UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

DESARROLLO DE UN SISTEMA DE RECOLECCIÓN DE DATOS SOBRE MEDIOS DIGITALES PARA SU ANÁLISIS CIBERMÉTRICO EN TWITTER

Trabajo Fin de Máster

Máster Universitario en Gestión de la Información

Autor: José Pablo Pascual Yarritu

Tutor: Nuria Lloret Romero

Curso 2016 - 2017

Agradecimientos

Mi agradecimiento a mi tutora, Nuria Lloret, y a Marga Cabrera, fundadora y directora del Observatorio de nuevos medios, por dejarme participar en este proyecto tan ilusionante y por su interés e implicación durante el desarrollo del trabajo realizado.

Mi gratitud también a todos los integrantes del proyecto del Observatorio, porque sin su contribución no hubiese sido posible llevar a cabo este trabajo, y en especial a Benjamín Arroquia, por su gran desempeño y dedicación en todo momento.

Resumen

La revolución digital que se ha producido en los últimos años ha resultado ser una gran oportunidad para proyectos nuevos que no disponen de demasiados recursos para llevar a cabo sus funciones. Este hecho ha afectado a prácticamente todos los sectores, desde comercios on-line hasta negocios de consumo colaborativo. No obstante, ha beneficiado de manera notable a un sector en particular: el periodismo.

El cambio en la forma de consumir información es hoy en día evidente. Los periódicos, la televisión y la radio ya no son los medios más utilizados para mantenerse informado. Ahora predomina Internet y más concretamente la Web, donde las redes sociales se erigen como las soberanas de la difusión de información.

Del mismo modo que se usa el índice de audiencia en la televisión y el número de ejemplares vendidos en la prensa escrita para calcular aproximadamente la situación de un medio informativo, en las redes sociales se contabilizan las interacciones de los usuarios. Éstas se registran en las bases de datos de cada red social y son accesibles, en mayor o menor medida, para poder estudiarlas y cuantificarlas.

Este proyecto pretende reunir una serie de iniciativas digitales de carácter divulgativo, los llamados nuevos medios, y elaborar un repositorio capaz de mostrar y comparar la situación de los mismos a partir de los datos generados en la red social Twitter. Para ello, se ha implantado un sistema que integra varios elementos y que es capaz de extraer, depurar y tratar datos de Twitter de manera autosuficiente, es decir, sin la intervención manual de un administrador.

El repositorio es accesible mediante la página web del Observatorio de nuevos medios, donde se representan los datos de forma visual para facilitar la interpretación de los mismos por parte de los usuarios.

Palabras clave

Twitter, medios digitales, sistema de recolección de datos, análisis cibermétrico, gestión de información, redes sociales, explotación de datos masivos.

Summary

The digital revolution that has occurred in recent years has proved to be a great opportunity for new projects that do not have too many resources to carry out their functions. This fact has affected all sectors, from e-commerce to collaborative consumer businesses. However, it has benefited notably to one sector in particular: journalism.

The change in the way in which information is consumed nowadays is evident. Newspapers, television and radio are no longer the most used media to keep informed. Now the Internet predominates and more specifically the Web, where social networks are established as the sovereigns of the dissemination of information.

In the same way that the audience index is used for the television and the number of copies sold are used for the press to calculate the situation of an information medium, user's interactions are measured in the social networks. These interactions are recorded in the databases of each social network and they are accessible, to a greater or lesser extent, to study and quantify them.

This project tries to collect a series of digital initiatives of divulgative character, the so-called nuevos medios, and to elaborate a repository able to show and to compare the situation of the same ones from the data generated in Twitter social network. For this, it has been implemented a system that integrates several elements and is able to extract, debug and treat Twitter data in a self-sufficient way, that is, without the manual intervention of an administrator.

The repository is accessible through the website of the Observatorio de nuevos medios, where the data are represented in a visual way to facilitate the interpretation of the same by the users.

Key words

Twitter, digital media, data collection system, cybermetric analysis, information management, social networks, big data.

Resum

La revolució digital que s'ha produït en els últims anys ha resultat ser una gran oportunitat per a projectes nous que no disposen de massa recursos per a dur a terme les seues funcions. Este fet ha afectat pràcticament tots els sectors, des de comerços online fins a negocis de consum col·laboratiu. No obstant això, ha beneficiat de manera notable a un sector en particular: el periodisme.

El canvi en la forma de consumir informació és hui en dia evident. Els periòdics, la televisió i la ràdio ja no són els mitjans més utilitzats per a mantindre's informat. Ara predomina Internet i més concretament la Web, on les xarxes socials s'erigixen com les sobiranes de la difusió d'informació.

De la mateixa manera que s'utilitza l'índex d'audiència en la televisió i el nombre d'exemplars venuts en la premsa escrita per a calcular aproximadament la situació d'un mitjà informatiu, en les xarxes socials es comptabilitzen les interaccions dels usuaris. Estes es registren en les bases de dades de cada xarxa social i són accessibles, en major o menor mesura, per a poder estudiar-les i quantificar-les.

Aquest projecte pretén reunir una sèrie d'iniciatives digitals de caràcter divulgatiu, els cridats nuevos medios, i elaborar un repositori capaç de mostrar i comparar la situació dels mateixos a partir de les dades generades en la xarxa social Twitter. Per a això, s'ha implantat un sistema que integra diversos elements i que és capaç d'extraure, depurar i tractar dades de Twitter de manera autosuficient, és a dir, sense la intervenció manual d'un administrador.

El repositori és accessible per mitjà de la pàgina web del Observatorio de nuevos medios, on es representen les dades de forma visual per a facilitar la interpretació dels mateixos per part dels usuaris.

Paraules clau

Twitter, mitjans digitals, sistema de recol·lecció de dades, anàlisi cibermétric, gestió d'informació, xarxes socials, explotació de dades massives.

Índice de contenidos

Índice de figuras

1 Introducción

1.1 Origen

La idea de este proyecto surge tras la necesidad de contar con un directorio web de nuevos medios, inexistente hasta hoy. La definición de un nuevo medio no es fácil de determinar, ya que actualmente no existe un consenso claro y definitivo sobre su concepto. A través de este proyecto, llamado Observatorio de nuevos medios, se pretende hallar su definición mediante entrevistas a expertos, en las cuales tratan de explicar qué es un nuevo medio para ellos. Aunque aún no se tiene una explicación precisa de lo que son, parece ser que hay unas características comunes: son proyectos de carácter divulgativo que utilizan cualquier canal, principalmente digital, para difundir sus contenidos. Como ejemplo de nuevos medios digitales, se encuentran algunos tan populares como: eldiario.es, Okdiario o El Confidencial (figura 1), entre muchos otros.

Figura 1. Logos de algunos de los nuevos medios más reconocidos en España.

La vía principal de comunicación hoy en día es sin duda Internet y, como tal, la gente utiliza cada vez más este medio para informarse sobre cualquier temática, en detrimento de los medios en formato papel, ya sean periódicos, revistas u otras publicaciones. Se está produciendo un cambio en cuanto al método de consumo de la información, y por eso se ha estimado preciso contar con un listado de los nuevos medios. Así pues, tras comprobar que no existía un repositorio común que albergase tal información, se puso en marcha este proyecto, el Observatorio de nuevos medios.

Desde un primer momento se pensó en un directorio que fuera dinámico, es decir, que tuviera datos actualizados sobre los nuevos medios, de modo que se pudiera hacer un seguimiento de los mismos. Además, la clave de todo pasaba porque la actualización de los datos fuera, en gran medida, un proceso automático o al menos semi-automático. Asimismo, se necesitaba también una fuente de datos que fuera dinámica, de forma que nutriera de manera constante al repositorio. La solución a esto último se encuentra en las redes sociales. Las redes sociales constituyen una fuente inmensa de datos que se retroalimenta a partir de la información que proporcionan sus propios usuarios. Si a esto se le añade que permiten acceder a esta información y extraerla, se convierten en una de las mejores opciones para la descarga masiva y continua de datos. Como los nuevos medios utilizan frecuentemente estas herramientas para difundir y hacer visibles sus contenidos, la nutrición de datos al Observatorio está asegurada.

Una de las principales funcionalidades con las que se deseaba dotar al directorio de nuevos medios era la capacidad de comparar los medios entre sí. Para ello, se debían tratar los datos obtenidos de las redes sociales para conseguir algún tipo de métrica que hiciese posible la comparación. Además, estas métricas debían ser fácilmente comprensibles por cualquier visitante del directorio, por lo que se hacía indispensable representar los datos en forma de gráficas, mapas y otros elementos visuales. Por último, también se deseaba que el usuario pudiese interactuar con el Observatorio, de modo que desde un primer momento se propuso la idea de elaborar rankings de los medios a partir de filtros de métricas, temáticas, etc. y que fuese el propio visitante quien decidiera "contextualizar" la situación de los medios.

1.2 Fuente de datos

Como ya se ha mencionado, la principal fuente de datos son las redes sociales, aunque no es la única. A los datos obtenidos a partir de éstas hay que añadir otros que se obtienen tras una labor de investigación manual. Hay información muy útil sobre los nuevos medios que no se puede conseguir, al menos directamente, a través de las redes sociales. Datos como la cobertura geográfica, es decir, información que indica si un medio es de ámbito internacional, nacional, local, etc. o las temáticas, si se trata de medios sobre política, cultura, deporte, etc. son ejemplos de ello.

Actualmente, la base principal de datos del Observatorio es Twitter, aunque con el tiempo se pretenden incorporar otras redes sociales como Facebook, Instagram y Youtube. De ésta última ya se ha comenzado a descargar datos, sin embargo aún está pendiente su integración. El motivo de la elección de Twitter como primera red social desde la que obtener datos sobre nuevos medios se debe por la gran presencia de los mismos en ella. La rapidez con la que una cuenta pude informar a sus seguidores y la facilidad que ofrece para que una publicación se haga viral son quizás los puntos fuertes por los que los nuevos medios hacen uso de esta red social. Por otra parte, Twitter es una de las redes sociales que mayor facilidad y menos trabas pone para extraer datos. Aunque como todas, hay que seguir una serie de pasos y requisitos para acceder a su información. Existe documentación detallada de su API o Interfaz de Programación de Aplicaciones y de las buenas prácticas para conseguir una extracción de datos correcta y continua.

Figura 2. Logos de Twitter (izquierda) y Youtube (derecha).

La información que se muestra en el Observatorio de nuevos medios no es directamente la que se obtiene de Twitter. Entre la extracción de datos de la red social y la información que aparece en el directorio web de nuevos medios hay una labor de depuración y tratamiento de datos. Twitter ofrece una gran cantidad de datos, de los cuales se seleccionan los que realmente son útiles para el proyecto y se desechan los demás. De esta forma se optimiza el espacio, es decir, a la hora de almacenar esta información en una base de datos ocupa menos. Esto puede parecer exagerado a primera vista, pero si se tiene en cuenta que se almacenan datos por cada medio, que actualmente el directorio cuenta con más de 1.600 medios y que de cada medio se obtienen datos sobre sus publicaciones, sobre las interacciones sobre el propio medio, seguidores, localizaciones, etc. se convierte en una tarea casi obligatoria.

Twitter ha resultado ser una buena apuesta para obtener datos pero es prioritario incorporar otras redes sociales, ya que ahora mismo el Observatorio depende totalmente de ella. Los últimos balances indican que Twitter está siendo menos utilizado, por esto se debe la importancia de contar con distintas fuentes de información.

1.3 Funcionamiento

Existen varios procesos desde que se descubre un nuevo medio hasta que aparece en el Observatorio y es visible por el usuario final. Esta serie de procesos definen un flujo de trabajo o funcionamiento que hacen posible este proyecto. Así pues, el procedimiento seguido es el siguiente (figura 3):

Figura 3. Flujo de trabajo del Observatorio de nuevos medios

En primer lugar, es necesario una labor de investigación para encontrar nuevos medios que cumplan los requisitos y condiciones del Observatorio. Estos requisitos se explican en detalle en el apartado *condiciones* dentro de esta misma sección. En el caso en que un nuevo medio los cumpla, se procede a la búsqueda de información valiosa sobre el mismo para completar una ficha general para todos los medios. Esta ficha contiene campos como las cuentas de los medios en las redes sociales, el correo electrónico, la fecha de creación, las temáticas, la cobertura geográfica, etc. El resultado de la labor de investigación se refleja en dicha ficha completada con el conjunto de nuevos medios encontrados.

Una vez se ha recopilado la información documental de los medios, ésta se almacena en la base de datos asignada para ello. Es entonces en este punto cuando comienza a funcionar el sistema de extracción automática de datos de Twitter. Este sistema accede a la base de datos donde se ha insertado la ficha con los medios y obtiene las cuentas de Twitter de los mismos. Con esta información y las peticiones correspondientes a la API de Twitter ya es posible la descarga de datos. Mientras se produce la extracción da comienzo el proceso de depuración, que se encarga de discriminar los datos considerados como relevantes de aquellos que no lo son para que, inmediatamente después, se almacenen en una nueva base de datos.

En el momento en que acaba la descarga (tres veces al día), el sistema realiza la explotación de los datos con el fin de obtener unas métricas generales para todos los medios y así poder compararlos. Este cálculo se realiza diariamente y tiene en cuenta las interacciones realizadas en el mes en el que se ejecuta este proceso. De esta forma al final se tienen almacenadas unas métricas globales de los medios para cada mes.

La página web del Observatorio, por su parte, hace la petición de los datos a través de un web service, del que se hablará en el apartado *Estructura del sistema*, y obtiene tanto los datos extraídos de Twitter como los que tienen su origen en la labor de investigación. Sin embargo, no basta con mostrar los datos directamente. De nada sirve todo lo mencionado hasta este momento si finalmente cualquier visitante del Observatorio no puede analizar ni interpretar fácilmente los datos ofrecidos. Es por esto que la representación y visualización de datos es tan importante como la obtención de los mismos. Así pues, se emplean elementos visuales tales como gráficas y mapas para facilitar el análisis de los datos por parte del usuario.

El último proceso es el que denota si el trabajo realizado ha conseguido llegar a buen puerto o falla en alguno de los procedimientos anteriores. Este consiste en que el usuario final sea capaz de obtener conocimiento y conseguir ideas a partir de la información que ve. Por ejemplo, sería deseable que un usuario que formara parte de algún nuevo medio incluido en el Observatorio pudiera sacar conclusiones de cuál es la situación de su proyecto y cuál la de su principal competencia.

1.4 Enfoque

El proyecto del Observatorio de nuevos medios pretende ser una herramienta útil para todo aquel que desee saber más acerca de los nuevos medios digitales. Periodistas, investigadores o incluso miembros que formen parte de algún nuevo medio pueden sacar provecho de la información que se facilita en la página web.

El Observatorio está en continuo desarrollo. Ahora mismo incluye unas métricas que permiten al menos conocer de forma general la situación de los nuevos medios disponibles. Sin embargo, estos cálculos son solo un pequeña parte de toda la información que se podría extraer. Además, conforme pasa el tiempo se obtienen más datos, y tener más datos significa mayor conocimiento.

El enfoque del Observatorio en el futuro va en la línea de ser capaz de definir la situación de un nuevo medio de forma específica, de modo que el visitante pueda sacar conclusiones acerca de qué medios están teniendo éxito, cuáles no, qué factores debe tener en cuenta un medio para mejorar su influencia, etc. Además, los medios podrían consultar al mismo Observatorio los aspectos que deben mejorar si quieren subir posiciones en el ranking, por ejemplo. Como se estima que más adelante se incorporarán otras redes sociales a parte de Twitter (figura 4), es probable que la cantidad de información reflejada en la página web sea bastante elevada, por lo que quizás habrá medios que prefieran un informe en lugar de invertir tiempo analizando los datos por sí mismos.

Figura 4. Redes sociales a incorporar en el Observatorio.

1.5 Estructura del sistema

La estructura del sistema que permite relacionar todos los procesos englobados en el flujo de trabajo o funcionamiento del Observatorio se puede dividir en cuatro grandes elementos: el software o middleware de gestión de nuevos medios, el sistema de extracción y explotación de datos de redes sociales o kernel, el web service y la página web del observatorio. A continuación se hace una breve explicación de la función de cada uno de ellos:

Software de gestión de nuevos medios: se trata de una página web que actúa de intermediario entre los agentes encargados de la labor de investigación, es decir, de recopilar información manual de los nuevos medios, y la base de datos. En el proyecto del Observatorio trabajan personas que no tienen un perfil técnico, por lo que se hace necesario facilitar en todo lo posible la incorporación de nuevos medios a la base de datos. Esto último define la razón de ser de este elemento del sistema. Asimismo, también permite la edición y la eliminación de los medios ya almacenados. En pocas palabras, consiste en una herramienta de gestión de la base de datos.

Sistema de extracción y explotación de datos de redes sociales: es el componente de la estructura encargado de los datos, tanto de su descarga como de su tratamiento y explotación. Representa el elemento de unión entre la información introducida desde la herramienta de gestión y la extracción de datos de Twitter, ya que necesita el nombre de las cuentas de los nuevos medios en las redes sociales para poder realizar peticiones a las mismas.

Este sistema está siempre vivo, es decir, se inicia automáticamente a las horas que se le indica y él mismo tiene la autonomía para realizar la extracción, el tratamiento, la explotación y el almacenamiento de los datos. Además, tal y como está desarrollado, se pueden configurar aspectos como el número de procesos de descarga y explotación de datos al día o las horas de ejecución de dichos procesos.

Web service: es la pieza intermedia entre la página web del Observatorio de nuevos medios y los datos. Así pues, su función es intercambiar datos entre aplicaciones. Por temas de seguridad, no es muy recomendable que una página o cliente web tenga acceso directo a las bases de datos. Sin embargo, con el web service se consigue una especie de barrera en la que el cliente no sabe cómo está estructurado el sistema ni los datos que éste contiene, únicamente sabe que si hace una petición correcta obtendrá los recursos solicitados. Además,

se le han añadido características para optimizar su rendimiento, como la inclusión de motores de bases de datos en memoria, para que el proceso de petición y obtención de datos sea lo más rápido posible.

Página web del Observatorio: es la parte visible del sistema. Mediante peticiones al web service obtiene toda la información de los nuevos medios, tanto de aquella que se consigue con la labor de investigación como la que tiene su origen en Twitter. Utiliza librerías de visualización para representar la información en forma de gráficas y mapas (figura 5) y permite al usuario interactuar con los datos por medio de filtros y rankings. A parte de todo esto, incluye diversos contenidos relacionados con el Observatorio, como las entrevistas a periodistas, las condiciones que un nuevo medio debe cumplir para ser considerado como tal, etc.

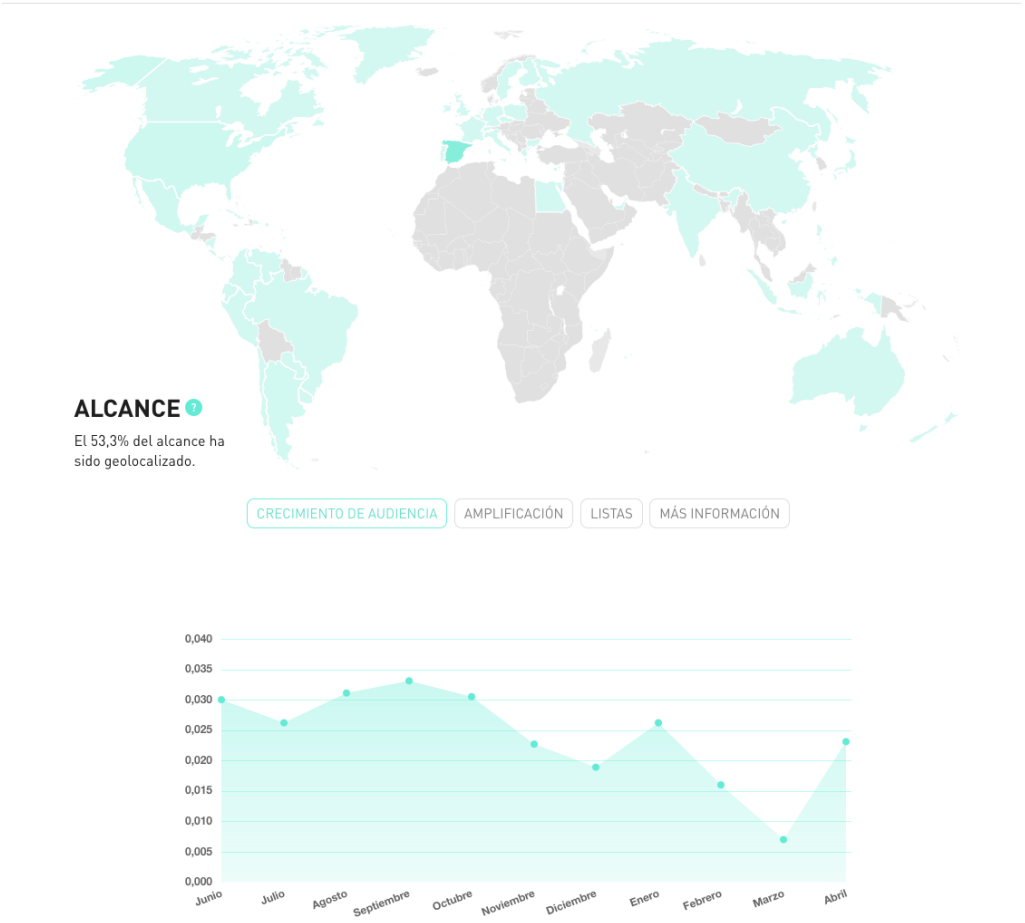

Figura 5. Ficha de un nuevo medio. Datos representados en mapas y gráficas.

1.6 Condiciones

No todos los medios digitales son considerados como nuevos medios por el Observatorio. Para que formen parte del mismo deben cumplir los requisitos y condiciones que se listan a continuación:

- Tienen que estar escritos en español en su mayor parte (más del 60%).
- Deben ser y haber nacido digitales, sin importar el canal que empleen: página web, aplicación, etc.
- Deben publicar contenidos de forma periódica.
- Deben tener al menos seis meses de antigüedad.
- Sus contenidos deben ser informativos sobre cualquiera de sus temáticas.
- Tienen que estar registrados como empresas independientes o al menos con intención de serlo en un futuro muy próximo.
- Deben indicar quién o quiénes están detrás de sus proyectos.
- Deben ser activos en las redes sociales. Por el momento, al menos en Twitter y Facebook.

Todos estos criterios se tienen en cuenta en la labor de investigación, donde cada candidato a ser incluido en el Observatorio se analiza para comprobar su "validez". Sin embargo, este proyecto está vivo y en continuo proceso de mejora, por lo que no se puede afirmar que estos requisitos sean definitivos. De hecho, algunos medios que en un principio no se consideraron como nuevos medios por el incumplimiento de alguno de los criterios sí resultaron serlo al preguntar a los expertos su opinión. Es el caso por ejemplo de Vice, una compañía canadiense que produce y distribuye contenidos en inglés. Sin embargo, cuentan con ediciones locales por todo el mundo donde los contenidos se distribuyen en el idioma local. En España existe una de esas ediciones y se denomina Vice España. Este medio publica su información en español y dispone de cuenta propia en Twitter, de ahí que finalmente se haya aceptado en el Observatorio.

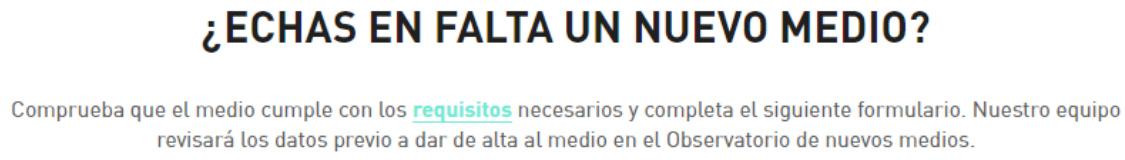

Figura 6. Cada nuevo medio es analizado previamente a su inserción en el Observatorio.

2 Justificación

No hay lugar a dudas de que los últimos avances han cambiado, y siguen haciéndolo, los hábitos de las personas en su día a día. Y es que lo que antes se presuponía imposible ahora es real e incluso normal. Uno de esos hábitos que se han visto modificados en los últimos años es el método de consumo de información. La forma tradicional de obtener información, a excepción de la radio y la televisión, se conseguía mediante la lectura de prensa en formato papel: periódicos, revistas, etc. donde la frecuencia de actualización de contenidos es como mínimo diaria. Hoy en día aún se emplea este canal de acceso a la información, aunque parece que poco a poco se está viendo eclipsado por otro más inmediato: Internet.

Internet es un nexo que permite, a cualquier persona con conexión, acceder y distribuir información de manera inminente, ya sea mediante la web, el correo electrónico, etc. Así pues, un gran número de medios informativos se están adaptado a este canal por las ventajas que ofrece en comparación con la prensa impresa, por ejemplo. De hecho, muchos de ellos aún mantienen el formato en papel porque todavía tienen tirón. Sin embargo, hay otros casos que han nacido directamente en formato digital y que están siendo fundamentales en este proceso de cambio: los nuevos medios.

