

14-1-2016

Caso Práctico

Integración y monitorización de
fuentes de información



Javier García Pérez - 100290698
José Manuel Fernández Ruiz - 100290892

MÁSTER EN INGENIERÍA INFORMÁTICA
CURSO 1
INTEGRACIÓN DE SISTEMAS INFORMÁTICOS
UC3M

Índice

1.	Introducción	1
2.	Resumen ejecutivo	2
2.1.	Objetivo	2
2.2.	Alcance	2
2.3.	Solución diseñada.....	3
2.4.	Aspectos claves	3
2.5.	Conclusiones.....	4
3.	Análisis y diseño del problema de integración propuesto	5
3.1.	Análisis.....	5
3.1.1.	Definición del sistema	5
3.1.2.	Fuentes de datos	9
3.1.3.	Modelo de datos	11
3.2.	Diseño.....	12
3.2.1.	Diseño de clases	12
3.2.2.	Diseño de casos de uso reales.....	23
3.2.3.	Interfaces de usuario.....	26
3.2.4.	Diseño del modelo de datos.....	34
3.2.5.	Especificación del control de versiones	37
4.	Prueba de concepto	39
4.1.	Máquina virtual	39
4.2.	Abrir ATMOSPHERESPAIN	41
4.3.	Visualizaciones	41
4.3.1.	Opinión social	42
4.3.2.	Comparar provincias	47
4.3.3.	Gráficos de provincia.....	52
4.4.	Descarga de datasheets	58
4.5.	Contaminantes	60
5.	Conclusiones	63
	ANEXO I: Cantidad de estaciones por provincia.....	64
	ANEXO II: Elementos químicos y variables medidas	65
	ANEXO III: Temas, hashtags y cuentas Twitter.....	66
	ANEXO IV: Fuentes de datos	67

Índice de Tablas

Tabla 1 Casos de uso	24
Tabla 2 Estación atributos	35
Tabla 3 Contaminante atributos	36
Tabla 4 Tweet atributos	37
Tabla 5 Tweet atributos	37
Tabla 6 Núm. Estaciones por provincia	64
Tabla 7 Elementos químicos y variables medidas.....	65
Tabla 8 Temas, hashtags y cuentas de tweets	66

Índice de ilustraciones

Ilustración 1 Datos aire Galicia	9
Ilustración 2 Datos aire Madrid.....	10
Ilustración 3 Datos aire Madrid PDF	10
Ilustración 4 Diagrama E-R de la base de datos	11
Ilustración 5 Aire.java	13
Ilustración 6 Datos aire Galicia.....	13
Ilustración 7 Datos aire Castilla-La Mancha	14
Ilustración 8 AGUIMES.csv	15
Ilustración 9 estaciones.csv.....	16
Ilustración 10 Personas.java	17
Ilustración 11 Twitter widget	18
Ilustración 12 Arquitectura aplicación web	19
Ilustración 13 Provincia_Servlet.java	20
Ilustración 14 Opcion_Servlet.java.....	20
Ilustración 15 Page_Servlet.java	20
Ilustración 16 Datasheet_Servlet.java	21
Ilustración 17 Contaminante_Servlet.java	21
Ilustración 18 Provincia.java	21
Ilustración 19 Opcion.java.....	22
Ilustración 20 HashtagFreq.java.....	22
Ilustración 21 Datasheet.java.....	22
Ilustración 22 Contaminante.java	23
Ilustración 23 C00.....	24
Ilustración 24 C01.....	24
Ilustración 25 C02.....	25
Ilustración 26 C03.....	25
Ilustración 27 C04.....	25
Ilustración 28 C05.....	26
Ilustración 29 Página de inicio. Parte 1	26
Ilustración 30 Página de inicio. Parte 2	27
Ilustración 31 Página de inicio. Parte 3	27
Ilustración 32 Selección de provincia 1	28
Ilustración 33 Selección de provincia 2	28

Ilustración 34 Visualización_1	29
Ilustración 35 Visualización_2	30
Ilustración 36 Visualizacion_3	31
Ilustración 37 Datasheets.....	32
Ilustración 38 Contaminantes	33
Ilustración 39 Error.....	33
Ilustración 40 Navegabilidad.....	34
Ilustración 41 Colecciones MongoDB.....	34
Ilustración 42 Estación JSON	35
Ilustración 43 Contaminante JSON.....	35
Ilustración 44 Tweet JSON.....	36
Ilustración 45 Provincia JSON.....	37
Ilustración 46 Control de versiones.....	38
Ilustración 47: Parámetros Configuración VirtualBox	39
Ilustración 48: Proceso de importación	40
Ilustración 49: Arrancar Máquina Virtual.....	40
Ilustración 50: Acceso Visualizaciones	43
Ilustración 51: Selección Provincia Opinión Social	44
Ilustración 52: Tweets	44
Ilustración 53: Paginación Tweets.....	44
Ilustración 54: Gráfico Frecuencia de HASHTAGS	45
Ilustración 55: Gráfico Sentimiento de Tweets.....	45
Ilustración 56: Gráfico Evolución Hashtags	46
Ilustración 57: Contaminación Vs Opinión Social.....	46
Ilustración 58: Comparación Provincias	48
Ilustración 59: Selección Contaminante.....	48
Ilustración 60: Mapa temporal provincia 1	49
Ilustración 61: Mapa temporal provincia 2	49
Ilustración 62: Menú selección de gráficos	49
Ilustración 63: Gráfico barras verticales.....	50
Ilustración 64: Gráfico barras horizontales	50
Ilustración 65: Gráfico lineal	50
Ilustración 66: Gráfico combo	51
Ilustración 67: Gráfico de área	51
Ilustración 68: Gráfico escalonado	51
Ilustración 69: Selección provincia.....	53
Ilustración 70: Selección Contaminante para mapa temporal.....	53
Ilustración 71: Mapa temporal Madrid NO2.....	53
Ilustración 72: Menú Gráficos una provincia	53
Ilustración 73: Gráfico de Barras Vertical Madrid	54
Ilustración 74: Gráfico Barras Horizontales Madrid	54
Ilustración 75: Gráfico Lineal Madrid	54
Ilustración 76: Gráfico Área Madrid	55
Ilustración 77: Gráfico Escalonado Madrid	55
Ilustración 78: Gráfico Barras Verticales CO y TOL	56

Ilustración 79: Gráfico Barras Horizontales CO y TOL	56
Ilustración 80: Gráfico lineal CO y TOL.....	56
Ilustración 81: Gráfico Área CO y TOL.....	57
Ilustración 82: Gráfico Escalonado CO y TOL	57
Ilustración 83: Menú Descarga Datasheets.....	58
Ilustración 84: Descarga Datasheets	58
Ilustración 85: Selección Provincia Datasheets.....	59
Ilustración 86: Mapa de estaciones de calidad de aire	59
Ilustración 87: Nombre de las estaciones de calidad de aire.....	60
Ilustración 88: Descarga Datasheets	60
Ilustración 89: Menú superior Contaminantes	61
Ilustración 90: Enlace Contaminantes.....	61
Ilustración 91: Selección de Contaminante.....	61
Ilustración 92: Ranking por Contaminante	62
Ilustración 93: Información Contaminante	62
Ilustración 94 PM10 y PM25	65

1. Introducción

El siguiente documento recoge la memoria de la realización del caso práctico de la asignatura Integración de Sistemas Informáticos del Máster en Ingeniería Informática de la Universidad Carlos III de Madrid. Este caso práctico tratará sobre el análisis y estudio de la contaminación atmosférica en toda España entre los años 2014 y 2015 en base a los datos recogidos por las estaciones de calidad del aire del país y a la opinión social de la gente a través de Twitter.

En este documento se mostrará, entre otras cosas, un resumen ejecutivo con el objetivo, alcance, principales aspectos claves del trabajo, etc., un análisis y diseño del problema de integración propuesto, unas visualizaciones y guía sobre la prueba de concepto, y finalmente, unas conclusiones sobre los objetivos alcanzados y problemas que se han tenido a lo largo de la elaboración del caso práctico.

Este trabajo ha sido realizado por el grupo 2 compuesto por los alumnos José Manuel Fernández Ruiz y Javier García Pérez.

A lo largo de la memoria se han añadido una serie de anexos con información complementaria sobre la realización del caso práctico.

2. Resumen ejecutivo

Este apartado recoge una panorámica del trabajo realizado. Los puntos a destacar de este apartado son:

- **Objetivo:** recoge las metas a conseguir.
- **Alcance:** la extensión que tendrá la solución propuesta.
- **Solución diseñada:** resumen sobre la solución diseñada.
- **Aspectos claves:** principales características que proporciona la solución diseñada.
- **Conclusiones:** breve resumen sobre lo principal de los apartados anteriores.

2.1. Objetivo

El objetivo del caso práctico es abordar el tema de la contaminación atmosférica en España desde diferentes puntos de vista para su posterior análisis. Para esta meta se han usado datos validados de las estaciones de calidad del aire de las diferentes provincias del país y opiniones sobre la contaminación realizadas por las personas en Twitter.

De esta manera, se pretende hacer una comparación de los niveles de contaminación reales extraídos de las estaciones con lo que realmente percibe y expresa la gente mediante sus mensajes en Internet. Se mostrarán niveles de todas las provincias de España según los datos oficiales de las estaciones de aire y de la misma manera, para estos lugares, comprobaremos si las personas también perciben y expresan este hecho mediante su opinión online. Ej. En Madrid el nivel medido de NO₂ es muy alto y se observan quejas en twitter de como el gran volumen de tráfico afecta al aire de la ciudad.

Toda esta información se mostrará en unos dashboards intuitivos y completos, que permitirán analizar una provincia de forma individual o comparándola con otra.

2.2. Alcance

Para lograr los objetivos anteriores se partirá de los datos recogidos por las estaciones de aire de España durante el año 2015, aunque se usarán datos de 2014 para todas esas estaciones que todavía no hayan validado sus datos de 2015. A la hora de mostrar la información en la solución se usarán medidas diarias.

Se empezará con un total de 331 estaciones de calidad del aire con sus respectivas medidas para las fechas mencionadas. Estas estaciones abarcan todo el territorio español salvo las provincias de las CCAA de Andalucía y Extremadura, y las ciudades autónomas de Ceuta y Melilla, debido a la falta de documentación compatible con la que se pudiera trabajar y extraer datos. Se pueden consultar el número de estaciones de aire recopiladas por provincia en la tabla del ANEXO I: *Cantidad de estaciones por provincia*. Los contaminantes medidos por las estaciones y que se analizarán en la solución final se pueden ver en el ANEXO II: *Elementos químicos y variables medidas*. Matizar que cada estación de calidad de aire no mide todos los contaminantes listados, si no que la cantidad que mida dependerá de su ubicación y función.

Por otro lado, en cuanto la opinión social, se recopilarán tweets entre 2014-2015 en base a mensajes publicados con los temas, hashtags y cuentas listados en el ANEXO III: *Temas, hashtags*

y cuentas Twitter. Para la prueba conceptual únicamente se usarán tweets entre el 1 de noviembre y 31 de diciembre debido a que la API de Twitter únicamente permite recuperar mensajes con antigüedad de una semana máximo.

2.3. Solución diseñada

Se ha diseñado una aplicación web que se lanza en un servidor Tomcat con Ubuntu como SO y desde la cual se puede acceder a diferentes visualizaciones en forma de gráficos y mapas de la contaminación atmosférica de las provincias de España.

Los datos usados para los gráficos y funcionalidades estarán guardados en colecciones de una base de datos MongoDB, mientras que los mapas se crearán a partir de sus CSVs de datos vía CartoDB, posteriormente el *iframe* usado para mostrar el mapa en la web también se almacenará en MongoDB.

2.4. Aspectos claves

La aplicación web permitirá realizar a los usuarios las siguientes funcionalidades durante su visita:

- Ver los datos de contaminación atmosférica de una provincia en base a la información extraída de las estaciones de calidad del aire. Esta información se mostrará en mapas y gráficos con evolución en el tiempo. Con la posibilidad de centrarse en un contaminante concreto en los mapas.
- Comparar la contaminación de dos provincias en base a la información extraída de las estaciones de calidad del aire. Esta información de la comparativa se mostrará en una misma página con mapas y gráficos del estilo de la primera funcionalidad, pero en este caso, se compararán contaminantes en vez de una visión general.
- Observar la opinión social sobre la contaminación en una provincia. Se mostrarán los tweets recogidos de esa provincia, relacionados con la contaminación, y gráficos en los que se podrá apreciar en porcentaje el sentimiento de los tweets (positivo, negativo y neutro) y como han ido evolucionando los hashtags con más menciones a lo largo del tiempo estudiado para esa provincia. Además, desde la misma sección se permitirá comparar la relación de evolución de los hashtags con la de los contaminantes de esa provincia.
- Posibilidad de descargar en formato CSV los datos de las medidas de las estaciones de calidad del aire ya preprocesados y limpiados en un formato estándar para todas las estaciones de todas las provincias. Esta característica está destinada a un perfil de usuario más técnico.
- Informarse sobre los riesgos a la salud de los contaminantes mostrados en la web según los niveles de exposición y, en base al contaminante elegido, ver el top 5 de provincias con niveles más altos del él.

2.5. Conclusiones

La solución diseñada permitirá visualizar en una aplicación web la contaminación atmosférica en España desde dos puntos de vista: el social y el medido por las estaciones de calidad del aire.

La información se orientará a nivel provincial y abarcará el tiempo transcurrido entre los años 2014 y 2015.

Los datos se mostrarán de forma visual mediante el uso de mapas y gráficos que permitirán ver la información de forma detallada con la posibilidad de ver medidas para fechas concretas.

3. Análisis y diseño del problema de integración propuesto

Este punto recoge de forma concisa el análisis y diseño del sistema desarrollado. En este apartado se busca obtener una especificación detallada del sistema que servirá como base para la explicación del diseño.

3.1. Análisis

Como se ha mencionado anteriormente, el principal objetivo de este apartado es obtener una especificación detallada del sistema que se va a desarrollar. Posteriormente, este análisis servirá de base para el diseño del sistema.

En esta fase de análisis se pretende definir el problema a resolver. Se trata de captar las necesidades que debe resolver y modelar el problema utilizando distintas técnicas, en función de las características del caso práctico, para posteriormente en el diseño del sistema resolverlo

3.1.1. Definición del sistema

Este apartado define la aplicación al completo. Todas las funcionalidades que debe cumplir y su alcance, la identificación del entorno tecnológico y sus dependencias.

3.1.1.1. Funcionalidad y alcance

Este apartado determina de forma detallada la funcionalidad y alcance del sistema que se introdujo en los puntos del apartado 2. *Resumen ejecutivo*.

El sistema es una aplicación web desarrollada con Java Servlets + MongoDB que mediante el uso de medidas de estaciones de calidad del aire y tweets sobre la contaminación en España recuperados por APIs y web crawlers permite ver esta información de forma visual mediante mapas y gráficos que proporcionan a los usuarios finales la posibilidad de ver el estado de la contaminación en las diferentes provincias de España desde dos puntos de vista diferentes:

- Estaciones de calidad del aire: mediante las medidas realizadas por estas estaciones los usuarios podrán ver la evolución de los principales contaminantes atmosféricos en las provincias españolas.
- Opinión social: mediante la recopilación de tweets en base a temas, hashtags y mensajes de cuentas relacionadas con la contaminación los usuarios podrán ver el estado de la contaminación en las provincias españolas en base a la opinión de la gente de estas zonas.

Como funcionalidades principales el sistema proporcionará a los usuarios las siguientes posibilidades:

- Visualización del estado de una provincia en base a las medidas recopiladas por sus estaciones de calidad del aire. Se usarán mapas creados con la herramienta CartoDB que se apoyará de los datos almacenados en ficheros CSV, cuyo contenido será obtenido por web crawlers que recogerán las medidas de las estaciones de aire de

los respectivos ayuntamientos de España. Estos mapas mostrarán la evolución del contaminante elegido a lo largo de un periodo de tiempo, para ello se usarán diferentes tonalidades de colores para indicar si el nivel del contaminante es muy alto, alto, medio, bajo o muy bajo en las diferentes zonas de la provincia donde haya estaciones de calidad del aire. También, para esta funcionalidad, se usarán diferentes tipos de gráficos creados con las librerías JavaScript de Google Charts, que usarán los datos de las medidas guardados en una colección de MongoDB. Estos gráficos mostrarán las medias mensuales de los contaminantes medidos en la provincia seleccionada por el usuario. De esta manera, se le da al usuario una visualización del estado general de toda la provincia mediante el mapa y, a la vez, de forma más detallada, usando los gráficos.

- Visualización del estado de una provincia en base a la opinión social. Recopilación de tweets usando la API de Twitter y posteriormente clasificados por provincias y sentimiento con la API de Meaningcloud, estos mensajes y sus metadatos serán almacenados en MongoDB. Los usuarios podrán visualizar los tweets relacionados con la contaminación de la provincia elegida. Además, podrán ver la frecuencia de las palabras claves que más se repiten en los tweets de esa provincia, el porcentaje de tweets que son negativos, positivos y neutros y, por último, ver la evolución en un periodo de tiempo de los hashtags principales relacionando estos con la evolución de los contaminantes medidos en esa provincia.
- Comparador de provincias en base a las medidas recopiladas por sus diferentes estaciones de aire. Usando las mismas herramientas de la primera visualización el usuario podrá hacer énfasis en un contaminante y comparar su evolución en dos provincias seleccionadas mediante mapas y gráficos de las mismas características que la primera visualización.
- Descargar los datasheets con las medidas de las estaciones en formato CSV con la misma estructura para todas las provincias de España. Destina a usuarios u organizaciones con un perfil más técnico. Se les permite de esta manera acceder a los datos de las estaciones de calidad ya limpios y preprocesados, en un formato estructurado para que así puedan trabajar con los datos de la contaminación en las provincias de España en un formato común, ahorrándoles la tarea de tener que limpiar y preprocesar ellos mismos los datos extraídos de los diferentes ayuntamientos de España.
- Visualización de los riesgos para la salud de los diferentes contaminantes analizados en la web y top 5 de provincias con los niveles más altos en estos contaminantes. Los usuarios podrán informarse de los riesgos para la salud según los niveles y tiempo de exposición de los contaminantes que se han analizado en las visualizaciones anteriores. Además, mediante gráficos de Google Charts los usuarios podrán conocer el top 5 de provincias cuya media de las medidas del contaminante elegido son las más altas.

En cuanto al alcance, el trabajo abarca toda España a nivel provincial. Las medidas y tweets mostrados serán de 2015, aunque se usarán medidas de estaciones de aire de 2014 cuando los datos de 2015 no hayan sido validados por sus respectivos ayuntamientos. De esta manera se permitirá a los usuarios poder analizar la contaminación en España provincia

a provincia, con datos reales validados de las estaciones de calidad del aire y con opiniones de los habitantes de las provincias analizadas.

Las fuentes de los datos, así como la forma de extracción de estos y las herramientas usadas, se explicarán más adelante en el apartado 3.1.2. *Fuentes de datos* del análisis.

3.1.1.2. *Entorno tecnológico*

El siguiente apartado describe las herramientas software usadas para el desarrollo del caso práctico, así como las características hardware necesarias para hacer funcionar correctamente la VM con la solución.

3.1.1.2.1. Componentes software

A continuación, se describen las herramientas y componentes software usados para desarrollar la solución final:

- **Apache Tomcat**



Es el servidor de aplicaciones sobre el que se desplegará la solución a mostrar. La versión utilizada es la 8.0.29 y el puerto el 8080.

- **CartoDB**



Herramienta usada para la creación de los mapas de las provincias con la evolución de los contaminantes durante un periodo de tiempo medidos por sus estaciones. Utiliza los ficheros CSV limpios y preprocesados con las medidas de las estaciones de aire que han sido recopiladas de las diferentes fuentes de datos.

Los mapas realizados han sido exportados mediante un *iframe* y almacenado este en MongoDB para su posterior uso en la aplicación web.

- **Google APIs**



Se usan la API de Google Maps Geocoding y Google Charts. La primera servirá para automatizar el proceso de clasificación de las estaciones de calidad de aire por provincia en base a sus coordenadas o ciudad a la que pertenecen.

La segunda es una librería JavaScript utilizada para crear gráficos relacionados con las medidas de las estaciones de aire y los tweets. Gráficos de barras, columnas, tarta, etc.

- **Meaning Cloud**



Esta herramienta software será utilizada para clasificar los tweets por provincias y sentimiento. Se usará la herramienta *Topics Extraction* para clasificar los tweets almacenados en MongoDB por provincias según su texto del mensaje y metadatos, para ello

se buscarán entidades que sean lugares de España y conceptos que tengan que ver con la contaminación. A partir de los tweets clasificados se asignarán un sentimiento (positivo, negativo o neutro) en base a su texto con la herramienta *Sentiment Analysis*.

La aplicación web al estar desarrollada en Java usa las APIs proporcionadas por Meaning Cloud para que actúen como interfaces de comunicación para utilizar estas herramientas software descritas anteriormente, se usa la versión 2.0 para ambas.

- **MongoDB**



Es el tipo de base de datos usada en el caso práctico. Esta base de datos será la encargada de almacenar todos los datos usados y recopilados de las diferentes fuentes de información (medidas de estaciones de aire, tweets sobre la contaminación, *iframes* de CartoDB, ...). En el apartado 3.2. *Diseño* se hará una explicación de la estructura de las colecciones creadas para trabajar de manera eficiente con los datos extraídos de internet.

Se usa la versión 3.0.8 de MongoDB y como interfaz de comunicación de la aplicación web con la base de datos se utiliza el driver de MongoDB para Java versión 3.1.0. El puerto usado es el 27017.

- **Twitter API**



API de Twitter utilizada para recopilar los mensajes de twitter relacionados con la contaminación en base a temas, hashtags o mensajes de cuentas. Se usará la versión 4.0.4 de la librería en Java.

3.1.1.2.2. Componentes hardware

Para montar la VM entregada con la solución se recomienda aplicar los siguientes requisitos mínimos en Oracle VM VirtualBox para garantizar su correcto funcionamiento:

- 2 procesadores
- 6 GB de RAM
- 20 GB de disco duro
- Conexión a internet

3.1.1.3. Dependencias

La solución final dependerá de varios subsistemas para su correcto funcionamiento. El primero es el servidor de aplicaciones, del cual habrá que garantizar su disponibilidad para asegurarse del correcto funcionamiento de la aplicación web donde se muestran las visualizaciones.

