* Realizar los análisis descriptivos o predictivos, según sea el caso.

Posterior a eso, si existe el ingreso de información nueva al conjunto de datos se debe garantizar la adaptación de la nueva información al formato en el que se encuentra la base de datos existente. Se aplican criterios de  *Tidy Data* para optimizar el modelamiento, la visualización y el análisis.

<http://www.redalyc.org/html/925/92502902/>

<https://elartedemedir.com/blog/preparar-los-datos-y-convertirlos-en-informacion/>

**Herramientas para visualización de componentes visuales mediante mapas y gráficos.**

**….**

**3) Búsqueda de soluciones creativas**

Se puede optar que el desarrollo de un software sea por medio de programación modular para dividir los problemas del manejo de la información (agrupamiento y clasificación) y para las tareas de consulta.

Así pues se buscan soluciones para los distintos módulos que compondrán la solución del proyecto.

Para el manejo de la base de datos y el tratamiento de la información se planea convertir los datos de texto a objetos para realizar operaciones de consulta y registro sobre ellos. También el modelo de los datos puede manipularse en diferentes formatos de acuerdo a la necesidad y se tiene algunas opciones:

* Alternativa 1:Manejarlos como archivos de bases de datos embebidas al proyecto solución, ya sea en formato .mdf o .sdf según se vea cuál tiene mejores beneficios para el problema a solucionar.
* Alternativa 2: Manejarlos como formato .JSON para el intercambio de datos.
* Alternativa 3: Manejarlos como formato .CSV para la consulta de la información y su posterior manipulación.
* Alternativa 4: Los datos también podrían intentar manipularse de forma directa, texto plano para su consulta y demás requerimientos.

Por otra parte las ideas para manejar las consultas podrían desarrollarse con….

* Alternativa 1: Una conexión con Excel, haciendo una tabla dinámica que esté enlazada al programa que al momento de buscar algo en el programa, Excel arroje el resultado que necesita el usuario y se visualice en la pantalla con los gráficos.
* Alternativa 2: Descargar toda la base de datos en visual y al momento de hacer una consulta, visual se encargará de mostrar los resultados.
* Alternativa 3: Manejarlos como formato .CSV para la consulta de la información y su posterior manipulación.

**4) Transición de las ideas a los diseños preliminares.**

Lo primero que haremos será descartar las soluciones que no son factibles a la hora de resolver el problema, o que demandan una complejidad muy altas y sería un obstáculo para la eficiencia del programa.

La alternativa 4 es la primera que se va a descartar, puesto que manejar cinco mil archivos en un texto plano es ineficiente a la hora de hacer las consultas.

La alternativa 2, de manejar los datos en archivos .JSON también se descarta debido a que para realizar las clasificaciones para las consultas en un archivo semiestructurado hace que necesariamente se tenga que leer todo el archivo para filtrar la información, y esto una y otra vez.

**En la parte de manejar las consultas:**

La alternativa 2 se descarta, puesto que meterle esa base de datos a visual haría que se volviera más lento de lo usual, además de que toca volver a cargar la información en el propio visual y luego si hacerle las consultas, por lo que se gastaría mucho tiempo, memoria y no sería eficaz.

Luego de descartar las ideas no factibles, se hizo una revisión de las ideas restantes y se encontró lo siguiente:

Para estructurar la base de datos:

Alternativa 1: Estructurar la información usando una base de datos resulta más manejable debido a que se pueden realizar filtro y la información puede ser clasificada y retornada. De igual forma se puede usar un lenguaje de consulta declarativo para buscar los datos y esto hace que no se tenga que pensar en crear métodos para la búsqueda de la información.

Alternativa 3:

-Manejo por tabulaciones.

-Se separan por punto y coma y es más fácil tener acceso.

-No es tan complicado, puesto que el formato siempre va a ser el mismo

Para la consulta;

Alternativa 1:

-Usando Excel, se ahorra tiempo de búsqueda.

-Solo sería ingresar el nombre de lo que se quiere buscar y excel automáticamente arroja un resultado, el cual se visualizará en la interfaz gráfica.

-Su manipulación es fácil y no requiere de conocimientos avanzados.

-Admite la actualización en la base de datos y no entorpece la consulta.

**Paso 5. Evaluación y selección de la mejor solución.**  
*Criterios*  
   
Deben definirse los criterios que permitirán evaluar las alternativas de solución y con base en este resultado elegir la solución que mejor satisface las necesidades del problema planteado. Los criterios que escogimos en este caso son los que enumeramos a continuación. Al lado de cada uno se ha establecido un valor numérico con el objetivo de establecer un peso que indique cuáles de los valores posibles de cada criterio tienen más peso, es decir cuanto más puntos tenga será el adecuado para realizar la solución.

Criterio A: Precisión a la hora de cargar los archivos, el método entrega una solución:

-[1] Rápida.  
 -[2] Eficaz.

-[3] Concisa.

Criterio B: Tiempo, el método entrega una solución:

-[1] Lineal.

-[2] Sin retraso.

Criterio C: Complejidad: el método entrega una solución:

-[1] Fácil de entender.

-[2] Fácil de modificar.

Evaluación.

Según la evaluación con los criterios anteriores en las alternativas se obtiene la siguiente tabla:

Para el manejo de datos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Criterio A | Criterio B | Criterio C | Total |
| Alternativa 1 | Rápida 3 | Lineal  2 | Fácil de entender  3 | 7 |
| Alternativa 3 | Concisa  4 | Sin retraso  4 | Fácil entender/modificar  5 | 13 |

Para las consultas:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Criterio A | Criterio B | Criterio C | Total |
| Alternativa 2 | Rápida 2 | Lineal  1 | Fácil de entender  3 | 6 |
| Alternativa 3 | Concisa  3 | Sin retraso  3 | Fácil entender/modificar  5 | 11 |

Selección

De acuerdo con la evaluación anterior selecciona la Alternativa 3 en ambas ocasiones, ya que obtuvo la mayor puntuación de acuerdo con los criterios definidos y porque este método brinda más facilidades al momento de acceder a un vehículo determinado en una bahía determinada.

Diagrama de clases