La mayoría de nuevos medios empiezan siendo proyectos pequeños con muchas ganas de trabajar para mantener informados a sus lectores. Es por ello que hacen mucho uso de las redes sociales para difundir sus contenidos con la intención de fidelizar lectores y darse a conocer a aquellos que no saben de su existencia. De entre los nuevos medios que más éxito han cosechado, por ejemplo, en España destacan eldiario.es, El Mundo Today, El Español o El Confidencial.

Al constatar que no existía aún un directorio que englobara los nuevos medios digitales y que había una necesidad de encontrar de manera fácil y rápida información acerca de ellos, se decidió por desarrollar este proyecto. Asimismo, se pensó que podría ser interesante y útil, tanto para los nuevos medios como para cualquier persona interesada en el periodismo, elaborar una evaluación de los mismos según la interacción de los usuarios de las redes sociales en las que publican contenidos. De esta forma, se pretende ayudar a estos pequeños proyectos dándoles a conocer cuál es su situación y cuál es la de los demás medios con la intención de resultar una herramienta valiosa para guiarles en su camino hacia el éxito informativo.

3 Objetivos

Dentro del global del proyecto realizado, este trabajo se centra en la parte más técnica. Sin embargo, no tiene lógica que se definan los objetivos técnicos si no se define la finalidad común del Observatorio de nuevos medios. Por tanto, se distinguen dos tipos de objetivos: los generales, que aluden a la finalidad global del proyecto y los específicos, que hacen referencia a aquellos más puntuales pero fundamentales para la consecución de los primeros. Así pues, los objetivos se definen a continuación:

Objetivos generales

- Analizar e investigar el perfil y características de los nuevos medios escritos en español. Desde un principio, la intención del proyecto ha sido la de conocer detalladamente qué y qué tipos de nuevos medios en español existen en el mundo. Para ello, ha sido necesario definir un rango de perfiles con el que poder agruparlos y diferenciarlos. Por ejemplo, las temáticas o el ámbito territorial. Asimismo, el propósito de la idea ha ido encaminado hacia el análisis cuantitativo de los nuevos medios, es decir, a hacer posible la medición de los mismos para conocer mejor el contexto en el que se mueven. Gracias a los datos extraídos de Twitter, se han podido calcular una serie de métricas globales para todos ellos. De esta forma, se da un paso más al frente en el camino hacia la comprensión y ubicación de los medios.

- Convertirse en una herramienta útil para conocer y analizar la situación de los nuevos medios. Este objetivo persigue transmitir la información que se tiene de la forma más efectiva posible, de modo que cualquier visitante del Observatorio se sienta cómodo y a gusto utilizando la herramienta y satisfecho por la información proporcionada. Se pretende que el usuario final, ya sea periodista, investigador, miembro de algún nuevo medio o simplemente alguien con interés en el tema, pueda obtener conocimiento y sacar conclusiones a partir de los datos facilitados. Por ejemplo, se conseguirá el objetivo cuando el visitante que forme parte de algún nuevo medio pueda comprender de manera fácil y rápida cuál es la situación de su proyecto y cuál la de su principal competencia. Con intención de cumplir este objetivo, se han empleado elementos visuales como gráficas y mapas para representar los datos y facilitar su interpretación.

Objetivos específicos

- Diseñar un sistema capaz de coordinar todos los procedimientos relacionados con la gestión de datos. Cada proceso de esta parte del proyecto necesita algo del anterior, de modo que si por algún motivo se rompe la cadena, el sistema deja de funcionar correctamente. Por ello es tan importante conectar los diferentes procedimientos que se llevan a cabo durante la manipulación de datos. Las conexiones entre procesos se han realizado mediante desarrollos específicos. Se trata del software de gestión de nuevos medios, del sistema de extracción y explotación de datos de redes sociales y del web service. Elementos que se han definido en el apartado *Estructura del sistema*.

- Automatizar el proceso de extracción, tratamiento y transmisión de los datos. Este es uno de los requisitos indispensables que se buscan cumplir desde que se concibió la idea del Observatorio. El objetivo es que el sistema sea completamente autónomo y que no necesite más que algún control rutinario para comprobar su funcionamiento. Para ello, se ha programado un horario de descarga (tres veces al día) y otro de tratamiento de datos, de modo que a las horas especificadas el sistema se enciende por sí solo y realiza su función.

- Localizar el mayor número posible de usuarios que interactúan con las cuentas de los nuevos medios en Twitter. La intención de obtener esta información se debe a que gracias a ella es posible detectar áreas de influencia de los medios incluidos en el Observatorio. Sin embargo, la localización en muchos casos no se consigue de forma directa a través de Twitter. Es cierto que hay usuarios que sí tienen activada la ubicación. En esos casos, las coordenadas geográficas se incluyen cada vez que éstos publican o interactúan en la red social. Pero son muchos los que no la tienen activada y es entonces cuando hay que recurrir a otras opciones. Se ha optado por extraer el nombre de la localidad que los usuarios introducen en su perfil y transformarlo en coordenadas geográficas mediante un servicio web. De esta forma, se pueden representar después en el mapa que aparece en la página del Observatorio.

- Realizar el desarrollo teniendo en cuenta la escalabilidad del proyecto. No se quiere llevar a cabo una implementación particular por cada red social que se desee añadir al sistema en un futuro. El propósito es elaborarlo de la forma más configurable posible y que al mismo tiempo se puedan extraer datos de diferentes fuentes.

4 Metodología

4.1 Análisis del problema y búsqueda de soluciones

La primera labor que se llevó a cabo tras conocer la idea del Observatorio de nuevos medios fue analizar la manera en que se iba a proceder para conseguir darle forma a la misma. El proyecto en sí cuenta con varios integrantes y cada uno de ellos tiene un perfil y unas funciones concretas dentro del Observatorio. Por ello, ha sido necesario definir un esquema de trabajo con la intención de conseguir una forma de armonizar y encajar las labores de todos.

Para lograr elaborar un directorio de nuevos medios era indiscutible que se necesitaba una investigación previa que determinase las características comunes de los propios medios y así poder comenzar con la búsqueda de los mismos en la Red. Como este paso es el origen de todo, fue el primero en iniciarse. Los encargados de este papel son los integrantes del proyecto con un perfil relacionado a la documentación y su labor se refleja en una ficha con información sobre los nuevos medios encontrados. Para que los medios sigan una descripción común, se ha elaborado una plantilla con los mismos campos para todos ellos. Por otra parte, como la idea del Observatorio contemplaba desde un principio la creación de una página web, los integrantes especializados en desarrollo web se pusieron manos a la obra. De esta forma, los dos extremos del proyecto comenzaron a construirse. Sin embargo, faltaba una parte intermedia que se ocupara de la extracción, combinación, tratamiento y transmisión de los datos. Esta parte del proyecto es la que ocupa este documento y la que se trata de describir a continuación.

En primer lugar, se han definido los elementos necesarios para lograr el objetivo de la parte del proyecto encargada de los datos, que no es otro que servir la información extraída tanto de la labor de investigación como de las redes sociales a la página web del observatorio. Se trata de cuatro elementos necesarios para relacionar todo el "espacio" existente entre las fichas derivadas de la investigación documental y la página web. Estos son: el software de gestión de nuevos medios, el sistema de recolección y tratamiento de datos de Twitter, el almacenamiento y el web service.

El software de gestión es una herramienta necesaria para almacenar las fichas que completan los integrantes responsables de encontrar nuevos medios. En este caso, es mucho más

práctico utilizar una base de datos para guardar la información que una serie de hojas o fichas de trabajo, ya que la cantidad de datos a manejar es bastante grande y además se requiere que otras aplicaciones puedan acceder a ellos. Se trata de una página web en la que los usuarios encuentran varias funcionalidades para interactuar con la base de datos dirigida a guardar la información de los nuevos medios. Entre estas funcionalidades destacan la de importar medios de forma masiva a través de un fichero en formato CSV (Comma-Separated Values), la de inserción manual, la de modificación y borrado o la que permite buscar un medio por su nombre. En definitiva, es un instrumento que permite gestionar una base de datos sin necesidad de tener conocimientos técnicos sobre el lenguaje SQL (Structured Query Language), cosa que les viene muy bien a los que en este caso se ocupan de la labor de investigación.

El siguiente elemento a tener en cuenta es el más complejo de todos y el que más procesos tiene relacionados. Es el sistema responsable de extraer, depurar, tratar y explotar los datos que al final se muestran en la web del Observatorio. Se ha desarrollado de forma que él mismo, es decir, de forma automática, busca las cuentas de Twitter incluidas en la base de datos con la información de las fichas de los nuevos medios y con ellas realiza las peticiones a la API del propio Twitter. Conforme se realiza la extracción de datos, los depura para seguidamente almacenarlos en diferentes colecciones en una nueva base de datos. El proceso de depuración persigue optimizar el espacio que ocupa la información extraída mediante la discriminación de la que se considera útil de la que no. Una vez depurada, se almacena en distintas colecciones según sea el tipo de petición que realiza a Twitter para obtenerla. Por ejemplo, los datos obtenidos a partir de una petición sobre seguidores se guardan en una colección dirigida a ello mientras que los adquiridos a partir de una sobre menciones se almacenan en una colección destinada a guardar ese otro tipo de información.

La extracción de datos de Twitter se lleva a cabo tres veces al día, aunque se puede configurar la frecuencia e incluso las horas en las que deben comenzar los procesos. Asimismo, hay otro proceso que se ejecuta al finalizar el día que se encarga del cálculo de métricas a partir de los datos incluidos en las colecciones existentes tras la extracción y de almacenar éstas en una nueva colección. Finalmente, se obtiene un conjunto de métricas (figura 7) para cada medio y con rotación mensual, es decir, unas estadísticas mensuales, que permiten un seguimiento de los nuevos medios en el tiempo. Sin embargo, se desea que estos datos puedan visualizarse en la página web del Observatorio, por lo que la siguiente cuestión a

resolver ha sido buscar la forma más óptima de comunicar las dos bases de datos con las que cuenta el proyecto con la propia web.

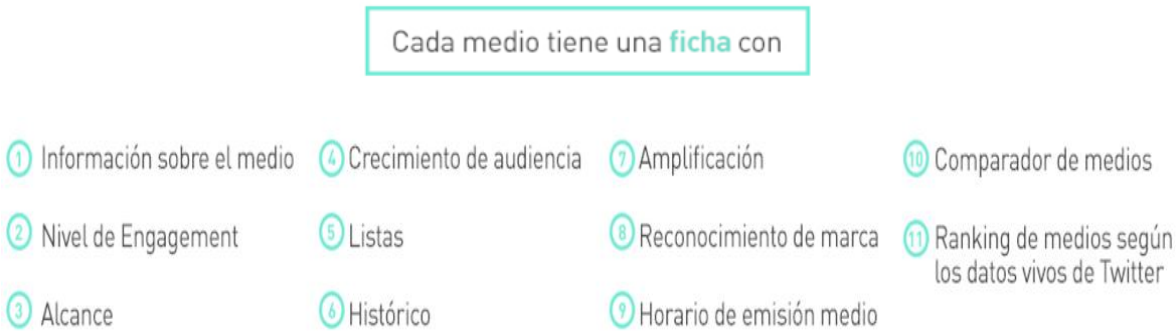

Figura 7. Métricas de los nuevos medios en el Observatorio.

La solución que ha contado con mayor aprobación para transmitir información entre página web y bases de datos ha sido la de incluir un web service (servicio web). Este elemento es fundamental en el proyecto, ya que sin acceso a los datos la página web del Observatorio carece de sentido. Se sitúa como una barrera entre las bases de datos y la web y hace las veces de una recepción, de modo que el cliente, la página web en este caso, solicita unos datos y el recepcionista hace las consultas al sistema, las bases de datos. Si la solicitud y la consulta son correctas, el web service, o recepcionista en su caso, devuelve la información requerida para que el cliente la utilice conforme desee. En el Observatorio de nuevos medios, la respuesta devuelta por el web service incluye tanto la información que tiene origen en la investigación como la que proviene del tratamiento de datos de Twitter. Debido a que las peticiones pueden llegar a ser consultas que requieren un gran consumo de recursos, ya sea por la cantidad de información solicitada o por el propio número de peticiones, se ha optado por mejorar el rendimiento del web service empleando bases de datos en caché. De esta forma, los datos consultados se guardan en memoria durante un tiempo determinado, lo que significa que cualquier usuario que haga la misma consulta recibirá la respuesta de manera inmediata, ya que el web service únicamente tiene que consultar a la memoria guardada y no a las bases de datos "convencionales".

Finalmente, cuando los datos son accesibles por la página web, son los desarrolladores web quienes los representan en elementos visuales tales como mapas y gráficos. Es muy importante aquí la comunicación entre los integrantes que se encargan de los datos y los

responsables de la visualización, ya que el formato en que el web service sirve la información debe ser adecuado o al menos adaptable al formato que requieren las gráficas de la web.

4.2 Investigación e instalación de la tecnología necesaria

4.2.1 Recopilación de información

La fase más compleja y duradera en el desarrollo de este proyecto ha sido, y aún es, la relacionada con la investigación. Antes de comenzar a implementar y de crear un sistema como el del Observatorio, se necesita buscar información sobre la manera más adecuada de conseguir el objetivo. En algunas ocasiones, si no la mayoría, darse cuenta de cuál es la forma correcta de trabajar conlleva probar a hacer los mismos pasos de diferentes modos. Pero esto no ocurre solo al comenzar un proyecto, sino que sucede continuamente durante todas las fases del mismo. En algunas ocasiones, se desconoce el uso de alguna herramienta o de su propia existencia y en otras, no basta aunque ya se conozca y se haya empleado anteriormente porque puede haber sufrido actualizaciones, haber quedado desfasada o simplemente no recordar su manejo, por ejemplo. De ahí que la formación continua sobre las tecnologías existentes y su uso sea parte fundamental de cualquier proyecto que necesite un soporte tecnológico.

Figura 8. La formación continua es vital en proyectos de tecnologías de información (TI).

El primer paso a la hora de recopilar información para cubrir las necesidades técnicas del Observatorio ha sido la búsqueda de servidores donde poder alojar el proyecto. El problema

a resolver aquí ha consistido en prever las características y dimensiones adecuadas para la idea que se tenía en mente, ya que según la capacidad de éstas el precio del servidor aumenta o disminuye. Asimismo, ha sido necesario una formación específica y, en gran medida, autodidacta en materias de gestión de servidores, de herramientas fundamentales para el acceso a los sistemas, de administración de sistema operativo Linux, de paquetes y servicios básicos y esenciales para el correcto funcionamiento del proyecto, etc.

Quizá el asunto que menos problemas ha ocasionado ha sido la elección del lenguaje de programación a emplear para el desarrollo de las herramientas propias: el software de gestión de nuevos medios, el sistema de recolección de datos de redes sociales y el web service. Es cierto que no tiene por qué utilizarse un único lenguaje, pero se ha decidido utilizar Python. El motivo de su empleo viene ligado a las competencias de los desarrolladores, ya que éstos comparten experiencia en su uso, y a la existencia de un framework muy potente escrito en dicho lenguaje. Sobre éste último se hablará en el apartado *Software específico* de esta misma sección, por lo que basta con conocer por ahora que es la herramienta con la que se ha desarrollado el software de gestión de nuevos medios.

Python es un lenguaje muy utilizado y cuenta con un gran número de librerías que facilitan al desarrollador la implementación de código, ya que disponen de una serie de funciones que, al ser llamadas, realizan unas acciones determinadas. La investigación sobre qué librerías son las adecuadas para conseguir lo que se desea es una tarea complicada que implica un tiempo considerable, sobre todo si no se sabe lo que se quiere buscar. De esta forma, se ha explorado en Internet para dar con librerías y tecnologías fundamentales para el desarrollo del proyecto, tales como Tweepy (descargas de Twitter), Geopy (localización de usuarios), Web.py (web service), etc.

Figura 9. Logo del lenguaje de programación Python.

Dado el carácter del proyecto, es indiscutible la necesidad de contar con bases de datos para almacenar la información. Sin embargo, es importante conocer cuáles son las que mejor se adaptan al Observatorio. Se ha decido utilizar una base de datos relacional para guardar la información documental de los nuevos medios. En concreto, se ha hecho uso de MySQL, ya que es uno de los servicios que venían con el sistema operativo Linux Debian. Por otra parte, la elección de la otra base de datos incluida en el proyecto se debe al formato en que se extraen los datos de las redes sociales. Generalmente, éstas devuelven la información en JSON (JavaScript Object Notation), por lo que bases de datos no SQL como MongoDB, que almacenan documentos en representaciones binarias de JSON (BSON) parecen ser apropiadas. También se ha introducido una base de datos llamada Redis que no se dedica a almacenar persistentemente la información, sino a mantenerla durante un tiempo limitado en memoria. Su inclusión en el proyecto se ha propiciado tras investigar el modo más rápido de transferir datos entre un web service y las aplicaciones que hacen uso de éste.

4.2.2 Servidores

El Observatorio es un proyecto web que busca ser accesible por cualquier persona con interés en los nuevos medios. Que sea accesible al mundo implica que el proyecto se aloje en un servidor, de modo que el cibernauta pueda entrar mediante una IP o un dominio asociado a ella. Se ha decidido contratar dos servidores, uno para albergar la solución final y otro donde hacer las pruebas antes de aplicar los cambios en el definitivo y donde guardar las copias de seguridad. Las características de los servidores son diferentes. Como el que se utiliza de pruebas (“servidor de pre” a partir de ahora) es accesible solo por los desarrolladores, no necesita demasiados recursos, únicamente espacio en disco para almacenar las copias de seguridad. El que almacena la solución definitiva (“servidor de producción” a partir de ahora), por su parte, necesita mejores recursos ya que además de almacenar grandes volúmenes de información en las bases de datos tiene que soportar el acceso de todos los visitantes.

A la hora de contratar los servidores, se puede elegir qué sistema operativo instalarles. Se ha optado por un Linux Debian sin interfaz de escritorio, es decir, con acceso por terminal. La elección de Linux en su versión Debian se debe a que es software libre y gratuito, además de que los desarrolladores se sienten a gusto trabajando con esa distribución. Que no

disponga de interfaz de escritorio no se debe solo por un uso familiarizado por parte de los programadores, sino porque su inclusión consume más recursos.

Figura 10. Logo de la distribución Debian de Linux.

A parte del software específico que se define en el apartado siguiente, se han tenido que instalar otros paquetes esenciales como el servidor web Apache o el CMS WordPress, y realizar algunas configuraciones para hacer funcionar el proyecto. Por ejemplo, se han creado demonios o programas en segundo plano que tienen la función de iniciar y parar los desarrollos propios realizados, como el web service o el sistema de extracción de datos. También sirven para iniciar automáticamente estos servicios cuando se produce un reinicio después de una caída del servidor. Asimismo, se ha configurado un firewall o barrera de seguridad para controlar los accesos a los servidores, de modo que se han establecido unas IPs fiables con acceso libre y otras dudosas que se han denegado. Por otra parte, se ha programado una tarea (cron) para que semanalmente se lleven a cabo las copias de seguridad de las bases de datos y se traspasasen al servidor de pre.

Figura 11. Algunos de los paquetes esenciales instalados en los servidores.

4.2.3 Software específico

Django

Django es un framework de desarrollo web de código abierto, escrito totalmente en Python. El principal objetivo de cualquier framework es hacer más sencilla la creación de sitios web complejos. En el caso de Django, se hace especial hincapié en el re-uso, la conectividad y extensibilidad de componentes, el desarrollo rápido y el principio No te repitas (DRY, del inglés Don't Repeat Yourself). Python está presente en todo el framework, incluso en configuraciones, archivos, y modelos de datos.

Al tratarse de un framework escrito en Python, Django permite al desarrollador implementar código de forma ágil. El resultado se refleja en menos líneas de código y, en consecuencia, menos posibilidades de que aparezcan bugs o errores. De ahí que haya expertos que consideren que "fomenta el desarrollo rápido y el diseño limpio y pragmático".

Django también se vincula al diseño MVC (Modelo-Vista-Controlador), por lo que la estructura interna del sitio web está visiblemente ordenada. Como ejemplo, el código orientado a establecer el modelo de datos es completamente independiente al que está destinado al aspecto externo de las páginas. Por otra parte, cuenta con el soporte de una gran comunidad de desarrollo y dispone de una documentación muy completa, incluso para principiantes.

La decisión de emplear un framework como Django en lugar de un CMS (Content Management System) como WordPress o Drupal se debe, en primer lugar, al carácter tan específico que tiene su función, ya que se emplea para la gestión y administración de los nuevos medios almacenados en las bases de datos del Observatorio, y en segundo por la mejor formación de los desarrolladores en el manejo de Python que en PHP, lenguaje predominante en los CMS más conocidos.

Figura 12. Logo del framework Django.

MongoDB

MongoDB (Mongo en adelante) es una de las bases de datos no relacionales de código abierto más conocidas del mercado. Se trata de una base de datos orientada a documentos. Esto significa que en vez de almacenar los datos en registros, como es el caso de las bases de datos relacionales, los guarda en documentos. El formato en que se almacenan estos documentos es en BSON, que consiste en una representación binaria de JSON.

Mongo no sigue un esquema, es decir, los documentos de una misma colección – similar a una tabla convencional – pueden tener diferentes campos, cosa inimaginable en una base de datos relacional. Por otra parte, las consultas se realizan utilizando el lenguaje JavaScript en lugar de SQL. Esto es bastante razonable, ya que Mongo viene por defecto con una consola construida sobre este lenguaje desde la que se realizan las peticiones. No obstante, existen drivers para poder administrar Mongo con el lenguaje deseado. Oficialmente a día de hoy existen drivers para Java, C#, PHP, Python, etc.

Esta base de datos es especialmente útil para desarrollos que necesitan escalabilidad. Incluye opciones de replicación y sharding – partición de los datos de una base de datos según una clave designada por el administrador – muy útiles para balancear la carga de datos entre servidores.

El motivo del empleo de Mongo en este proyecto prácticamente lo han determinado las redes sociales en general. Éstas devuelven los datos en documentos JSON, por lo que la opción más rápida y sencilla es almacenarlos en colecciones dentro de Mongo. Por otro lado, la documentación existente sobre la forma de extraer datos de Twitter suele incluir ejemplos empleando esta base de datos.

Figura 13. Logo de la base de datos MongoDB.

Redis

Redis es un motor de base de datos en memoria y de código abierto que almacena la información en pares de clave-valor. Aunque se puede emplear como una base de datos durable o persitente, generalmente se suele utilizar para guardar datos durante un tiempo limitado. Su popularidad se debe en gran medida a su gran velocidad, ya que conserva la información en memoria, pero también a su flexibilidad y fácil uso.

La presencia de Redis en el proyecto del Observatorio de nuevos medios es esencial para agilizar la transmisión de información entre las bases de datos del sistema y la página web. En la página inicial del Observatorio se hace una petición de unos pocos indicadores sobre cada medio almacenado para mostrar un avance de la información con la que se cuenta: rankings, mapa de localización de medios, etc. Ahora mismo hay más de mil quinientos medios en el repositorio, por lo que la consulta realizada es de un tamaño considerable. Esto significa que, si no se aplican soluciones, el tiempo que tardaría el sistema en resolver la petición sería suficientemente alto como para que el visitante dejara de mostrar interés en la web. Con Redis funcionando, los datos devueltos por la primera consulta se almacenan en memoria, por lo que los siguientes visitantes que entren a la página obtendrán la información de manera inmediata.

Los datos almacenados por Redis suelen tener un tiempo limitado, es decir, pasado un tiempo establecido se borran. Esto tiene sentido, pues si la información guardada varía la función de mantener unos datos "desfasados" no tiene lógica. Por ello, para que ningún usuario sea el que hace la primera consulta (cuando no existen datos en Redis) se ha programado una tarea automática en el Observatorio para que realice la petición diariamente y de madrugada

.

Figura 14. Logo de la base de datos en memoria Redis.

Tweepy

Tweepy es una librería de Python para acceder a la API de Twitter y utilizarla. Es bastante intuitiva y fácil de manejar y existe una gran cantidad de documentación en la web con ejemplos e información acerca de su uso y sus características.

Gracias a las funciones que incluye Tweepy en su core, interactuar con la API de Twitter es más factible. Simplemente se debe analizar con detenimiento la documentación de la propia librería y tener bien claro qué datos se quieren obtener y cómo lograr hacerse con ellos.

Geopy

Si se necesita conocer cuáles son las coordenadas geográficas de alguna dirección, ciudad, país o simplemente un punto en el mapa y se sabe manejar el lenguaje de programación Python, Geopy es la solución. En el caso del Observatorio, se ha empleado para conseguir la ubicación de aquellos usuarios de Twitter que han interactuado de alguna forma con alguno de los nuevos medios almacenados y que no han habilitado el posicionamiento GPS al hacerlo. Sin tener datos sobre coordenadas, la única posibilidad de localizarlos ha sido transformando con Geopy las localizaciones que aparecen en sus perfiles, como “Madrid” o “Valencia”, a coordenadas representables en un mapa.