Por otro lado, la solución final también tendrá una dependencia con la base de datos donde se almacena la información extraída de las fuentes de información. Al igual que con el servidor de simulaciones, habrá que garantizar su disponibilidad, pero también la

integridad de sus datos almacenados, para que así, toda la información que se muestre en la aplicación web sea correcta y esté exenta de fallos.

3.1.2. Fuentes de datos

Este apartado tratará sobre el origen de los datos usados y de las herramientas y procedimientos usados para extraerlos desde sus fuentes.

Para la información de las medidas de las estaciones de calidad del aire se han recogido los datos de páginas web de ayuntamientos y comunidades autónomas. En el *ANEXO IV: Fuentes de datos* se pueden ver todas las páginas web que han servido de fuente de datos para esta tarea.

En cuanto a la información de la opinión social se han usado mensajes extraídos de twitter, para ello se ha usado la API de Twitter para que recopilará tweets en base a los temas, hashtags y cuentas recogidas en el *ANEXO III: Temas, hashtags y cuentas Twitter*.

3.1.2.1. Extracción de los datos

Para la extracción de los datos se usan varias herramientas como web crawlers, pdf readers y csv readers, que son los encargados de recoger las medidas para que luego se puedan pasar a la base de datos desde un formato estándar.

Se usará un web crawler para recoger las medidas de las páginas web donde se muestren directamente a la vista del público, esta herramienta se encargará de coger los valores y luego estos serán escritos en un fichero CSV con un formato estándar mediante la librería de Java llamada [JavaCSV](#). La siguiente imagen muestra un ejemplo para las provincias gallegas donde se puede usar un web crawler para obtener los valores de las medidas.

Consulta de datos de Calidad del Aire																
Redes	Datos de la Estación Santiago - San Caetano															
Estaciones	NO ₂ -NO ₂ -NO _x	NO ₂ -O ₃	SO ₂ -CO	PM10	PM25	BEN	PM10	PM25	SO ₂ ug/m ³	NO ₂ ug/m ³	NO ₂ ug/m ³	NO _x ug/m ³	CO ug/m ³	O ₃ ug/m ³	PM10 ug/m ³	PM25 ug/m ³
Parámetros de la Estación																
Santiago - San Caetano	1	T. 1	T. 2	T. 4	T. 0.1	T. 79	T. 14	T. 13	T. 0.01	T. 0	N. 13	T.				
28/12/15 01:00	1	T. 1	T. 2	T. 3	T. 0.1	T. 73	T. 21	T. 14	T. 0.01	T. 0	N. 14	T.				
28/12/15 02:00	1	T. 1	T. 2	T. 3	T. 0.1	T. 74	T. 23	T. 13	T. 0.01	T. 0	N. 13	T.				
28/12/15 03:00	1	T. 1	T. 2	T. 3	T. 0.1	T. 74	T. 22	T. 13	T. 0.01	T. 0	N. 13	T.				
28/12/15 04:00	1	T. 1	T. 2	T. 4	T. 0.1	T. 79	T. 20	T. 15	T. 0.01	T. 0	N. 15	T.				
28/12/15 05:00	1	T. 1	T. 2	T. 4	T. 0.1	T. 72	T. 15	T. 14	T. 0.01	T. 0	N. 14	T.				
28/12/15 06:00	1	T. 1	T. 3	T. 5	T. 0.1	T. 72	T. 16	T. 15	T. 0.01	T. 0	N. 15	T.				
28/12/15 07:00	1	T. 1	T. 6	T. 8	T. 0.1	T. 67	T. 16	T. 12	T. 0.01	T. 0	N. 12	T.				
28/12/15 08:00	1	T. 2	T. 6	T. 9	T. 0.1	T. 65	T. 14	T. 13	T. 0.01	T. 0	N. 11	T.				
28/12/15 09:00	1	T. 2	T. 8	T. 11	T. 0.1	T. 60	T. 10	T. 11	T. 0.01	T. 0	N. 11	T.				
28/12/15 10:00	1	T. 2	T. 9	T. 11	T. 0.1	T. 68	T. 13	T. 10	T. 0.01	T. 0	N. 10	T.				
28/12/15 11:00	1	T. 2	T. 9	T. 11	T. 0.1	T. 68	T. 13	T. 10	T. 0.01	T. 0	N. 10	T.				
28/12/15 12:00	1	T. 3	T. 13	T. 17	T. 0.1	T. 68	T. 13	T. 10	T. 0.01	T. 0	N. 10	T.				
28/12/15 13:00	1	T. 3	T. 19	T. 22	T. 0.1	T. 63	T. 15	T. 12	T. 0.01	T. 0	N. 12	T.				
28/12/15 14:00	1	T. 3	T. 17	T. 21	T. 0.1	T. 63	T. 17	T. 12	T. 0.01	T. 0	N. 12	T.				
28/12/15 15:00	1	T. 4	T. 22	T. 27	T. 0.1	T. 55	T. 19	T. 13	T. 0.01	T. 0	N. 13	T.				
28/12/15 16:00	1	T. 4	T. 21	T. 26	T. 0.1	T. 56	T. 18	T. 12	T. 0.01	T. 0	N. 12	T.				
28/12/15 17:00	1	T. 4	T. 28	T. 33	T. 0.1	T. 50	T. 17	T. 11	T. 0.01	T. 0	N. 11	T.				
28/12/15 18:00	1	T. 4	T. 28	T. 33	T. 0.1	T. 50	T. 18	T. 12	T. 0.01	T. 0	N. 12	T.				
28/12/15 19:00	1	T. 4	T. 30	T. 35	T. 0.1	T. 49	T. 18	T. 15	T. 0.01	T. 0	N. 15	T.				
28/12/15 20:00	1	T. 2	T. 24	T. 27	T. 0.1	T. 52	T. 16	T. 14	T. 0.01	T. 0	N. 14	T.				
28/12/15 21:00	1	T. 2	T. 23	T. 26	T. 0.1	T. 53	T. 18	T. 12	T. 0.01	T. 0	N. 12	T.				
28/12/15 22:00	1	T. 2	T. 20	T. 22	T. 0.1	T. 54	T. 16	T. 12	T. 0.01	T. 0	N. 12	T.				
28/12/15 23:00	2	T. 2	T. 16	T. 18	T. 0.1	T. 59	T. 17	T. 13	T. 0.01	T. 0	N. 13	T.				
Realizar Consulta	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Las consultas se realizarán eligiendo uno o más parámetros de una estación concreta para la red seleccionada.																
Se podrá definir como máximo un período de consulta de un mes tanto para horas como para medidas diarias.																

Ilustración 1 Datos aire Galicia

La ejecución del web crawler, y posterior escritura de los valores a un fichero CSV, se automatizará para que se ejecute mensualmente en aquellas páginas donde el formato de muestra de las medidas de calidad del aire sea compatible con este.

Por otro lado, hay muchas páginas de provincias donde estos datos únicamente se dan mediante la descarga de un fichero CSV o PDF. Para estos casos, se automatizarán las descargas de estos ficheros para que se realicen mensualmente y se usará la librería de java para leer y escribir en CSV mencionada anteriormente y en el caso del que el fichero descargado esté en formato PDF se usará la librería [PDFBox](#) que transformará el contenido a formato de texto desde el cual se partirá para transformarlo al formato CSV estándar que se usará en el futuro. En las imágenes siguientes se muestra un ejemplo donde habría que realizar esta tarea.

The screenshot shows a web page titled 'calidad del aire'. On the left, there's a sidebar with links like 'Contaminación atmosférica', 'La Contaminación Acústica', 'El Sistema Integral de la Calidad del Aire', and 'Consulta de datos' (which is currently selected). Under 'Consulta de datos', there are links for 'Mapa de la red', 'Informes', 'Representación gráfica', 'Sistema de Predicción', 'Contaminación acústica', 'Boletín meteorológico diario', 'Datos históricos' (selected), 'Cuestionario de satisfacción', 'Preguntas frecuentes', 'Acciones ciudadanas', 'Publicaciones', 'Servicios al ciudadano', and 'Calidad y gestión ambiental'. The main content area has a section titled 'Descarga de datos' with instructions: 'Junto a la descarga de los ficheros solicitados (horarios o diarios), se podrá descargar/consultar el archivo de interpretación de su estructura y posterior tratamiento. INTPHORA-DIA'. It also includes fields for 'Empresa/Institución:' and 'E-mail(*)', and dropdown menus for 'Archivos de contaminación atmosférica' (with options 'Valores medios diarios y horarios' and 'Valores medios mensuales'), and 'Archivos de contaminación acústica' (with option 'Valores mensuales'). At the bottom right is a blue 'Enviar' button.

Ilustración 2 Datos aire Madrid



Sistema de Vigilancia de la Calidad del Aire del Ayuntamiento de Madrid Red de Calidad del Aire 2801

Informe Anual de Valores Medios Mensuales.

Fecha: 2015

Estación: 28079004. Plaza de España

Fecha	SO2	CO	NO	NO2	VV	DV	TMP	HR	LL
	µg/m³	mg/m³	µg/m³	µg/m³	m/s	-	°C	%	l/m³
Enero	12	0.7	115	63	1.00	6	5.7	80	0.45
Febrero	7	0.3	27	46	1.50	7	7.6	80	0.55
Marzo	8	0.4	42	54	1.72	6	11.8	72	1.06
Abril	6	0.3	17	46	1.09	6	15.5	70	1.17
Mayo	6	0.3	15	42	1.24	6	20.6	67	0.00
Junio	5	0.2	11	48	0.68	98	25.0	63	1.53
Julio	5	0.3	11	48	0.46	150	30.0	62	0.48
Agosto	6	0.2	10	33	0.53	132	26.0	72	0.06
Septiembre	7	0.3	24	47	0.46	151	20.7	78	0.00
Octubre	7	0.4	47	52	0.38	177	15.8	-	1.60
Noviembre	10	0.6	96	61	0.40	153	11.7	-	0.85
Diciembre	-	-	-	-	-	-	-	-	-
Máximo	18	1.0	209	106	1.72	206	30.0	80	1.60
Mínimo	5	0.2	10	33	0.21	6	5.7	62	0.00
Promedio	8	0.4	52	54	0.81	91	16.4	72	0.65

Ilustración 3 Datos aire Madrid PDF

Todos estos valores de las medidas que han sido recopilados con las herramientas descritas anteriormente se escribirán en ficheros CSV con un formato estándar (uno por estación de medida). Estos ficheros CSV servirán para luego agruparlos por provincias y crear los mapas en CartoDB y para introducirlos en la colección de MongoDB para usar sus valores en futuros gráficos y análisis. En el apartado de diseño se describirá en detalle este procedimiento.

En cuanto a la opinión social, como ya se ha mencionado antes se va a usar la API de Twitter para recoger los tweets relacionados con la contaminación. Este proceso se automatizará para que se ejecute semanalmente. En este proceso se recogerán todos los tweets relacionados con la contaminación y se guardarán en una colección en MongoDB su mensaje y metadatos. Una vez guardados se usarán las herramientas de Meaning Cloud descritas en el apartado de 3.1.1.2.1. *Componentes software* para:

- Clasificar los tweets almacenados por provincias en base a la información guardada en la colección de MongoDB.
- Asignar a los tweets clasificados por provincias un sentimiento de negativo, positivo o neutro en base al contenido del mensaje.

Todo este proceso se describirá en detalle y paso a paso en el apartado del diseño, en el cual se verán las funcionalidades usadas para conseguir este objetivo.

3.1.3. Modelo de datos

En esta tarea se detalla el modelo de datos de la aplicación, es decir, toda la información que se almacenará en el sistema y que formará el punto de partida para la base de datos.

A modo resumen se guardarán los valores tomados por cada una de las estaciones de medida, la información y límites sobre los contaminantes analizados, los tweets recogidos, los tweets clasificados por provincia y los iframes de los mapas generados por CartoDB.

Aunque la base de datos que se usa no es relacional, se usa MongoDB, se adjunta este diagrama entidad-relación, sin atributos, para entender la estructura de la base de datos.

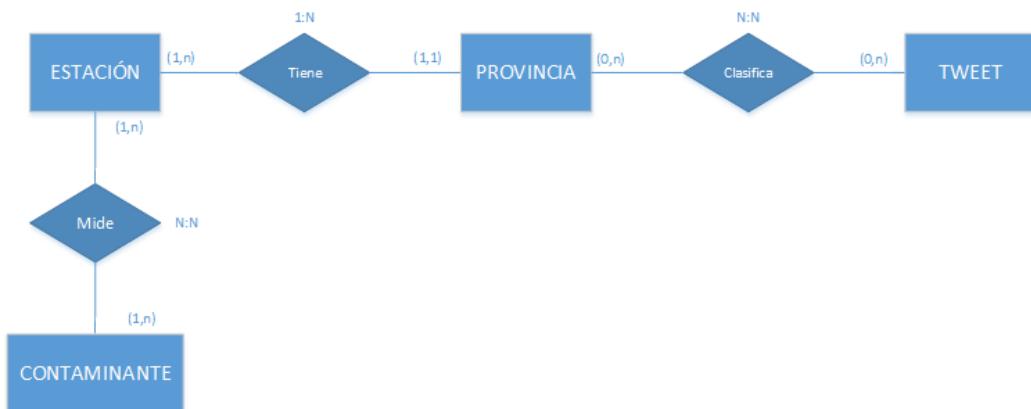


Ilustración 4 Diagrama E-R de la base de datos

Una estación únicamente puede pertenecer a una provincia, mientras que una provincia puede albergar varias estaciones. Un contaminante puede ser medido por varias estaciones, a la vez que una estación puede medir varios contaminantes. En cuanto a los tweets, en base a su contenido, un tweet puede estar relacionado con varias o ninguna provincia, mientras que una provincia puede tener clasificados varios o ningún tweet.

En el apartado de 3.2.4. *Diseño del modelo de datos* se describirán en detalle los atributos de cada una de las entidades y la estructura estándar que tendrá el documento en las colecciones de MongoDB.

3.2. Diseño

En este apartado se va a resolver el problema descrito e introducido en la sección de 3.1. Análisis. Con este apartado se busca definir completamente la arquitectura del sistema y su funcionamiento.

3.2.1. Diseño de clases

A lo largo de este apartado se explicará la funcionalidad aportada por cada una de las clases del sistema desarrollado.

Se ha dividido este apartado en dos partes: extracción de datos y aplicación web. En el primero se hablará de las clases que intervienen en el proceso de recopilación de la información de las fuentes de datos y en el posterior proceso de limpieza y preprocesado. En el segundo se hablará de la funcionalidad de las clases que realizan la tarea de visualización de la información en la aplicación web diseñada para los usuarios finales.

3.2.1.1. Extracción y almacenamiento de los datos

Este apartado engloba la definición de las clases y funcionalidades relacionadas con la extracción de los datos de las diferentes fuentes consultadas. Estos datos son los valores medidos por las estaciones de aire de las diferentes provincias de España, cuyos orígenes se pueden consultar en el *ANEXO IV: Fuentes de datos*. Otra fuente de datos ha sido Twitter, los mensajes recogidos responden a las pautas marcadas en el *ANEXO III: Temas, hashtags y cuentas Twitter*.

A lo largo de este apartado se definirán tres procesos principales: la extracción de los datos de las fuentes, la limpieza y preprocesado de estos para su uso posterior y su introducción en la base de datos MongoDB.

El apartado se divide en dos subapartados, uno para la recopilación de los datos de las estaciones de calidad del aire y otro para la de los mensajes de Twitter.

3.2.1.1.1. Calidad del aire

La funcionalidad de este proceso se centra en una clase llamada Aire, la cual se apoyará en la API de Google Maps Geocoding, en la librería JavaCSV y en web crawlers.

Antes de empezar con la explicación de la funcionalidad de la clase se proporciona una caja UML del contenido de esta, para que de esta manera la explicación siguiente sea lo más clara posible.

```

Aire

-client : MongoClient
-database : MongoDBDatabase
-collection : MongoCollection<Document>
-limits : MongoCollection<Document>
-limits : List<Document>
-formatoFecha : SimpleDateFormat

+Aire()
-insertarEstaciones() : void
-CSVgrande() : void
-optimizarDB() : void
-cogerCabezas(file : CsvWriter) : void
-parametrosMedicion(meidas : CsvReader) : String[]
-cogerMedidasCSV(file : CsvWriter, mediciones : String[]) : void
-insertarLimites() : void
-cogerMedidasDecimal(media : Document, mediciones : String[]) : Document

```

Ilustración 5 Aire.java

Antes de empezar a recolectar las medidas de las estaciones de calidad del aire, la clase, en su primera ejecución, almacenará en la base de datos los contaminantes a analizar, sus límites y sus efectos para la salud. Estos valores se han obtenido de las páginas listadas en el *ANEXO II: Elementos químicos y variables medidas*. Este proceso será realizado por la función *insertarLimites()* y los valores recogidos se insertarán en la colección *Contaminante* la cual tiene la estructura mostrada en la Ilustración 43 *Contaminante JSON*.

Las páginas consultadas para la extracción de las medidas de las estaciones de calidad del aire de las provincias de España están listadas en el *ANEXO IV: Fuentes de datos*. En esta lista encontramos dos tipos de páginas, las que muestran directamente en ellas los datos de las medidas y las que solo dan links de descarga a ficheros CSV con las medidas.

Para las primeras usaremos la librería [Jsoup](#), la cual se encargará de capturar el HTML donde estén las medidas, generalmente será buscar los datos dentro de las etiquetas *<table>*, *<td>* y *<tr>* de las páginas que se le pase al crawler. Estos datos capturados serán escritos en un fichero CSV, uno por cada estación, cuyo formato se explicará más adelante.

Consulta de datos de Calidad del Aire													
Redes		Datos de la Estación Lugo											
Estaciones		Datos	NO-NO ₂ -NOx	NO ₂ -03	SO ₂ -CO	PM10	SO ₂ µg/m ³	NO µg/m ³	NO ₂ µg/m ³	NOX µg/m ³	CO mg/m ³	O ₃ µg/m ³	PM10 µg/m ³
Red Xunta de Galicia													
Lugo													
Parámetros de la Estación													
SO ₂ - Fluorescencia ultravioleta													
NO - Quimioluminiscencia													
NO ₂ - Quimioluminiscencia													
NOX - Quimioluminiscencia													
Periodo													
Horarios													
Fecha de Comienzo													
01/12/2015													
Fecha de Fin													
31/12/2015													
Realizar Consulta													
Las consultas se realizarán eligiendo uno o más parámetros de una estación en concreto para la red seleccionada.													
Se podrá definir como máximo un periodo de consulta de un mes tanto para datos horarios como para medias diarias.													
01/12/15 00:00	3	V	40	V	17	V	79	V	0.26	V	1	V	20
01/12/15 01:00	2	V	36	V	17	V	72	V	0.24	V	1	V	20
01/12/15 02:00	2	V	24	V	18	V	55	V	0.18	V	1	V	23
01/12/15 03:00	2	V	10	V	17	V	31	V	0.14	V	1	V	23
01/12/15 04:00	2	V	8	V	16	V	28	V	0.13	V	1	V	11
01/12/15 05:00	2	V	7	V	15	V	25	V	0.12	V	1	V	10
01/12/15 06:00	2	V	15	V	16	V	40	V	0.12	V	1	V	12
01/12/15 07:00	3	V	25	V	16	V	54	V	0.13	V	1	V	20
01/12/15 08:00	2	V	39	V	20	V	80	V	0.13	V	1	V	21
01/12/15 09:00	2	V	25	V	14	V	52	V	0.12	V	3	V	13
01/12/15 10:00	2	V	26	V	17	V	58	V	0.15	V	5	V	15
01/12/15 11:00	3	V	35	V	18	V	72	V	0.23	V	6	V	38
01/12/15 12:00	4	V	28	V	15	V	57	V	0.14	V	8	V	42
01/12/15 13:00	7	V	20	V	15	V	46	V	0.13	V	12	V	28
01/12/15 14:00	8	V	17	V	20	V	45	V	0.13	V	15	V	23
01/12/15 15:00	8	V	12	V	21	V	39	V	0.11	V	17	V	30
01/12/15 16:00	7	V	10	V	26	V	42	V	0.11	V	14	V	18
01/12/15 17:00	8	V	39	V	41	V	101	V	0.15	V	3	V	22
01/12/15 18:00	5	V	78	V	42	V	163	V	0.32	V	1	V	22
01/12/15 19:00	6	V	62	V	36	V	131	V	0.29	V	1	V	32
01/12/15 20:00	6	V	41	V	33	V	97	V	0.26	V	1	V	32
01/12/15 21:00	5	V	37	V	35	V	91	V	0.28	V	1	V	32
01/12/15 22:00	4	V	39	V	29	V	90	V	0.32	V	1	V	26
01/12/15 23:00	3	V	32	V	24	V	73	V	0.32	V	1	V	23
01/12/15 23:59	?	v	+	v	+	v	+	v	+	v	+	v	+

Ilustración 6 Datos aire Galicia

Para el segundo tipo de páginas, las que sólo proporcionan links de descargas, se automatizará la descarga de los ficheros donde están las medidas. Para ello se obtendrán los

links de descargas también con Jsoup. Una vez descargados estos ficheros se procederá a su lectura. La lectura dependerá del formato, si el fichero es un CSV o Excel se usará la librería JavaCSV, en caso de ser un PDF se usará la librería [PDFBox](#). Tras este paso se pasará a su limpieza y transformación al formato estándar de fichero CSV que se explicará ahora. En resumen, primero se leerá el fichero descargado para obtener sus medidas y luego se escribirá esta información en un fichero CSV con el formato establecido.

Datos validados mensuales (xls)		
ENERO 2013	ENERO 2014	ENERO 2015
FEBRERO 2013	FEBRERO 2014	FEBRERO 2015
MARZO 2013	MARZO 2014	MARZO 2015
ABRIL 2013	ABRIL 2014	ABRIL 2015
MAYO 2013	MAYO 2014	MAYO 2015
JUNIO 2013	JUNIO 2014	JUNIO 2015
JULIO 2013	JULIO 2014	JULIO 2015
AGOSTO 2013	AGOSTO 2014	AGOSTO 2015
SEPTIEMBRE 2013	SEPTIEMBRE 2014	SEPTIEMBRE 2015
OCTUBRE 2013	OCTUBRE 2014	OCTUBRE 2015
NOVIEMBRE 2013	NOVIEMBRE 2014	
DICIEMBRE 2013	DICIEMBRE 2014	

Informe meteorológico anual de la Red de Control y Vigilancia de Castilla-La Mancha 2008

Ilustración 7 Datos aire Castilla-La Mancha

Los ficheros CSV resultantes tras la limpieza y preprocesado de los datos se han generado con la librería JavaCSV y hay uno por cada estación. Todos estos ficheros siguen un formato estándar para todas las estaciones de la provincia. Estos ficheros tendrán el nombre de la estación a la que pertenecen y como columnas la fecha, hora (si aplica) y tantas columnas de contaminantes como mida. A continuación, se pone un ejemplo para el CSV de la estación de Aguimes en Gran Canaria:

	A	B	C	D	E	F	G	H
1	Fecha	Hora	SO2	NO	NO2	PM10	PM25	O3
2	01/01/2014	0:00:00	3	1	15	43	13	58
3	01/01/2014	1:00:00	3	1	8	38	8	61
4	01/01/2014	2:00:00	3	1	6	25	5	60
5	01/01/2014	3:00:00	3	1	5	24	6	61
6	01/01/2014	4:00:00	3	1	6	28	9	64
7	01/01/2014	5:00:00	3	1	3	23	6	66
8	01/01/2014	6:00:00	3	1	5	22	7	65
9	01/01/2014	7:00:00	3	1	15	22	7	61
10	01/01/2014	8:00:00	3	1	7	18	6	72
11	01/01/2014	9:00:00	3	1	4	16	4	76
12	01/01/2014	10:00:00	3	1	3	19	5	76
13	01/01/2014	11:00:00	3	1	4	19	4	75
14	01/01/2014	12:00:00	3	1	6	21	4	75
15	01/01/2014	13:00:00	3	1	4	22	5	79
16	01/01/2014	14:00:00	3	1	6	21	6	77
17	01/01/2014	15:00:00	3	1	4	20	6	76
18	01/01/2014	16:00:00	3	1	10	18	6	66
19	01/01/2014	17:00:00	3	1	22	16	6	50
20	01/01/2014	18:00:00	3	1	13	15	6	61
21	01/01/2014	19:00:00	3	1	9	15	6	65
22	01/01/2014	20:00:00	3	1	3	16	6	73
23	01/01/2014	21:00:00	3	1	2	14	5	75
24	01/01/2014	22:00:00	3	1	8	14	4	63
25	01/01/2014	23:00:00	3	1	29	16	5	45
26	02/01/2014	0:00:00	3	1	31	20	6	41
27	02/01/2014	1:00:00	6	5	49	19	6	26
28	02/01/2014	2:00:00	3	1	11	20	6	59
29	02/01/2014	3:00:00	3	1	4	20	6	64
30	02/01/2014	4:00:00	3	1	7	20	6	64

Ilustración 8 AGUIMES.csv

Además de generarse un CSV por estación, se tendrá otro donde se recogerá el nombre de la estación, tendrá que ser el mismo que el nombre del fichero que recoja sus datos, y la longitud y latitud de esta. Por ejemplo, para el caso anterior, en este nuevo fichero tendremos una entrada para Aguimes, con su longitud y latitud, pero aparte existirá otro fichero llamado Aguimes.csv que será el que recoge las medidas por fechas. El formato del nuevo CSV es el siguiente:

	A	B	C
1	Nombre	Longitud	Latitud
2	Aguimes	-15.44503	27.89764
3	Arafo	-16.39972	28.33722
4	Arinaga	-15.38715	27.86907
5	Arrecife	-13.54758	28.97699
6	Barranco Hor	-16.3581	28.39341
7	Buzanada	-16.65302	28.07255
8	Caletillas	-16.36193	28.37671
9	Casa Cuna	-16.27769	28.45102
10	Castillo del F	-15.46116	27.80128
11	Centro de Ar	-13.8524	28.50026
12	Ciudad Depo	-13.55586	28.96787
13	Costa Teguis	-13.51655	28.99031
14	El Rio	-16.52369	28.14509
15	Galletas	-16.65591	28.00774
16	Granadilla	-16.57758	28.11249
17	Igualero	-16.37199	28.38054
18	Jinamar	-15.41631	28.03512
19	Los Realejos	-16.57067	28.3831
20	Medano	-16.53605	28.04733
21	Mercado Cer	-15.43282	28.13373
22	Parque de la	-13.8537	28.50276
23	Pedro Lezcar	-15.41609	28.02985
24	Playa del Ing	-15.56385	27.76365
25	San Agustin	-15.54188	27.7726
26	San Isidro	-16.55986	28.08001
27	Tajao	-16.47163	28.11137
28	Telde	-15.41466	28.02632
29	Parque de Sa	-15.41185	28.00365
30	Torre Caño	-16.2619	28.46217

Ilustración 9 estaciones.csv

Con estos dos tipos de ficheros CSV ya se puede realizar el proceso de inserción en la colección de MongoDB desde la función `insertarEstaciones()`. Esta función recorrerá cada línea del fichero `estaciones.csv`, cogerá el valor de la columna `Nombre` y buscará el fichero CSV correspondiente a ese nombre, que será el que recoge todas las medidas de los contaminantes realizadas por la estación. Por otro lado, con las columnas `Longitud` y `Latitud` obtendrá la ciudad, provincia y país al que pertenece la estación usando la API de Google Maps Geocoding. Una vez obtenidas las medidas y la localización exacta de la estación se generará un documento JSON que se insertará en la colección de MongoDB llamada `Estación`, la estructura de esta colección se puede apreciar en la Ilustración 42 Estación JSON. Tras la inserción en la base de datos con la función `optimizarDB()` se crearán una serie de índices, por ejemplo el de provincias, para mejorar la eficiencia en las consultas y agregaciones que se hagan posteriormente.

Una vez introducida cada estación, con sus medias y ubicación en la base de datos, se procede a la creación de un CSV por provincia con todas las medidas de sus estaciones. Estos CSV nuevos serán los que se usarán para crear los mapas en CartoDB. Este proceso se realizará en la función `CSVGrande()`, la cual recorrerá la colección `Estación` y agrupará sus JSON por provincias mediante una agregación para generar un CSV por cada una, el cual será mandado a CartoDB donde se creará un mapa dinámico temporal por cada contaminante medido por la estación. El iframe de los mapas creados se guardará en la colección `Provincia`, cada uno con su provincia correspondiente. La estructura de la colección `Provincia` es la mostrada en la Ilustración 45 Provincia JSON.

3.2.1.1.2. Twitter

La funcionalidad de este proceso se centra en una clase llamada *Personas*, la cual se apoyará en la API de Meaning Cloud para los procesos de clasificación de tweets por provincias y sentimiento.

Antes de empezar con la explicación de la funcionalidad de la clase se proporciona una caja UML del contenido de esta, para que de esta manera la explicación siguiente sea lo más clara posible.



Ilustración 10 Personas.java

En primer lugar, para la extracción de tweets se usa la API de Twitter. Este proceso se desarrollará en el método *quejasTwitter()*. En este método se recogerán todos los tweets relacionados con los temas, hashtags y cuentas recogidas en el *ANEXO III: Temas, hashtags y cuentas Twitter*. Una vez recopilados todos se pasa a la limpieza de los tweets devueltos.

El objeto del tweet en sí tiene muchos atributos, pero solo se guardará la id, el mensaje, la fecha, el usuario, los hashtags del mensaje, el idioma, la localización puesta en el perfil del usuario, la localización desde donde se mandó el tweet y la geolocalización desde donde se mandó el tweet. Estos tres últimos atributos pueden que sean nulos en el objeto tweet, de ahí que intentemos recopilar tantos para luego poder clasificar el mensaje en una provincia.

Cada tweet limpiado es añadido a una lista, luego esta lista se inserta a la colección *Tweet* de MongoDB, previa comprobación de si la ID de ese tweet ya existe en la colección. La estructura de la colección *Tweet* se puede apreciar en la *Ilustración 44 Tweet JSON*.

Una vez realizada esta primera inserción se procede a hacer una segunda limpieza con el método *limpiarTweets()*, este método recorrerá la colección y borrará todos los tweets cuyo atributo *place*, *localización* o *geo* sea lugares fuera de España.

Tras la limpieza, preprocessado y almacenamiento de los mensajes en la colección *Tweet* se procede a su clasificación por provincias usando el método *clasificarTweets()*. Aquí se usará la API de Meaning Cloud llamada Topics Extraction la cual primero comprobará si el mensaje tiene alguna palabra relacionada con el medio ambiente, si esto no es así se procede a borrar ese tweet de la colección. Si el mensaje tiene contenido relacionado con el medio ambiente, por orden, en base a los atributos *contenido*, *geo*, *place* y *localización*, intentará ubicar al tweet en una provincia de España. Si no se consigue ubicar, o la ubicación es de fuera de España, se borra el tweet de la colección. Si se consigue ubicar se llama al

método *pasarContenidoTweet()* y se almacena el id y usuario del tweet en la colección *Provincia* para la provincia o provincias correspondientes (puede que el tweet se queje de dos sitios diferentes). La estructura de la colección *Provincias* se puede apreciar en la *Ilustración 45 Provincia JSON*.

Tras finalizar este proceso de clasificación de tweets, únicamente deberían quedar en la colección *Tweet* los tweets, con todos sus atributos, que han sido clasificados. Y en la colección *Provincia* esos mismos tweets, pero únicamente almacenando la *id* y el *usuario*. Estos dos atributos, más un script, serán los únicos necesarios para mostrar el mensaje íntegro en la aplicación web, mientras que los demás atributos permanecerán en la colección *Tweet* y serán accesibles si se busca por la *id* del tweet deseado.



Ilustración 11 Twitter widget

Con la clasificación resuelta, el último paso es asignar sentimiento a los tweets almacenados en la colección *Provincia*. Para esto se usará la API de Meaning Cloud de Sentiment Analysis, la cual devolverá si el contenido del mensaje del tweet es positivo, negativo o neutro. Este resultado se añadirá como atributo del documento en la colección de *Provincia*.

3.2.1.2. Aplicación web

La aplicación web desarrollada sigue una arquitectura MVC (modelo vista controlador). Para adaptar la solución a esta arquitectura se ha usado la tecnología Java Servlets.

Se ha optado por la arquitectura MVC debido a su sencillez, su reducción de dependencias, su reutilización y por la flexibilidad que proporciona ante futuros cambios.

Las siguientes imágenes muestran como son los diferentes niveles de la arquitectura elegida para la aplicación web.

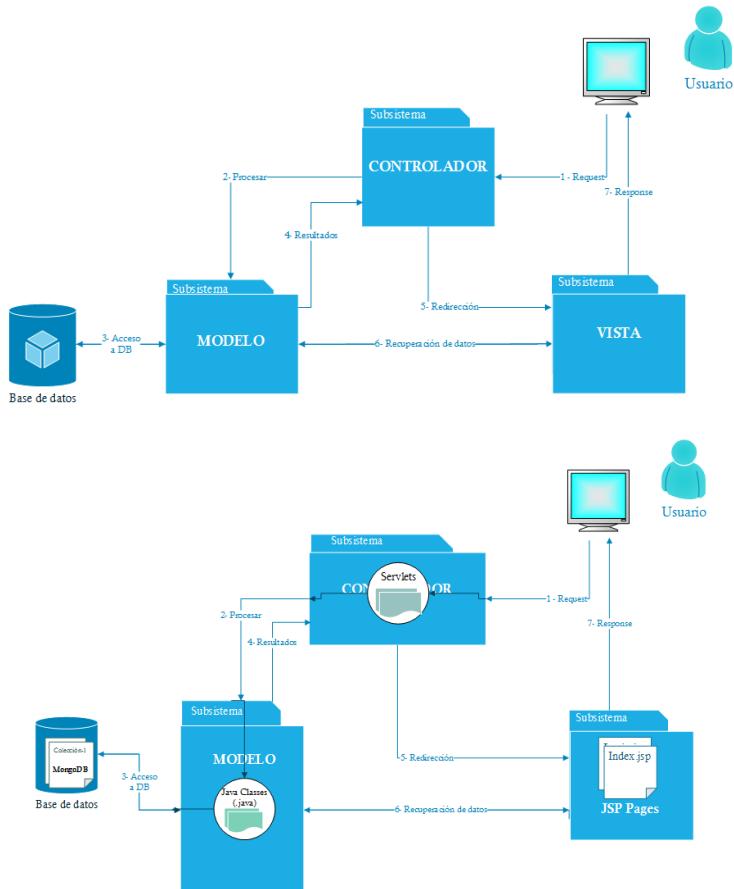


Ilustración 12 Arquitectura aplicación web

A continuación, se definen las clases y la funcionalidad que aportan en cada uno de los subsistemas de la imagen anterior.

3.2.1.2.1. Vista

En este apartado se definen las clases que componen el subsistema vista de la aplicación web desarrollada para las visualizaciones de los datos sobre la contaminación.

- **Index.jsp:** Página de inicio de la aplicación web. Proporciona links directos a todas las visualizaciones de la contaminación desarrolladas.
- **Provincia.jsp:** Página de selección de provincia para cualquiera de las tres visualizaciones que proporciona la aplicación web.
- **Opcion1.jsp:** Página que muestra la información de la contaminación desde un punto de vista social. Muestra los tweets, su sentimiento, hashtags más populares y la evolución de estos para la provincia elegida por el usuario.
- **Opcion2.jsp:** Página que muestra la información de la comparación sobre la contaminación entre dos provincias. Muestra gráficos y mapas en base a los datos recopilados por las estaciones de calidad del aire.
- **Opcion3.jsp:** Página que muestra de forma general la información sobre la contaminación en una provincia. Muestra gráficos y mapas en base a los datos recopilados por las estaciones de calidad del aire.

- **Datasheets.jsp:** Página que muestra en mapas la geolocalización de todas las estaciones de calidad del aire para una provincia y proporciona links de descarga a ficheros CSV con los valores de las medidas reales de cada estación, en un formato estándar para todas las provincias de España.
- **Contaminantes.jsp:** Página que muestra la información y los riesgos para la salud de los contaminantes analizados en la web. Además, muestra gráficos de las cinco ciudades con valores más elevados para el contaminante elegido por el usuario.
- **Error.jsp:** Página de error por defecto que muestra la aplicación web ante algún suceso que no puede controlar.

3.2.1.2.2. Controlador

En este apartado se definen los servlets que componen el subsistema controlador y que son los encargados de tratar las peticiones de los usuarios en la aplicación web.

El nombrado de los servlets seguirá el siguiente formato '*Nombre_Servlet.java*', donde nombre deberá ser una palabra identificativa y relacionada con su funcionalidad. A continuación, se listan los servlets usados en la aplicación web:

- **Provincia_Servlet.java:** Servlet encargado de tramitar la petición de un usuario para seleccionar una provincia y poder usar una de las tres visualizaciones.



Ilustración 13 Provincia_Servlet.java

- **Opcion_Servlet.java:** Servlet encargado de tramitar la petición de un usuario para mostrarle los datos de la visualización elegida para la/s provincia/s seleccionada/s.

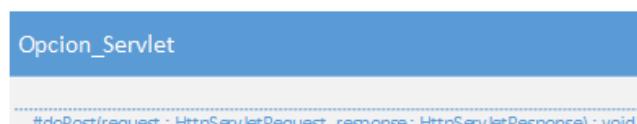


Ilustración 14 Opcion_Servlet.java

- **Page_Servlet.java:** Servlet encargado de tramitar y realizar la tarea del cambio de página de tweets mostrados en la visualización_1 (opinión social). Máximo cinco tweets por página.



Ilustración 15 Page_Servlet.java

- **Datasheet_Servlet.java:** Servlet encargado de tramitar la petición de un usuario para ver los datasheet en CSV en un formato estándar de todas las estaciones de calidad del aire de cuyos datos se han usado en la aplicación web.



Ilustración 16 Datasheet_Servlet.java

- **Contaminante_Servlet.java:** Servlet encargado de tramitar la petición de un usuario para ver la información sobre los contaminantes analizados en la web y ver el top cinco de ciudades con el nivel más alto para todos los contaminantes.



Ilustración 17 Contaminante_Servlet.java

3.2.1.2.3. Modelo

En este apartado se definen las clases que componen el sistema modelo y que son las encargadas de realizar la funcionalidad de la petición que les llega desde el servlet invocada por el usuario.

- **Provincia.java:** Clase encargada de devolver la lista de provincias disponibles para seleccionar en base a la visualización elegida por el usuario. Recoge las provincias de la colección de MongoDB *Estación* si se trata de una visualización con relación a las estaciones de aire y devuelve las provincias de la colección *Provincia* si se trata de la visualización de Twitter. Esto es debido a que puede que para una provincia no haya datos sobre las estaciones de calidad del aire pero sí existan tweets, o viceversa.

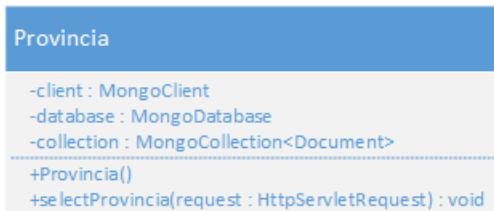


Ilustración 18 Provincia.java

- **Opcion.java:** Clase encargada de devolver los resultados de la visualización en base a la opción y provincia elegida por el usuario. Recoge los datos almacenados en la colección *Estación* si se trata de una de las visualizaciones de las estaciones de calidad aire, sino los recogerá de las colecciones *Tweet* y *Provincia*.

```

Opcion
-client : MongoClient
-database : MongoDB
-collectionTweet : MongoCollection<Document>
-collectionTweetProv : MongoCollection<Document>
-collectionAire: MongoCollection<Document>
-----
+Opcion()
+opcion3(request : HttpServletRequest) : void
+opcion2(request : HttpServletRequest) : void
+opcion1(request : HttpServletRequest) : void
-limpiarMedias(old : List<Document>, aux : String) : List<Document>
-mapContaminantes() : Map<Integer, String>
-quitarTildes(input : String) : String
-insertarPalabraDate(palabra : String, hashDate : HashMap<Date, List<HashtagFreq>>, fecha : String) : void
-insertarPalabra(palabra : String, hashTag : HashMap<String, Integer>) : void
-sortByComparator(unsortMap : Map<String, Integer>, order : boolean) : Map<String, Integer>

```

Ilustración 19 Opcion.java

- **HashtagFreq.java:** Clase auxiliar usada para cuando el usuario elige la visualización_1. Esta clase sirve para construir la lista de frecuencia con los hashtags que más se repiten y que luego se mostrará en un gráfico.

```

HashtagFreq
-hashtag : String
-size : Integer
-----
+get()/set() de los atributos

```

Ilustración 20 HashtagFreq.java

- **Datasheet.java:** Clase encarga de devolver la lista de ficheros CSV y los mapas de geolocalización de las estaciones de calidad del aire. También es la encargada de realizar la descarga del fichero CSV de la estación seleccionada por el usuario. Recoge datos de las colecciones *Estación* y *Provincia*, y recorre el directorio donde están los datasheet para sólo devolver los ficheros CSV disponibles.

```

Datasheet
-client : MongoClient
-database : MongoDB
-collection : MongoCollection<Document>
-collectionCarto : MongoCollection<Document>
-----
+Datasheet()
+mostrarFiles(request : HttpServletRequest) : void
+descargarFiles(response : HttpServletResponse, csvName : String) : void

```

Ilustración 21 Datasheet.java

- **Contaminante.java:** Clase encargada de devolver la lista con la información y riesgo de los contaminantes para la salud y el top cinco de ciudades con niveles más altos para los contaminantes analizados en la página. Recoge los datos de la colección *Contaminante*.

```

Contaminante

-client : MongoClient
-database : MongoDBDatabase
-collection : MongoCollection<Document>
-collectionAire: MongoCollection<Document>
+Contaminante()
+infoContaminantes(request : HttpServletRequest) : void

```

Ilustración 22 Contaminante.java

3.2.2. Diseño de casos de uso reales

El propósito de esta actividad es especificar el comportamiento de la aplicación web para cada caso de uso, mediante objetos o subsistemas de diseño que interactúan, y determinar las operaciones de las clases e interfaces de los distintos subsistemas de diseño.

A continuación, se adjunta una tabla en la que se muestra cada caso de uso identificado con sus respectivas clases participantes y al componente que corresponden.

Caso de uso	Clase asociada	
	Nombre	Componente
C00: Acceso a la aplicación web	index.jsp	Vista
C01: Acceso a visualización social	provincia.jsp opcion1.jsp Provincia_Servlet.java Opcion_Servlet.java Page_Servlet.java Provincia.java Opcion.java HashtagFreq.java	Vista Vista Controlador Controlador Controlador Modelo Modelo Modelo
C02: Acceso a visualización de comparador de provincias	provincia.jsp opcion2.jsp Provincia_Servlet.java Opcion_Servlet.java Provincia.java Opcion.java	Vista Vista Controlador Controlador Modelo Modelo
C03: Acceso a visualización de gráficos de provincia	provincia.jsp opcion3.jsp Provincia_Servlet.java Opcion_Servlet.java Provincia.java Opcion.java	Vista Vista Controlador Controlador Modelo Modelo
C04: Ver datasheets	datasheets.jsp Datasheet_Servlet.java Datasheet.java	Vista Controlador Modelo

C05: Ver contaminantes	contaminantes.jsp Contaminante_Servlet.java Contaminante.java	Vista Controlador Modelo
-------------------------------	---	--------------------------------

Tabla 1 Casos de uso

Ahora se describirán cómo interactúan entre sí los objetos identificados en la tabla anterior para realizar cada uno de los casos de uso de la aplicación web, desde un punto de vista técnico. Para ello se utilizarán de diagramas de secuencia.

3.2.2.1. C00: Acceso a la aplicación web

A continuación, se exponen los diagramas de secuencia del caso de uso C00 adaptado al nombrado y diseño de las clases del sistema.

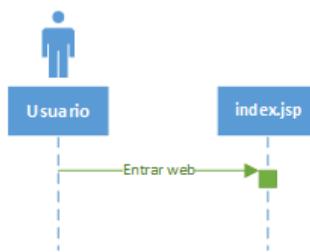


Ilustración 23 C00

3.2.2.2. C01: Acceso a visualización social

A continuación, se exponen los diagramas de secuencia del caso de uso C01 adaptado al nombrado y diseño de las clases del sistema.