Web.py

Web.py es un framework web para Python muy simple y potente. Existen otros frameworks como Django o web2py que son más robustos pero en este caso era necesario elegir la forma más rápida de desarrollar un sitio web, de ahí la elección de web.py. El sitio web que se ha desarrollado con web.py es el web service, cuya funcionalidad es simplemente hacer consultas a las bases de datos del sistema, relacionar los datos obtenidos y devolverlos en el formato acordado con los desarrolladores de la página web del Observatorio. Como no es necesario incluir más complejidad, no se ha optado por utilizar Django, por ejemplo.

4.3 Exploración de la API de Twitter

Una gran parte del tiempo que se ha requerido para la creación del proyecto del Observatorio se ha empleado en investigar y aprender el funcionamiento de Twitter, tanto para el usuario general como para los desarrolladores. Conocer cuáles son las formas posibles de interacción en esta red social y qué características propias tienen cada una de ellas han sido los primeros pasos en esta fase de formación.

Twitter dispone de una API (Interfaz de Programación de Aplicaciones) que permite al interesado extraer e introducir información. Está pensada para desarrolladores, de modo que es necesario implementar un mínimo de código para poder empezar a darle uso. Aunque en la actualidad existen herramientas que permiten la obtención de datos de Twitter simplemente introduciendo unos parámetros vía web, como una palabra clave, el nombre de una cuenta, etc., si se precisa de una metodología de extracción particular, como es el caso del Observatorio de nuevos medios, es inevitable realizar un esfuerzo en materia de desarrollo.

Existe infinidad de documentación en la web sobre cómo utilizar Twitter para tratar con sus datos, empezando por la propia página oficial de Twitter para desarrolladores. En ella se explican todas las funcionalidades: las peticiones disponibles, los datos que de ellas se extraen, formato y estructura de la información proporcionada, normas de uso, etc. Asimismo, es de gran utilidad como guía para los primeros pasos, ya que describe detalladamente cómo crear una aplicación (necesaria para utilizar la API) y cómo obtener las credenciales para autenticarse, además de incluir una serie de ejemplos con código predefinido para orientar al usuario sobre la forma en que se utilizan los servicios de la API. Por otra parte, hay también mucha documentación externa a la oficial. Ésta suele ser más especializada, como la explicación del empleo de la API con determinados lenguajes de programación o el modo de extraer de manera continua unos datos en concreto.

La cantidad de datos que es posible obtener a partir de Twitter es, como en la mayoría de las redes sociales, limitada en el tiempo. Twitter concede recopilar su información durante un tiempo determinado, después del cual cierra sus puertas temporalmente. Sin embargo, pasados unos minutos la descarga se reanuda y así sucesivamente. Hay formas de intentar evitar estos parones, como la creación de nuevas aplicaciones y el empleo de varias IPs desde donde realizar extracciones. En el caso del Observatorio, se ha empleado una única

aplicación para utilizar la API. No obstante, se ha optimizado el desarrollo para que, aunque haya limitaciones, las descargas se lleven a cabo como se desea. La clave está en controlar los datos que se obtienen, de forma que cuando se produce un cese temporal el sistema reconoce cuál ha sido el último conjunto de datos descargado para comenzar por el siguiente al reiniciarse la extracción.

Por otro lado, cada conjunto de datos obtenido de Twitter (una publicación) contiene una retahíla de información (figura 15) que no siempre es relevante en su totalidad, es decir, que puede que no todos los campos incluidos sean útiles. Esto generalmente depende del fin para el que se quieren emplear, sin embargo, muchos de ellos figuran por ser campos de control de la propia red social. Por ello, aunque la depuración de datos sea una tarea enrevesada es aconsejable realizarla para conseguir optimizar el espacio que ocupan si al final van a ser almacenados en una base de datos, como en el caso de este proyecto.

Figura 15. Datos asociados a una única publicación en Twitter.

Así pues, la metodología de implantación de la extracción de datos en el proyecto ha consistido, en primer lugar, en documentarse y formarse acerca de las posibilidades que ofrece Twitter y su API. Tras la parte teórica se ha creado una aplicación, imprescindible para acceder a los datos, y se ha comenzado a realizar las primeras peticiones a modo de prueba. Una vez obtenidos los primeros conjuntos de datos, se han analizado y se han seleccionado aquellos campos que se consideran de utilidad. Alguno de éstos incluso han sido empleados como parámetros en otras consultas a la API para obtener mayor volumen de información. Tras tener claro qué peticiones debían realizarse y qué datos se iban a almacenar, ya era posible comenzar con el desarrollo del sistema.

4.4 Desarrollo y puesta en marcha

4.4.1 Implementación en servidor de pruebas

Todas las dependencias necesarias para el desarrollo del Observatorio han sido instaladas primero en el servidor de pre. Éste, al ser accesible solo por los desarrolladores, es idóneo para realizar la implementación de código y hacer las pruebas pertinentes. También es de gran utilidad cuando se busca incorporar alguna nueva funcionalidad y se requiere de un tiempo de control antes de aceptar su integración con el proyecto definitivo. Como ejemplo se tienen los datos que se extraen de Youtube. Al igual que con Twitter, se ha desplegado un "plugin" para el sistema que descarga información de Youtube de manera continua. Esta información aún no se muestra en el Observatorio porque se encuentra todavía en proceso de seguimiento, pero sí se almacena en las bases de datos del servidor de pre para utilizarla cuando sea necesario.

Dado que el servidor de pruebas es inaccesible por el usuario común, ha resultado ser de gran utilidad para la propia formación de los desarrolladores. Sobre todo en materia de gestión de sistemas y redes, transmisión e integración de datos entre aplicaciones, seguridad y control de procesos.

4.4.2 Solución de errores

El método de trabajo a la hora de programar e implementar código sigue una pauta continua. Este patrón consiste en realizar pequeños desarrollos y comprobar, por cada uno de ellos, su funcionamiento. Pocas veces una implementación funciona a la primera y por ello la mayoría de software orientado a la programación dispone de herramientas que facilitan al desarrollador la localización de los posibles errores, indicando generalmente la línea donde se ha producido el problema (figura 16).

Figura 16. Error en Django.

La implementación realizada para el Observatorio cuenta con un sistema de logs por cada elemento del mismo. De tal modo que se listan en ficheros todos los avisos, excepciones y errores que se producen en el software de gestión de nuevos medios, en el sistema de extracción de datos de redes sociales y en el web service. Estos archivos generalmente indican, además de los propios problemas, la hora y fecha en que se producen, por lo que sirven de gran ayuda para el seguimiento de los distintos procesos en acción.

4.4.3 Lanzamiento en servidor de producción

Cuando ya se han llevado a cabo los desarrollos y se han solucionado todos los posibles errores en el servidor de pre, es momento de, por fin, migrar el sistema creado al servidor de producción. Éste sí es accesible por el usuario final, por lo que se debe estar seguro de que todo está preparado para funcionar con normalidad.

El primer paso en la implantación del proyecto en el servidor de producción ha sido instalar sobre éste las mismas dependencias que están presentes en el servidor de pre. Hecho esto, se han migrado todos los desarrollos y todas las configuraciones tal y como están en el de pruebas a excepción de la IP, los dominios y enlaces, que son propios de cada servidor. Tras verificar que el sistema está preparado para ponerse en marcha, se lanzan los procesos que activan tanto la descarga de datos, como el tratamiento y la transmisión de los mismos. A partir de entonces, el sistema debe poder funcionar por sí solo, completando el fichero de logs con cualquier aviso importante.

Como la parte visible del Observatorio de nuevos medios es la página web, ésta está preparada para ponerse en modo mantenimiento cuando se produce algún problema en alguna de las partes del sistema que impide una correcta transmisión de los datos.

4.4.4 Testeo y mantenimiento

Una vez el sistema está en marcha y funciona correctamente, se debe realizar una labor de testeo y mantenimiento continua, ya que el que marche bien en un momento determinado no asegura que lo vaya a hacer en otro más lejano (figura 17). El mantenimiento consiste en revisar los ficheros de logs, en comprobar las últimas descargas de datos, en asegurarse de la creación de copias de seguridad de las bases de datos, etc.

La inserción de nuevos medios al Observatorio es una tarea continua, por lo que también forma parte de la fase de mantenimiento. Así como la modificación y eliminación de los mismos, ya que hay algunos que cesan su actividad o cambian de nombre de cuenta, por ejemplo.

Figura 17. Página de error de la web del Observatorio.

5 Desarrollo

5.1 Configuración de los servidores

El proveedor de alojamiento contratado para disponer de los servidores dedicados ha sido Kimsufi, la filial económica de la marca contrastada OVH. Ofrece servidores baratos que generalmente se usan para formación en sistemas. No tiene el mismo soporte que el que pueden ofrecer otras compañías con productos más caros, pero es una buena opción si lo que se desea es dar unos primeros pasos en materias de administración de servidores.

Al contratar un servidor con Kimsufi, se puede especificar el sistema operativo que se desea emplear. Tras la instalación, el proveedor proporciona los datos de acceso (IP) a la máquina. Como se ha optado por no incluir una interfaz de escritorio, el acceso a la terminal se hace mediante SSH. Una vez dentro, es posible comenzar con la configuración, ya que el sistema operativo en ese momento contiene solo los paquetes y configuraciones por defecto.

El usuario que se proporciona al instaurar el servidor es el "root" o super usuario, es decir, el que tiene permisos de administrador del sistema. Como se requería que los servidores dedicados del Observatorio fueran accesibles por varios usuarios, se ha creado uno por cada integrante que tuviera que hacer uso de él, concediendo más o menos permisos según fuesen las funciones a ejercer. Tras esto, se ha iniciado el proceso de instalación de paquetes, algunos específicos para el proyecto. Como ya se ha mencionado anteriormente, Python es el lenguaje de programación elegido para el desarrollo de las herramientas necesarias para el funcionamiento del Observatorio. Este lenguaje ya viene instalado en la distribución Debian, sin embargo, se precisan varias de sus librerías que no se incluyen por defecto. Para instalarlas, el propio Python dispone de un administrador de paquetes llamado Pip. A continuación se definen algunas de las dependencias más importantes añadidas con este programa:

- Django: Framework para creación de webs complejas.
- Tweepy: librería para acceder y utilizar la API de Twitter.
- Web.py: framework de creación de webs sencillas.
- Pymongo: contiene herramientas para trabajar con MongoDB.
- Redis-py: cliente para interactuar con Redis.
- Geopy: librería para transformación a coordenadas geográficas.

- Python-mysqldb: requerida para conectar Django con MySQL.
- Django_mongodb_engine: precisa para conectar Django con MongoDB.
- Schedule: programación de tareas.
- Pysftp: conexión sftp entre servidores remotos (para las copias de seguridad).

Tras haber incorporado al servidor las librerías imprescindibles de Python para el proyecto, se ha procedido a instalar las bases de datos. Se han necesitado tres: MySQL, MongoDB y Redis, aunque la primera viene casi de la mano con Debian, pues su instalación la sugiere la propia distribución. Del modelo de datos tanto para MySQL como para Mongo se hablará en los apartados *Software de gestión interna* y *Sistema de extracción y explotación de datos* respectivamente. En cuanto a Redis, como se ha empleado para mantener la información durante un tiempo no tiene un modelo muy complicado pero aun así se explicará en el apartado *Transmisión de los datos*.

Por otro lado, se ha programado una tarea semanal en el servidor de producción para realizar las copias de seguridad de las bases de datos de MySQL y Mongo. Para ello se ha implementado un script en Python que realiza los backups de las mismas y las guarda en una carpeta temporal. Tras esto, las comprime y las transfiere vía SFTP a un directorio del servidor de pre destinado a ser su almacén. Con el propósito de no acumular demasiadas copias de seguridad, el propio script comprueba la fecha de las mismas y conserva solo las creadas hace menos de treinta días. Desde el comienzo del desarrollo se ha hecho mucho hincapié en salvaguardar la información, sobre todo una vez que el sistema de extracción de datos de Twitter ha estado en marcha, ya que una pérdida de información en un proyecto como el del Observatorio, que se especializa en datos, puede ser fatal.

Otro elemento básico del proyecto y de los servidores en general es el servidor web. Se ha instalado Apache y se ha configurado para servir las páginas creadas para el Observatorio, que son la propia página web, desarrollada en WordPress y la página de gestión interna, implementada en Django. Para que Apache sea capaz de suministrar el contenido de ésta última debe apuntar al arhivo wsgi.py dentro de Django y habilitar el módulo WSGI (mod_wsgi) si no lo está de fábrica.

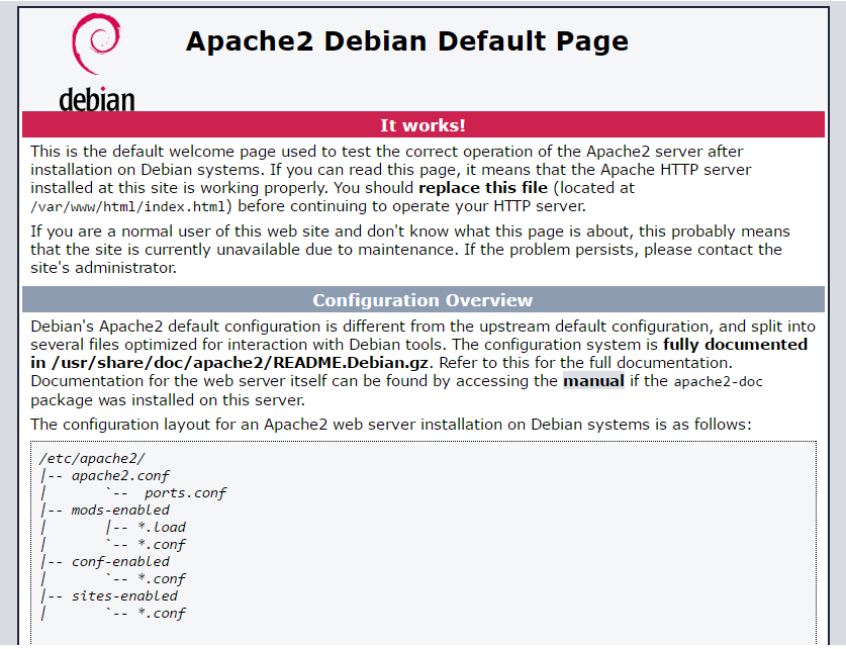

Figura 18. Página por defecto del servidor web Apache en Debian.

Para facilitar el acceso de los usuarios a la página web, se ha comprado un dominio para la IP del servidor de producción: nuevosmedios.es. El servidor de pre, como no está pensado para ser accesible por alguien externo a los integrantes del proyecto del Observatorio, solo es accesible mediante su IP. Por otra parte, la página de gestión interna del servidor de producción también responde al dominio nuevosmedios.es pero mediante el puerto 9000. No obstante, a esta página solo pueden entrar un grupo reducido de IPs. El control de accesos a una determinada IP y a un puerto o puertos en concreto se puede administrar mediante un firewall. Existen muchos para Linux en el mercado, como APF (Advanced Policy Firewall), Shorewall o CSF (ConfigServer Security & Firewall). En el caso de los dos servidores empleados para este proyecto se ha utilizado el CSF, quizás por ser uno de los más conocidos y utilizados.

En la configuración del firewall CSF se pueden establecer qué puertos van a permanecer abiertos y cuáles cerrados. Esto es importante cuando los servicios instalados necesitan comunicarse por algún puerto determinado. Si dicho puerto está cerrado, la comunicación no podrá existir. Asimismo, CSF permite denegar IPs de varias formas. Acepta que se incluyan las propias IPs o el país de origen de las mismas en su lista negra, por ejemplo. Con esto se evita en lo posible que los intrusos “conocidos” puedan ocasionar problemas al servidor.

5.2 Software de gestión interna

5.2.1 Organización interna de la aplicación

Como ya se ha hecho referencia anteriormente, el software de gestión interna se ha desarrollado con Django, el framework escrito en Python que facilita la creación de páginas webs complejas. Tanto es así, que con solo un comando genera ya un proyecto con la estructura incluida (figura 19).

```
                django-admin startproject miproyecto.
miproyecto/
    __init__.py
    manage.py
    settings.py
    urls.py
    wsgi.py
```

Figura 19. Estructura predefinida de un proyecto Django.

A continuación se definen las funciones de cada uno de los archivos generados que se muestran en la figura 19:

- __init__.py. Se trata de un fichero cuyo propósito no es otro que indicar que el directorio en el que se encuentra, *miproyecto* en el ejemplo de la figura, debe ser tratado como una carpeta con paquetes Python. Su contenido puede estar vacío o incluso puede albergar código que inicie algún proceso.

- Manage.py. Es una herramienta de línea de comandos empleada para tareas administrativas, como la sincronización entre el modelo de datos y la base de datos configurada o la ejecución de la consola de desarrollo, un entorno virtual independiente que ofrece Django para probar los desarrollos.

- Settings.py. En este archivo se encuentran todas las configuraciones del proyecto. Se pueden añadir nuevas e incluso modificar las que ya están presentes. Entre éstas,

destacan la que establece las propiedades de la base de datos con la que se quiere sincronizar el modelo o la que fija las rutas de los directorios que almacenan los archivos estáticos, las imágenes y los templates.

- Urls.py. Este fichero contiene las URLs del proyecto, de manera que se especifica qué es lo que se va a mostrar cuando un usuario emplee una de las URLs incluidas. Asimismo, las rutas incorporadas indican las ubicaciones de las apps de Django, que se definen más adelante.

- Wsgi.py. Este archivo es el que permite el despliegue del proyecto Django en un servidor web. Por ello es tan importante habilitar el módulo wsgi. Sin él, Apache no puede lanzarlo para que sea accesible.

En referencia a la consola de desarrollo o entorno virtual independiente, es una herramienta muy útil para ver los progresos en el desarrollo web realizado, ya que dispone incluso de servidor web (figura 20) para comprobar los resultados de la implementación. No obstante, en el proyecto del Observatorio no se ha empleado por contar ya con un servidor donde poder hacer las pruebas. Es por esto que desde el comienzo de la primera implementación de código se ha lanzado el proyecto con el módulo wsgi de Apache.

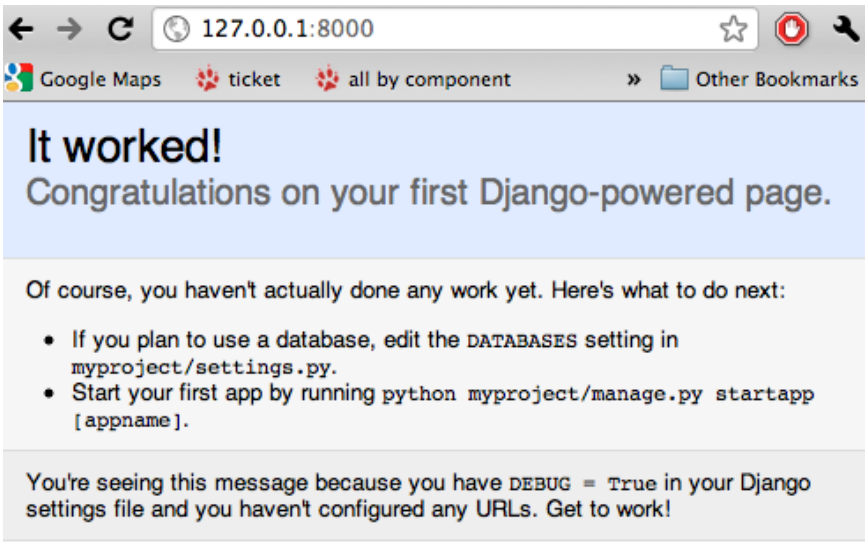

Figura 20. Página índice del servidor web de Django.

La estructura del proyecto Django definida hasta ahora es la básica. Con ella aún no es posible comenzar a ver los resultados en un navegador web. Toda web necesita de, al menos, algo de contenido HTML. Y si, además, se requiere una buena apariencia y una mejor funcionalidad se tienen que establecer estilos y funciones que hagan de la página un proyecto completo. Nada de esto aparece en la estructura mostrada, por lo que falta añadir nuevos elementos.

Un proyecto Django se conforma por un conjunto de aplicaciones más las configuraciones de las mismas. Una aplicación en Django es un conjunto de modelos y vistas de una funcionalidad concreta del proyecto. Por ejemplo, un sistema de comentarios sería una aplicación y una interfaz de administración otra y la combinación de las dos podría constituir un proyecto. Es muy sencillo crear una aplicación con este framework, ya que dispone de un comando que establece automáticamente la estructura básica de la misma. Esto para el desarrollador se traduce en un ahorro considerable de tiempo, ya que no debe ir paso por paso creando la estructura interna. Así pues, para crear una nueva aplicación en Django se debe permanecer en el proyecto deseado y lanzar el comando auto-generador de la misma (figura 21).

```
                python manage.py startapp miaplicacion
miaplicacion/
    __init__.py
    admin.py
    models.py
    tests.py
    views.py
```

Figura 21. Estructura de una aplicación Django.

El conjunto de archivos generado se definen a continuación:

- __init__.py. De la misma forma que aparece este fichero en el momento de crear un proyecto Django, también lo hace en sus aplicaciones. Su función es la misma, es decir, transmitirle al propio framework que el directorio en que se encuentra contiene paquetes indicados para ser interpretados por el lenguaje de programación Python.

- Admin.py. Éste es el administrador de Django. Permite añadir, modificar o eliminar registros de la base de datos. Está creada para que lo empleen los administradores del sitio mediante una interfaz clara y sencilla (figura 22).

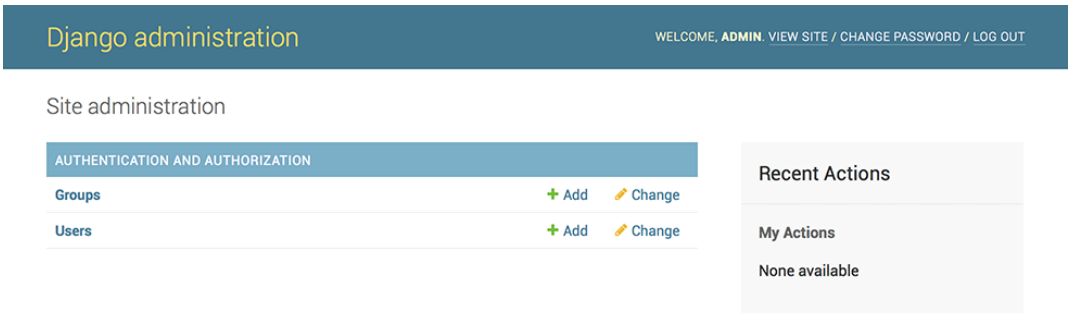

Figura 22. Administración de Django.

- Models.py. Este archivo contiene el modelo de datos de la aplicación, entendiendo este modelo como la estructura de tablas y relaciones que se aplican en la base de datos. Con el archivo manage.py que se ha definido en esta misma sección se puede ejecutar una función que sincroniza las entidades incluidas en models.py con la base de datos establecida en el fichero settings.py, de modo que se crean dichas tablas y relaciones de manera automática. Las propiedades de cada tabla, como los nombres de los campos, el tipo de dato que almacenan, etc. también se establecen en este archivo.

- Test.py. Django recomienda que todo tipo de desarrollo pase por un proceso de testeo antes de nada con la intención de prevenir posibles bugs o errores. Incluye sobre todo casos de pruebas unitarias o unidades de código.

- Views.py. En este archivo es donde se establece la lógica de la aplicación. Contiene métodos de Python que reciben la solicitud web por parte del usuario y devuelve una respuesta en forma de contenido, errores, imágenes, XML, etc. Durante el procesamiento de la solicitud se suelen realizar consultas a la base de datos, regresando la respuesta por medio de plantillas o templates.

Externamente a esta estructura, Django recomienda incluir dos elementos más en cada aplicación generada. Se trata de los archivos forms.py y urls.py. Éstos se explican a continuación:

- Forms.py. Este archivo se emplea para definir los formularios, los cuales permiten la entrada de datos para su procesamiento, bien para crear nuevos registros, modificar los ya existentes o incluso para realizar búsquedas. Se incluyen los diferentes campos estableciendo sus propiedades y reglas de validación, resultando éstas últimas muy útiles para el control de la información introducida.

- Urls.py. La presencia de este fichero en cada aplicación es algo meramente organizativo. Ya existe un urls.py en el directorio raíz del proyecto, sin embargo, para no sobresaturarlo con URLs se recomienda distribuirlas y estructurarlas por aplicaciones, de modo que el archivo urls.py de cada aplicación se ocupa únicamente de las rutas relativas a la aplicación en la que se aloja.

Para que una aplicación Django funcione correctamente, es necesario que los ficheros individuales que se acaban de definir estén relacionados de alguna manera. Por ejemplo, la función realizada por una vista (views.py) requiere la relación entre varios elementos, como los modelos, para consultar información almacenada; los formularios, si se necesita hacer alguna transacción en la base de datos; las URLs para dirigir las rutas o incluir parámetros en ellas y los templates o plantillas HTML para mostrar la solución a una solicitud dada. Para ello, al comienzo de cada fichero se importan los archivos y funciones pertinentes con el propósito de llevar a cabo una conexión apropiada entre los mismos.