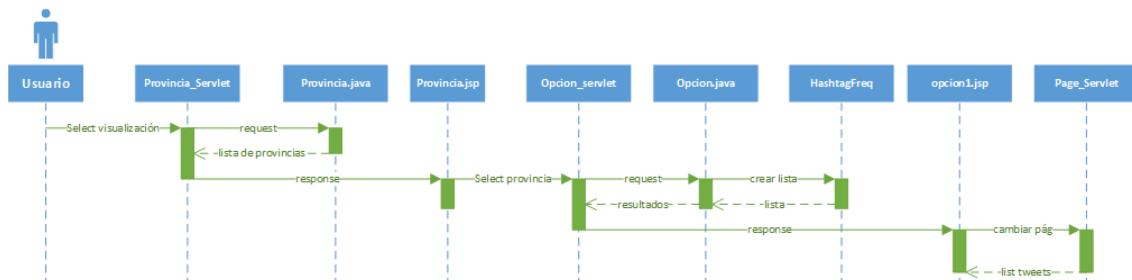


Ilustración 24 C01

3.2.2.3. C02: Acceso a visualización de comparador de provincias

A continuación, se exponen los diagramas de secuencia del caso de uso C02 adaptado al nombrado y diseño de las clases del sistema.

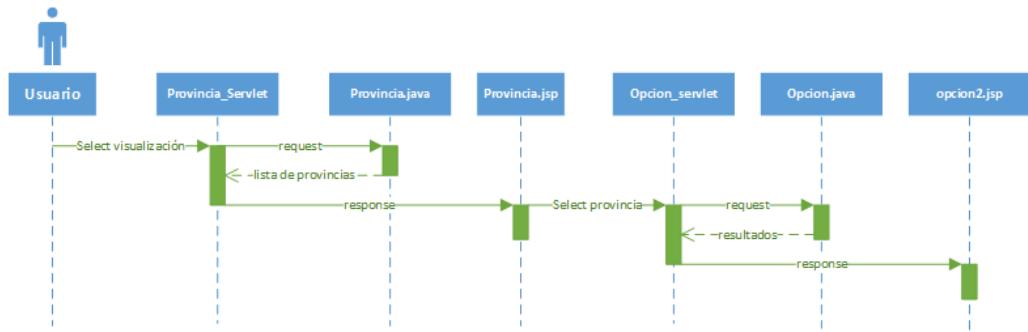


Ilustración 25 C02

3.2.2.4. C03: Acceso a visualización de gráficos de provincia

A continuación, se exponen los diagramas de secuencia del caso de uso C03 adaptado al nombrado y diseño de las clases del sistema.

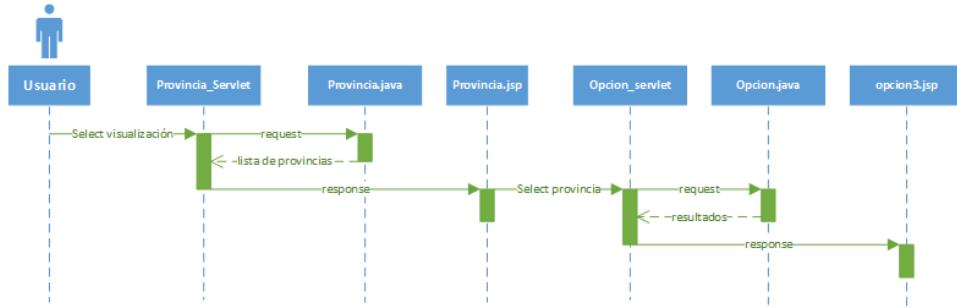


Ilustración 26 C03

3.2.2.5. C04: Ver datasheets

A continuación, se exponen los diagramas de secuencia del caso de uso C04 adaptado al nombrado y diseño de las clases del sistema.

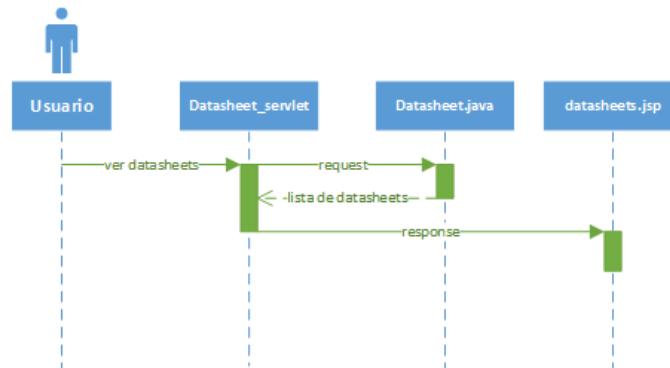


Ilustración 27 C04

3.2.2.6. C05: Ver contaminantes

A continuación, se exponen los diagramas de secuencia del caso de uso C05 adaptado al nombrado y diseño de las clases del sistema.

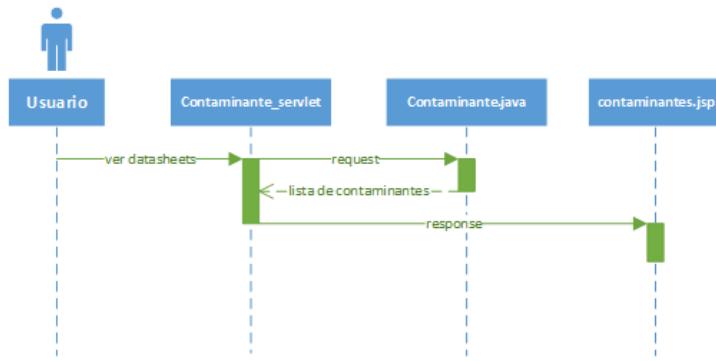


Ilustración 28 C05

3.2.3. Interfaces de usuario

El objetivo de esta apartado es realizar el diseño detallado del comportamiento de las interfaces de usuario de la aplicación web.

También será revisada la navegación entre las ventanas, los elementos que la forman, características, disposición y eventos relacionados con las partes que las componen.

Las imágenes que se presentan en este apartado muestran el aspecto real que tendrán las interfaces de usuario en la aplicación web.

3.2.3.1. Página de inicio

Es la página de inicio que se mostrará a todos los usuarios al entrar en la web. Está dividida en tres partes principales. El diseño de esta página viene recogido en la clase *index.jsp*.

La primera parte, la superior, está compuesta por un menú horizontal desde el que se pueden acceder a todos los servicios de la web. También, debajo de este menú, se cuenta con un texto explicativo sobre la aplicación y un botón que lleva a la parte dos de la página de inicio.



Ilustración 29 Página de inicio. Parte 1

La segunda parte de la página de inicio está compuesta de tres botones con imágenes de relleno que redirigen cada una a la visualización que indica su frase en la parte inferior.



Ilustración 30 Página de inicio. Parte 2

Por último, la tercera parte de la que se compone esta página de inicio se compone de dos botones que redirigen respectivamente a los datasheet con las medidas de las estaciones de aire y a la información sobre los contaminantes analizados en la aplicación web.

DESCARGAR LOS DATASHEET
LAS MEDICIONES DE LAS ESTACIONES DE CALIDAD DE AIRE UTILIZADAS

ATMOSPHERE SPAIN le da la oportunidad de descargar los datasheets limpios y preprocesados con los datos reales de los contaminantes medidos por las estaciones de calidad del aire

DATASHEETS

NUMEROSES CONTAMINANTES
INFÓRMENSE SOBRE LOS CONTAMINANTES MOSTRADOS

ATMOSPHERE SPAIN le ayuda a informarse sobre los riesgos para la salud de los contaminantes atmosféricos mostrados en las visualizaciones. Además, podrá ver el top 5 de provincias a la cabeza en cada contaminante.

CONTAMINANTES

Ilustración 31 Página de inicio. Parte 3

3.2.3.2. Selección provincia

Se accederá a ella cuando se elija una de los tres tipos de visualizaciones. Esta página mostrará un formulario donde habrá que elegir la provincia para la que se quiere ver la información. En el caso de que el usuario haya elegido la visualización_2 (comparador de provincias) deberá elegir dos. El código que recoge esta interfaz está en la clase *provincia.jsp*.



Ilustración 32 Selección de provincia 1



Ilustración 33 Selección de provincia 2

Tras elegir la provincia y pulsar en el botón '*Enviar consulta*' se redirigirá a la visualización elegida con los datos de la provincia seleccionada por el usuario.

3.2.3.3. Visualización_1 (*Opinión social*)

Visualización que recoge todos los análisis realizados en base a los tweets de la provincia seleccionada. El contenido de esta interfaz está recogido en la clase *opcion1.jsp*.

Esta página está dividida en dos partes verticales. La de la izquierda muestra tres gráficos: frecuencia de hashtags, sentimiento de tweets y evolución diaria de los hashtags principales, en este último se permitirá ampliar haciendo clic en el link de abajo, esto desplegará unos gráficos que permitirán comparar la evolución de los contaminantes con la de los hashtags principales y así ver la relación por fechas. La parte de la derecha es una columna donde se mostrarán todos los tweets recogidos para esa provincia, por cuestiones de tamaño y estética se ha añadido una paginación para esta columna, y sólo se mostrarán cinco tweets por página. Además, debajo de cada tweet, mediante una línea de color (rojo, verde o gris) se podrá ver el sentimiento que tiene asignado ese tweet.

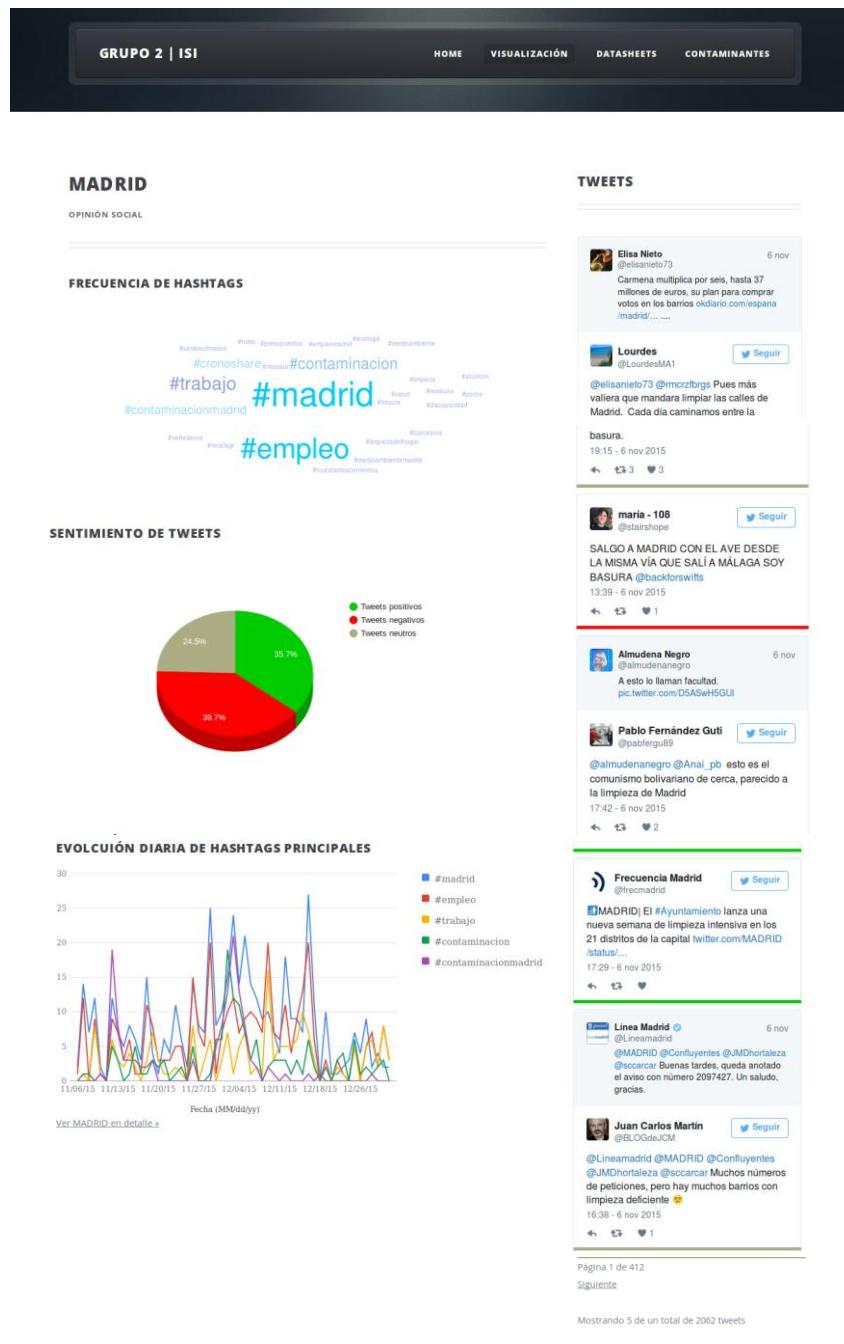


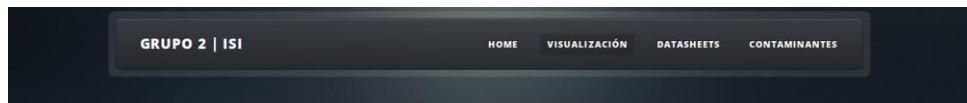
Ilustración 34 Visualización_1

3.2.3.4. Visualización_2 (Comparar provincias)

Visualización que recoge los análisis de comparar dos provincias en base a las medidas de sus estaciones de calidad del aire. El contenido de esta interfaz está recogido en la clase *opcion2.jsp*.

Dividida en dos partes. En la parte superior se mostrará un menú para elegir el contaminante sobre el que se quieren ver los resultados y dos mapas de la evolución de este en cada una de las dos provincias seleccionadas anteriormente.

En la segunda parte se mostrarán una serie de gráficos en los que se comparará la evolución del contaminante seleccionado en la parte superior para estas dos provincias.



ALAVA VS BARCELONA

EVOLUCIÓN DEL NO₂ EN LAS PROVINCIAS DE ALAVA Y BARCELONA

NO₂ PM10 SO₂ PM2.5 CO O₃ BEN NO_x

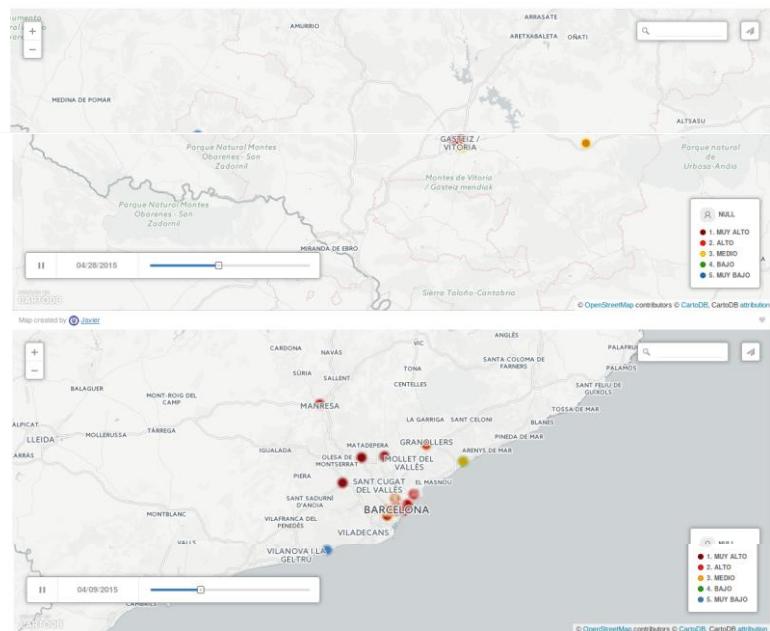


GRÁFICO DE COMPARATIVA DE CONTAMINANTE

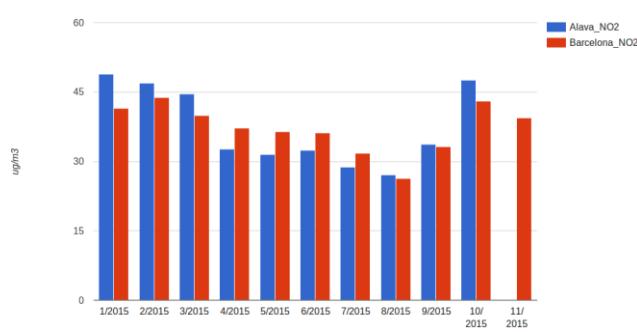
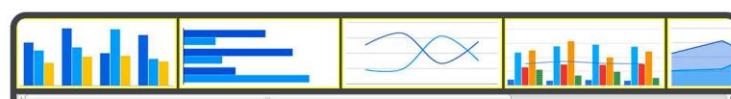


Ilustración 35 Visualización_2

3.2.3.5. Visualización_3 (Gráficos provincia)

Parecida a la visualización 2 pero únicamente para una provincia. En esta visualización se verá de manera general un análisis de todos los contaminantes medidos por las estaciones de calidad del aire de la provincia seleccionada por el usuario. El contenido de esta interfaz está recogido en la clase *opcion3.jsp*.

La estructura de colocación de mapas y gráficos es idéntica a la de la visualización vista anteriormente.

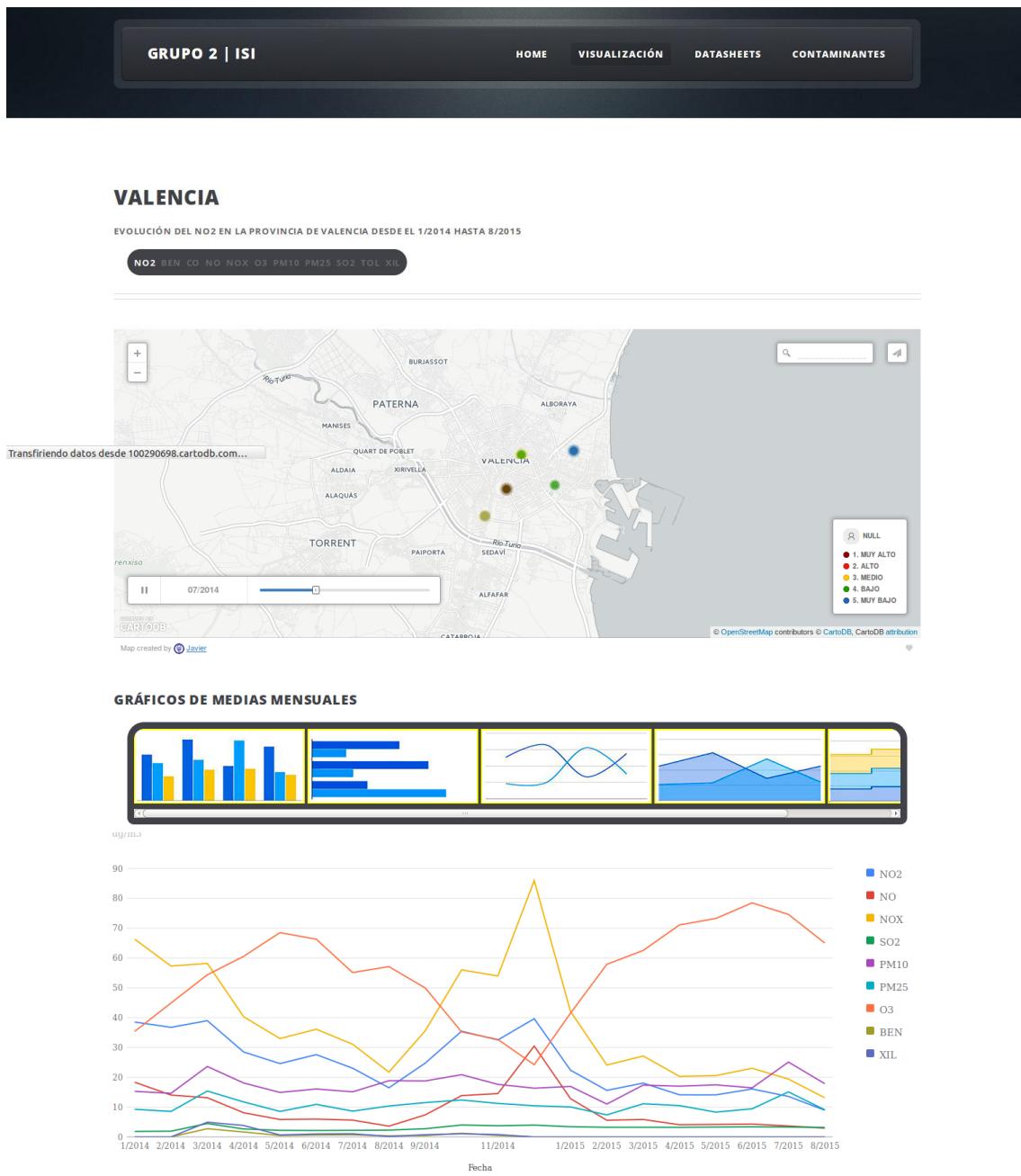


Ilustración 36 Visualización_3

3.2.3.6. Datasheets

Página que dará la posibilidad al usuario de descargar en un formato estándar y estructurado los valores medidos por cada una de las estaciones de calidad del aire usadas en la web. El contenido de esta interfaz está recogido en la clase *datasheets.jsp*.

Está compuesta por un menú lateral izquierdo en el que se podrá seleccionar la provincia deseada y una parte derecha donde se mostrará en un mapa la geolocalización de cada una de las estaciones y la lista de los ficheros CSV disponibles para descargar.

The screenshot shows a web interface for 'ALAVA'. At the top, there's a navigation bar with 'GRUPO 2 | ISI' on the left and 'HOME', 'VISUALIZACIÓN', 'DATASHEETS' (which is highlighted in blue), and 'CONTAMINANTES' on the right. Below the navigation bar, a sidebar on the left lists various Spanish provinces: ALAVA, ALBACETE, ALICANTE, ASTURIAS, AVILA, BARCELONA, BURGOS, CANTABRIA, CASTELLON, CIUDAD REAL, CUENCA, GERONA, GUADALAJARA, GUIPUZCOA, ISLAS BALEARES, LA CORUNA, LA RIOJA, LAS PALMAS, LEON, LUGO, MADRID, MURCIA, NAVARRA, OURENSE, PALENCIA, PONTEVEDRA, PROVINCIA DE LERIDA, SALAMANCA, SANTA CRUZ DE TENERIFE, SEGOVIA, SORIA, TARRAGONA, TERUEL, TOLEDO, VALENCIA, VALLADOLID, VIZCAYA, and ZARAGOZA. The main content area is titled 'ALAVA' and contains the text 'DESCARGUE LAS MEDIDAS REALIZADAS POR LAS SIGUIENTES ESTACIONES DE CALIDAD DEL AIRE'. It features a map of Gasteiz/Vitoria with several orange dots indicating station locations. Below the map, there are links to download CSV files for each station: VALDEREJO.csv, LOS HERRAN.csv, LLUDIO.csv, ELCIEGO.csv, AV.GASTEIZ.csv, AGURAIN.csv, 3 DE MARZO.csv, and FARMACIA.csv. A small note at the bottom of the map area says 'Map created by javier © OpenStreetMap contributors © CartoDB, CartoDB attribution'.