Hasta ahora, se ha hecho referencia en numerosas ocaciones a las plantillas HTML o templates, sin embargo aún no se ha descrito su funcionamiento ni su ubicación en un proyecto Django. Según la documentación de este framework, se aconseja que los archivos HTML se alojen en un directorio externo a las aplicaciones, es decir, que no se dividan por funcionalidades del proyecto. La ruta escogida en el caso del Observatorio ha sido la carpeta raíz del proyecto. Aquí se ha creado un directorio *Templates* que organiza los HTML en subcarpetas, una por cada aplicación. Sin embargo, para que Django sepa dónde se encuentran se debe especificar la ruta en el archivo de configuración settings.py.

En cuanto a la estructura de las plantillas HTML, Django cuenta con una funcionalidad que facilita, y de gran manera, las labores del desarrollador. Utiliza un sistema de bloques de contenido (figura 23) que ayuda a organizar el esqueleto de un archivo HTML, de modo que, por ejemplo, se puede crear una plantilla base común para todos los HTML del proyecto e ir añadiendo o quitando bloques según interese. En otras palabras, el contenido dentro de los bloques de este archivo base aparecerá en todos los HTML que lo importen, teniendo la posibilidad de modificarlo en mayor o menor medida según se desee. Esto se ve reflejado luego al navegar por la web, ya que aunque cada página tenga un contenido diferente, sigue un patrón similar con elementos comunes.

```
1  {% extends 'base.html' %}
2
3  {% load i18n %}
4
5  {% block style_css %}
6    <link rel="stylesheet" type="text/css" href="{{STATIC_URL}}css/jquery-ui.css">
7    <link rel="stylesheet" type="text/css" href="{{STATIC_URL}}css/fondoweb.css">
8    <link rel="stylesheet" type="text/css" href="{{STATIC_URL}}css/pestañas.css">
9    <link rel="stylesheet" type="text/css" href="{{STATIC_URL}}css/paginacionmedios.css">
10   <link rel="stylesheet" href="{{ STATIC_URL }}css/busca_medios.css">
11
12 {% endblock %}
13
```

Figura 23. Ejemplo del sistema de bloques HTML en Django.

En la figura anterior, además de apreciarse un bloque de contenido HTML, se puede observar unas etiquetas de importación de hojas de estilos. Estos archivos, junto con los JavaScripts y las imágenes conforman el contenido estático del proyecto. Para organizar los contenidos, se ha creado un directorio para los mismos en la carpeta raíz. Al igual que con los templates, Django necesita que se le especifique la ruta de los estáticos en el archivo de configuración. De esta manera, cuando se quiere importar un estilo, un JavaScript o una imagen se hace con la ruta relativa (STATIC_URL) que se advierte en la figura 23.

5.2.2 Estructura de la página web

La página web creada con Django sigue una estructura sencilla. Se debe recordar que es esto mismo lo que se quiere conseguir, es decir, una herramienta fácil y manejable con la que los integrantes del Observatorio encargados de añadir, modificar y eliminar nuevos medios puedan interactuar con la base de datos a través de una interfaz web intuitiva. Así pues, a continuación se detallan paso a paso las páginas que conforman este software de gestión, sus funcionalidades y los procesos que con ellas se ejecutan.

En primer lugar, el usuario que accede a la herramienta se encuentra con una página inicial que le da la bienvenida (figura 24). En el recuadro del centro, debería aparecer información acerca de cómo registrarse y qué soluciones ofrece la página web una vez el usuario está registrado. No obstante, como se trata de una herramienta interna no se ha puesto demasiado énfasis en este aspecto y aún está por decidir qué contenido se incluirá. Por otra parte, en el panel de navegación localizado en la parte superior se aprecia un sistema de login y una opción para realizar el registro en la aplicación. Ésta última esta deshabilitada por el momento, ya que como los registros de nuevos medios son comunes a todos los integrantes, únicamente es necesario un mismo usuario para todos. Sin embargo, en el caso que se necesitara manejar diferentes conjuntos de datos se podría activar, de modo que el visitante rellenaría unos datos personales (nombre, apellidos, usuario y contraseña) que se almacenarían en la tabla usuarios de la base de datos.

Figura 24. Página de bienvenida al usuario antes de su autenticación.

El sistema de autenticación que aparece en la parte superior derecha de la figura actúa como barrera entre los visitantes a la página y la herramienta que administra la base de datos. De este modo, aunque se habilite la página para ser accesible por cualquier persona, si ésta no está registrada no se le permite el acceso. El proceso que sigue esta funcionalidad para decidir si un usuario entra o no consiste simplemente en una consulta a la tabla usuarios de la base de datos. Si existe el usuario y la contraseña vinculada al mismo coincide con la introducida, el sistema lo acepta y continúa. De lo contrario permanece en la página de bienvenida.

Así pues, una vez el usuario se valida, la página de bienvenida modifica su contenido en algunos aspectos (figura 25). El recuadro incluye ahora otro texto, notificando a la persona correspondiente que el proceso de autenticación se ha realizado correctamente. De igual modo que antes de la validación, el texto debería resumir en pocas palabras las acciones que puede llevar a cabo ahora que el sistema lo ha reconocido. Al final del mismo se incluye el botón *Comenzar* en azul con el que se accede a la herramienta en sí, es decir, la que permite administrar la base de datos.

Figura 25. Página de inicio tras la autenticación del usuario.

Por otro lado, el sistema de autenticación ha desaparecido y ahora se muestra el nombre del usuario validado junto con un botón *Salir* en verde. Al hacer click en éste, el sistema vuelve directamente a la página de la figura 24, esperando de nuevo que el usuario se autentique. Esta funcionalidad permanece en todas las páginas mientras el usuario siga validado.

Haciendo referencia a las aplicaciones Django, las páginas de bienvenida y el sistema de autenticación vistos conforman una de ellas.

Inmediatamente después de pulsar el botón *Comenzar* de la figura 25, se muestra una nueva página (figura 26) con varias pestañas, una de ellas activada. Se trata de la pestaña *Buscar*, la cual permite consultar nuevos medios a la base de datos utilizando el título (o parte de él) de los mismos (figura 27).

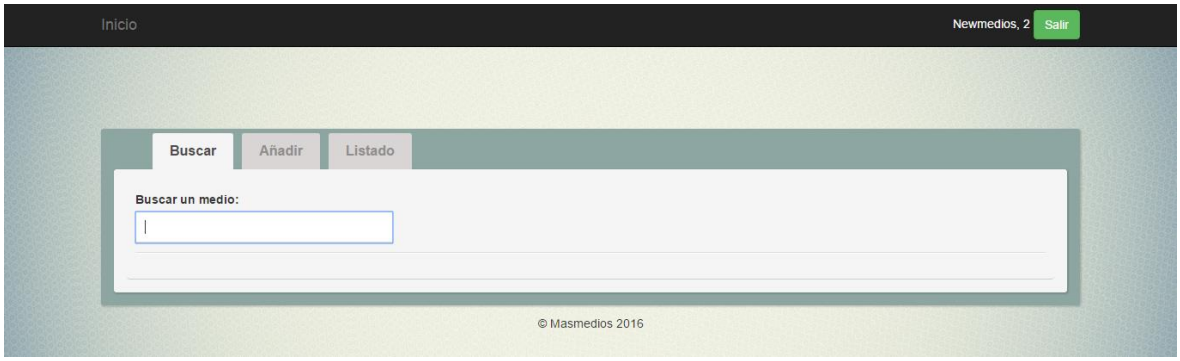

Figura 26.Página principal de la herramienta de gestión.

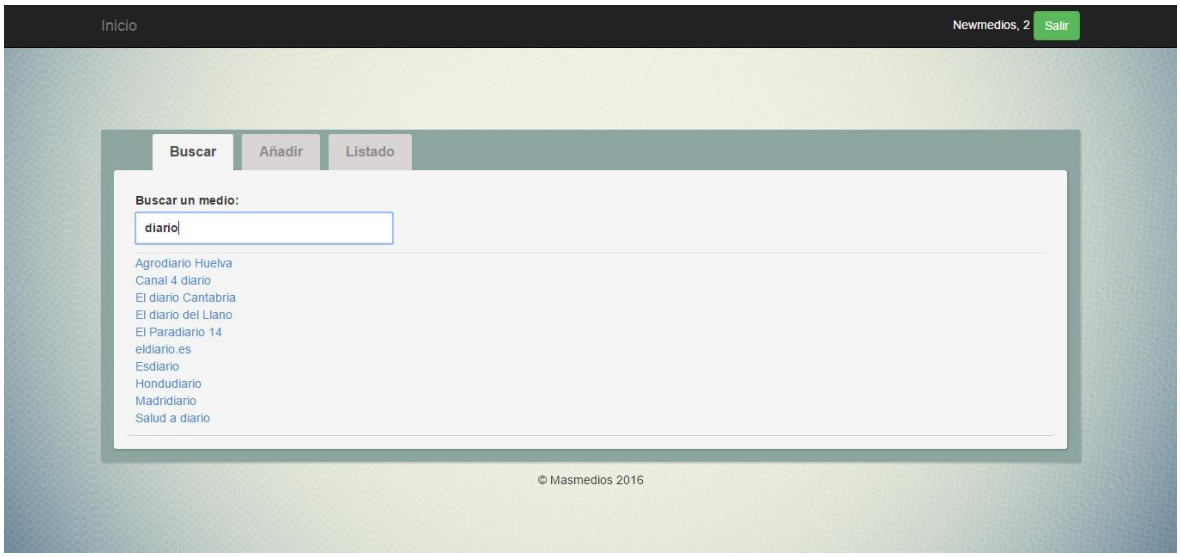

Figura 27. Funcionalidad de la pestaña Buscar.

El panel de búsqueda es un formulario de Django. El contenido que se le introduce se transmite a la vista, la cual realiza la petición a la base de datos a partir del modelo especificado. La consulta se hace a la tabla que contiene todos los registros de nuevos

medios, concretamente al campo título y al campo id, y busca todos los registros que contengan el texto insertado en la búsqueda. Como resultado aparece un listado de links con los títulos que encajan con dicho texto. Todo este proceso se realiza de forma dinámica gracias a AJAX (JavaScript Asíncrono y XML).

La construcción de los links que aparecen como resultado de la consulta a la base de datos se lleva a cabo con los ids obtenidos de la propia petición. Se ha especificado una URL en Django a la que se le pasa como parámetro esos ids, de forma que al pulsar sobre cualquiera de las que aparecen en el listado, el sistema redirige a la página de modificación del medio seleccionado (figura 28). En ella aparece un formulario con todos los campos relacionados con un nuevo medio, sin embargo, alguno de los mismos ya están completados. Esto es importante para conocer qué información está presente en la base de datos sobre ese medio en concreto. Así pues, si se desea modificar algún dato o completar campos vacíos simplemente se deja rellenada la ficha como se requiere y se pulsa sobre el botón *Modificar medio* situado en la esquina inferior izquierda del formulario. Seguidamente, la herramienta redirige de nuevo hacia la página principal con la pestaña *Buscar* activada.

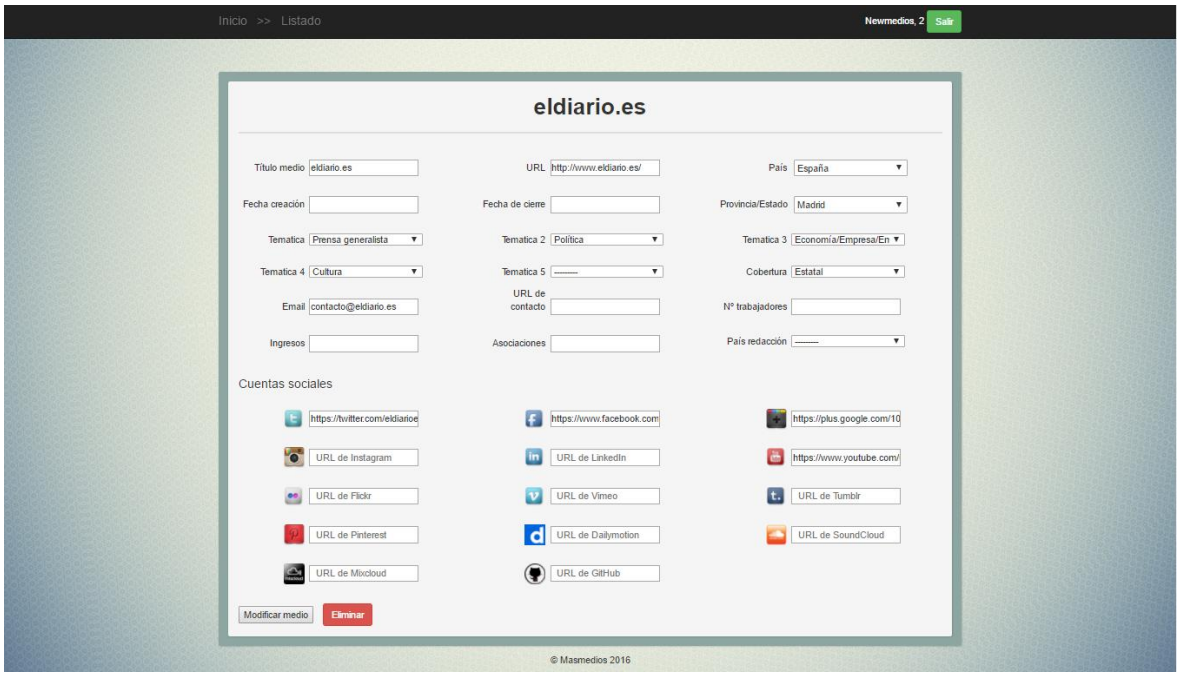

Figura 28. Página de modificación de un nuevo medio.

Si se observa la figura 28, destaca también el botón *Eliminar* rojo situado justo a la derecha del de *Modificar medio*. Hay que tener especial cuidado con este botón, ya que permite

suprimir el medio de la base de datos. Pero no solo lo elimina de la base de datos MySQL donde se encuentra la información documental, sino que también se liquidan todos los datos que se hayan podido descargar de las redes sociales sobre ese medio en particular. Para evitar equivocaciones y grandes sustos se ha añadido una doble confirmación, de manera que al pulsar sobre el botón *Eliminar* del formulario de modificación aparece una ventana emergente o pop-up (figura 29) donde se requiere una confirmación de la acción. De nuevo, si se decide eliminar el medio, la herramienta redirige a la página principal con la pestaña *Buscar* activada.

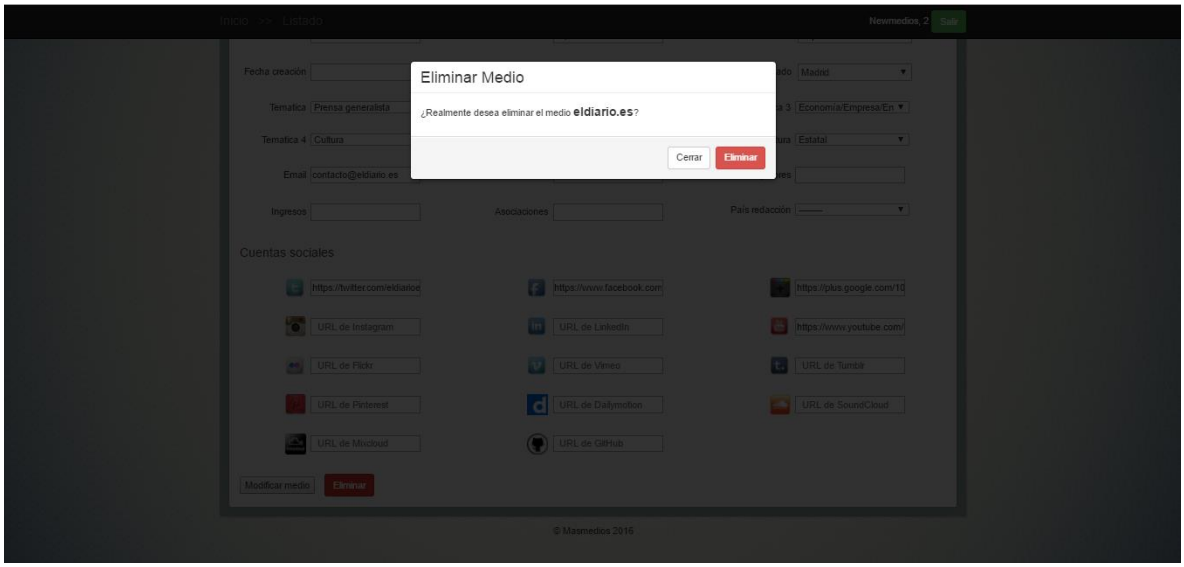

Figura 29. Ventana emergente para confirmación de la eliminación de un medio.

La siguiente pestaña a tener en cuenta en la página principal es la de *Añadir*. Si se pulsa sobre ella, inmediatamente cambia el contenido de la página. Esto se logra utilizando JavaScript, concretamente su biblioteca JQuery. En esta sección se encuentra todo lo necesario para realizar inserciones en la base de datos. El contenido se divide en dos partes, una opción para importar un fichero en formato CSV (figura 30) y un formulario igual al de la página de modificación pero esta vez con todos los campos vacíos. La opción del fichero permite realizar una inserción masiva de nuevos medios en la base de datos. De esta forma, se evita tener que introducir manualmente y uno a uno cada registro. Para evitar que se introduzcan datos erróneos, esta utilidad realiza una comprobación de los datos, de modo que si detecta algún problema la transacción no se lleva a cabo. Asimismo, aparece una ventana emergente indicando en qué línea del archivo seleccionado se encuentra el error.

Figura 30. Opción de añadir un fichero CSV para inserción masiva de datos.

Se ha elaborado una plantilla CSV para que los integrantes encargados de la labor de investigación puedan completar la información de los nuevos medios tal y como el sistema los acepta. No obstante, se hará referencia a esto en el siguiente apartado.

Siguiendo con el contenido de la pestaña *Añadir*, hay un formulario (figura 31) que permite la introducción de los nuevos medios uno a uno. Este formulario contiene todos los campos de información que tienen vinculados los medios en base de datos. Por ello, a continuación se hace una breve definición de cada uno de ellos:

- Título medio: nombre por el que es conocido el nuevo medio. Obligatorio.
- URL: enlace a la página oficial del medio en cuestión. Obligatorio.
- País: país de origen del medio. Obligatorio.
- Provincia/estado: provincia o estado de origen del medio. Depende del campo País, de forma que su contenido varía dinámicamente. Por ejemplo, si se completa España en el campo País, en el de Provincia solo se pueden seleccionar provincias españolas y no de otra nación. Obligatorio.
- Fecha de creación: fecha en el que nace el nuevo medio. No confundir con la fecha de creación de cuentas en redes sociales.
- Fecha de cierre: fecha en el que un nuevo medio cesa sus funciones.
- Temáticas: temática o temáticas sobre las que tratan los contenidos del medio. Este campo puede contener varias opciones.
- Cobertura: ámbito geográfico de los contenidos del medio: local, internacional, etc.
- Email y URL de contacto: información para contactar con el medio.
- Número de trabajadores, ingresos y asociaciones: información extra sobre el entorno del medio.

- País redacción: país desde donde redacta un medio. No confundir con país de origen. Un nuevo medio puede ser cubano, por ejemplo, pero redactar desde Estados Unidos.
- Cuentas sociales: se trata de las URLs de las cuentas del medio en las redes sociales.

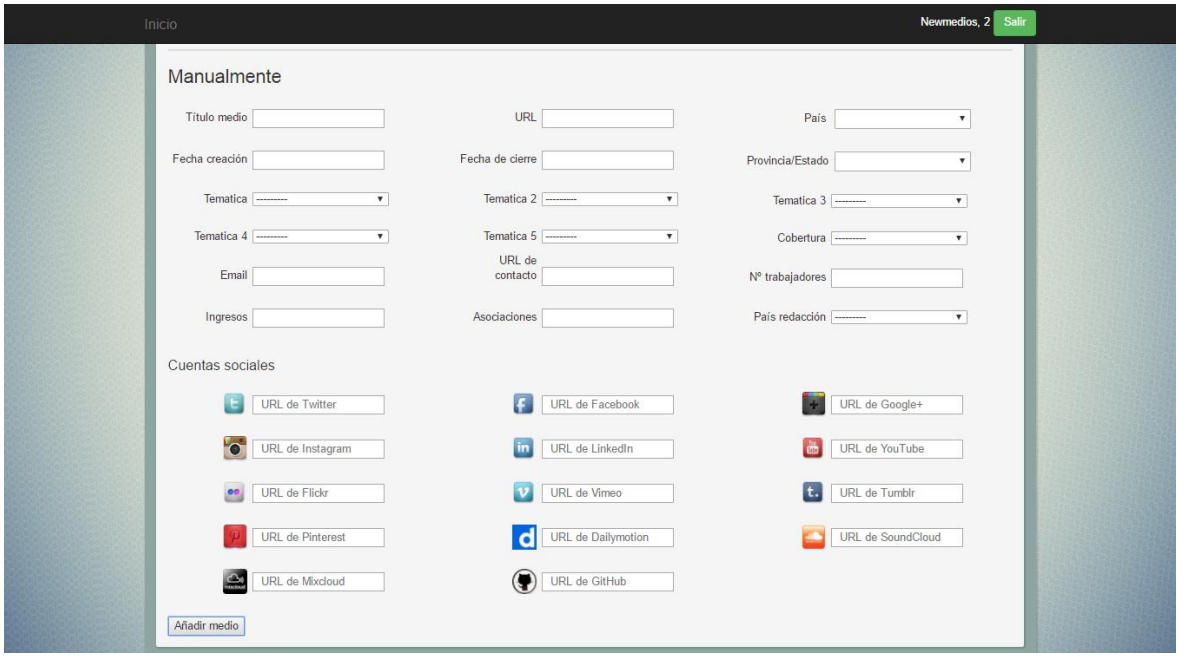

Figura 31. Formulario de inserción manual de un nuevo medio.

En la definición de los campos del formulario se ha hecho referencia al carácter obligatorio de alguno de ellos. Éstos son el título, la URL, el país y la provincia o estado. Si no se completan estos datos, al pulsar sobre el botón *Añadir medio* situado en la parte inferior el sistema detecta que se han incumplido las reglas de validación y no deja llevar a cabo la transacción. Además, aparece una ventana emergente indicando los campos obligatorios (figura 32).

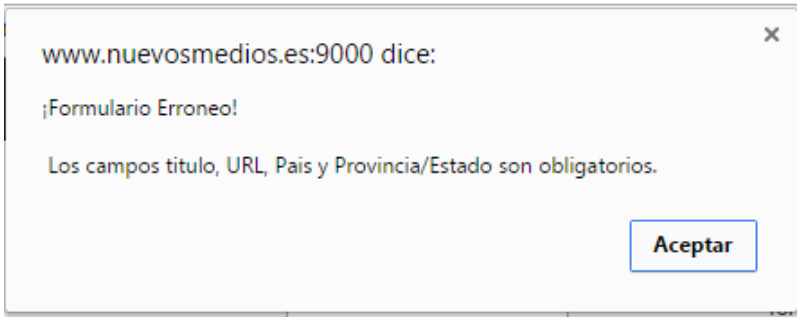

Figura 32. Ventana emergente indicando los campos obligatorios en el formulario de inserción.

Si se completan apropiadamente los campos del formulario de inserción de medios, la herramienta redirige a la página principal. Donde todavía aparece una pestaña *Listado* de la cual aún no se ha hablado. El origen de esta pestaña es anterior a la de *Buscar* y tras la incorporación de ésta última su empleo ha bajado considerablemente. Se trata de un listado de los nuevos medios presentes en la base de datos ordenado alfabéticamente (figura 33) que incluye unos pocos datos sobre los medios y dos acciones: editar y eliminar. Si se pulsa el botón *editar*, la herramienta redirige a la página de modificación de un medio mostrada en la figura 28 mientras que si se hace click en el de *eliminar*, aparece directamente la ventana emergente de confirmación de la acción. La ventaja de este botón reside en lo dinámico de su proceso, pues utilizando AJAX se logra que, sin necesidad de refrescar la página, desaparezca inmediatamente el nuevo medio del listado.