Ilustración 37 Datasheets

3.2.3.7. Contaminantes

Esta página le permitirá al usuario informarse sobre los contaminantes mostrados en la web y conocer sus efectos para la salud. El contenido de esta interfaz está recogido en la clase *contaminantes.jsp*.

La página está dividida en dos partes, al igual que la del apartado anterior, contará con un menú lateral izquierdo para elegir el contaminante a mostrar. En la parte derecha se

mostrará un gráfico con las cinco provincias con niveles más altos de esos contaminantes y los efectos para la salud de estar expuesto a ciertas cantidades de él.

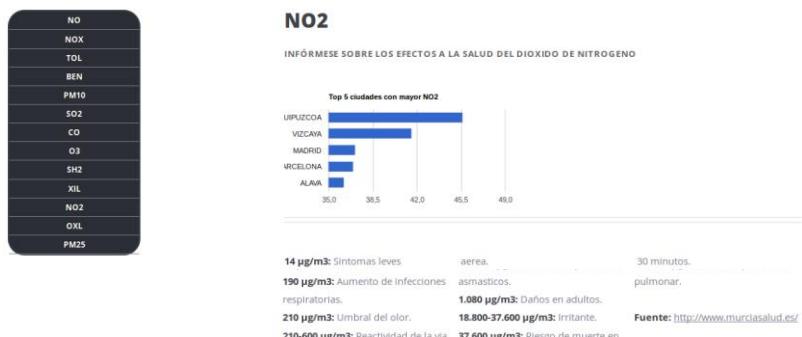


Ilustración 38 Contaminantes

3.2.3.8. Página de error

Página simple de error que aparecerá cuando alguna petición del usuario no pueda realizarse correctamente u ocurra un error interno en la aplicación web. El contenido de esta interfaz está recogido en la clase *error.jsp*.

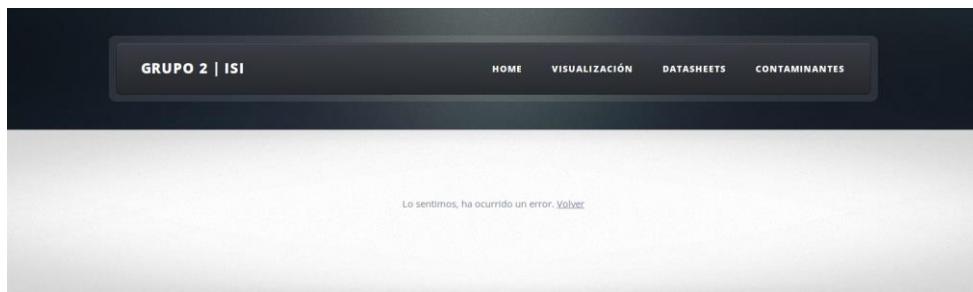


Ilustración 39 Error

3.2.3.9. Navegabilidad

El objetivo de esta apartado es definir los flujos entre los distintos formatos de interfaz de pantalla descritos anteriormente. Para ello se adjunta el mapa de navegación considerado para esta aplicación web.

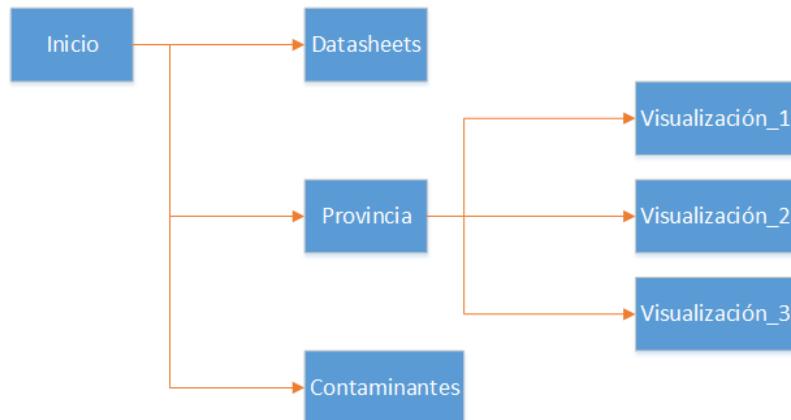


Ilustración 40 Navegabilidad

3.2.4. Diseño del modelo de datos

En esta actividad se define la estructura física de datos que utilizará el sistema a partir del modelo descrito en el apartado 3.1.3. *Modelo de datos*.

El sistema gestor de base de datos que usará la solución final es MongoDB. La aplicación web y las herramientas de recopilación de datos interactuarán con ella a través del driver de MongoDB para Java.

Se ha optado por un sistema gestor de base de datos no relacional debido al deseo de tener una escalabilidad alta sin que afecte en exceso al rendimiento y al no tener que depender de restricciones de claves ajenas entre tablas (colecciones en nuestro caso).

Las entidades especificadas en el apartado del análisis son las colecciones de la base de datos MongoDB creada. En la siguiente tabla se muestra el propósito básico de cada una de las colecciones creadas.

Colección	Descripción
Estación	Almacena toda la información relacionada con una estación de calidad del aire
Contaminante	Almacena la información sobre los límites y riesgos de los contaminantes
Tweet	Almacena todos los tweets recuperados con la API de Twitter
Provincia	Almacena información relacionada con la provincia como tweets y mapas

Ilustración 41 Colecciones MongoDB

A continuación, se expone la estructura estándar de los documentos que se almacenarán en las colecciones listadas anteriormente:

- **Estación**

```

▼ object {7}
  ► _id {1}
    Estacion : 'Nombre'
  ▼ Localizacion {2}
    type : Point
  ▼ coordinates [2]
    0 : 'Longitud'
    1 : 'Latitud'
  Ciudad : 'Ciudad'
  Provincia : 'Provincia'
  Pais : 'País'
  ▼ Medidas [11]
    ▼ 0 {n}
      ▼ Fecha {1}
        $date : 'yyyy-MM-ddTHH:mm:ss.000+0000'
        SO2 : 'Valor'
        NO2 : 'Valor'
        ... : 'Valor'
        BEN : 'Valor'
    ► ... {n}
    ► m {n}

```

Ilustración 42 Estación JSON

La siguiente tabla muestra una breve descripción de cada uno de los atributos listados en el JSON anterior:

Nombre	Descripción
_id	ID asignada por defecto por el SGBD
Estación	Nombre de la estación de calidad del aire
Localización	Localización tipo 'Point' con las coordenadas de la estación
Ciudad	Ciudad donde está la estación
Provincia	Provincia donde está la estación
País	País donde está la estación
Medidas	Array con la fecha y valores de los contaminantes medidos

Tabla 2 Estación atributos

● Contaminante

```

▼ object {9}
  _id : 'Contaminante'
  Nombre : 'Nombre'
  MA : 'Muy alto'
  A : 'Alto'
  M : 'Medio'
  B : 'Bajo'
  MB : 'Muy bajo'
  Unidades : 'Unidades de medida'
  info : 'Sintomas y efectos según la exposición y cantidad del contaminante'

```

Ilustración 43 Contaminante JSON

La siguiente tabla muestra una breve descripción de cada uno de los atributos listados en el JSON anterior:

Nombre	Descripción
_id	Nomenclatura química del contaminante
Nombre	Nombre completo del contaminante
MA	Valor por el cual se categoriza como muy alto la medida del contaminante
A	Valor por el cual se categoriza como alto la medida del contaminante
M	Valor por el cual se categoriza como medio la medida del contaminante
B	Valor por el cual se categoriza como bajo la medida del contaminante
MB	Valor por el cual se categoriza como muy bajo la medida del contaminante
Unidades	Unidades en las que se realizan las medidas (ug/m3 o mg/m3)
info	Información sobre los síntomas y efectos a la salud por la exposición a ciertos niveles del contaminante

Tabla 3 Contaminante atributos

- **Tweet**

```

▼ object {9}
  _id : 'id_tweet'
  usuario : 'usuario'
  contenido : 'mensaje'

▼ hashtag [n]
  0 : '#hashtag1'
  ... : '#hashtag...'
  n : '#hashtagN'

▼ fecha {1}
  $date : yyyy-MM-ddTHH:mm:ss.000+0000
  lang : es

▼ place {2}
  ciudad : 'ciudad'
  pais : 'pais'
  localizacion : 'localizacion usuario'

▼ geo {2}
  type : Point
  ▼ coordinates [2]
    0 : 'longitud'
    1 : 'latitud'
  
```

Ilustración 44 Tweet JSON

La siguiente tabla muestra una breve descripción de cada uno de los atributos listados en el JSON anterior:

Nombre	Descripción
_id	ID del tweet dada por Twitter
usuario	Nombre del usuario
contenido	Mensaje del tweet
hashtag	Array con todos los hashtags del mensaje

fecha	Fecha de publicación del tweet
lang	Idioma del tweet
place*	Localización desde donde se mandó el tweet
localización*	Localización que tiene el usuario en su perfil
geo*	Geolocalización desde donde se mandó el tweet

Tabla 4 Tweet atributos

(*)- Los atributos marcados con * pueden ser no existir en el documento, esto dependerá de cómo el usuario del tweet tenga configurado su perfil para que se pueda acceder o no a esos atributos.

- **Provincia**

```

▼ object {4}
  _id : 'provincia'
  Geo : 'iframe de geolocalización de estaciones'
  ▼ Mapas {n}
    NO2 : 'iframe de mapa con evolución del contaminante'
    ... : 'iframe de mapa con evolución del contaminante'
    CO : 'iframe de mapa con evolución del contaminante'
  ▼ tweets [m]
    ▼ 0 {3}
      id_tweet : 'id_tweet'
      user : 'usuario'
      feeling : 'sentimiento'
    ▶ ... {3}
    ▶ m {3}

```

Ilustración 45 Provincia JSON

La siguiente tabla muestra una breve descripción de cada uno de los atributos listados en el JSON anterior:

Nombre	Descripción
_id	Nombre de la provincia
Geo	Iframe de CartoDB con la geolocalización de las estaciones de esa provincia
Mapas	Iframes de CartoDB con la evolución de los contaminantes de esa provincia
tweets	Array con todos los tweets de esa provincia

Tabla 5 Tweet atributos

3.2.5. Especificación del control de versiones

Se ha usado un repositorio Git para realizar un control de versiones sobre toda la documentación generada durante el caso práctico (memoria, código, base de datos, ...).

Con el uso del repositorio se ha conseguido una disponibilidad total de los elementos del caso práctico. Además, con el control de versiones se ha conseguido proteger los ficheros

ante cualquier cambio que pudiera afectar a la integridad de la solución. La siguiente imagen muestra un ejemplo del log de cambios en el repositorio.

The screenshot shows a Git commit history and a detailed view of a specific commit. The commit history lists 174 commits across various branches, with most being 'WEB APP' changes. A detailed view of commit 4e15865c92d512925cef23efc7639ea2060dbd4e is shown, which was authored by javie and committed by javie on 24/12/2015 at 13:03:15. The commit message indicates the completion of contaminant visualization and mentions changes to the meaningcloud API key. The detailed view also shows work items and modified files.

Git COMMITS

Branch: master (showing 20 of 174 commits)

20 | 50 | 100 | All

- WEB APP (SHA a16982e00889453fd66b00e95e1ffcc8a92e8ddf) by isi on 1/1/2016, 22:26:59
- WEB APP (SHA 6e7c7731c445784c2a12db6c75d0ad614d1f22fe) by isi on 12/31/2015, 16:58:16
- ISI (SHA 6704c3089fd2381dddef7067f01b5bd947713b71) by isi on 12/31/2015, 11:20:39
- WEB APP (SHA 3e761c12fb3c55d60b7d0935ea2d6638590542be) by javie on 12/30/2015, 12:14:32
- AIRE (SHA 8a24803faccfedc96496572d82a59b456412646a1) by javie on 12/29/2015, 14:07:14
- ADAPTACION A CAMBIO EN ESTRUCTURA DE BBDD (SHA 9d4855faf6eeafa05205e0330fec1d79bb11ced) by isi on 12/28/2015, 18:59:42
- WEB APP (SHA af7a3dbdff60eb35e7fd3063a3508066bc870bf4) by isi on 12/24/2015, 17:52:19
- WEB APP (SHA a36a3086a4d4b88bf5d092855890f0c4a327eb97) by isi on 12/24/2015, 17:47:31
- WEB APP (SHA 64348763b9ff96982d62ac54c866fe7e4d2d4840) by javie on 12/24/2015, 17:37:46
- WEB APP Y MEANINGCLOUD (SHA 4e15865c92d512925cef23efc7639ea2060dbd4e) by javie on 12/24/2015, 13:03:15

Git COMMIT 4e15865

WEB APP Y MEANINGCLOUD

-Terminada visualización de contaminantes y efectos
-Cambio de la clave del API de meaningcloud

commit: 4e15865c92d512925cef23efc7639ea2060dbd4e
parent: 099fe0fd0f0777cb56449671ac9ac24ddcd6f73

authored by javie (javie@JAVIER-CASA) on 24/12/2015, 13:03:15
committed by javie (javie@JAVIER-CASA)

Work Items

[Link Work Item](#)

There are no linked items.

4e15865c92d512925cef23efc7639ea2060dbd4e

3 archivos modificados

- DATOS/src/main/java/com/isi/master/meaningcloudAPI/sentimentanalysis/SentimentClient.java
- DATOS/src/main/java/com/isi/master/meaningcloudAPI/topicsextraction/TopicsClient.java
- Visualizacion/WebContent/contaminante.jsp

Ilustración 46 Control de versiones

Los cambios aplicados habrá que indicarlos con un título que sea orientativo de la zona del sistema al que afecta, seguidamente de la enumeración de los pasos donde se indique específicamente lo que se ha cambiado (clases, métodos, etc).

4. Prueba de concepto

Este apartado describirá la prueba de concepto desarrollada incluyendo instrucciones para la ejecución de la prueba de concepto en la máquina virtual presentada.

En este punto se incluirá los pasos para montar la máquina virtual y las instrucciones para interactuar con ella. De forma complementaria, una vez montada la máquina virtual, se explicarán las visualizaciones a las que se puede acceder en la aplicación y como interactuar con ellas.

4.1. Máquina virtual

Para la prueba de concepto se ha creado una máquina virtual que contiene todas las aplicaciones y todos los datos necesarios para la correcta realización de la prueba de concepto.

En primer lugar, la máquina virtual creada está exportada en un fichero de tipo OVA mediante el programa Oracle VM VirtualBox (VirtualBox en adelante). Como la máquina virtual ha sido creada con este programa es necesario tener instalado VirtualBox en el ordenador en el que se desee ejecutar la prueba de concepto.

Si no dispone del programa a continuación se incluye un enlace para realizar la descarga en caso de que fuera necesario: <https://www.virtualbox.org/wiki/Downloads>.

Una vez se dispone del programa instalado para montar la máquina virtual es suficiente con hacer doble click sobre el archivo tipo OVA que se ha entregado, o abrir VirtualBox y dar en *Archivo/Importar servicio virtualizado*. Lo primero que se realiza es la configuración de los parámetros que utiliza la máquina virtual tal como numero de procesadores, cantidad de RAM, etc. Por defecto aparecerá la siguiente configuración:

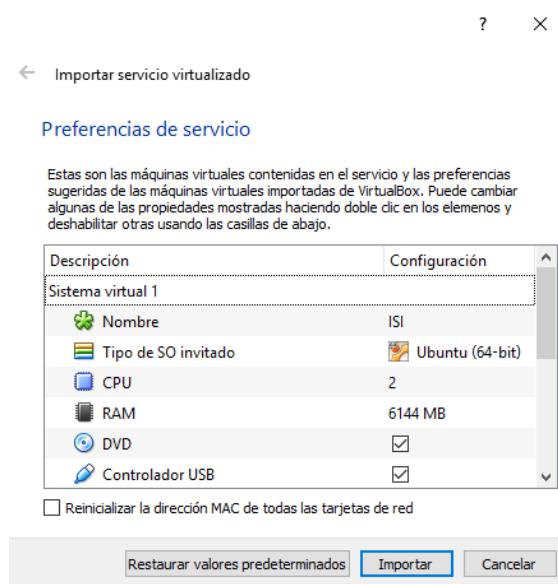


Ilustración 47: Parámetros Configuración VirtualBox

Se recomienda siempre y cuando el ordenador de destino tenga los recursos suficientes no modificar la configuración por defecto, debido a que se ha configurado de esta manera determinada para que el uso de la máquina virtual y de la aplicación sea fluido. Si dispone de

una máquina más potente puede mejorar estos parámetros de la máquina virtual. Por lo tanto, para comenzar la importación de la máquina virtual es suficiente con pulsar sobre el botón “Importar” que se podía observar en la imagen anterior.

Mientras dura el proceso de importación se mostrará una ventana como la de la siguiente imagen, en la cual se puede seguir el estado de la importación.

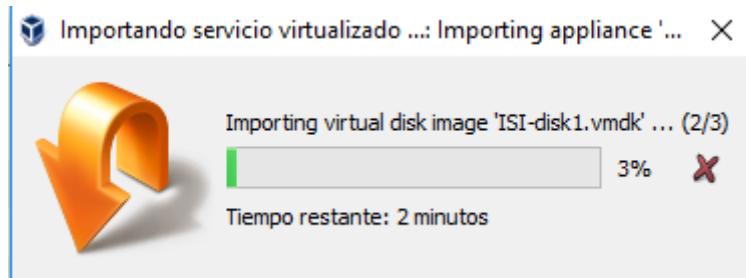


Ilustración 48: Proceso de importación

Una vez termina el proceso de importación, la máquina virtual, con la aplicación, está preparada para ser utilizada. En la ventana de VirtualBox aparecerá lo siguiente:

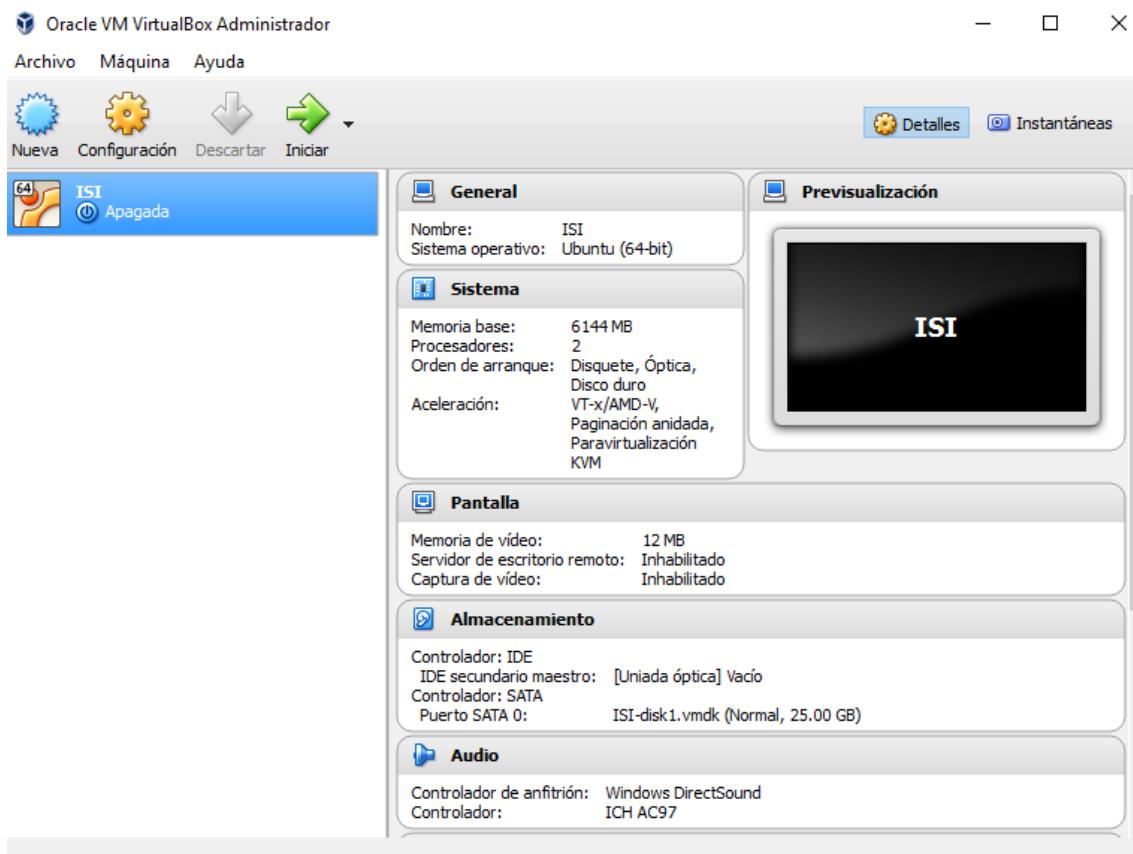


Ilustración 49: Arrancar Máquina Virtual

Se puede observar que ya se tiene la máquina virtual, pero se encuentra apagada, por lo tanto, para poder ejecutar la prueba de concepto solo faltaría inicial la máquina virtual pulsando sobre el botón “Iniciar” que se observa en la imagen anterior y la máquina virtual se pondrá en funcionamiento.

Durante la iniciación de la máquina virtual será necesario introducir la contraseña del usuario por defecto de Ubuntu con el cual se accede, esta es: “*isi2015*”.

Con todos los pasos anteriores se dispondría de la máquina virtual en correcto funcionamiento para realizar la prueba de concepto.

4.2. Abrir ATMOSPHSPAIN

ATMOSPHSPAIN es la aplicación que el grupo 2 ha creado para permitir la visualización de todo el trabajo realizado sobre la contaminación a nivel nacional.

Se trata de una aplicación web que se encuentra corriendo sobre un servidor Tomcat y que usa MongoDB como base de datos. Para que el usuario pueda acceder a la aplicación es necesario que el servidor de Tomcat esté en funcionamiento, así como la base de datos, y por ello el grupo ha añadido el servidor Tomcat y MongoDB como un servicio más de la máquina virtual para conseguir que ambos estén en funcionamiento al arrancar el SO.