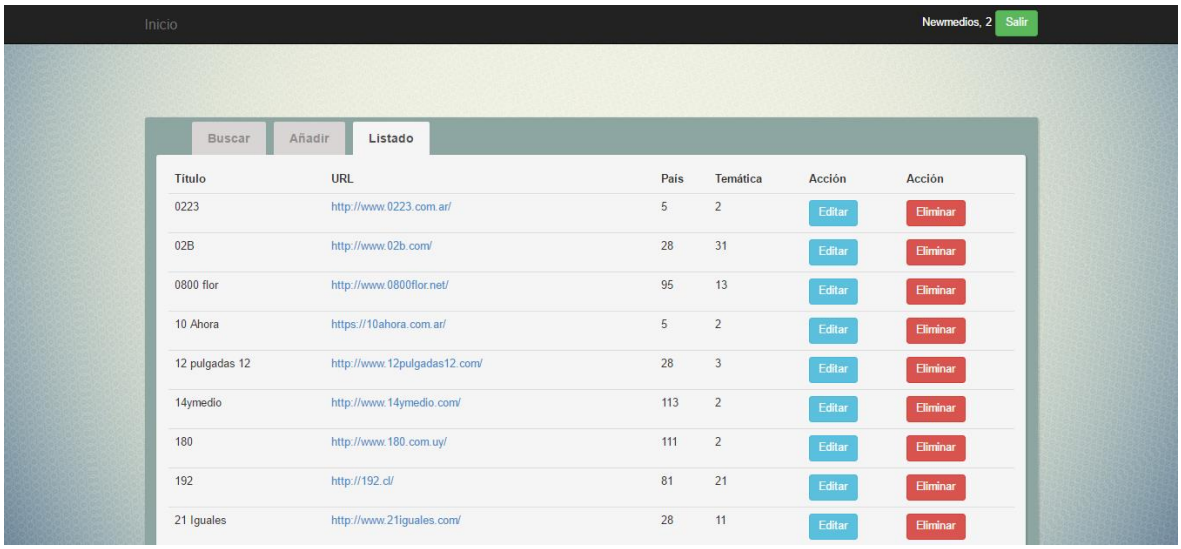

Figura 33. Contenido de la pestaña Listado.

Debido a que la base de datos cuenta ahora mismo con más de mil quinientos registros de medios, ha sido necesario paginar de algún modo el listado, ya que si no la página no tendría fin. De esta forma, mediante JavaScript se ha desarrollado una función que muestra 25 resultados por cada índice del paginador (figura 34).

Figura 34. Paginador en la pestaña Listado.

5.2.3 Plantilla y modelo de datos

El modelo de datos que se ha definido con Django es el presente en MySQL. De hecho, las tablas, sus atributos y relaciones se establecen en el archivo models de las aplicaciones del propio framework y se crean tras hacer una sincronización entre éstas y la base de datos especificada.

Ya se ha hecho referencia a los campos que incluye el registro de un nuevo medio en los formularios del apartado anterior y se ha podido observar que existen algunos que se rellenan con información de países, temáticas, coberturas y provincias. Estos datos están presentes en diferentes tablas y se relacionan entre sí por medio de campos de control. A continuación se detallan los atributos y relaciones entre tablas:

- Tabla países: se trata de una tabla con tres campos, el id, el nombre del país y su código ISO. El contenido de esta tabla se obtuvo mediante la búsqueda en Internet, donde se pueden encontrar archivos SQL con los atributos mencionados de todos los países del mundo. El código ISO ha resultado muy útil para emplear APIs de obtención de coordenadas de países. El id es clave principal de la tabla países y clave foránea en la tabla medios y provincias.

- Tabla provincias: la información de todas las provincias y estados del mundo se ha extraído de la misma fuente que la de países, por lo que la relación de estas dos tablas venía establecida desde el principio. Contiene tres campos: el id de provincia, el id del país al que pertenece y el nombre propio. De esta manera, el id de provincia es la clave principal y el de país la foránea que apunta a la tabla países. Asimismo, el id de provincia es clave foránea de la tabla medios.

- Tabla temáticas: esta tabla tiene simplemente dos campos, el id y el nombre de la temática. Los datos incluidos son propios y no definitivos, ya que pueden surgir nuevos temas. El id es la clave principal de la misma tabla y foránea en la de medios.

- Tabla coberturas: también de elaboración propia. Contiene un id y el nombre del ámbito geográfico. El id es clave principal de la tabla y foránea en la tabla medios.

- Tabla medios: los campos de la tabla medios son los que se han definido al explicar los formularios de la herramienta de gestión. Contiene un id que es clave principal y

que representa un elemento fundamental para relacionar los datos que proceden de redes sociales con los que se introducen con la herramienta. Asimismo, incluye claves foráneas apuntando a las tablas que se han detallado antes.

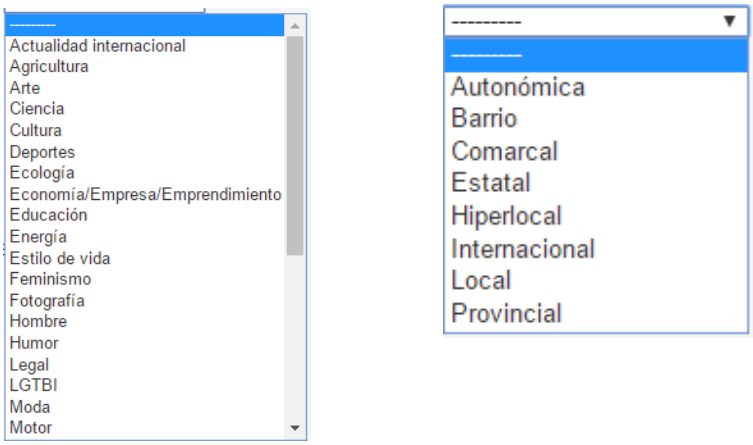

Figura 35. Información contenida en base de datos de temáticas (izq.) y coberturas (dcha).

Teniendo en cuenta el modelo de datos definido, se debe ser muy cuidadoso a la hora de elaborar la ficha o plantilla de importación de nuevos medios, ya que hay datos que deben coincidir exactamente con lo que existe en la base de datos. Los campos de la plantilla son los mismos que están presentes en la tabla medios, sin embargo, mientras los campos con clave foránea se completan con ids en base de datos, en la ficha se completan con los nombres de países, por ejemplo. Sería muy confuso rellenar manualmente en la ficha un campo de país con su id en vez de su nombre y muy poco fiable si se introdujera directamente, ya que tendría que ser exactamente el mismo texto que el que se almacena en el campo nombre de la tabla paises. Para solucionar esto, se han implantado unos desplegables (figura 36) con los nombres exactos de los países, provincias, temáticas y coberturas en base de datos. De esta forma, al importar el fichero CSV el sistema coge el valor España (por ejemplo) del campo país y lo busca en el campo nombre de la tabla países. Como coinciden exactamente, se obtiene el id de España y queda almacenado en la tabla medios. De igual modo se procede con provincias, coberturas y temáticas.

Por otra parte, para evitar otros posibles errores se ha definido el campo provincia del CSV como dependiente del campo país. De este modo se consigue la misma funcionalidad que en el formulario de Django, es decir, el campo provincia tiene un contenido dinámico que varía en función del país seleccionado.

Figura 36. Desplegables en la plantilla de nuevos medios.

5.3 Sistema de extracción y explotación de datos

5.3.1 Requisitos y condiciones

Antes de empezar a desarrollar código, lo primero que se ha llevado a cabo es una investigación de los requisitos que exige Twitter para poder utilizar su API. El primero y más lógico es que el usuario o los usuarios que vayan a extraer datos de la red social estén registrados, es decir, que dispongan de una cuenta de Twitter. No obstante, éste no es el único. Twitter solo proporciona datos a aquellos usuarios que se autentican como desarrolladores y, para ello, ha habilitado una sección solo para éstos (figura 37).

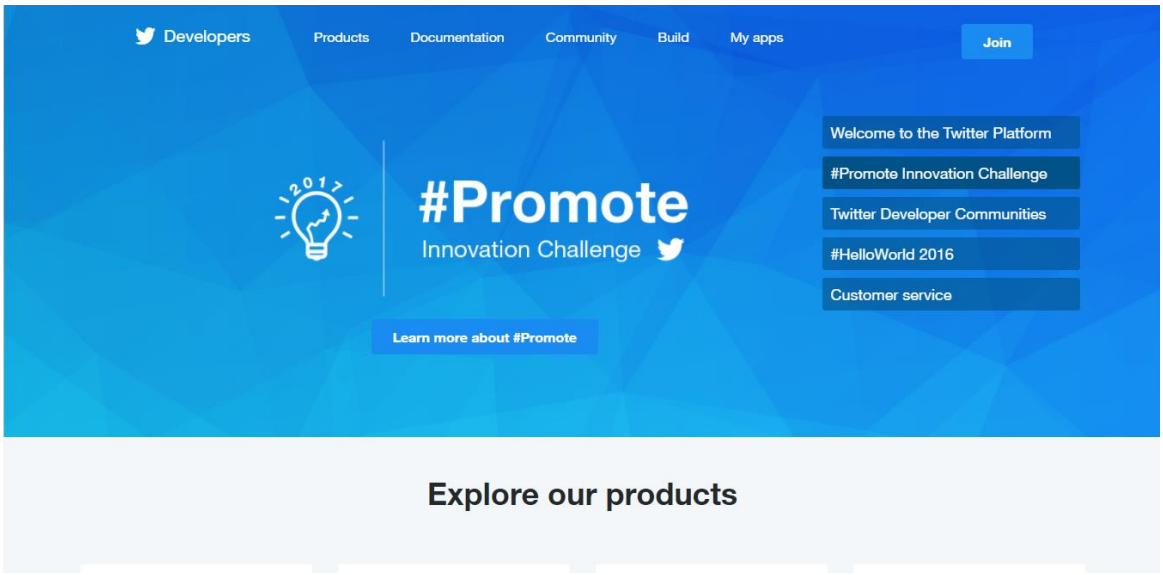

Figura 37. Página de desarrolladores de Twitter.

En la página de desarrolladores de Twitter se encuentra todo tipo de información sobre cómo interactuar con su API, tanto para extraer datos como para introducirlos. También incluye un apartado para aplicaciones, ya que Twitter solo permite realizar acciones con su API a través de ellas. Para acceder a esta sección, previamente se debe estar autenticado como usuario de la red social.

Una vez dentro, si aún no se ha creado ninguna aplicación, la página estará vacía y solo contendrá una opción para crear una nueva. Al pulsar sobre esta opción, el usuario debe introducir datos sobre el nombre de la aplicación, una descripción y una página web donde encontrar información al respecto. Tras haber leído las condiciones y estar de acuerdo con ellas, es posible finalizar la creación de la nueva aplicación. De esta forma, la página volverá a la sección de aplicaciones. Sin embargo, ahora aparecerá la que se acaba de crear (figura 38).

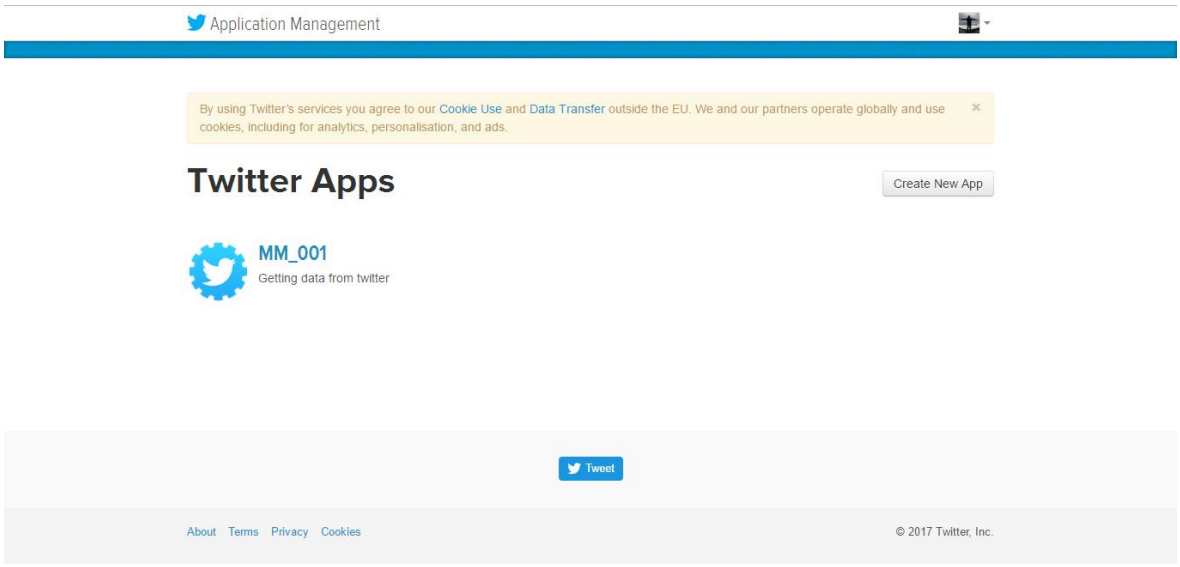

Figura 38. Sección de aplicaciones de Twitter.

Dentro de la aplicación creada están los detalles de la misma. Los datos más importantes para el desarrollador de aplicaciones se encuentran dentro de la pestaña *Keys and Access Tokens.* Aquí se hallan las claves que permiten autenticarse al usuario para poder acceder a la API de Twitter. Concretamente, se trata del *Consumer Key* y del *Consumer Secret* y del *Access Token* y el *Access Token Secret*.

El siguiente paso en el camino hacia la interacción con la API de Twitter ha consistido en efectuar varias pruebas de descarga de tuits utilizando para ello diferentes librerías de Python. Finalmente se ha elegido Tweepy por resultar más fácil su uso y por contar previsiblemente con más documentación y ejemplos que otras. Durante las pruebas realizadas, se ha podido comprobar que el flujo de extracción de datos no es continuo, sino que Twitter establece unos *rate limits* (límites de tarifa) que dependen del tipo de petición (figura 39), de modo que se permite la descarga de un número limitado de tuits en intervalos de quince minutos.

Tipo	Tweepy	Familia	Request per 15 m	Observaciones
GET statuses/user_timeline	api.user_timeline	statuses	300	Con 10 llega a 300 peticiones más rápido.
GET statuses/retweets/:id	api.retweets(id)	statuses	60	
GET search/tweets	api.search()	search	450	Búsqueda con mayor capacidad de peticiones.

Figura 39. Tabla de peticiones a Twitter y límites.

Por otra parte, se debe tener en cuenta las respuestas que se pueden recibir de Twitter, ya que una variación en las mismas puede originar un problema y afectar gravemente al proceso de descarga. No todas las consultas se resuelven de la misma forma. Se pueden producir errores de diferentes características, de hecho se puede dar el caso en que la misma petición efectuada para dos medios diferentes resulte correcta para uno y problemática para el otro. Esto se puede deber, por ejemplo, a que un usuario de Twitter, la cuenta de un nuevo medio en este caso, puede hacer privados algunos elementos que otros tienen públicos. Aunque en la documentación oficial se puede encontrar una información más detallada sobre las posibles respuestas de error de Twitter, en la figura 40 se especifican aquellas que se han encontrado durante el proceso de extracción de este proyecto.

Tipo	Texto	Observaciones
401	Unauthorized	No permite consultar dicha cuenta
404	Not found	No encuentra cuenta

Figura 40. Tabla con los errores de Twitter más frecuentes en el Observatorio.

Una vez se han hecho las pruebas pertinentes y se ha comprendido mejor el funcionamiento de la API de Twitter, se ha procedido a estudiar la forma de llevar a cabo una descarga continua y autosuficiente. En los siguientes apartados se detallan los elementos y los procesos que se han desarrollado con la intención de crear un sistema capaz de conseguir ese objetivo.

5.3.2 Organización interna del sistema

La estructura del sistema de extracción de datos de redes sociales o kernel está formada por tres partes distintas y relacionadas entre sí: el daemon o demonio, el administrador de plugins y los propios plugins. A continuación se definen de manera específica las características de cada uno de ellos:

Demonio

El demonio consiste en un script de Python con el que se realizan las instancias que activan los procesos del sistema. En él se detallan algunas funciones como la que crea y pone en marcha los hilos o threads, que permiten ejecutar distintos procesos a la vez compartiendo el mismo espacio de memoria; la que define las horas de ejecución de dichos hilos; la que establece un directorio y unos archivos dedicados a almacenar los logs y las que activan los procesos de Twitter, tanto la descarga como el cálculo de métricas.

Con motivo organizativo y con la intención de abstraer el demonio de los datos específicos de este proyecto, se ha creado un archivo de configuración, de modo que cuando el daemon se inicia, éste lo importa para llevar a cabo las instancias conforme se haya establecido.

El archivo de configuración engloba varios elementos fundamentales, como son los paths o rutas donde se encuentra el administrador de plugins y los mismos plugins, las claves de

autenticación de la aplicación de Twitter para utilizar su API, las propias de Youtube que, aunque aún no se han incorporado los datos al proyecto, el proceso para su extracción sigue funcionando, el Schedule o las horas a las que se inician los procesos de descarga y de cálculo de datos, los parámetros de conexión a la base de datos MySQL y MongoDB y la configuración específica de los logs, como el nombre de los ficheros o el formato de los avisos.

En cuanto al horario programado para iniciar los procesos, se han establecido hasta tres horas distintas para la descarga de datos de Twitter, concretamente a las 4:00, las 14:00 y las 22:00. Youtube, por su parte, solo cuenta con una descarga diaria y se produce a las 14:00. Esta diferencia en la frecuencia de activación de procesos de una u otra red social viene dada por el comportamiento de las mismas, pues mientras que en Twitter la publicación de tuits es muy elevada en el tiempo, en Youtube la difusión de videos no lo es tanto, al menos a gran escala. Asimismo, también se lleva a cabo un proceso diario encargado de calcular las métricas de Twitter para los medios incluidos en la base de datos. Éste comienza a las 23 horas.

Administrador de plugins

Del mismo modo que se ha importado desde el daemon el archivo de configuración, éste llama al administrador de plugins cuando se inicia el proceso de descarga de datos de Twitter y de Youtube. Cuando se habla de plugins, se hace referencia a las redes sociales que se tienen en cuenta para la extracción de datos. De esta forma, Twitter es un plugin, Youtube es otro y cualquier otra red social que se incorpore conformaría uno nuevo. El kernel se ha estructurado así para evitar desarrollos múltiples, es decir, eludir en lo posible la recreación de un mismo código para cada red social. De la manera en que está organizado, si se quiere añadir una nueva fuente de datos simplemente habría que agregar una función nueva de llamada en el administrador de plugins y un archivo independiente con el código para la descarga.

El administrador de plugins define una clase con los atributos de los plugins. Su actividad principal consiste en realizar la consulta de todos los registros de la base de datos MySQL donde se encuentra la información documental de los nuevos medios y llamar a las diferentes funciones de descarga y explotación de datos incluidas en los propios plugins. En concreto para Twitter, se llama a las funciones del plugin que hacen las peticiones a la API para

obtener los tuits publicados por los propios medios, las menciones que éstos reciben, la información de sus seguidores y las listas en las que participan. A parte de éstas, también se llama a la función de cálculo de índices o métricas.

Para llevar a cabo las conexiones a las bases de datos y establecer funciones útiles para las transacciones a realizar, se ha creado un módulo independiente del administrador de plugins y de los mismos plugins. Este módulo, llamado DAO (Objeto de Acceso a Datos), contiene varios archivos. Uno de ellos define el modelo de datos utilizado en MySQL, el cual es importado por otro archivo encargado de realizar la conexión con la base de datos relacional y de definir funciones para la manipulación de sus datos, tales como la que efectúa la búsqueda de las URLs de las cuentas de Twitter de los medios o la que a partir de esas URLs extrae el screenname o nombre de cuenta para ser empleado en las peticiones a la API. Por otro lado, se encuentra el archivo que maneja la conexión con la base de datos Mongo e incluye algunas funcionalidades como la creación de colecciones, la búsqueda de uno o varios documentos, consultas de comprobación, búsqueda del primer y último elemento almacenado en una colección, contador de coincidencias, etc.

Plugins

El último elemento del kernel es el de los plugins. Dentro del plugin de Twitter existen dos archivos, uno dirigido a la interacción con su API y otro al cálculo de métricas. El primero define una clase para la conexión con la API. Para ello, utiliza la librería Tweepy y las credenciales que figuran en el archivo de configuración mencionado antes. Asimismo, se definen una serie de excepciones por si la conexión falla o se produce algún problema, en cuyo caso quedaría reflejado en los logs. A su vez, se inicia la conexión con Mongo y se establecen las colecciones que van a salvaguardar las respuestas de las consultas. Sobre éstas se hará mención en el apartado *Modelo de datos* de esta misma sección. Seguidamente, se recuperan los campos requeridos de los medios en MySQL, como el id (que es el campo que relaciona un medio en las dos bases de datos empleadas) o el nombre de la cuenta de Twitter, necesario para hacer las consultas a la API y se realizan las peticiones. Para que el proceso de extracción sea eficiente y no duplique información ya existente, se han incorporado una serie de comprobaciones. De éstas y de todos los procesos relacionados con la extracción de datos se hablará con precisión en el apartado siguiente.

El segundo archivo, dedicado al cálculo de métricas, contiene una función por cada una de ellas. Asimismo, también contiene otras como la que junta todas en un mismo JSON o la que consulta a Mongo los datos almacenados en el último mes. Sobre la obtención de estas métricas se hará especial hincapié en el apartado *Explotación de datos* de esta sección.

5.3.3 Extracción y depuración de datos

Como ya se ha hecho referencia, la extracción y depuración de los datos se lleva a cabo en uno de los archivos de la parte del plugin. Dentro del mismo, tras la conexión a la API de Twitter y a la base de datos Mongo y después de tomar los campos id y la cuenta de Twitter de los medios, se procede a realizar la primera consulta, la que extrae los tuits del timeline de los medios. El timeline de Twitter es la página principal del usuario, aquella en la que aparecen sus últimas publicaciones. Con la intención de evitar que una y otra vez se descarguen los mismos datos, se llevan a cabo una serie de acciones. En primer lugar, se busca en Mongo el id del último tuit almacenado sobre el medio que se va a consultar. Si no existe aún un id, se realiza la consulta sin más, ya que significa que aún no existen tuits almacenados para ese medio. La primera búsqueda (figura 41) se hace para un total de 20 ítems o tuits, aunque este valor se puede cambiar en el archivo de configuración. Por el contrario, si sí que existe quiere decir que ha habido ya una consulta previa a ese medio, por lo que se debe tener especial cuidado en no almacenar información duplicada. Para ello, se ha establecido que, si en la consulta a Twitter existe un tuit con un id que coincida con el último insertado en Mongo, la petición se reanude fijando el parámetro *since_id* de Tweepy (figura 42), el cual modifica la consulta para extraer aquellos tuits más actuales que el id especificado. Por otro lado, si los tuits extraídos no contienen un id que coincida con el último guardado de Mongo, se realizan nuevas consultas hasta que aparezca.

```
# primera busqueda, sera limitada
statuses = tweepy.Cursor(self.__api.user_timeline, id=cuenta).items(self.__first_down)
```

Figura 41. Primera consulta sobre el timeline de un medio a la API de Twitter.

```
statuses = tweepy.Cursor(self.__api.user_timeline, since_id=last_id, q=query).items()
```

Figura 42. Consulta al timeline de un medio utilizando el parámetro since_id.

Cuando en una consulta se obtiene un tuit con el mismo id que el último almacenado de Mongo, no solo se guardan aquellos más actuales, sino que también se actualizan los que ya existían previamente en la base de datos. Para ello se emplea el parámetro *max_id* en las consultas (figura 43), de modo que se obtienen tuits más antiguos que el id especificado (último tuit en Mongo). Es necesario poner un límite a esta petición, ya que si no podría estar descargando datos de un medio hasta llegar a su primera publicación, así que se recuperan y actualizan aquellos tuits que tienen una fecha de creación inferior a un día desde la fecha en que se realiza la consulta.

```
old_statuses = tweepy.Cursor(self.__api.user_timeline, id=cuenta, \
                             max_id=last_inserted_id).items()
```

Figura 43. Consulta sobre el timeline de un medio con el parámetro max_id.

Los tuits resultantes de la petición a Twitter sobre el timeline de un medio se almacenan en dos colecciones diferentes: Tweet_medio y Twitter_main_user. Una se dedica a guardar documentos sobre los tuits que aparecen en el timeline de los medios, es decir, sus últimas publicaciones y la otra se ocupa de la información del propio medio, ya que en los mismos tuits del timeline se recogen también estos datos. No obstante, para que no se almacene un documento con información del medio por cada publicación del timeline, se ha puesto un límite de un recurso al día.

No todos los metadatos de los tuits extraídos a partir de la consulta del timeline son útiles para el proyecto. De hecho, con la intención de optimizar espacio, se realiza una depuración de los mismos previa a su almacenamiento en la base de datos Mongo. En la figura 44 se muestra el conjunto de datos presente en un tuit del timeline. El listado de la izquierda resume el conjunto de atributos total mientras que el de la derecha muestra de forma expandida el atributo *user* presente (en forma recogida) en el anterior. Aun así, tantos son los campos que incluye el atributo *user* que la figura 44 solo alcanza a mostrar una parte de ellos. En el apartado siguiente sobre el *modelo de datos* se especifican los atributos que se han aceptado finalmente como válidos para almacenarlos. No obstante, como ejemplo de campos que se han descartado están los que tienen que ver con el diseño de la página principal del usuario, los campos de control del propio Twitter (con excepción del id), el conjunto de entities o aquellos con valor booleano que no se estiman necesarios.

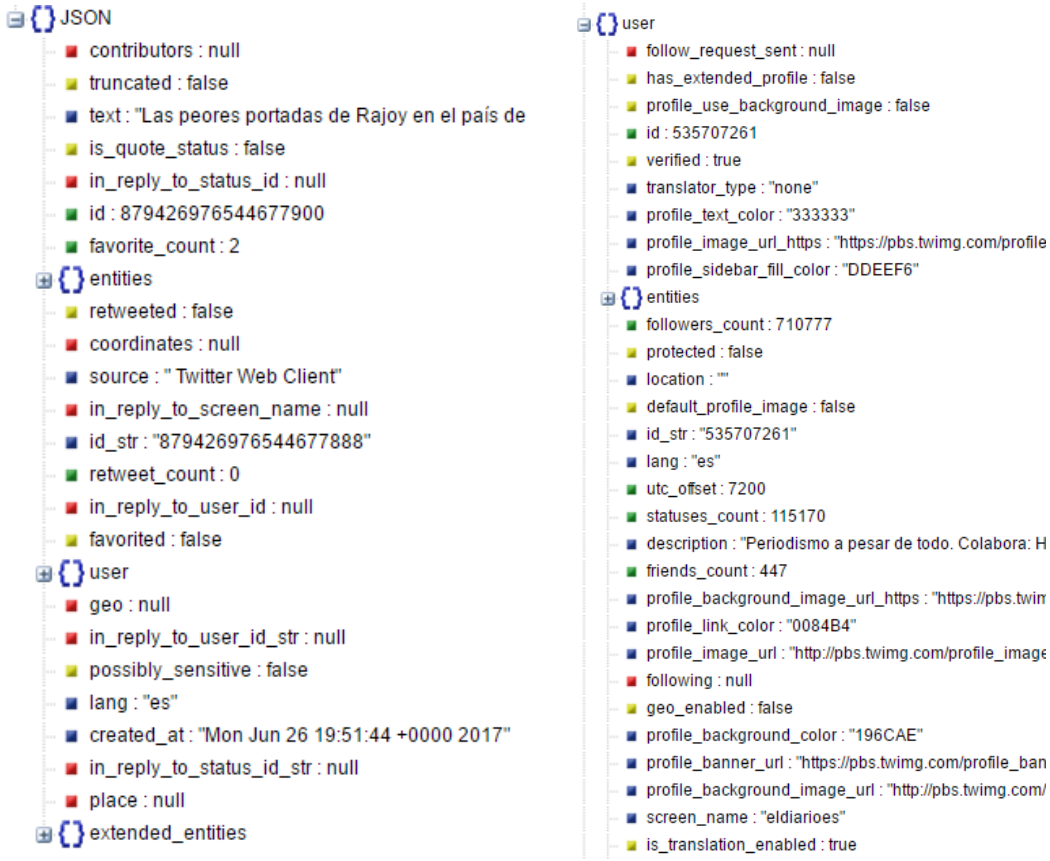

Figura 44. Resumen de los atributos de un tuit del timeline (izq.) y parte del campo user extendido (dcha.).

Por otra parte, al igual que se desechan los campos de control de Twitter, se añaden unos propios con la misma función. De esta forma, los documentos de algunas colecciones contienen un atributo insert_at que define la hora y fecha en que se insertan en base de datos y prácticamente todas ellas contienen un campo id_in_sql que sirve de nexo de unión tanto con MySQL como con otras colecciones del propio Mongo. Asimismo, hay otros atributos como hash_name en la colección que guarda los documentos extraídos de las consultas por hashtag o index_anyo e index_mes en la colección de las métricas que su inclusión se debe a que facilitan los procesos de consulta en base de datos.

Tras haber finalizado las consultas a la API de Twitter sobre los tuits del timeline de los medios y haber almacenado en Mongo los recursos depurados, comienza una segunda función que se encarga de hacer lo propio pero con los tuits que hacen mención a los medios. Para ello, se emplea la consulta sobre hashtags (figura 45). De igual modo que con las peticiones del timeline, cuando se trata de la primera búsqueda por hashtag se extraen hasta un máximo de 20 tuits. No obstante, cuando en base de datos ya hay documentos presentes

en la colección de hashtags Tweet_hash, se busca cuál es el último id almacenado y se empieza a consultar datos a Twitter desde ese valor con la ayuda del parámetro *since_id*. Es importante incluir aquí un límite, ya que los nuevos medios más relevantes pueden conseguir muchas menciones diarias. El límite establecido para las consultas a Twitter se ha fijado en cien tuits, aunque este valor se puede modificar en el fichero de configuración.

```
if last_id:
    statuses = tweepy.Cursor(self.__api.search, since_id=last_id, q=query).items()
else:
    statuses = tweepy.Cursor(self.__api.search, q=query).items()
```

Figura 45. Consultas a Twitter por hashtag.

A parte de que los datos obtenidos a partir de las peticiones sobre hashtags son clave para la obtención de algunas métricas como el engagement o el reconocimiento de marca, también son esenciales para llevar a cabo otras consultas a Twitter en un futuro. Especialmente importante es el campo screen_name, pues su valor coincide con el nombre de la cuenta de los usuarios que mencionan a los medios. Así pues, empleando este atributo como parámetro en una nueva consulta de timeline a Twitter se podría extraer mayor información sobre ellos.

La siguiente consulta realizada a la API de Twitter trata de obtener datos sobre los seguidores del medio. Esta información es fundamental para conseguir lograr el objetivo de geolocalizarlos y de calcular la métrica del alcance localizado. Entre los datos que se recogen de las peticiones se encuentra el de geo_enabled. Éste puede tener un valor True o False dependiendo de si el usuario tiene activado el acceso a su ubicación o no. En los casos en que es True, se incluye un atributo geometry en el documento con las coordenadas latitud y longitud del seguidor y otros campos como la dirección, el nombre completo del país o su abreviatura. Por el contrario, cuando es False no se proporciona ninguna coordenada geográfica, por lo que hay que recurrir a Geopy (figura 46) para transformar el valor del campo location. Este proceso se ejecuta durante la extracción de los datos sobre seguidores y previamente a su almacenamiento, de forma que se añade a los documentos (aquellos con un geo_enabled False) un campo geometry con una estructura idéntica al que aparece en los tuits localizados. Las direcciones geolocalizadas se almacenan en la colección Localizaciones con la intención de evitar que se hagan múltiples peticiones a Geopy para una misma dirección.

```
try:
    location = self.__geolocator.geocode(self.__address, exactly_one=True)
except Exception as e:
    location = None
```

Figura 46. Consultas a Geopy por dirección.

Como se aprecia en la figura 47, en la consulta de followers a Twitter se añade el parámetro *count*, el cual determina el número de usuarios que se extraen por petición. Se ha establecido un valor de 200, que es el máximo posible. Por otro lado, para que no se produzcan duplicidades en el almacenamiento de los mismos se llevan a cabo comprobaciones en Mongo mediante el control de los ids. Sin embargo, se debe tener muy en cuenta que un usuario puede ser seguidor de un medio en un momento dado y en cualquier otro decidir dejar de serlo. Para evitar guardar seguidores que ya no lo son, se ha realizado una nueva petición sobre seguidores a Twitter, pero esta vez solicitando únicamente los ids (figura 48). De esta forma, si alguno de los ids de los seguidores del medio en Mongo no se encuentra en la lista de ids obtenida, se elimina, garantizando así la alineación de los datos.

```
users_creciente = tweepy.Cursor(self.__api.followers, id=cuenta['screen_name'], count=200).items()
```

Figura 47. Consulta a Twitter sobre los seguidores de un medio.

```
# busca y si no esta el follower lo elimina de la lista de mongo
follow_ids = tweepy.Cursor(self.__api.followers_ids, id=cuenta['screen_name'], count=200).items()
```

Figura 48. Consulta a Twitter sobre ids de seguidores del medio.

5.3.4 Modelo de datos

El almacenamiento de los recursos extraídos de Twitter es esencial para conseguir conjuntos de datos históricos con los que poder calcular métricas, conocer y comparar la situación de los nuevos medios en un determinado momento e incluso predecir tendencias, descubrir patrones, realizar clusters, etc. Por este motivo, todos los datos procedentes de la red social se guardan, una vez depurados, en distintas colecciones de Mongo, lo cual define una estructura o modelo de datos. Aunque la base de datos empleada para ello sea no relacional y los documentos de sus colecciones no tengan que seguir una estructura idéntica, la

información se ha organizado en diferentes colecciones según la consulta empleada para la obtención de la misma. De este modo, en principio todos los documentos de una misma colección poseen la misma estructura. Con el propósito de describir el modelo de datos detalladamente, a continuación se presentan y definen las colecciones utilizadas y sus características.

- Twitter main user: esta colección se dedica a almacenar documentos con información específica sobre los medios. Se guarda un documento por día de cada medio. De esta forma se tiene la evolución de las características de los medios en el tiempo, lo que resulta muy útil para hacer un seguimiento del número de tuits publicados al mes o para la obtención de métricas como el crecimiento de audiencia. Los atributos más significativos que incluyen los documentos dentro de esta colección son los siguientes:

    - Id: identificador del medio en Twitter.
    - Id_in_sql: identificador del medio en MySQL. Se emplea como nexo de unión.
    - Insert_at: fecha en la que se inserta el documento. Muy útil para consultas de información en un rango de tiempo. Se incluye durante la depuración de datos.
    - Screen_name: nombre de la cuenta del medio en Twitter.
    - Description: descripción del medio en Twitter.
    - Profile_image_url_https: URL del logo del medio.
    - Statuses_count: número total de tuits publicados hasta el momento.
    - Friends_count: número de usuarios seguidos por el medio hasta el momento.
    - Followers_count: número de seguidores del medio hasta el momento.
    - Favourites_count: número de likes o me gusta que ha recibido el medio.
    - Listed_count: número de listas de Twitter en la que el medio es miembro.
    - Created_at: fecha de creación de la cuenta de Twitter del medio.

- Tweet_medio: almacena documentos con información sobre las publicaciones de los medios, tanto tuits propios como retuits. Dentro de esta información se incluyen algunos contadores de interacción necesarios para calcular, por ejemplo, la métrica

del engagement total. A continuación se listan los atributos más destacados de los documentos de esta colección:

- Id: identificador del tuit.
- Id_in_sql: identificador del medio en MySQL. Se emplea como nexo de unión.
- Screen_name: nombre de la cuenta del medio que ha publicado el tuit.
- Text: texto incluido en la publicación. Si es un retuit comienza con RT.
- Retweet_count: número de veces que se ha retuiteado la publicación.
- Favorite_count: número de likes o me gustas del tuit.
- Created_at: fecha de creación del tuit. Fundamental para las consultas de los tuits publicados en un determinado rango de tiempo.
- User_mentions: este atributo solo se incluye cuando en el texto del tuit aparece al menos una mención a otra cuenta de Twitter. Se trata de un diccionario clave-valor en el que la clave indica el id del usuario mencionado y el valor el screenname o nombre de cuenta del mismo.

- Tweet_hash: esta colección almacena los documentos resultantes de la extracción de datos mediante la consulta de hashtags, de modo que la información incluida hace referencia a las publicaciones de usuarios (tuits o retuits) que mencionan a las cuentas de los medios. Como dichas publicaciones contienen información de los usuarios que interactúan con los medios, el almacenamiento de esta colección es interesante para conservar datos, como el nombre de la cuenta, que permitirían aumentar la información que se tiene de los mismos con nuevas consultas en un futuro. Así pues, los atributos presentes en los documentos de esta colección son:

  - Id: identificador de la publicación.
  - Id_in_sql: identificador del medio en MySQL. Se emplea como nexo de unión.
  - Screen_name: nombre de la cuenta del usuario propietario de la publicación que menciona al medio. Clave para llevar a cabo nuevas peticiones en el futuro.

  - Text: texto incluido en la publicación. Si es un retuit comienza con RT.
  - Retweet_count: número de veces que se ha retuiteado la publicación.
  - Favorite_count: número de likes o me gustas que ha recibido el tuit.
  - Created_at: fecha de creación de la publicación.
  - User_mentions: este atributo aparece en todos los documentos, ya que todos tienen al menos una mención al medio. Del mismo modo, se trata de un diccionario clave-valor en el que la clave es el id del medio y el valor el nombre de la cuenta del mismo.
  - Hash_name: este atributo se añade durante la depuración de los datos. Se ha incluido para que las consultas a Mongo por mención sean más sencillas.

- Twitter_user: almacena documentos con información sobre los usuarios que son seguidores de los medios. La diferencia entre los datos de esta colección y la anterior se debe al tipo de petición realizada para la extracción de los mismos, ya que mientras los anteriores se obtienen mediante la consulta por hashtag, los de Twitter_user se adquieren mediante la de followers. Es importante no confundir entre un seguidor y un usuario que hace mención, pues éste último no tiene por qué ser follower. Los datos presentes en esta colección permiten extraer la localización de los usuarios que interactúan con los medios, de forma que se pueden calcular métricas como la del alcance geolocalizado. Los atributos más importantes en los documentos de esta colección son los siguientes:

  - Id: identificador del usuario.
  - Id_in_sql: identificador del medio en MySQL. Se emplea como nexo de unión.
  - Insert_at: fecha de inserción de los datos del usuario en Mongo. Es un dato significativo para realizar consultas por fechas.
  - Screen_name: nombre de la cuenta del usuario.
  - Hash_name: nombre de la cuenta del medio.
  - Location: localidad del usuario que está presente en su perfil de Twitter. Este atributo es fundamental, ya que si no está activado el acceso a su ubicación, las coordenadas del usuario se obtienen a partir de este campo.

    - Geometry: este campo es opcional. Aparece cuando el usuario tiene activado el acceso a su ubicación, de forma que se incluyen sus coordenadas geográficas. No obstante, para aquellos que no tienen activada la localización se ha añadido un campo con el mismo nombre durante el procesamiento de datos, de modo que se han incluido las coordenadas obtenidas tras la transformación del valor de Location con Geopy.
    - Geo_enabled: atributo que indica si el usuario tiene activa o no su localización.

- Localizaciones: en esta colección se almacenan documentos con información sobre localizaciones y coordenadas. Su presencia agiliza el proceso de transformación de coordenadas, ya que si la localización de un usuario ya existe en Mongo no se requiere volver a generarla. Los atributos existentes en los documentos de esta colección son:

    - Geometry: se trata de un array con los valores de las coordenadas longitud y latitud.
    - Type: indica el tipo de geometría.
    - Properties: es un diccionario que incluye varias claves: el nombre de provincia, el nombre del país, la abreviatura del país y la dirección. El valor de ésta última coincide con el del campo location de la colección anterior.

- Twitter_indice: se trata de la colección que contiene los documentos con las métricas calculadas de los medios. Aunque su cálculo es diario, solo existe un documento de métricas por mes. Esto se debe a que se va actualizando día a día hasta que se cambia de mensualidad, en cuyo caso se crea un nuevo documento. A continuación se listan los atributos de los documentos presentes en la colección:

    - Id_twitter: identificador del medio en Twitter.
    - Id_in_sql: identificador del medio en MySQL. Se emplea como nexo de unión con Mongo.
    - Screen_name: nombre de la cuenta del medio en Twitter.
    - Description: descripción del medio en Twitter.

- Index_anyo: indica el año del documento. Útil para realizar búsquedas.
- Index_mes: indica el mes del documento. Práctico para consultas.
- Insert_at: fecha de inserción del documento en Mongo.
- Created_at: fecha de creación de la cuenta de Twitter del medio.
- Profile_img: URL del logo del medio en Twitter.
- Friends_count: número total de usuarios seguidos por el medio.
- Statuses_count: número total de publicaciones en Twitter.
- Tweets_mes: número de publicaciones en el mes actual.
- Tweets_per_day: número de publicaciones al día en el mes actual.
- Followers_count: número total de seguidores del medio.
- Listed_subscribed: número total de listas en las que el medio es miembro.
- Reach_potencial: métrica del alcance total.
- Alcance_geo: métrica del alcance geolocalizado. Se trata de un array de arrays cuyos valores son la abreviatura del país y el porcentaje de interacciones.
- Engagement_total: métrica del engagement acumulado.
- Engagement_per_follower: engagement medio por seguidor.
- Engagement_per_tweet: engagement medio por tuit.
- Amplification: métrica de la amplificación.
- Horario_emision: franja horaria donde el medio publica más tuits.
- Reconocimiento_marca: métrica del reconocimiento de marca.
- Current_crecimiento_audi: métrica del crecimiento de audiencia actual.
- Pos_rank_local: posición del medio con respecto al total de medios del mismo país en cuanto a reconocimiento de marca.
- Total_rank_local: número de medios total en el país del medio.
- Pos_rank_global: posición del medio con respecto al total de medios existentes en cuanto a reconocimiento de marca.
- Total_rank_global: número total de medios existentes.

En todos los documentos insertados en las colecciones de Mongo es común la presencia del atributo "_id". Este figura como la clave principal y puede tener el formato que de desee. No obstante, si no se especifica ninguno, Mongo utiliza un valor por defecto en hexadecimal.

5.3.5 Explotación de datos

El cálculo de las métricas o explotación de datos es el último proceso que se lleva al cabo en el día. Tras las tres descargas diarias de datos de Twitter, la base de datos de Mongo se nutre de información nueva y lista para ser tratada. El resultado del tratamiento da como lugar la creación o actualización, depende del momento del mes, de un documento con datos que permiten medir en mayor o menor medida la situación de un medio.

El proceso calcula las métricas teniendo en cuenta las fechas de inserción (insert_at) de los documentos que se han almacenado en las colecciones definidas en el apartado anterior. Si el primer día del mes, un medio ha publicado tuits o ha recibido alguna mención, el proceso recorre los documentos correspondientes y calcula las métricas, creando un documento con las mismas dentro de la colección Twitter_indice. Si sigue publicando y recibiendo interacciones durante el mismo mes, se va actualizando el documento día a día hasta llegar al principio de la siguiente mensualidad, en cuyo caso se creará un nuevo documento y se repetirá el procedimiento. En el caso en que un medio no reciba ni publique ningún tuit en un mes determinado, no se creará el documento para ese mes.

Una vez se tienen las métricas totales de los distintos medios en un mismo mes, es posible realizar comparaciones y comprobar la situación de cada uno de ellos. Hasta ahora se ha hecho mención en varias ocasiones a las métricas calculadas para los medios, sin embargo aún no se han explicado detalladamente. A continuación, se define el origen y la finalidad de todas y cada una de ellas:

Reach o alcance potencial

El alcance potencial se define como el número total de usuarios únicos que han podido leer tuits o retuits que mencionan al medio correspondiente. La fórmula para su cálculo consiste en la suma del número de cuentas únicas que han mencionado al medio más la suma de los seguidores de esas cuentas. Para hallar esta métrica se ha realizado una consulta a la colección Tweet_hash de Mongo y se han recuperado los documentos del mes actual. Una vez obtenidos los datos, se han recorrido controlando que las cuentas no se repitieran y se ha llevado a cabo un contador de los documentos y de los seguidores totales de la cuentas. Al finalizar el bucle, se han sumado los documentos y los seguidores.

Alcance geolocalizado

Está métrica tiene en cuenta las menciones y los hashtags que publican los seguidores que mencionan al medio y saca el porcentaje de estas interacciones por país. En la colección Tweet_hash se guardan los documentos con información sobre los usuarios que mencionan a los medios y en la colección Twitter_user se almacena la información sobre los seguidores del medio, incluida su localización. Se realiza una consulta conjunta para extraer las coordenadas de los seguidores que han mencionado al medio y se aplica un contador por cada país que interactúa con el medio. Finalmente, si se compara este valor con el del total de interacciones localizadas se obtiene un porcentaje nacional. El resultado del alcance es un array de arrays con las abreviaturas de los países y el porcentaje de interacción desde los mismos. Al final del mismo se incluye el porcentaje total de interacciones localizadas que se deriva de la relación entre el número de menciones localizadas sobre el total de interacciones publicadas.

Engagement total / engagement por seguidor / engagement por tuit

El engagement total determina el número total de interacciones que recibe un medio durante un mes. Esta métrica es fruto de sumar el número de retuits, el número de likes y el número de menciones recibidas. El número de menciones a un medio se obtiene del cálculo de la métrica del alcance mientras que la suma de retuits y la de likes proviene de recorrer la colección Tweet_medio. Se trata de sumar los retuits y likes de cada uno de los tuits publicados por el medio en un mes.

Por otro lado, el engagement por seguidor es el resultado de dividir el engagement total entre el número de seguidores del medio. De esta forma se obtiene la cantidad de interacción media que recibe el medio por seguidor. Del mismo modo, el engagement por tuit es el valor resultante entre la división del engagement total y el número de publicaciones del medio. Así se obtiene la interacción media por cada tuit publicado.

Amplificación

La amplificación es el número de retuits simples que recibe el medio más el número de retuits que contienen algún tipo de modificación: un texto, menciones, etc. El número de retuits se obtiene de la función que calcula el engagement mientras que los retuits

modificados se consiguen recorriendo la colección Tweet_hash. Los documentos de esta colección que incluyen un campo is_quote_status con valor True son los que tienen su origen en retuits modificados. Llevando a cabo un contador de ese atributo se puede hallar el número total de retuits alterados. La suma de este valor y el del número de retuits simples proporciona la solución.

Horario emisión

El horario de emisión del medio indica la franja horaria usual en el que el medio publica más tuits. Para calcular este valor, se han establecido tres rangos de horas: por la noche (0:00 – 8:00), por la mañana (8:00 – 16:00) y por la tarde (16:00 – 24:00). La colección Tweet_medio recoge todos los tuits publicados por el medio, por lo que se consultan los del mes actual y se aplican tres contadores, uno por cada rango de horas. El contador de mayor valor define la franja horaria media de publicación del medio.

Reconocimiento de marca

El reconocimiento de marca tiene en cuenta el número de menciones al medio más el número de menciones a la marca del medio. Una mención a la marca se realiza mediante hashtag (#) y no tiene por qué coincidir con el nombre de la cuenta. Por ejemplo, los hashtags #coke o #destapalafelicidad conformarían menciones a la marca de la compañía Coca Cola. Por el momento se han tenido en cuenta los que incluyen hashtags con el nombre exacto de la cuenta, sin embargo se prevé ampliar las posibilidades en el futuro. Así pues, tanto el número de menciones "normales" (@) como el número de menciones por hashtag se obtienen a partir de la colección Tweet_hash, aunque la primera se halla en la función que calcula el alcance potencial. El número de menciones por hashtag se obtiene comprobando la presencia del nombre de la cuenta del medio en el atributo user_mentions_id.

Crecimiento de audiencia

El crecimiento de audiencia es la relación entre el número nuevo de seguidores que ha conseguido un medio en un mes con respecto a los que tenía en el mes pasado. El resultado es un porcentaje que mide el aumento o bajada de seguidores en el mes actual. La colección

que almacena los documentos con la información del número total de seguidores del medio es la de Twitter_main_user. Se hace la consulta y se obtienen el primero y el último documento de un mes, de forma que se saca la diferencia entre el número de seguidores más reciente y el mismo al principio de mes. Tras esto, se compara la diferencia hallada con el total que se tenía al comenzar la mensualidad.

5.4 Transmisión de los datos

5.4.1 Proceso interno

El elemento del proyecto que posibilita la transmisión de la información almacenada en las bases de datos a la página web es el web service o servicio web. Este componente se encarga de recibir consultas, de procesarlas y de devolver recursos fieles a la petición realizada. La manera en que se realizan las consultas es mediante URLs con parámetros incluidos, siendo los valores de éstos los que marcan la respuesta a recibir. Al tratarse de un servicio que solo admite métodos GET, se pueden hacer consultas sobre las bases de datos pero no inserciones ni modificaciones.

Se han habilitado dos tipos de consulta para la obtención de información. Esto se debe en gran parte a la organización de la página web del Observatorio para mostrar los datos, la cual se detalla en el apartado siguiente. No obstante, para comprender el funcionamiento del web service se adelanta que en este sitio existe, por una parte, una página inicial donde se requieren algunos datos de cada uno de los medios almacenados y la capacidad de poder filtrar por el valor de los mismos y por otra, una ficha detalle por cada medio donde se necesita la información total del medio seleccionado.

La diferencia entre las dos páginas del sitio web en relación a los datos que muestran salta a la vista ya que, mientras la página principal hace una consulta a todos los nuevos medios existentes, la ficha detalle solo realiza la petición a uno de ellos. El consumo de recursos también es muy desigual, pues aunque la primera solamente solicita unos pocos datos, el número de registros que debe proporcionar es, hoy por hoy, más de mil quinientas veces mayor que la segunda. Este problema se ha tenido en cuenta y su solución se define en el apartado *Optimización del proceso*.

Así pues, tal y como se ha mencionado antes, se han especificado dos modelos de consulta. Para realizar aquella que proporciona información acerca de todos los medios se debe completar la barra de dirección de un navegador con el contenido que se muestra en la figura 49. Ésta, a su vez, permite la inclusión de varios parámetros con los que realizar filtros de información, tales como país (pais), provincia (prov), temática (tema) y cobertura geográfica (cobg). Se pueden acumular e incluso se permite filtrar con varios valores para un mismo parámetro (figura 50).

Figura 49. Consulta web para obtención de datos de todos los medios.

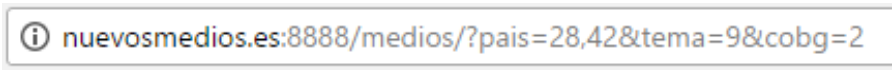

Figura 50. Consulta con varios parametros de flitrado.