Una vez explicado lo anterior, para que un usuario pueda acceder a la aplicación y realizar la prueba de concepto es suficiente con abrir al navegador web Firefox, en el cual aparece como página de inicio la aplicación ATMOSPHSPAIN.

4.3. Visualizaciones

En este apartado se incluirán las diferentes visualizaciones que permite ATMOSPHSPAIN y como el usuario puede interactuar con ellas para conseguir de una manera efectiva la información sobre la contaminación.

Cuando se accede a la aplicación se puede observar que en el menú de *Visualización* existen tres posibilidades:

- **Opinión Social:** en esta sección el usuario podrá observar la percepción que tienen los usuarios de Twitter sobre la contaminación para la provincia deseada comparada con la contaminación real.
- **Comparar Provincias:** en esta sección el usuario podrá comparar la contaminación para dos provincias diferentes a través de mapas temporales y gráficos.
- **Gráficos de una provincia:** en esta sección el usuario podrá observar la contaminación de los diferentes contaminantes para una única provincia a través de mapas y gráficos.

En la sección de *visualizaciones* de la aplicación se pueden ver los datos recogidos de Twitter en la primera de las opciones y los datos de las medidas recogidas de las estaciones de calidad del aire sobre la contaminación en las dos opciones restantes. Ambos datos están procesados y puestos a disposición de los usuarios de forma que puedan extraer información de interés.

4.3.1. Opinión social

En la sección de opinión social, a la cual se puede acceder directamente desde la página de inicio, se muestra la opinión recogida de los usuarios de Twitter sobre la contaminación existente en su provincia y una comparativa con los datos oficiales de contaminación.

Para que el usuario de la aplicación pueda obtener una gran cantidad de información a partir los tweets que se han recogido en esta sección, además de mostrar los tweets se muestran una serie de análisis realizados a estos y que se explican a continuación.

En primer lugar, los tweets relacionados con la contaminación en la provincia se muestran en la columna de la derecha de la aplicación y estos han sido obtenidos en base a los temas, hashtags y mensajes de cuentas listados en el *ANEXO III: Temas, hashtags y cuentas Twitter*. Como el número de tweets puede ser muy elevado para mostrarlos en una sola columna se ha incluido una paginación de los ellos, mediante la cual el usuario puede ir avanzando en la aplicación y leyéndolos poco a poco. Se mostrarán un máximo de 5 mensajes por página.

De forma complementaria a mostrar los tweets, se ha realizado un análisis de sentimiento sobre dichos tweets con el objetivo de conocer si su opinión sobre la contaminación es positiva, negativa o neutra. El resultado de este análisis de sentimiento será mostrado en la parte inferior de cada tweet con una barra de diferentes colores como se podrá observar en el ejemplo que se realiza en esta sección.

En la columna de la izquierda se muestran tres análisis relacionados con los tweets para que los usuarios puedan extraer la información más importante de un solo vistazo. Los análisis que se realizan a los tweets y se muestran en los gráficos son:

- Análisis de hashtag mediante una Word-Cloud para conocer cuáles son los más utilizados para esa provincia. De esta manera el usuario de la aplicación puede conocer lo más comentado en relación con la contaminación para la provincia que ha seleccionado.
- Análisis de sentimiento para conocer si los tweets contienen contenido positivo, negativo o neutro. De esta manera el usuario de la aplicación conocerá el porcentaje de los tweets en los cuales se habla negativamente de la contaminación (quejas) y en cuales de forma positiva a manera global en la provincia.
- Evolución diaria de hashtag para conocer a lo largo del tiempo los tipos de quejas que realizan los usuarios por twitter en referente a diversos temas relacionados con la contaminación.

De forma complementaria a continuación de los gráficos de análisis, se incluye un enlace que permite a los usuarios acceder a un gráfico de contaminación. De esta manera el usuario podrá comparar en una sola ventana la contaminación real recogida en los gráficos frente a la contaminación que perciben los usuarios de Twitter.

Este grafico mostrará a los usuarios la evolución temporal de un determinado contaminante para la provincia de la cual están consultando los tweets. Además, en la parte superior del grafico se incluye un menú horizontal que permite al usuario seleccionar el contaminante sobre el cual desea obtener el gráfico.

A continuación, se realiza una consulta sobre la opinión social que se recoge en ATMOSPHSPAIN de la provincia de Valencia. En esta consulta se mostrará como un usuario puede interactuar con esta sección y las visualizaciones que puede observar.

En primer lugar, desde la página de inicio se accede a *Visualizaciones->Opinión social*, o se pulsa en el botón comenzar de la página principal y se pulsa el ícono de opinión social, como se muestra en las siguientes imágenes:

The screenshot shows the main landing page of the ATMOSPHSPAIN website. At the top, there is a title: "ATMOSPHSPAIN: CONSULTA DATOS SOBRE LOS DIFERENTES CONTAMINANTES DEL AIRE EN ESPAÑA USANDO HERRAMIENTAS DE VISUALIZACIÓN". Below the title, there is a button labeled "VER LOS NIVELES POR PROVINCIAS, MOSTRAR LAS OPINIONES DE SUS HABITANTES, COMPARAR, ...". To the right of this button is a blue "EMPEZAR" button with a checkmark icon. At the bottom of the page, there is a navigation bar with tabs: "GRUPO 2 | ISI", "HOME", "VISUALIZACIÓN" (which is highlighted in white), "DATASHEETS", and "CONTAMINANTES". Under the "VISUALIZACIÓN" tab, there are three sub-options: "OPINIÓN SOCIAL", "COMPARAR PROVINCIAS", and "GRÁFICOS DE PROVINCIA".

The screenshot shows a sub-page titled "ELIJA UN MÉTODO DE VISUALIZACIÓN". The page text says: "DISPONE DE DIVERSAS MANERAS DE VER LOS DATOS DE CONTAMINACIÓN EN SU PROVINCIA: MAPAS, GRÁFICOS, OPINIÓN SOCIAL,...". Below this text, there are three main visualization options: 1) "OPINIÓN SOCIAL" (TWEETS Y GRÁFICOS) which shows a pie chart and some text about employment and pollution in Madrid; 2) "COMPARAR PROVINCIAS" (COMPARE PROVINCIAS CON MAPAS Y GRÁFICOS) which shows a map with a large "VS" symbol; 3) "GRÁFICOS DE UNA PROVINCIA" (GRÁFICOS Y MAPAS DE CONTAMINACIÓN) which shows a line graph with data for NO₂ and SO₂ levels from January to December.

Ilustración 50: Acceso Visualizaciones

Este paso abrirá redirigirá a una nueva ventana en la cual aparece un menú desplegable con todas las provincias de las cuales se puede conocer la opinión social. En este caso seleccionaremos Valencia para mostrar el ejemplo.

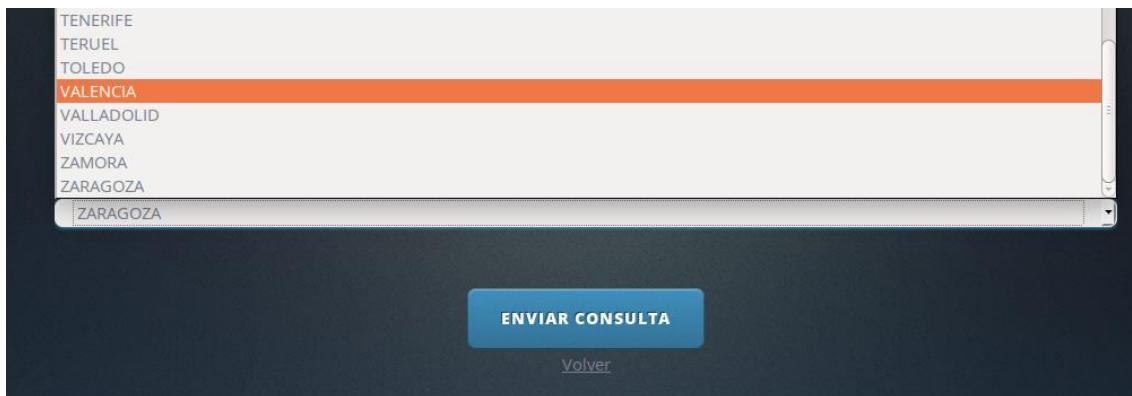


Ilustración 51: Selección Provincia Opinión Social

Una vez seleccionada la provincia deseada es suficiente con pulsar sobre el botón “Enviar Consulta” para poder acceder a toda la información sobre la provincia de seleccionada.

En la parte derecha se podrán observar los tweets recogidos sobre las quejas de los habitantes de la provincia de Valencia en relación con la contaminación que perciben. Como se podrá observar en la siguiente imagen, los tweets van acompañados de una franja roja en su parte inferior indicando que su análisis de sentimiento es negativo, por lo tanto, estos tweets son considerados como negativos. Esta franja se puede mostrar en rojo para los tweets negativos, en verde para los positivos o en gris para los neutros.

Ilustración 52: Tweets

Como se puede observar en la parte inferior de la columna que muestra los tweets existe una paginación para que los usuarios puedan ir viendo los tweets poco a poco.

Ilustración 53: Paginación Tweets

Una vez explicado cómo se muestran los tweets recogidos, en la parte izquierda se muestran los gráficos explicados anteriormente. A continuación, se incluyen los gráficos del ejemplo que se está realizando para la provincia de Valencia.

En primer lugar, un gráfico para mostrar cual son los hashtags que más se utilizan las personas para sus quejas en la provincia de Valencia:

FRECUENCIA DE HASHTAGS



Ilustración 54: Gráfico Frecuencia de HASHTAGS

Este gráfico incrementará gradualmente el tamaño de los hashtags que más se repiten para esa provincia

El siguiente gráfico que se muestra indica el porcentaje de tweets positivos, negativos y neutros después de haber realizado un análisis de sentimiento con la herramienta de Meaning Cloud explicada anteriormente en la memoria:

SENTIMIENTO DE TWEETS

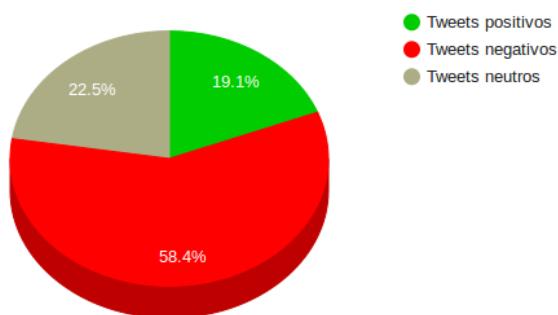


Ilustración 55: Gráfico Sentimiento de Tweets

El último gráfico sobre los tweets muestra la evolución diaria de los cinco hashtags más repetidos para la provincia de Valencia:

Evolución Diaria de Hashtags Principales

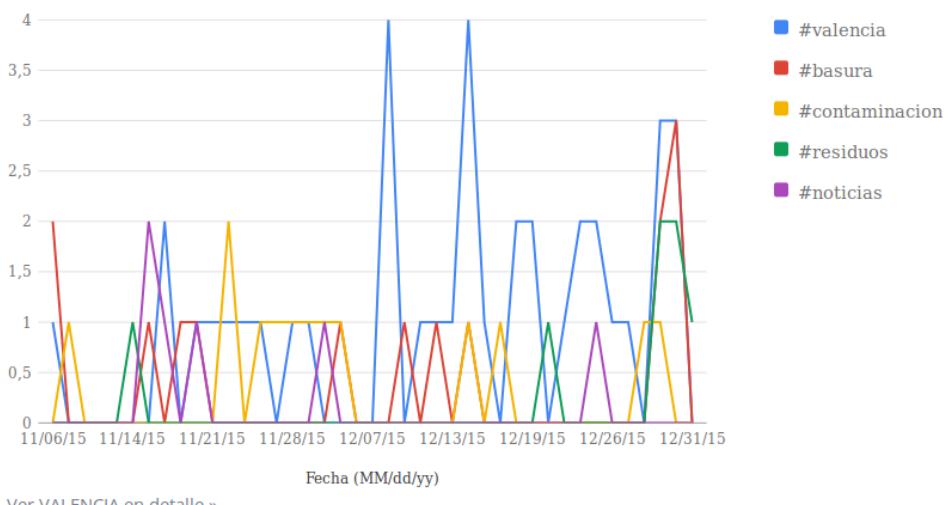


Ilustración 56: Gráfico Evolución Hashtags

Como se puede observar en la parte inferior de la imagen anterior existe un enlace “Ver VALENCIA en detalle >>”. Si el usuario que está realizando la consulta sobre opinión social desea visualizar un gráfico sobre el estado real de la contaminación para compararlo con el estado percibido en el estudio social, este enlace proporciona la opción de visualizar un gráfico que recoge dicha contaminación.

[« Ocultar](#)

NO2 BEN CO NO NOX O3 PM10 PM25 SO2 TOL XIL

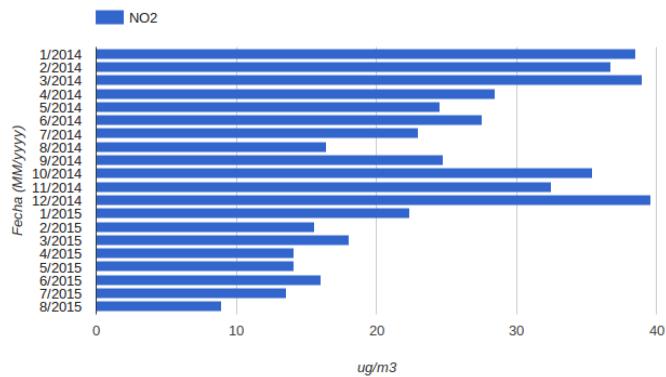


Ilustración 57: Contaminación Vs Opinión Social

En la imagen anterior se pueden observar varias cosas. En primer lugar, en la parte superior aparece un enlace “<< Ocultar” por si el usuario desea finalizar la visualización del gráfico. A continuación, se muestra un menú para que el usuario elija el contaminante sobre el cual desea obtener información.

Por último, se muestra el grafico para la provincia que se está estudiando y el contaminante seleccionado. De esta manera el usuario podrá comparar la contaminación existente para un mes en el gráfico y justo en la parte superior disponer de los gráficos de análisis sobre los tweets de los ciudadanos.

Al tratarse de una prueba de concepto realizada mientras se encuentra en desarrollo la versión final de la aplicación, se observa que el periodo de contaminación es anterior al periodo de tweets, esto será solucionado en la versión de la aplicación, debido a que en esta fase del proyecto no se dispone de la versión final de la aplicación.

4.3.2. Comparar provincias

En esta sección se permite comparar el estado de la contaminación de dos provincias diferentes. Con esta sección se consigue que los usuarios conozcan qué provincia está más contaminada y los contaminantes culpables de dicha contaminación respecto a otra. Se permite que un usuario pueda elegir la provincia menos contaminada para fijar su residencia, para viajar y, en general, para llevar a cabo la actividad deseada en base a la información mostrada en esta visualización.

Para acceder a ella, como ya ocurría en el apartado anterior, es elegir la opción de '*Comparador de provincias*' y seleccionar las provincias entre las cuales se quiere realizar la comparación.

Una vez seleccionadas dichas provincias la aplicación muestra en primer lugar por defecto una comparativa de estas provincias en base al NO₂, aunque en la parte superior muestra los contaminantes a elegir para analizar, y, por lo tanto, los contaminantes sobre los cuales se podrá realizar la comparación.

Ahora que se ha explicado cómo se realiza el proceso para seleccionar las provincias a comparar se procede a la explicación de la forma de comparar ambas provincias.

En primer lugar, se muestran dos mapas dinámicos temporales, uno para cada una de las dos provincias seleccionadas y el contaminante elegido. En estos mapas se puede observar la evolución del contaminante marcado para cada provincia a lo largo del tiempo. Con estos gráficos se consigue que el usuario pueda conocer qué provincia es la más contaminada, y, además, conocer en qué meses del año se produce esta contaminación. Con esto ATMOSPHERESPAIN consigue adaptarse a las necesidades de los usuarios que desean conocer y comparar la contaminación de las diferentes provincias no solo de forma general de todo el año sino mes a mes o día a día para conocer las épocas más contaminadas de cada una de las provincias.

En segundo lugar, se muestra un menú en el cual el usuario puede seleccionar qué tipos de gráficos comparativos se muestren para comparar ambas provincias. Los gráficos posibles son:

- Gráfico de barras verticales
- Gráfico de barras horizontales
- Gráfico lineal
- Gráfico combo (barras verticales + lineal con la media)
- Gráfico de área
- Gráfico escalonado

El contaminante mostrado por estos gráficos dependerá del seleccionado en la parte superior de los mapas. Una vez el usuario ha marcado los gráficos que quiere visualizar, estos se

muestran inmediatamente debajo, sin necesidad de recargar la página. De esta manera se facilita el uso de la aplicación y la visualización de la información que se muestra.

Los gráficos contienen el valor de contaminación para el contaminante que se haya seleccionado al comienzo de la comparativa en la parte superior de los mapas, o por defecto para el NO₂ si no se cambió este. De esta manera los usuarios pueden conocer toda la información posible sobre ese contaminante en las provincias seleccionadas.

Estos gráficos aportan un valor extra a los mapas debido a que en los mapas comparativos se muestra la información de la contaminación categorizada en los siguientes valores: Muy alto, Alto, Medio, Bajo y Muy Bajo, mientras que en los gráficos se muestran los valores exactos que han sido medidos y que pueden resultar muy útiles cuando los usuarios sean expertos en la materia.

A continuación, una vez se conoce el contenido general de la aplicación para comparar dos provincias se incluye un ejemplo realizado para comparar las provincias de Barcelona y Madrid.

Para comenzar con el ejemplo es necesario seleccionar ambas provincias en la página que la aplicación muestra al acceder a esta visualización, como se muestra a continuación. En el caso de elegir la misma provincia en los dos selectores y querer avanzar se redirigirá al inicio de la página:



Ilustración 58: Comparación Provincias

Cuando se ha realizado la selección la aplicación muestra en su parte superior un menú horizontal para que el usuario pueda elegir sobre qué tipo de contaminante desea realizar la comparativa. En el caso de Barcelona y Madrid se podría realizar una comparativa en base a los contaminantes que aparecen en la siguiente imagen:

BARCELONA VS MADRID

EVOLUCIÓN DEL NO₂ EN LAS PROVINCIAS DE BARCELONA Y MADRID

NO₂ PM10 SO₂ PM25 CO O₃ BEN TOL XIL NO

Ilustración 59: Selección Contaminante

Cuando se ha seleccionado el contaminante se muestran los mapas dinámicos temporales de ambas provincias como se muestra a continuación:



Ilustración 60: Mapa temporal provincia 1

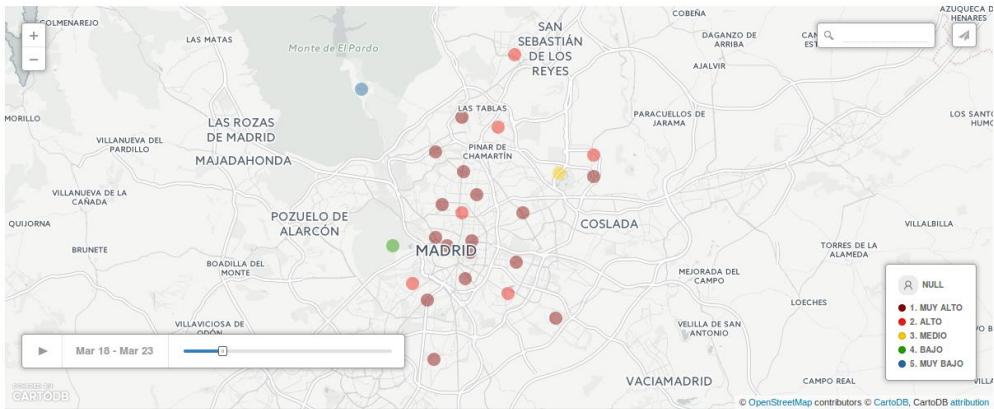


Ilustración 61: Mapa temporal provincia 2

Como se puede observar, en las imágenes anteriores se muestran los niveles del contaminante medido por las estaciones de calidad del aire que miden la contaminación en cada una de las provincias. En el mapa se marca con un color diferente en base al nivel de contaminación para el contaminante seleccionado.