El procedimiento interno en el web service comienza con la adquisición de los parámetros incluidos en la URL. Esto se realiza mediante la función *input* del framework web.py. Una vez se han obtenido, se consulta a la base de datos MySQL por el id, el título y el país de todos los registros de medios que coincidan con los valores introducidos. En el caso de que no exista ninguna coincidencia, la respuesta del web service es la que se muestra en la figura 51. Por el contrario, si sí que hay coincidencia, se realiza una consulta a Mongo por cada id extraído de MySQL, de modo que se solicitan los campos requeridos (figura 52) del documento más reciente de la colección que contiene las métricas calculadas de Twitter. Los resultados de las consultas de MySQL y Mongo se relacionan por medio del id mencionado y se almacenan primero en una lista de diccionarios Python para finalmente transformar éstos a JSON. Por último, el web service devuelve este conjunto de resultados con la estructura que se aprecia en la figura 52.

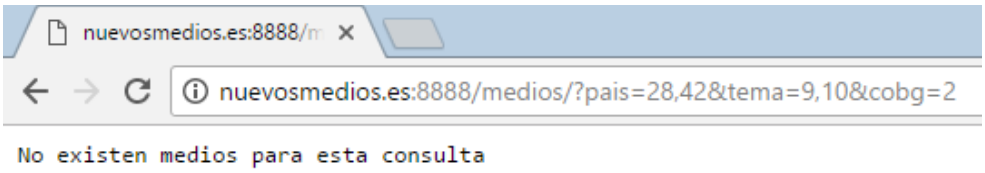

Figura 51. Respuesta del web service cuando no existen resultados para la consulta realizada.

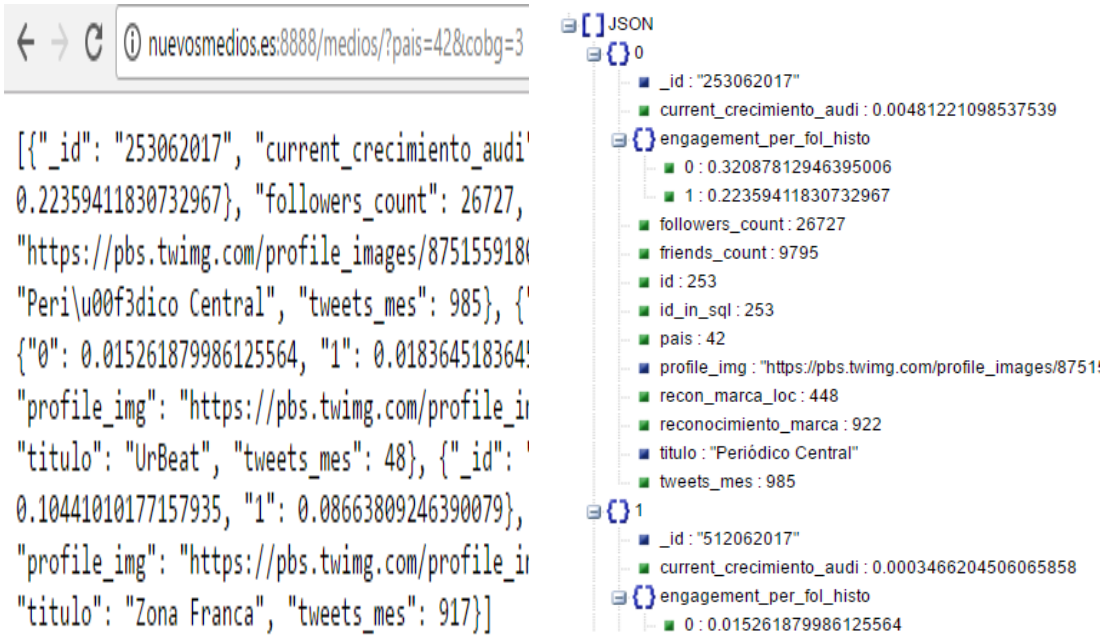

Figura 52. JSON devuelto por el web service (izq.). Resultado organizado (dcha.)

Si se presta atención a la figura 52, hay una métrica llamada *engagement_per_fol_histo* que contiene dos valores. El primero hace referencia al valor del engagement per follower del mes pasado mientras que el segundo se corresponde al del mes actual. Para conseguir este histórico, se ha realizado durante el mismo procedimiento una nueva consulta a Mongo con la intención de procesar los dos últimos documentos de la colección de métricas.

En cuanto al otro tipo de petición que acepta el web service, su funcionamiento es similar al que ya se ha visto. No obstante, la consulta es más sencilla (figura 53), ya que esta vez solo se pasa un parámetro que debe coincidir con el id de un medio en base de datos. Con este mismo id se relacionan los datos de MySQL y Mongo y se juntan en un mismo JSON, que finalmente se devuelve como resultado de la petición (figura 54).

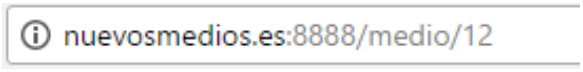

Figura 53. Consulta para obtener toda la información almacenada acerca de un medio.

Figura 54. JSON de un medio devuelto por el web service (izq.). Resultado organizado (dcha.).

Tal y como ocurre en la consulta de información a todos los medios, la petición de los datos de un solo medio también genera campos con históricos. Sin embargo, esta vez la dimensión de los mismos es mayor, pues engloban los datos de los últimos doce meses. La presencia aquí de esta información se debe a la existencia de varias gráficas de evolución temporal en la página detalle del Observatorio. Los datos específicos que incluyen un historial anual son los de amplificación, crecimiento audiencia y engagement por follower.

5.4.2 Optimización del proceso

Durante el desarrollo del web service y la validación de su funcionamiento, se tuvo constancia de que ya por entonces, cuando no se contaba con tantos registros de nuevos medios como en la actualidad, el tiempo que requería el proceso desde que se recibía una consulta a la devolución de los datos era excesivamente alto. Especialmente en el caso de la petición que solicita información acerca de todos los medios.

El problema estaba claro y el objetivo era indudable, era esencial conseguir optimizar el tiempo de procesamiento del servicio web. De lo contrario, pocos visitantes del Observatorio tendrían la paciencia suficiente para navegar por el sitio (figura 55). Es entonces cuando surge la idea de recurrir a las ventajas que ofrecen las bases de datos en memoria en cuanto a la rapidez en la carga de datos. De ahí que se optara por la inclusión de Redis en el proyecto. El resultado ha sido muy beneficioso y la elección de esta alternativa ha resultado ser un acierto, ya que una vez los datos de las consultas de almacenan en memoria, las respuestas a las mismas son inminentes. Para comprender cómo funciona Redis en el web service, a continuación se define su actividad en el proceso de las consultas. Para ello se va a suponer que no hay nada guardado aún en memoria y que dos usuarios hacen la misma petición.

Figura 55. Tiempo de carga de datos de la página principal del Observatorio sin Redis.

Para que el web service comience a funcionar, el primer usuario realiza una consulta. En ese momento, el proceso se activa y el sistema recoge los parámetros introducidos en la URL de la petición. Seguidamente, se conecta con Redis y busca una clave que contenga los datos que requiere la consulta. Como se ha supuesto que no existe aún información en memoria, la búsqueda no recupera ningún recurso, de modo que el web service procede a funcionar de manera normal, es decir, de la misma forma que lo hacía antes de incluir Redis en él. No obstante, antes de devolver la respuesta en formato JSON, ésta se guarda en memoria con una clave específica que identifica la consulta realizada y durante un tiempo determinado, que en este caso ha sido de un día. Al tratarse de la primera consulta, aún no se ha podido recuperar datos de Redis, por lo que la información tardara demasiado tiempo en ser recibida por parte del usuario.

Justo después de la primera consulta, un segundo usuario decide hacer la misma petición. Del mismo modo que antes, el web service toma los parámetros de la URL y comienza su proceso. Se conecta a Redis y busca una clave que coincida con la consulta realizada. Como esa petición ya se había realizado antes y no ha pasado aún un día desde entonces (tiempo en que se mantiene la clave en memoria), encuentra una coincidencia y la devuelve como

solución a la consulta. De esta forma, se evitan los procesos de extracción de datos de MySQL y Mongo, de nexos y relaciones entre los conjuntos de datos, conversiones, etc. Lo que se traduce en menor consumo de recursos y, por consiguiente, en mayor rapidez de carga (figura 56).

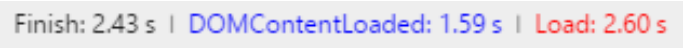

Figura 56. Tiempo de carga de datos de la página principal del Observatorio con Redis funcionando.

Aunque el problema se haya solventado de la forma indicada para la mayoría de visitantes que consumen los datos desde la web del Observatorio, tal y como se ha explicado el proceso, todavía existe un usuario (el primero en realizar una consulta) que debe esperar durante un tiempo excesivo para poder recibirlos. Esto se ha solucionado mediante la programación de una tarea diaria que ejecuta la primera consulta a las seis de la mañana, Horario Europeo Cental (UTC+1). La elección de la hora se debe al origen de la mayoría de visitantes del Observatorio, ya que al tratarse de medios escritos en español, la gran mayoría son españoles o latinoamericanos. Como la diferencia horaria entre España y los países latinoamericanos es de aproximadamente seis horas, se entiende que las seis de la mañana y las doce de la noche no son horas de mucha concurrencia en la web.

Por otra parte, el mantenimiento de los datos en memoria tiene una desventaja y es que durante el tiempo que se conservan no pueden variar aunque realmente lo hayan hecho. Por ejemplo, si se introduce un nuevo medio en el Observatorio, éste no aparecerá en la web hasta que los datos en Redis no se actualicen, es decir, hasta que deje de guardar los datos correspondientes y se produzca la primera petición del día.

5.5 Sitio web

5.5.1 Estructura

La página web del Observatorio de nuevos medios constituye la parte visual, informativa y accesible de este proyecto. Los encargados de su implementación han sido los desarrolladores y diseñadores web y aunque su labor es fundamental para lograr los objetivos, sus metodologías, su organización y las herramientas empleadas quedan al margen del núcleo de este trabajo. No obstante, sí se van a definir las partes de la web que tienen relación directa con los datos extraídos y tratados por el sistema en el que se basa este proyecto. De este modo, se puede hablar de tres páginas que actúan como clientes ante el servicio web que transmite la información: la página principal o home, el ranking y la página detalle o ficha de los medios. En los siguientes apartados se detallan las características propias de cada una de ellas.

5.5.2 Página principal

Si desde cualquier navegador se introduce la dirección *www.nuevosmedios.es* se accede a la página principal del Observatorio de nuevos medios (figura 57). Esta página es la que realiza la petición de un conjunto determinado de datos de todos los medios existentes al web service, de modo que también es la que más beneficio ha conseguido, en términos de rapidez, por la inclusión de Redis en el proyecto. Al entrar en ella, lo primero que se aprecia es un mapa del mundo con distintos números distribuidos a lo largo y ancho del mismo. Estos números representan la cantidad de nuevos medios que existen en el país donde están centrados, según la base de datos del Observatorio. Esta información se la proporciona el mismo web service, ya que éste devuelve un conjunto de datos con el id del país de origen por cada nuevo medio. Con ese id, es posible sacar el código ISO de las distintas naciones, que es lo que se requiere para conseguir las coordenadas mediante APIs de geolocalización de países.

Figura 57. Página principal del sitio web del Observatorio de nuevos medios.

El mapa permite la interacción por parte del usuario, de forma que si se pulsa sobre alguno de los círculos que contienen las cifras de medios aparece una ventana sobre el mismo con información adicional (figura 58). Esta ventana o pop-up, además de incluir un listado de los medios del país seleccionado y las miniaturas de los logos de algunos de ellos, ofrece dos opciones: ver el listado y ver por comunidades o estados. La primera redirige al ranking de medios, el cual se define en el apartado siguiente y la segunda realiza un zoom sobre la nación escogida, de manera que el círculo que antes mostraba el número total de nuevos medios en el país ahora se divide por provincias, indicando la cifra exacta de medios en cada una de ellas (figura 59). Para obtener las coordenadas de las provincias se ha empleado el mismo id de los países pero esta vez para encontrar las provincias relacionadas. De nuevo, se puede pulsar sobre los círculos de las provincias, sin embargo ahora solo ofrece la opción de ver el listado.

Figura 58. Ventana con información acerca de los medios de un país.

Figura 59. División de los nuevos medios españoles por provincias.

Por otro lado, si se sitúa la página debajo del mapa se observa un pequeño resumen de algunos de los datos que contiene el Observatorio (figura 60). En concreto, se trata de un contador del número de medios almacenados, del número de países de procedencia de los mismos y del número de temáticas existentes. Este conteo se realiza sobre el conjunto de datos que se recibe desde el web service.

Figura 60. Resumen de algunos de los datos presentes en el Observatorio.

5.5.3 Ranking de medios

El ranking de medios es la siguiente sección después de la de inicio en el panel de navegación de la web. De hecho, es tal su importancia en el Observatorio que es posible hacer uso de él en la misma página inicial. Como ranking que es, requiere tener información de todos los medios, por lo que la consulta que se lleva a cabo al web service es la misma que la que se emplea para la página principal.

El funcionamiento de esta utilidad es bastante intuitivo. El usuario puede interactuar con los medios mediante el empleo de los filtros que se aprecian en la figura 61. De esta forma, se pueden dividir los medios discriminando por temáticas, cobertura geográfica, países e incluso por nombre o parte de él. En cualquier caso, los filtros se pueden acumular. Por el contrario, si no se aplica ninguno el ranking muestra todos los medios existentes.

Figura 61. Opciones de filtrado del ranking de medios.

Justo debajo del panel de filtros de la figura 61, se halla un cuadro con el que el usuario también puede intervenir en el ranking. Éste incluye varias métricas de las que se han

extraído de Twitter y permite ordenar y comparar los medios de mayor a menor y viceversa según sean sus valores para cada una de ellas. Por defecto se encuentra activada la de reconocimiento de marca (figura 62), pero también se incluyen la de número de seguidores, ratio seguidores/seguidos, tuits mensuales y engagement global.

Figura 62. Ranking de medios del Observatorio.

El número de nuevos medios analizados es excesivamente alto para mostrarlos todos en la misma página. Por ello, se ha incluido un paginador que los expone de diez en diez y que facilita así la navegación por parte del usuario.

Cada uno de los medios que aparecen en el ranking se presenta en una especie de ficha donde se incluyen algunos de los datos relacionados con los mismos. En la figura 63 se aprecia dicha ficha, donde se puede observar el logotipo del medio, su nombre y los valores para el reconocimiento de marca, el número total de seguidores, la relación seguidores/seguidos y el número de tuits mensuales. No obstante, mientras la métrica que se resalta en un recuadro, en este caso el reconocimiento de marca, tiene un carácter dinámico y varía según se interactúa con los filtros, las que se sitúan en la parte inferior se disponen de manera uniforme.

Figura 63. Ficha de un nuevo medio en el ranking del Observatorio.

Todas las fichas de los medios en el ranking siguen la misma estructura. Incluso cuando se interactúa con los filtros, tanto los datos uniformes como el dinámico se representan de la misma manera. Sin embargo, hay una excepción. Cuando se selecciona en el panel la métrica de engagement global para que los registros se ordenen según su valor en relación a ésta, la información dinámica se muestra como se observa en la figura 64.

Figura 64. Ejemplo de fichas de medios ordenadas por engagement global.

Esta vez el recuadro que engloba la métrica dinámica puede aparecer de dos colores diferentes: en rojo o en verde. Asimismo, el valor se encuentra precedido de una flecha

orientada hacia abajo en el caso del recuadro rojo y hacia arriba en el caso del verde. Esto se debe al campo *engagement_per_fol_histo* al que se ha hecho referencia en el apartado sobre transmisión de los datos. Si se recuerda, éste es el único histórico que aparece en la consulta sobre todos los medios y está formado por dos valores, el engagement por seguidor actual y el engagement por seguidor del mes anterior. Así pues, cuando un nuevo medio tiene mayor engagement por seguidor que el que tenía hace un mes, la métrica se visualiza en verde como símbolo de mejora. Por el contrario, si la diferencia entre los valores es negativa se simboliza en rojo. En el caso de no existir diferencia, la ficha se visualiza de la misma forma que con las demás métricas dinámicas.

5.5.4 Página detalle

Si durante la navegación por la web del Observatorio se pulsa sobre las fichas de los medios en el ranking o sobre las imágenes de los logos que se incluyen en el mapa de inicio, se accede a la página detalle del medio seleccionado. En ella se muestra gran parte del total de información que se tiene sobre ese registro, es decir, tanto aquella que tiene su origen en la labor documental como la que procede de Twitter. Por ello, la consulta que se realiza al web service simplemente necesita como parámetro el id del medio en base de datos.

Una vez el usuario se ubica dentro de la ficha detalle de un medio, lo primero que distingue es el título y el logo del mismo, seguido de otra información como el país de origen, la cobertura geográfica, un resumen o slogan, las temáticas que trata, los enlaces a su web y a su cuenta de Twitter y datos sobre el número de seguidores, el número de tuits diarios y el ratio seguidores/seguidos (figura 65).

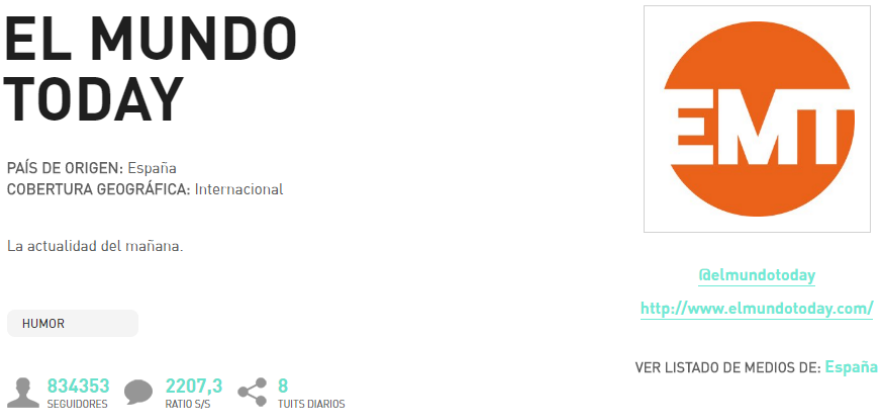

Figura 65. Primeros datos visibles en la ficha detalle de un medio.

De los datos mostrados en la figura 65, el país de origen, la cobertura geográfica, las temáticas y las URLs de la cuenta de Twitter y de la web del medio tienen origen en la labor de investigación. El resto de la información que aparece en la imagen y toda la que se va a ver a continuación procede de la red social.

Conforme el usuario continúa dirigiéndose al final de la página, se encuentra con más datos sobre el medio. En primer lugar se muestra un mapa de alcance. En él se plasma la métrica del alcance dividida por países, es decir, del total de interacciones sobre un medio se hallan los porcentajes del mismo según el país de origen de los seguidores o usuarios que han participado. De esta forma, según sea el valor del porcentaje, un país se muestra en un tono más claro o más oscuro en el mapa, manifestando más interacción cuanto más intenso es el color. Aun así, se puede obtener el valor de los porcentajes pasando el cursor sobre los distintos países involucrados (figura 66). En el web service, esta información viene dada en forma de array de arrays, de modo que cada elemento contiene el código ISO del país y su porcentaje de interacción. Asimismo, al final del array se incluye el porcentaje total de interacciones localizadas.

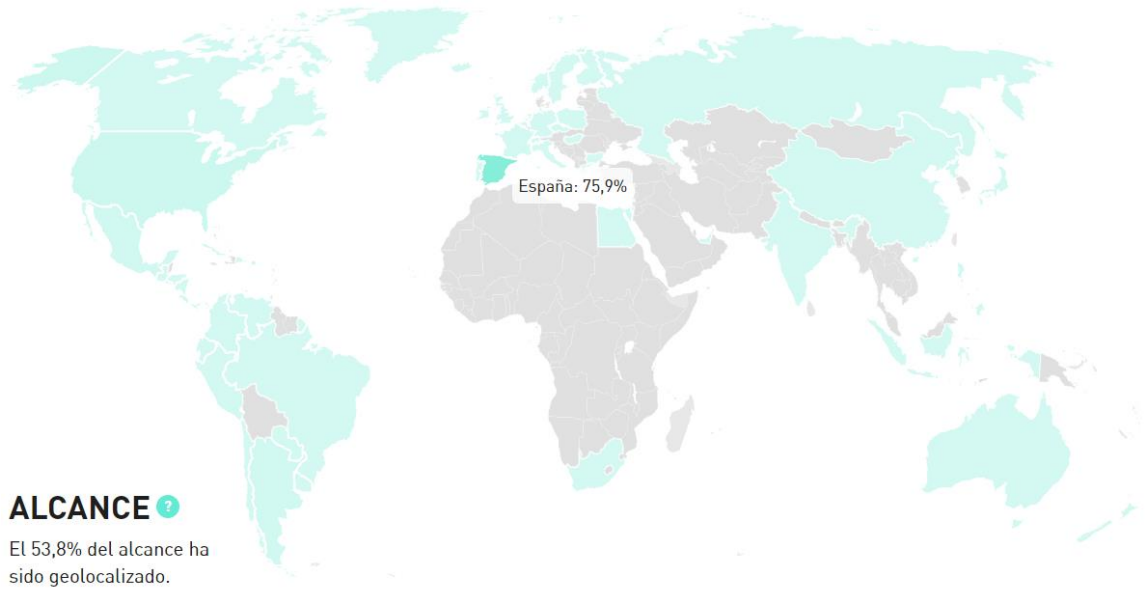

Figura 66. Mapa de alcance de la ficha del medio El Mundo Today.

Si se observa la figura del mapa, en la parte inferior izquierda aparece un texto donde se indica el porcentaje total de interacciones localizadas. Es prácticamente imposible conseguir

la situación geográfica de todas las interacciones, ya que no son pocos los usuarios que, además de no activar el acceso a su ubicación, no incluyen una localidad en su perfil. Por otro lado, la métrica del alcance tiene en cuenta todas las interacciones que ha sufrido el medio desde su inserción en el Observatorio y no solo del último mes, como sí que hacen las que se van a ver a continuación.

Debajo del mapa aparecen cuatro pestañas: crecimiento de audiencia, amplificación, listas y más información. Ésta última se encuentra activada por defecto y muestra un contenido con tres secciones (figura 67). La primera de ellas hace referencia al engagement, que es el resultado de la suma de likes, retuits y menciones que recibe un medio. Sin embargo, se incluyen dos valores correspondientes al engagement global y al engagement medio de cada tuit. Estos son fruto de dividir y redondear el engagement entre el número de seguidores y entre el número de tuits publicados respectivamente. Aun así, se pretende revisar la forma de representar estas métricas, ya que con solo valores numéricos resulta complejo comprender la situación que atraviesa el medio. Quizás con porcentajes o añadiendo alguna referencia sería más apropiado.

Figura 67. Pestaña Más información del detalle de El Mundo Today.

La siguiente sección dentro de la pestaña más información se enfoca en el reconocimiento de marca. Se recuerda que esta métrica se calcula con el número de menciones a la cuenta o

a la marca del medio. Esta vez, para que no figure un simple valor numérico, se ha añadido la posición del medio con respecto a los de su país y con respecto al global de medios almacenados en base de datos. En la última sección del contenido se incluye el horario de emisión, que no es más que la franja horaria media en la que un medio difunde sus publicaciones. Como el Observatorio contiene medios ubicados en husos horarios muy diferentes, se ha tomado el tiempo universal coordinado o UTC.

De derecha a izquierda, la siguiente pestaña es la de las listas de Twitter (figura 68). El contenido muestra ahora dos valores, uno para las listas del medio, en cuyo caso se incluye un desplegable con el nombre de las mismas y el enlace a cada una de ellas y otro para las listas en las que el medio es miembro, donde figura un enlace que redirige directamente a la página de listas de Twitter (el usuario debe autenticarse para acceder).

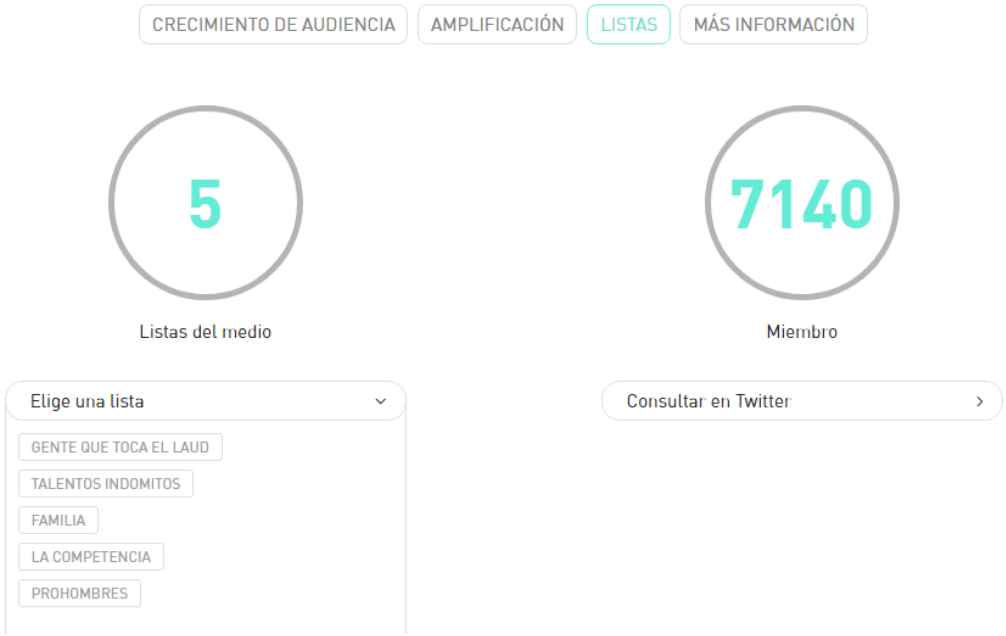

Figura 68. Pestaña Listas en la ficha detalle.