A continuación de los mapas, se muestra el menú para seleccionar los gráficos comparativos que se desea visualizar. Si el ícono del gráfico tiene un borde amarillo significa que está mostrándose, en caso de querer quitarlo vuelva a marcar el ícono:



Ilustración 62: Menú selección de gráficos

Los gráficos comparativos para el contaminante NO₂ el ejemplo de Barcelona y Madrid son los siguientes:

- **Gráfico de barras verticales:**

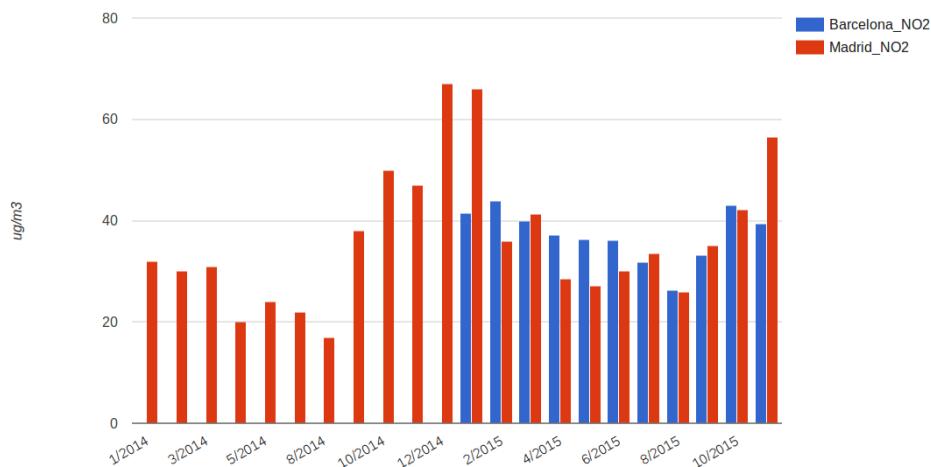


Ilustración 63: Gráfico barras verticales

- **Gráfico de barras horizontales:**

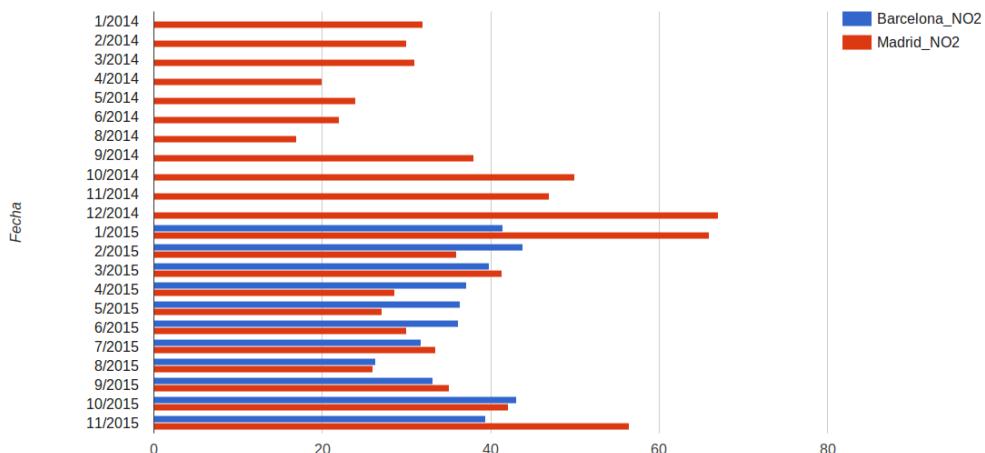


Ilustración 64: Gráfico barras horizontales

- **Gráfico lineal:**

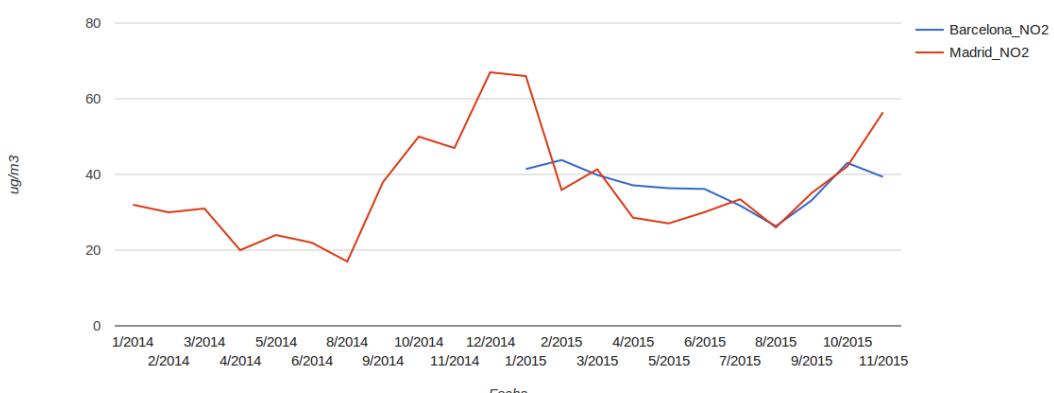


Ilustración 65: Gráfico lineal

- **Gráfico combo:**

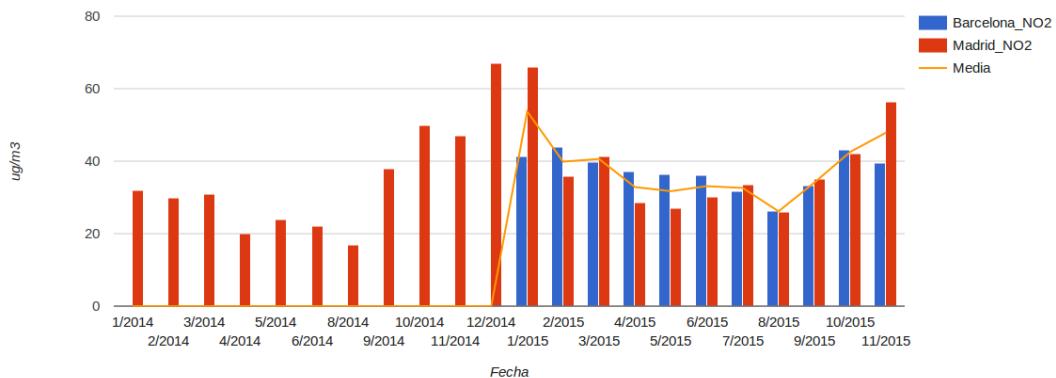


Ilustración 66: Gráfico combo

- **Gráfico de área:**

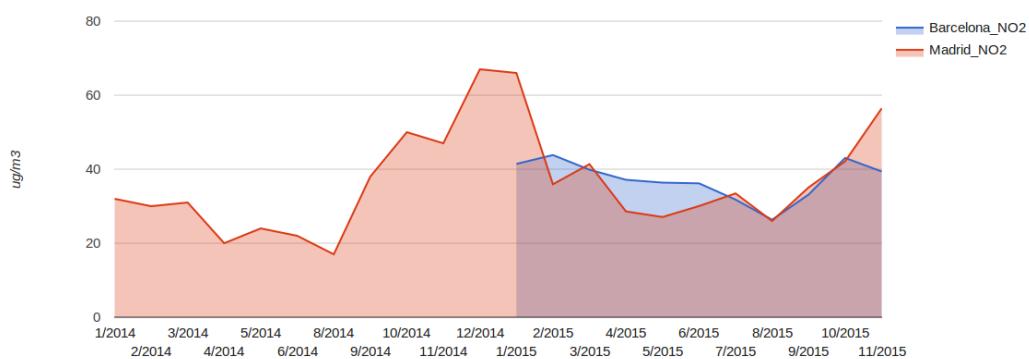


Ilustración 67: Gráfico de área

- **Gráfico escalonado:**

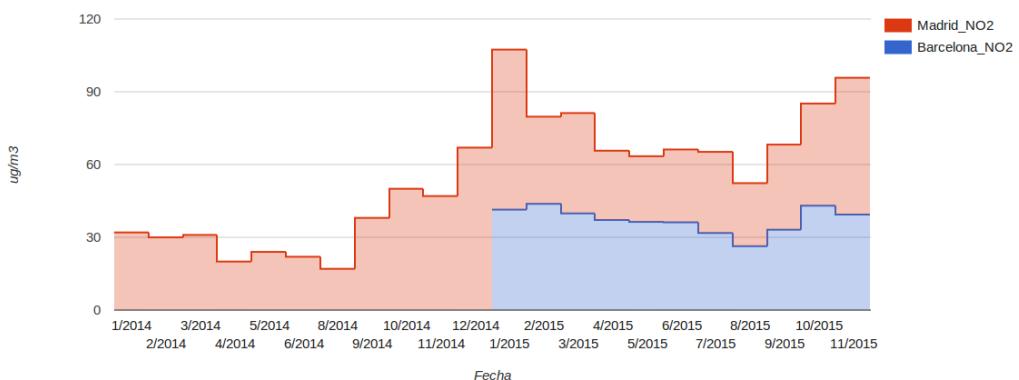


Ilustración 68: Gráfico escalonado

Se puede observar en todos los gráficos del ejemplo que para la provincia de Madrid existen datos desde enero de 2014 mientras que para la provincia de Barcelona solo desde enero de

2015. Esto no supone ningún problema a la hora de comparar ambas provincias, ya que para datos anteriores de una provincia que la otra no tiene únicamente se muestran estos como nulos, comparando sólo aquel periodo en el cual ambas provincias contienen datos.

4.3.3. Gráficos de provincia

En esta sección de ATMOSPHSPAIN se permite conocer el estado y la evolución de una determinada provincia para todos los contaminantes que sus estaciones de calidad de aire son capaces de medir. Con esta sección la aplicación consigue que sus usuarios puedan conocer el estado de una determinada provincia para todos los contaminantes de forma general.

En primer lugar, para poder conocer el estado de una determinada provincia es necesario seleccionarla en el menú de inicio y a continuación la aplicación nos mostrara un mapa dinámico temporal y una serie de gráficos de contaminación para esa provincia.

Al igual que para la visualización anterior, el contaminante por defecto es el NO₂. Para poder visualizar el mapa dinámico temporal del contaminante deseado es necesario seleccionar el contaminante en el menú horizontal superior.

Debajo del mapa temporal se muestra un segundo menú horizontal para seleccionar que tipo de gráficos se quieren mostrar. Al igual que ocurría en la sección anterior los posibles gráficos son los siguientes:

- Gráfico de barras verticales
- Gráfico de barras horizontales
- Gráfico lineal
- Gráfico de área
- Gráfico escalonado

Estos gráficos mencionados tienen una diferencia con respecto a los de la sección anterior, en la sección anterior se realizaba una comparativa en el grafico para un determinado contaminante de las dos provincias seleccionadas, y en este caso el grafico muestra una comparativa para todos los contaminantes medidos en la provincia seleccionada.

Con estos gráficos el usuario podrá visualizar el estado de todos los contaminantes de forma general de una sola vez para aportar un valor extra a los mapas temporales que únicamente podían mostrar la evolución de un único contaminante.

La integración de los mapas con los gráficos permite al usuario conocer la evolución temporal del contaminante deseado y además de forma complementaria la información de todos los contaminantes de la provincia en forma de gráficos para conocer rápidamente que contaminantes son más abundantes en la provincia.

A continuación, se muestra un ejemplo realizado para conocer el estado de la contaminación en la provincia de Madrid. Para realizar este ejemplo es necesario elegir la visualización de '*Gráficos de Provincia*' y después seleccionar la provincia en el selector:



Ilustración 69: Selección provincia

Una vez seleccionada la provincia se muestra un menú horizontal en el cual se puede elegir el contaminante del mapa temporal:

MADRID

EVOLUCIÓN DEL NO₂ EN LA PROVINCIA DE MADRID DESDE EL 1/2014 HASTA 11/2015

NO₂ NO PM10 PM25 O₃ SO₂ XIL BEN CO TOL

Ilustración 70: Selección Contaminante para mapa temporal

Seleccionando el contaminante deseado a continuación se muestra su correspondiente mapa temporal:

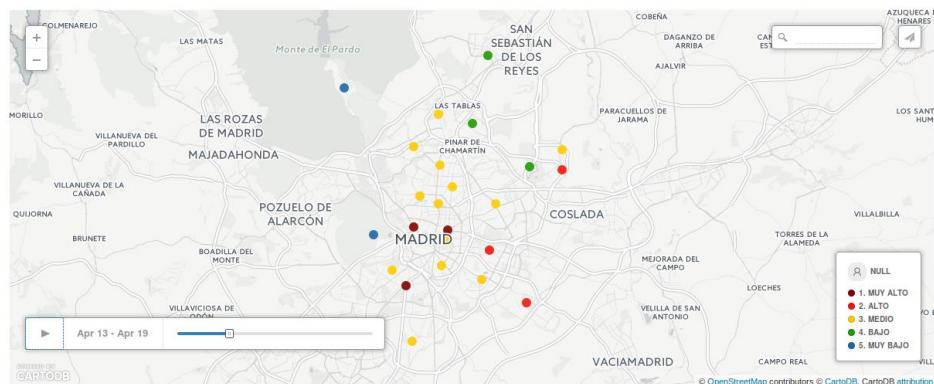


Ilustración 71: Mapa temporal Madrid NO₂

Debajo del mapa temporal se muestra un menú para seleccionar los gráficos que se desean visualizar, al igual que en la visualización anterior, si el borde es amarillo significa que el gráfico esta ya viéndose, si se desea quitar vuelve a pulsar el ícono del gráfico:

GRÁFICOS DE MEDIAS MENSUALES



Ilustración 72: Menú Gráficos una provincia

En base a los gráficos seleccionados del menú anterior el usuario podrá observar los siguientes gráficos:

- **Gráfico de barras verticales:**

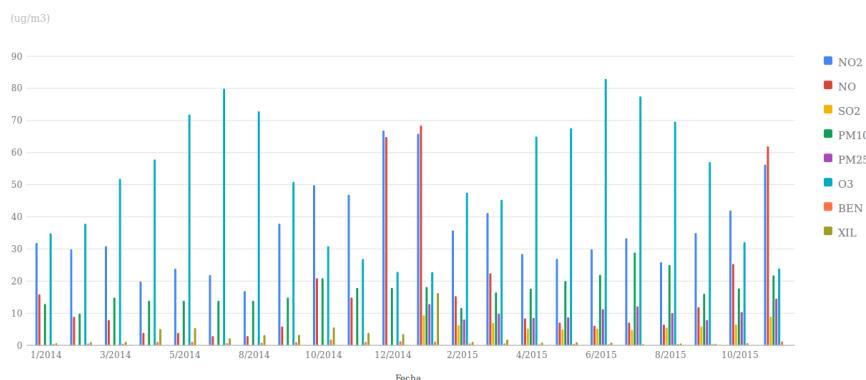


Ilustración 73: Gráfico de Barras Vertical Madrid

- **Gráfico de barras horizontales:**

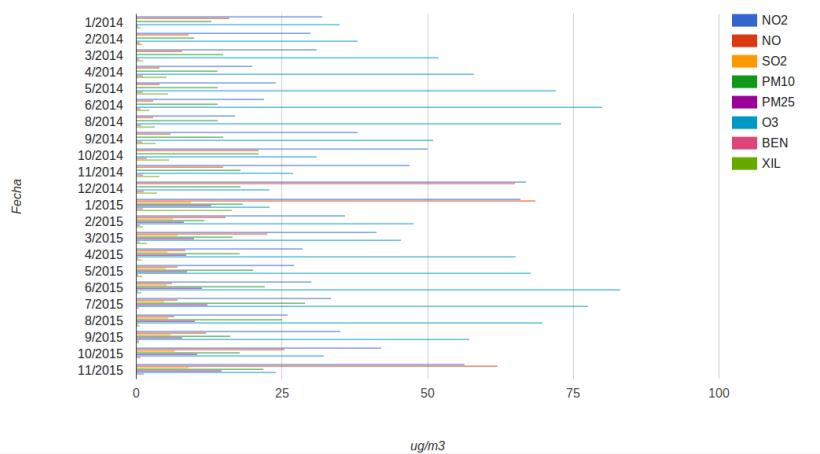


Ilustración 74: Gráfico Barras Horizontales Madrid

- **Gráfico lineal:**

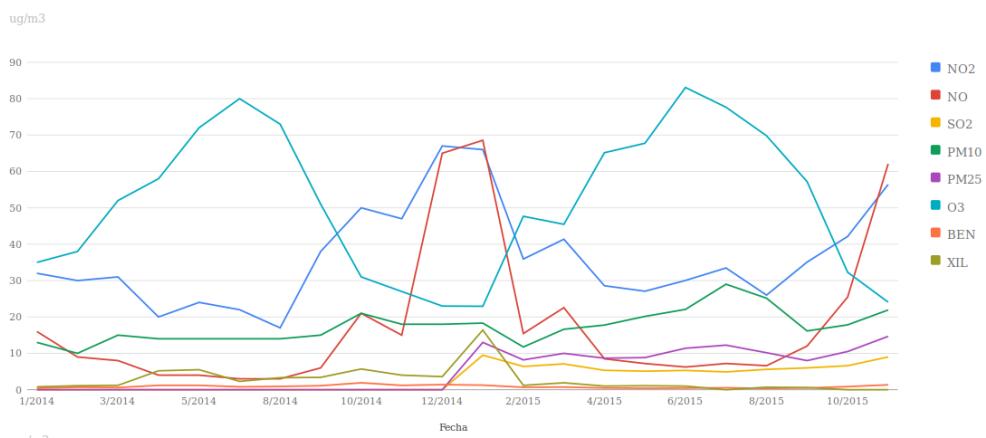


Ilustración 75: Gráfico Lineal Madrid

- **Gráfico de área:**

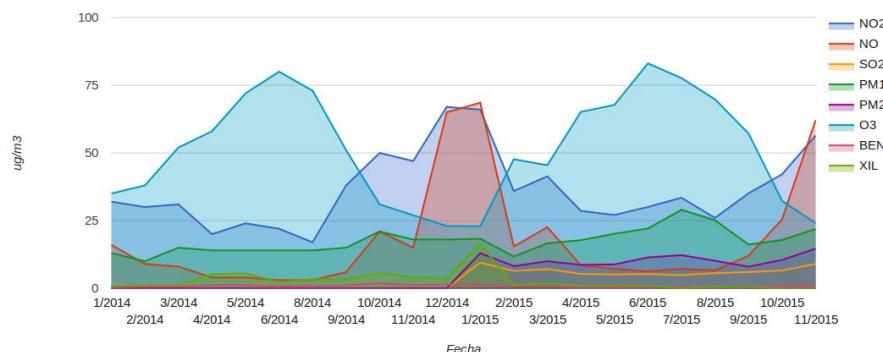


Ilustración 76: Gráfico Área Madrid

- **Gráfico escalonado:**

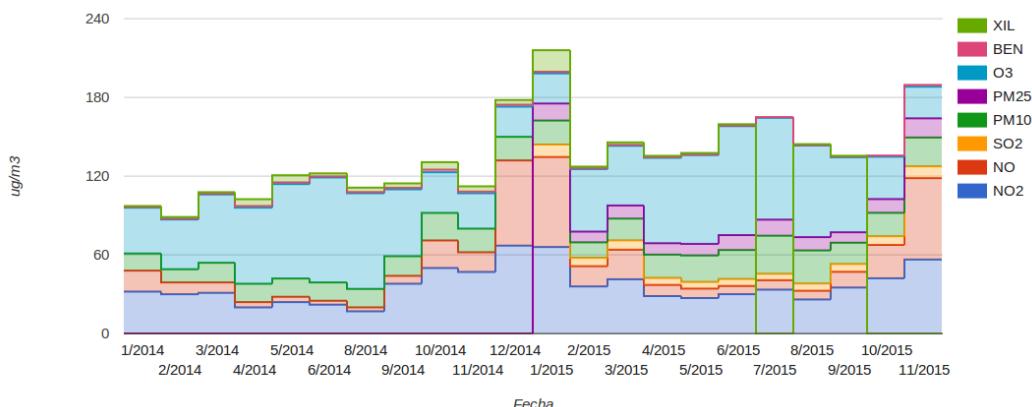


Ilustración 77: Gráfico Escalonado Madrid

Se puede observar si se fija con detenimiento en los gráficos y en los posibles mapas temporales para la provincia de Madrid que dos algunos contaminantes, como son el CO y el TOL, aparecen para mostrar su mapa, pero no en los mismos gráficos que los contaminantes mostrados anteriormente. Esto se debe a que el nivel de contaminación esta medido en diferentes unidades de medida.

Cuando ocurren en caso como el de Madrid donde diferentes contaminantes son medidos en diferentes unidades se muestran gráficos agrupando por la unidad de medida. De esta manera en las imágenes de los gráficos anteriores se han mostrado los contaminantes que han sido medidos en ug/m³ mientras que a continuación se incluyen los gráficos para los contaminantes que son medidos en mg/m³:

- **Gráfico de barras verticales:**

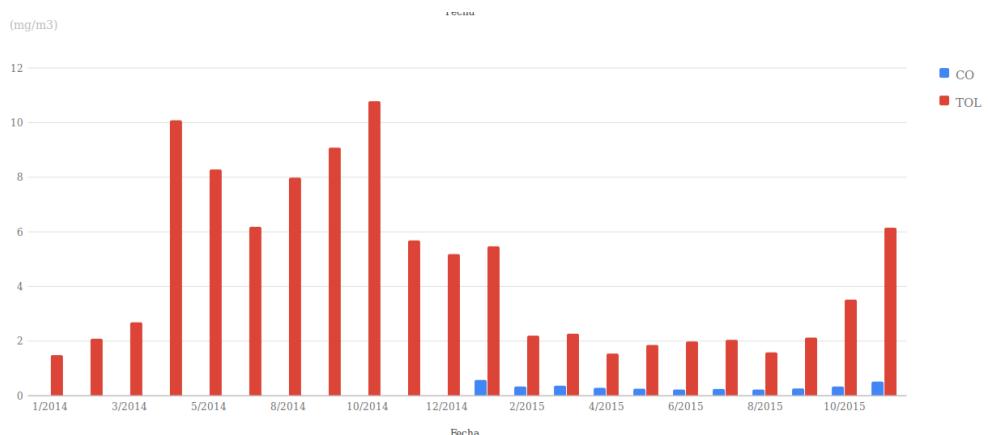


Ilustración 78: Gráfico Barras Verticales CO y TOL

- **Gráfico de barras horizontales:**

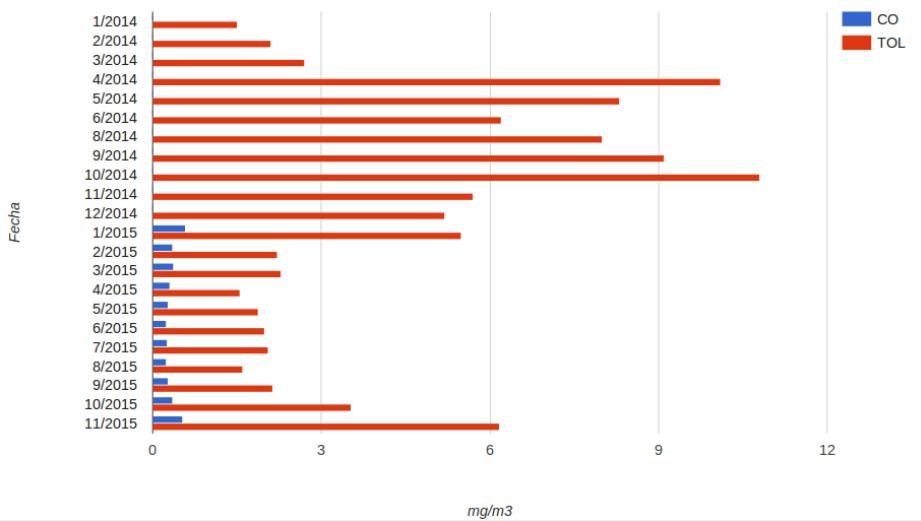


Ilustración 79: Gráfico Barras Horizontales CO y TOL

- **Gráfico lineal:**

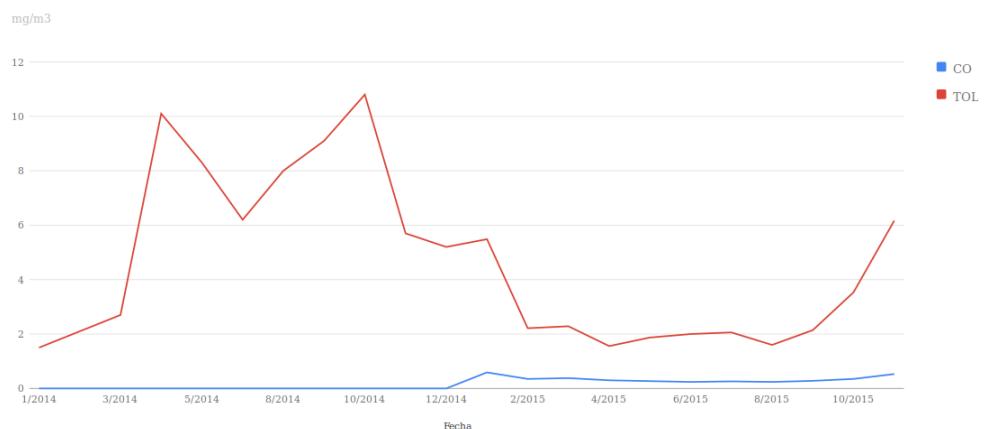


Ilustración 80: Gráfico lineal CO y TOL

- **Gráfico de área:**

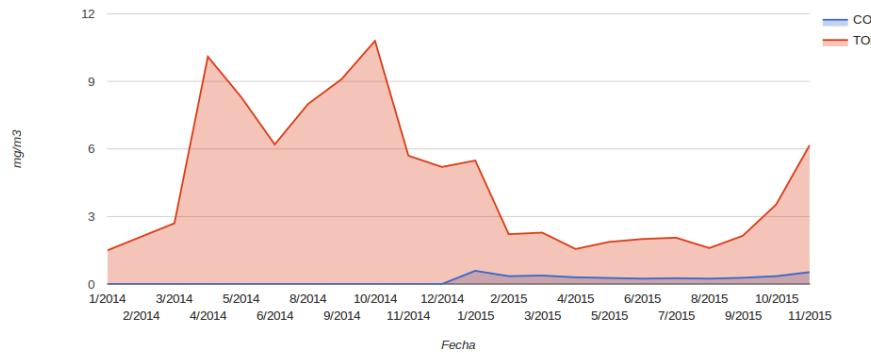


Ilustración 81: Gráfico Área CO y TOL

- **Gráfico escalonado:**

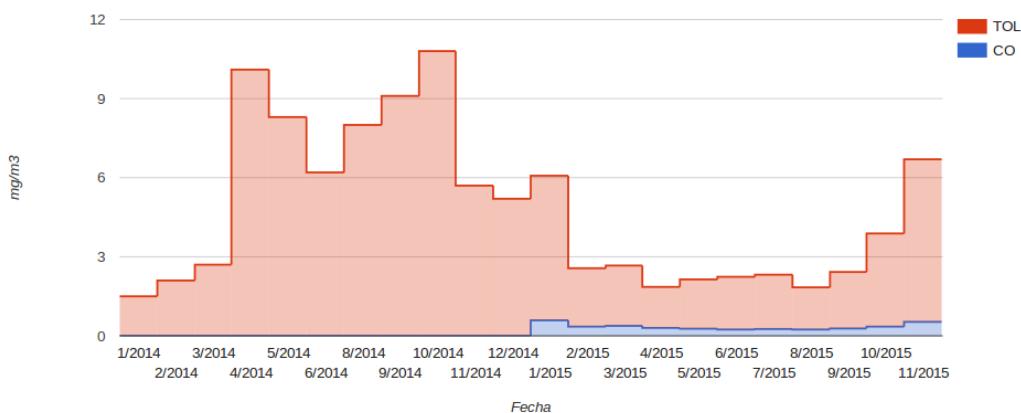


Ilustración 82: Gráfico Escalonado CO y TOL

Esta separación que se ha mostrado en gráficos diferentes ha sido realizada para facilitar la visualización de los usuarios. Esto se debe a que si se realiza la conversión de unos contaminantes a las unidades de medida de los otros y se intenta representar en un mismo grafico aquellos contaminantes medido en unidades más bajas no se aprecian en el grafico debido a la gran diferencia de valor con los otros contaminantes.

4.4. Descarga de datasheets

En este apartado se incluirá una descripción de la sección “DATASHEETS” de ATMOSPHSPAIN. Dentro de la aplicación se permite a los usuarios descargar los datos en formato CSV que la aplicación utiliza para realizar los mapas y gráficos. Estos ficheros CSV están todos en un formato estándar para las diferentes provincias, así de esta forma se facilita a los usuarios el trabajo con ellos y se les ahorra el trabajo de limpieza y preprocesado de los datos.

Esta sección ofrece a los usuarios la posibilidad de disponer de los datos de contaminación de todas las provincias de España a excepción de aquellas de Andalucía y Extremadura. Para acceder a la descarga de los datos de cada provincia se puede realizar de dos maneras:

- Menú superior:

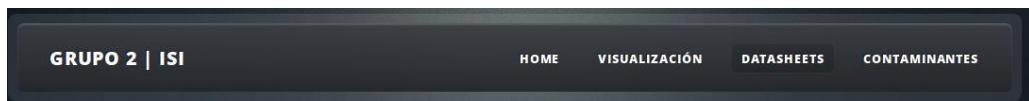


Ilustración 83: Menú Descarga Datasheets

- **Enlace directo:** Avanzando por la página principal que explica el contenido de la aplicación se puede observar un enlace directo a la sección de descarga de datasheets.

Ilustración 84: Descarga Datasheets

Siguiendo cualquiera de los pasos anteriores la aplicación lleva al usuario hasta la sección de descarga de datasheets.

En esta sección, en la columna de la izquierda aparece un listado de todas las provincias que tienen datasheets disponibles. En la parte derecha aparece un mapa de la provincia seleccionada con todas las estaciones de calidad de aire de esa provincia marcadas en el mapa, seguido de todos los enlaces a los ficheros de las estaciones de calidad de aire.

Estas estaciones aparecen marcadas en el mapa debido a que la aplicación permite a los usuarios descargar los datos de cada una de las estaciones que se encuentran en la provincia. Como se puede dar el caso de que los usuarios no se conozcan el nombre de la estación que se encuentra en su zona y por lo tanto que contiene los datos que desea, el mapa proporciona la ubicación de dichas estaciones. Por lo tanto, si un usuario selecciona sobre el mapa una de las estaciones se mostrará el nombre, la latitud y la longitud de la estación para que de esta manera el usuario pueda seleccionar la estación deseada de los enlaces que se encuentran debajo del mapa y que enlazan al fichero CSV.

Cuando el usuario conoce la estación de la cual desea descargar los datos y pulsa sobre su enlace correspondiente se realizará la descarga directa del fichero que contiene los datos para su posterior utilización.

Una vez se conoce el proceso necesario para descargar los datos desde la aplicación se incluye un ejemplo para descargar los datos de una estación de aire ubicada en la provincia de Madrid.

En primer lugar, es necesario acceder a la sección de descarga a través de cualquiera de las dos formas explicadas al comienzo de este apartado. Cuando el usuario se encuentra en la sección de descargas deberá seleccionar la provincia del listado de la izquierda:

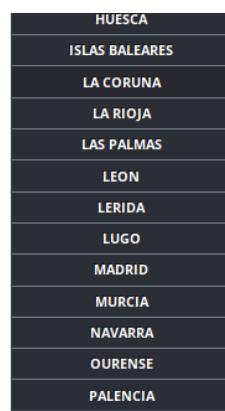


Ilustración 85: Selección Provincia Datasheets

Seleccionando Madrid en el menú izquierdo la aplicación mostrará el siguiente contenido: mapa de las estaciones de calidad del aire y enlaces de descarga de los ficheros CSV.

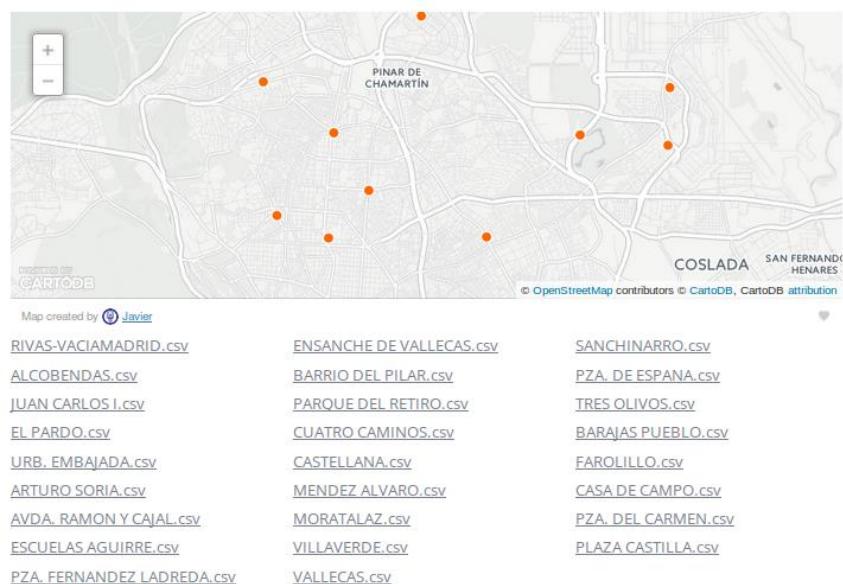


Ilustración 86: Mapa de estaciones de calidad de aire

Si el usuario conoce directamente el nombre de la estación podrá pulsar sobre su enlace correspondiente, en caso contrario, mediante el mapa el usuario podrá elegir la estación que se encuentra en la zona deseada y se mostrará lo siguiente:



Ilustración 87: Nombre de las estaciones de calidad de aire

Con el nombre que se puede observar en la imagen el usuario podrá seleccionarlo el CSV de la lista que aparece en la imagen anterior.

Una vez el usuario ha pulsado sobre el enlace de la estación se abrirá la ventana de descarga como se muestra en la siguiente imagen Y pulsando sobre el botón “Aceptar” comenzará el proceso de descarga

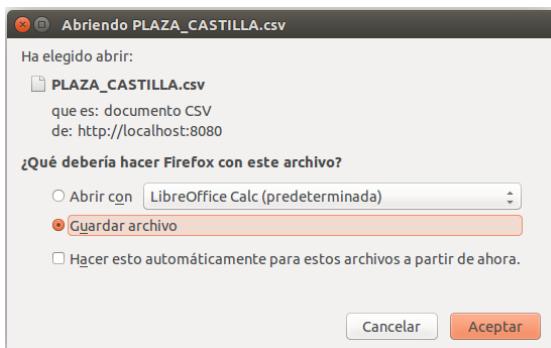


Ilustración 88: Descarga Datasheets

4.5. Contaminantes

En este apartado se explicará cómo funciona la sección “Contaminantes” de ATMOSPHSPAIN. Esta sección contiene información de interés para los usuarios sobre los contaminantes recogidos por las diferentes estaciones de aire, los cuales pueden consultarse en el *ANEXO II: Elementos químicos y variables medidas*. Dentro de esta información de interés para los usuarios de la aplicación se encuentra un ranking de las cinco provincias más contaminadas para el contaminante elegido y una breve descripción de los límites de cada contaminante y como los diferentes valores de un determinado contaminante influyen a la salud de las personas.

Para acceder a la información de los contaminantes se puede realizar de dos maneras:

- **Menú superior:**

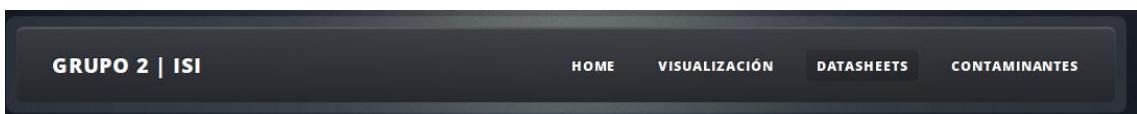


Ilustración 89: Menú superior Contaminantes

- **Enlace directo:** Avanzando por la página principal que explica el contenido de la aplicación se puede observar un enlace directo a la sección de contaminantes.

NUMEROSOS CONTAMINANTES

INFÓRMENSE SOBRE LOS CONTAMINANTES MOSTRADOS

ATMOSPHEREPAIN le ayuda a informarse sobre los riesgos para la salud de los contaminantes atmosféricos mostrados en las visualizaciones. Además, podrá ver el top 5 de provincias a la cabeza en cada contaminante.

CONTAMINANTES

Ilustración 90: Enlace Contaminantes

Una vez el usuario accede a la sección de Contaminantes, la aplicación le mostrará en la parte izquierda un menú vertical que contiene todos los contaminantes disponibles, y en la parte derecha un ranking con las cinco ciudades más contaminadas y una información sobre los contaminantes y los riesgos a la salud según los niveles.

De esta manera cuando un usuario selecciona uno de los contaminantes desde el menú lateral izquierdo la aplicación le muestra el ranking de las ciudades más contaminadas para ese elemento y a continuación la información de dicho elemento sobre los riesgos a la salud.

A continuación, se muestra un ejemplo para conocer la información sobre el elemento Tolueno (TOL en adelante). En primer lugar, es necesario seleccionar desde el menú lateral el Tolueno como se muestra en la siguiente imagen:

NO
NOX
TOL
BEN
PM10
SO2
CO
O3
SH2
XIL
NO2
OXL
PM25

Ilustración 91: Selección de Contaminante

Una vez seleccionado el TOL, la aplicación muestra el ranking de las ciudades más contaminadas para este elemento, en este caso son las que se pueden observar en la siguiente imagen:



Ilustración 92: Ranking por Contaminante

Debajo de este ranking la aplicación muestra la información relacionada con el Tolueno, así como los límites que pueden provocar daños para los seres humanos como se muestra a continuación:

8 mg/m³: Detección de olor.

188-377 mg/m³: Fatiga o dolor de cabeza. Probablemente no se produzca deterioro observable del tiempo de reacción o de la coordinación.

753 mg/m³: Irritación suave de los ojos y de la garganta.

377-1.130 mg/m³: Se pueden producir indicios perceptibles de incoordinación en períodos de exposición de hasta 8 horas.

1.507 mg/m³: Lagrimeo e irritación de ojos y garganta.

1.130-3.014 mg/m³: Se pueden esperar grandes indicios de incoordinación en períodos de exposición de hasta 8 horas.

2.260-3.014 mg/m³: Causa fatiga, náuseas, confusión y ataxia en exposiciones de 3 horas.

5650 mg/m³: Probablemente no es mortal durante períodos de exposición de hasta 8 horas.

15.067 mg/m³: Probablemente perjudicaría rápidamente al tiempo de reacción y a la coordinación. Exposiciones de una hora o más

pueden conducir a depresión del SNC y posiblemente a la muerte.

26.368 mg/m³: Se ha observado paresia, amnesia y estupefacción.

37.669 mg/m³: Causa anestesia general

37669-113006 mg/m³: En pocos minutos aparece la depresión del SNC, exposiciones más prolongadas pueden ser mortales.

Fuente: <http://www.murciasalud.es/>

Ilustración 93: Información Contaminante

5. Conclusiones

A nivel global se ha cumplido con los objetivos marcados al inicio del documento. Se ha conseguido mostrar de una manera visual la contaminación en las diferentes provincias de España desde un punto de vista social gracias a Twitter y uno real en base a las medidas realizadas por las estaciones de calidad del aire. Además, se han usado un gran número de fuentes de datos para esta tarea (numerosas webs de ayuntamientos, Twitter, ...), lo que ha permitido que los datos mostrados en la prueba de concepto sean lo más completos y veraces posibles.

Por otro lado, se han usado diferentes herramientas a la hora de integrar y trabajar con los datos, como son MongoDB, Apache Tomcat, CartoDB y APIs de Twitter, Google o Meaningcloud. Esto ha permitido al grupo facilitar el trabajo y manipulación de los datos.

Se ha realizado una prueba de concepto lo bastante completa como para comprobar todas las funcionalidades que se explican y recogen en esta memoria. De esta manera, cualquier usuario que use la aplicación web resultante puede hacerse una idea lo bastante completa de que se ha pretendido conseguir y de lo que se está mostrando.

Se han encontrado limitaciones a la hora de usar diferentes APIs como las de Twitter, Google o Meaningcloud, ya que en su versión gratuita tienen un límite de peticiones o de tiempo de uso. Esto, por ejemplo, ha conseguido que el número de tweets recogidos para la prueba de concepto no fuera tan mayor como se pretendía por parte del grupo.

Para finalizar, la realización de este caso práctico ha permitido a los alumnos comprender y ver las ventajas de como usando diversas herramientas se pueden integrar datos de diferentes fuentes para luego mostrar análisis y reportes de ellos en una única solución, en la cual se da una visión global del conjunto de ellos.

ANEXO I: Cantidad de estaciones por provincia

La siguiente tabla muestra el número de estaciones de calidad de aire por provincia cuyos datos se han recopilado para la solución del caso práctico.

Provincia	Cantidad	Provincia	Cantidad
A Coruña	6	La Rioja	5
Álava	8	Las Palmas	20
Albacete	1	León	5
Alicante	12	Lérida	1
Almería	0	Lugo	4
Asturias	5	Madrid	26
Ávila	1	Málaga	0
Badajoz	0	Melilla	0
Barcelona	24	Murcia	8
Burgos	6	Navarra	9
Cáceres	0	Ourense	2
Cádiz	0	Palencia	2
Cantabria	11	Pontevedra	4
Castellón	23	Salamanca	3
Ceuta	0	Santa Cruz de Tenerife	30
Ciudad Real	5	Segovia	1
Córdoba	0	Sevilla	0
Cuenca	1	Soria	2
Gerona	1	Tarragona	6
Granada	0	Teruel	1
Guadalajara	2	Toledo	3
Guipúzcoa	14	Valencia	25
Huelva	0	Valladolid	6
Huesca	3	Vizcaya	25
Islas Baleares	17	Zamora	1
Jaén	0	Zaragoza	2
Total	331		

Tabla 6 Núm. Estaciones por provincia

ANEXO II: Elementos químicos y variables medidas

La siguiente tabla muestra el número de elementos químicos y variables medidas por las estaciones de calidad del aire guardadas.

Abreviatura	Nombre	Unidades
BEN	Benceno	µg/m3
CO	Monóxido de carbono	mg/m3
NO	Monóxido de nitrógeno	µg/m3
NO2	Dióxido de nitrógeno	µg/m3
NOx	Óxidos de nitrógeno	µg/m3
O3	Ozono	µg/m3
OXL	Ortoxileno	µg/m3
PM10	Partículas en suspensión < 10 µm (*)	µg/m3
PM25	Partículas en suspensión < 2,5 µm	µg/m3
SH2	Sulfuro de Hidrógeno	mg/m3
SO2	Dióxido de azufre	µg/m3
TOL	Tolueno	mg/m3
XIL	Xileno	µg/m3

Tabla 7 Elementos químicos y variables medidas

(*)- PM10 y PM25 recogen información sobre partículas compuestas por polvo, cenizas, polen, cemento, hollín y metales pesados que hay en la atmósfera. El diámetro de cada partícula es el siguiente:

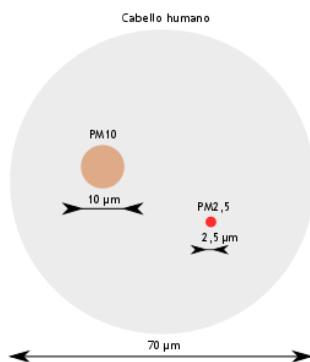


Ilustración 94 PM10 y PM25

Para el análisis de establecimiento de los límites de los contaminantes se ha usado como guía las siguientes páginas:

- <http://www.murciasalud.es/pagina.php?id=99958&idsec=2656#> - Murcia Salud
- <http://www.who.int/mediacentre/factsheets/fs313/es/> - OMS

ANEXO III: Temas, hashtags y cuentas Twitter

La siguiente tabla muestra en base a qué temas, hashtags y cuentas se han recopilado los mensajes usados en la solución final.

Temas	Hashtags	Cuentas
alergia contaminación	#airemadrid	ASURCAI
basura Barcelona	#airebarcelona	ecogestos
basura calle	#basura	EPmedioambiente
basura calles	#basuramadrid	FilterQueenES
basura ciudad	#calidadaire	
basura Madrid	#contaminaciobarcelona	
basura valencia	#contaminacion	
basureros	#contaminacionatmosferica	
calles limpias	#contaminacionmadrid	
CO2	#desperdicios	
contaminacion aire	#madridsincontaminacion	
contaminacion atmosférica	#medioambientemadrid	
huelga recogida basura	#malolorbcn	
limpieza Madrid	#pestebcn	
NO2	#residuos	
PM10		
papeleras Barcelona		
puntos limpios		
papeleras Madrid		
reciclaje		
recogida de basura		
SO2		
suciedad		
vertederos		

Tabla 8 Temas, hashtags y cuentas de tweets

ANEXO IV: Fuentes de datos

A continuación, se muestran los links donde se han extraído los datos de las medidas de las estaciones de calidad del aire para las diferentes provincias de España.

- Aragón: http://www.aragonaire.es/site_information.php?site_id=44014002
https://www.zaragoza.es/ciudad/risp/buscar_Risp
- Asturias: https://transparencia.gijon.es/set/medio-ambiente/calidad_aire_ultimos
- Canarias:
<http://www.gobiernodecanarias.org/medioambiente/calidaddelaire/ica.do>
- Cantabria: <http://www.airecantabria.com/historico.php>
- Castilla-La Mancha:
<http://pagina.jccm.es/medioambiente/rvca/estaciones/casetas/calleancha.htm>
<http://pagina.jccm.es/medioambiente/rvca/meteo.htm>
- Castilla y León: http://servicios.jcyl.es/esco/historicos_buscar.do
- Cataluña: <http://dtes.gencat.cat/icqa/start.do?lang=es>
- Comunidad Valenciana: <http://www.citma.gva.es/web/calidad-ambiental/datos-obtenidos-a-partir-de-la-rvvcca>
- Euskadi: <http://www.geo.euskadi.eus/calidad-aire-en-euskadi-2015/s69-geodir/es/>
- Galicia: http://www.meteogalicia.es/Caire/datos.action?request_locale=es
- Islas Baleares:
<http://www.caib.es/sacmicrofront/contenido.do?mkey=M145&lang=ES&cont=3184>
- La Rioja: <http://www.larioja.org/npRioja/default/defaultpage.jsp?idtab=439670>
- Madrid: <http://www.mambiente.munimadrid.es/>
<http://datos.alcobendas.org/dataset/contaminacion-atmosferica>
- Murcia: <http://sinclair.carm.es/calidadaire/>
- Navarra:
http://www.navarra.es/home_es/Temas/Medio+Ambiente/Calidad+del+aire/

Parte de los datos de calidad del aire de los links anteriores y otros se descubrieron usando el buscador siguiente <http://datos.gob.es/catalogo/>