Las siguientes dos pestañas muestran las métricas sobre amplificación y crecimiento de audiencia en forma de gráficas temporales. Éstas se construyen a partir de los campos históricos incluidos en la respuesta del web service. Se tratan de diccionarios clave valor en el que la clave "1" especifica el valor de la métrica actual y la clave "12" el valor de la misma un año antes (figura 69). De esta forma, los datos se representan en las gráficas tal y como se aprecia en las figuras 70 y 71. La amplificación de un medio es un valor entero, ya que

tiene en cuenta el número de retuits recibidos más el número de retuits modificados, es decir, aquellos que incluyen algún texto o contenido pero mencionan la misma URL del tuit retuiteado. Por otra parte, el crecimiento de audiencia se calcula como porcentaje, pues se trata de la relación entre el número nuevo de seguidores y el que se tenía el mes pasado.


```
"amplification_histo": {"1": 83007, "2": 116190, "3": 85737, "4": 75458, "5": 96590, "6": 114752,
                        "7": 83232, "8": 79320, "9": 95467, "10": 110205, "11": 139748, "12": 111355}
```


Figura 69. Historial de amplificación en la respuesta del web service.

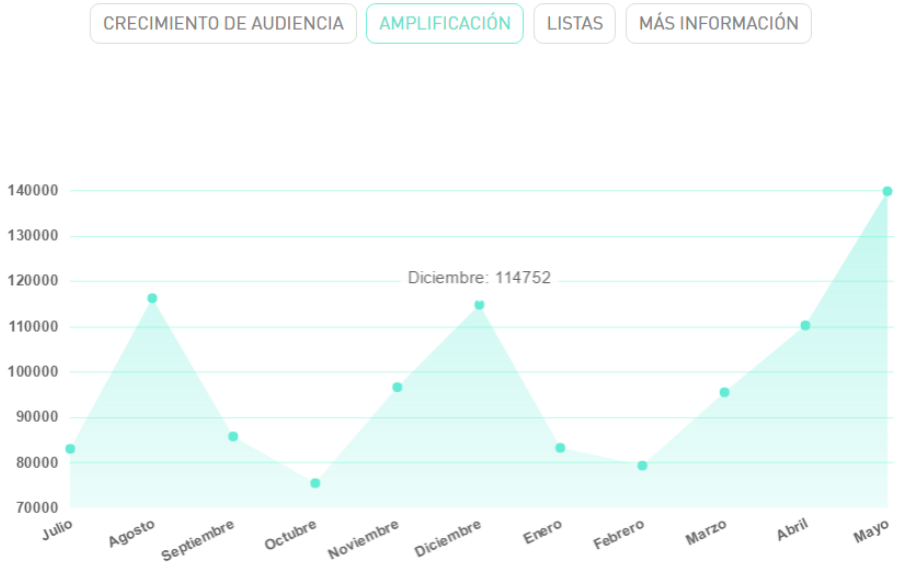

Figura 70. Gráfica sobre amplificación en la página detalle. Medio El Mundo Today.

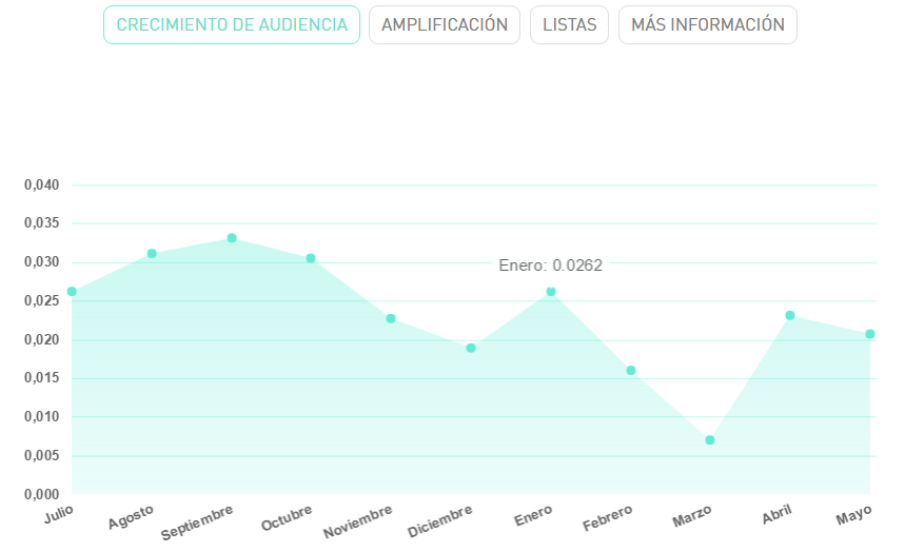

Figura 71. Gráfica sobre crecimiento audiencia en la página detalle. Medio El Mundo Today.

6 Conclusiones

El Observatorio de nuevos medios es un proyecto vivo y en creciente desarrollo. Aún se encuentra en una fase inicial, de modo que todavía existe mucho margen de mejora. Ahora mismo se nutre de una sola fuente de datos, lo cual implica una dependencia considerable hacia la misma. Se estima necesario la incorporación de nuevas redes sociales y por ello se marcó como objetivo prioritario el desarrollo de un sistema flexible y con garantías de ser escalable. Para cumplir este propósito, se ha estructurado el kernel de forma que la extracción de datos de una red social es independiente a la de las demás. No obstante, existe un administrador de plugins donde sí se deben indicar los procesos que se quieren ejecutar. Asimismo, el sistema se apoya sobre un archivo que establece las configuraciones de la descarga y permite añadir y modificar aspectos como las credenciales de acceso a las APIs, las horas de ejecución, el número de ítems a extraer por consulta, etc.

Para el correcto funcionamiento del sistema a nivel general, se requería una serie de elementos que se relacionaran entre sí para conseguir una sintonía entre el trabajo de todos los integrantes del proyecto. Este hito se ha logrado utilizando las bases de datos como puntos de referencia y el gestor de nuevos medios, el sistema de extracción de datos y el web service como herramientas intermedias. Hasta ahora está funcionando adecuadamente y si, por algún caso, algo fallara, los errores se quedan registrados en los logs, lo que facilitaría su resolución.

Uno de los objetivos planteados como fundamentales desde el inicio del proyecto ha sido la automatización de los procesos de extracción de datos. Tras realizar varias pruebas de descarga de información, se estudió la manera de conseguir ejecutar el proceso de forma automática. Afortunadamente, Python cuenta con una librería que permite exactamente esto: *Schedule*, de modo que se han predefinido un total de tres horas al día en las que se inicia el demonio del kernel de forma autosuficiente.

La ubicación de los usuarios que interactúan con los medios por medio de las redes sociales es un dato muy interesante y muy atractivo para los visitantes del Observatorio. Por ello, es uno de los propósitos marcados como más prioritarios. Para determinar la dificultad de su consecución, ha sido necesario analizar los datos que se incluyen en los tuits. De este estudio se extrajo que las coordenadas geográficas solo están presentes en las publicaciones de usuarios que tienen activado el acceso a su ubicación. En un principio, se localizaron

únicamente éstos, pero resultó que muy pocos tuits contenían coordenadas, por lo que la muestra era demasiado pequeña para representar la realidad. La clave aquí fue otra vez Python y concretamente el servicio de geocodificación GeoPy, de forma que pasándole las localidades de los perfiles de los usuarios (presentes en los tuits), devuelve las coordenadas geográficas de las mismas. Aunque no es totalmente fiable, pues un usuario puede completar su localidad con el valor que desee, sí se consigue una muestra mayor con la que poder interpretar la realidad.

Se confía en que el proyecto del Observatorio se establezca como una herramienta útil y de uso frecuente para que los propios medios que se incluyen en él puedan conocer su situación, compararla con la de los demás y saber qué están haciendo mejor y qué peor en su camino hacia el éxito informativo. Para conseguirlo, es necesario que tanto los datos proporcionados como la forma en que éstos se representan permitan al usuario entender las circunstancias por las que pasan los medios de una manera rápida y sencilla. Las métricas que se incluyen actualmente en la web son las que se han extraído hasta el momento, aunque se tienen presentes algunas más que ampliarían el conocimiento obtenido por parte del usuario. Por otro lado, se cuenta con varias gráficas y un mapa como elementos visuales de representación de datos. Sin embargo, sería interesante poder disponer de gráficas que comparasen varios medios a la vez o uno solo en intervalos de tiempo diferentes, por ejemplo. También está pendiente modificar la forma en que se ofrece la métrica del engagement, pues ahora mismo se muestra un número entero que no favorece su interpretación. Quizás un porcentaje sería lo más apropiado.

Al tratarse de un proyecto en el que han colaborado personas con diferentes perfiles profesionales, la comunicación ha sido esencial en todo momento. El trabajo en equipo ha resultado muy gratificante tanto en lo personal como en lo profesional. Ha sido muy útil para mejorar y renovar competencias transversales, especialmente la de la comunicación efectiva, cuyo incumplimiento es uno de los problemas más comunes durante el desarrollo de un proyecto. Por otro lado, los conocimientos técnicos adquiridos son considerables y se valoran muy positivamente. Echando la vista atrás, cuando el proyecto no era más que una idea y analizando el recorrido seguido hasta el presente, no hay duda del gran esfuerzo realizado en investigación y formación, sobre todo en temáticas como la administración de sistemas, el desarrollo de software específico, la gestión de bases de datos, la interacción con APIs de redes sociales o la transformación, tratamiento y transmisión de datos.

7 Futuras líneas de investigación

El volumen de información que se almacena en las bases de datos del Observatorio de nuevos medios es hoy por hoy suficientemente grande como para aplicar una serie de funciones que mejoren y faciliten la interpretación de los datos por parte del usuario. Entre ellas, las que se prevén implantar más temprano son el sistema de alertas y avisos y la posibilidad de extraer informes semestrales.

El sistema de alertas constituye un servicio eficaz para el control y seguimiento de los medios. Gracias a esta funcionalidad, el tiempo invertido en las comprobaciones manuales sobre la evolución de los registros se disminuye considerablemente, ya que es la propia herramienta la que se encarga de advertir al usuario cuando se presenta alguna singularidad en los datos extraídos de las redes sociales. Se deben definir los comportamientos que se consideran como peculiares, como la escasez de datos sobre un medio durante un periodo de tiempo determinado o los cambios drásticos en los valores de las métricas en un momento específico.

Los informes semestrales, por su parte, pretenden acotar el volumen de datos que se tiene sobre un medio. Esta funcionalidad adquiere sentido cuando el volumen de información sobre un medio es tal que se requiere la posibilidad de conocer su situación en un momento dado y no únicamente de manera global. Los documentos extraídos contienen un resumen con los principales datos del medio y pueden incluir como observaciones los avisos generados por el sistema de alertas.

Por otra parte, se contempla incorporar nuevas redes sociales como fuentes de datos del Observatorio. En la actualidad el proyecto se nutre de Twitter, pero ya se ha implementado y testeado la extracción de datos de Youtube. De hecho, de ésta última se cuenta con información acerca de los canales de los medios, de sus videos y sus listas de reproducción, de sus suscriptores e incluso de los usuarios que comentan sus publicaciones. Del mismo modo que con Twitter, se calculan métricas globales (engagement, reconocimiento de marca, alcance, etc.) a partir de los datos de las colecciones mencionadas.

Algunas de las redes sociales que se tienen en cuenta para su integración en un futuro próximo son Facebook, Google plus e Instagram. De ellas es importante obtener información sobre las interacciones que reciben las publicaciones de los medios, es decir, el número de

likes, +1, shares, menciones y hashtags, comentarios, etc., así como de los seguidores y subscripciones de cada uno de ellos. Con estos datos se podrían hallar métricas compatibles con las que ya se han calculado. Aun así, todavía se le puede sacar mucho partido a los datos que se extraen de Twitter. Tanto es así que se tienen en mente nuevas métricas, como los términos más relacionados con un medio, el análisis de sentimiento de sus publicaciones o las que se derivan de la estadística inferencial, con las que se podría aumentar el conocimiento de los visitantes sobre la situación de los nuevos medios digitales.

La cantidad de datos con los que se cuenta es lo bastante abundante como para empezar a invertir tiempo en análisis predictivos. Se tiene previsto comenzar con modelos de regresión lineal y árboles de regresión para conocer la situación aproximada de las diferentes métricas de un medio en los meses venideros. Asimismo, se pretende mejorar la capacidad de comparar medios del Observatorio, pues solo el ranking presente en la web lo permite. Para ello se ha pensado en incluir series temporales, de modo que se puedan contrastar, en una misma gráfica, las métricas de varios medios en el mismo intervalo de tiempo o una sola métrica de un mismo medio en distintos momentos, por ejemplo en diferentes años. Con ésta última funcionalidad se podrían detectar patrones, singularidades y tendencias fácilmente.

Otro posible cálculo que se podría añadir al proyecto es, por ejemplo, la relación entre los seguidores de un medio que interaccionan con el mismo con respecto del total de usuarios que siguen al medio. De esta forma se obtendría el porcentaje del total de usuarios que realmente reacciona ante las publicaciones del medio al que siguen.

Por otro lado, es posible llevar a cabo análisis muy específicos sobre el total de los medios incorporados. Es muy útil para examinar el comportamiento del público del nuevo periodismo en las redes sociales. Como ejemplo de métricas que podrían hacer esto posible se encuentra la comparación entre retuits y favoritos publicados por día en un determinado periodo de tiempo (un mes, varios meses, un año, etc.), que permitiría hacer un seguimiento sobre el tipo de interacción más frecuente o sobre las épocas donde más se interactúa y la distribución de publicaciones en un mismo mes, que podría deducir patrones de uso de la red social a nivel global.

8 Bibliografía

Cabrera Méndez, M. (2017). *El Observatorio de nuevos medios*. <http://www.nuevosmedios.es/que-es/> [Consulta: 19 de mayo de 2017]

Create a simple REST web service with python. <http://www.dreamsyssoft.com/python-scripting-tutorial/create-simple-rest-web-service-with-python.php> [Consulta: 10 de abril de 2017]

Django Software Foundation (2017). *Django documentation*. <https://docs.djangoproject.com/en/1.11/> [Consulta: 21 de mayo de 2017]

Dudler, R. *Git – la guía sencilla*. <http://rogerdudler.github.io/git-guide/index.es.html> [Consulta: 20 de septiembre de 2016]

Expósito, R. (2017). 'The Little Redis Book' en Castellano. <http://raulexposito.com/documentos/redis/> [Consulta: 3 de junio de 2017]

GeoPy Contributors (2015). *Welcome to GeoPy's documentation!* <https://geopy.readthedocs.io/en/1.10.0/> [Consulta: 30 de mayo de 2017]

Google Inc. (2017). *YouTube Developer Documentation*. <https://developers.google.com/youtube/documentation/> [Consulta: 23 de septiembre de 2016]

Karambelkar, B. (2015). *How to use Twitter's Search REST API most effectively*. <https://www.karambelkar.info/2015/01/how-to-use-twitters-search-rest-api-most-effectively./> [Consulta: 22 de octubre de 2016]

MongoDB, Inc. (2017). *Manage Users and Roles*. <https://docs.mongodb.com/manual/tutorial/manage-users-and-roles/> [Consulta: 19 de abril de 2017]

MongoDB, Inc. (2017). *Welcome to the MongoDB Docs*. <https://docs.mongodb.com/> [Consulta: 2 de junio de 2017]

Oracle Corporation (2017). *MySQL Documentation*. <https://dev.mysql.com/doc/> [Consulta: 27 de mayo de 2017]

Pascual Yarritu, J.P. (2015). *Modelo de introducción de datos de obra en el SIG de la Red de Alcantarillado de Valencia mediante aplicación web con visores cartográficos incorporados*. Trabajo Final de Grado. Valencia: Universidad Politécnica de Valencia.

Redis Labs (2015). *Documentation*. <https://redis.io/documentation> [Consulta: 3 de junio de 2017]

Roesslein, J. (2009). *Tweepy Documentation*. <http://docs.tweepy.org/en/v3.5.0/> [Consulta: 25 de mayo de 2017]

Schaul, T; Bayer, J; Wierstra, D; Yi, S; Felder, M; Sehnke, F; Ruckstieb, T y Schmidhuber, J. (2010). *PyBrain*. <http://pybrain.org/> [Consulta: 27 de junio de 2017]

SQLAlchemy (2017). *Connection Pooling*.
<http://docs.sqlalchemy.org/en/latest/core/pooling.html> [Consulta: 25 de septiembre de 2016]

Stack Exchange Inc. (2017). *Learn, Share, Build*. <https://stackoverflow.com/> [Consulta: 28 de marzo de 2017]

Swartz, A. *Welcome to web.py!* <http://webpy.org/> [Consulta: 27 de mayo de 2017]

Turi, G. *Online JSON Viewer*. <http://jsonviewer.stack.hu/> [Consulta: 12 de junio de 2017]

Twitter, Inc. (2017). *API Rate Limits*. <https://dev.twitter.com/rest/public/rate-limiting> [Consulta: 25 de mayo de 2017]

Twitter, Inc. (2017). *GET followers/list*.
<https://dev.twitter.com/rest/reference/get/followers/list> [Consulta: 28 de mayo de 2017]

Twitter, Inc. (2017). *See replys to a tweet.* <https://twittercommunity.com/t/see-replys-to-a-tweet/6953> [Consulta: 6 de noviembre de 2016]

Twitter, Inc. (2017). *Tweets*. <https://dev.twitter.com/overview/api/tweets> [Consulta: 25 de mayo de 2017]

Twitter, Inc. (2017). *Twitter Developer Documentation*. <https://dev.twitter.com/docs> [Consulta: 25 de mayo de 2017]

WebMining Consultores (2017). KDD: Proceso de Extracción de conocimiento. <http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/> [Consulta: 9 de marzo de 2017]

9 Anexos

9.1 Presentación del Observatorio en la Casa de América

El pasado 17 de mayo se presentó el Observatorio de nuevos medios en la Casa de América de Madrid de la mano de su fundadora, Marga Cabrera, el director de contenidos de la iniciativa, Jorge Morillo, y por el desarrollador técnico, Pablo Pascual.

Una breve presentación en la que se destacaron los objetivos del Observatorio, los rankings que se generan, el espacio de debate en torno a la definición de nuevo medio que supone el blog, o las futuras mejoras que se pretenden introducir si se logra obtener la financiación necesaria.

El proyecto fue muy bien recibido por los asistentes y por los seguidores en Twitter a través del hashtag de la jornada, #nuevosmedioslatam, mediante el cual también se recibieron sugerencias y propuestas de mejoras que serán estudiadas para su integración y seguir así mejorando el Observatorio de nuevos medios en español.

Figura 72. Presentación del Observatorio en la Casa de América.

Con Toño Fraguas como maestro de ceremonias, y aprovechando la celebración del Día de Internet, la presentación dio paso al debate posterior en torno a los nuevos medios en español.

Juan Luis Sánchez, co-fundador y director de eldiario.es; Rafa Ruiz, co-fundador y co-director de El Asombrario; Vicente Ferrer, director de contenidos de Vice España; y Cristina Pop, redactora jefa de Código Nuevo; fueron los protagonistas de este intercambio de ideas en los que se pusieron en valor características propias de los nuevos medios, las nuevas formas de dirigirse a la audiencia, la búsqueda de nichos específicos, la desmitificación de la precariedad y amateurismo en los nuevos medios, o la relevancia del vídeo en este terreno.

9.2 Presentación del Observatorio en DataBeers

El 30 de noviembre de 2016, Valencia acogió la segunda edición de DataBeers, un evento donde los participantes pueden transmitir sus experiencias y proyectos relacionados con el mundo de los datos en un ambiente informal y distendido. Allí estuvieron Jorge Morillo, director de contenidos, y Benjamín Arroquia, desarrollador técnico, para contar a los asistentes la relación entre los nuevos medios y los datos que dan forma al Observatorio.

Figura 73. Presentación del Observatorio en la segunda edición de DataBeers.

9.3 El blog del Observatorio

El blog del Observatorio de nuevos medios es una sección de la web que incluye las diferentes entrevistas realizadas a profesionales con experiencia en el sector del periodismo. La finalidad de esta iniciativa es identificar qué es un nuevo medio y qué no. Para ello, se valoran las reflexiones y criterios de los expertos con la intención de llegar a una definición consensuada.

Todos los entrevistados tienen que responder a las mismas tres preguntas, de modo que tienen que dar su parecer acerca de qué significa un nuevo medio para ellos, qué incidencia tiene la Red sobre los mismos y qué nuevo medio destacan por su aportación.

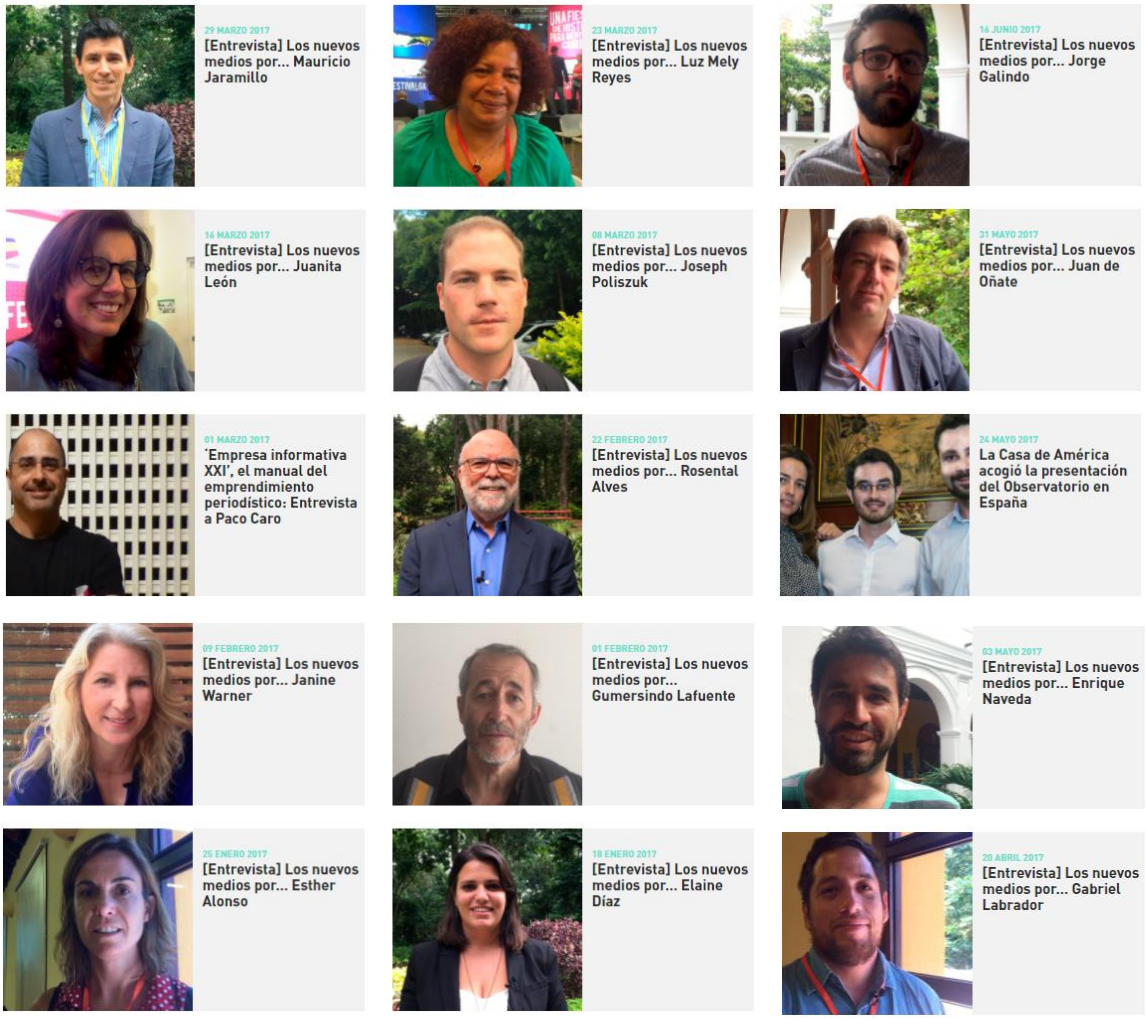

Figura 74. Entrevistas del Blog del Observatorio.