Datathon Madrid 2016

Participación en el concurso de datos abiertos

TuBarrioAlDetalle

Alejandro Fernández Carrera

Descripción del proyecto

El proyecto denominado TuBarrioAlDetalle responde a la necesidad por parte de los diferentes usuarios del portal de datos abiertos de Madrid, de visualizar de forma amistosa y sencilla los diferentes datasets que se publican en dicho portal de manera continuada.

Esta aplicación web (compatible en versión móvil y escritorio) permite a los ciudadanos ver la información agregada desde diferentes orígenes acerca de su barrio u otros barrios de Madrid y poder disfrutar de la actualización de los mismos gracias a los scripts automatizados que generan los datos finales.

Cada dataset ha sido analizado de forma específica y por ello, el trabajo ha resultado más que satisfactorio, dando lugar a datasets unificados que comparten los diferentes orígenes de datos que el portal ofrece para todos los ciudadanos o desarrolladores que deseen usarlos.

El objetivo principal del proyecto es poder dar a conocer las diferentes posibilidades de esta iniciativa cómo es la publicación de datos de forma abierta y transparente y posteriormente intentar un uso incentivado.

En este caso se ha realizado una puntuación para cada barrio de Madrid en base a diferentes KPI de cada dataset analizado. Además de poder dar la posibilidad de visualizar los recursos de forma más concreta.

Pulse aquí para ir a la aplicación web del proyecto

Valor del proyecto entregado

Las posibles formas de publicar los datos y actualización de los mismos es una pequeña hoja de doble filo que por una parte proporciona beneficios al igual que lo haría una biblioteca, plaza o sitio donde ir a buscar un determinado tipo de contenido, pero no todos ellos son del mismo tipo.

Algunos datos por sus diferentes propiedades deben ser alojados y utilizados de formas muy específicas y claras creando una barrera importante entre el usuario que posiblemente los use para un uso propio y/o aplicación concreta para un beneficio comunitario. Esta barrera es lo que normalmente denominamos ser usuario y ser usuario-experto, pero los datos y su composición no deberían definir este apelativo.

En este apartado TuBarrioAlDetalle permite romper esa barrera (mediante herramientas de automatización) y así el usuario, poder acoplarse o asimilar las alternativas que puede tener delante: utilizar los datos en un mapa, gráficos, representación semántica, sin unión con otros datasets de forma aparente, con o sin necesidad de técnicas de geocoding, etc.

Se ha realizado la limpieza y curación de datos de diferentes publicaciones del portal a modo de ejemplo de la utilidad de este tipo de proyectos y se ha plasmado toda la información en un mapa con significado propio.

Para darle la visibilidad ciudadana y no puramente técnica, se ha decidido una interpretación más llevadera con la premisa del barrio más adecuado para su visita y viabilidad en verano.

Este ejemplo ha sido elegido después del análisis de las fuentes de datos y la variedad y cantidad de los mismos. Es decir, el barrio que más variedad tenga de diversos eventos y localizaciones (prioritarias en ocasiones en meses de verano) alcanzará más o menos puntuación, siendo ésta desde un mínimo de 0 a máximo 10 puntos en el ranking visto a continuación (Fig. 1)

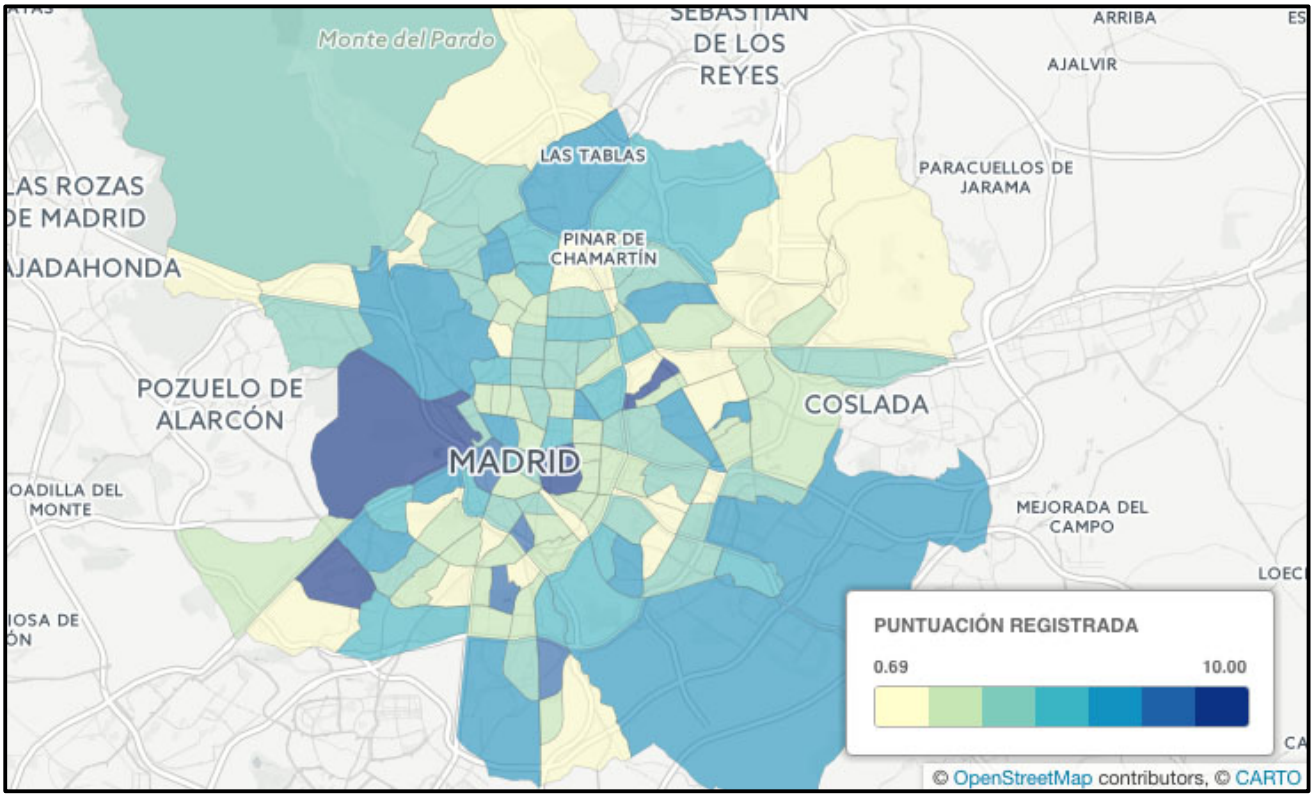

Fig. 1. Representación de los barrios (cloropeth) según puntuación obtenida

Por otra parte se han realizado dos visualizaciones básicas para la formación de la anterior por medio de herramientas de agregación de datos. Esta generación y curación de datos permite la unificación que hemos comentado anteriormente y por tanto el usuario puede ir a un nivel de detalle más amplio, ya que puede ver las características del recurso en cuestión o en su defecto navegar por el link provisto para visualizar más información añadida (Fig. 2) (Fig. 3).

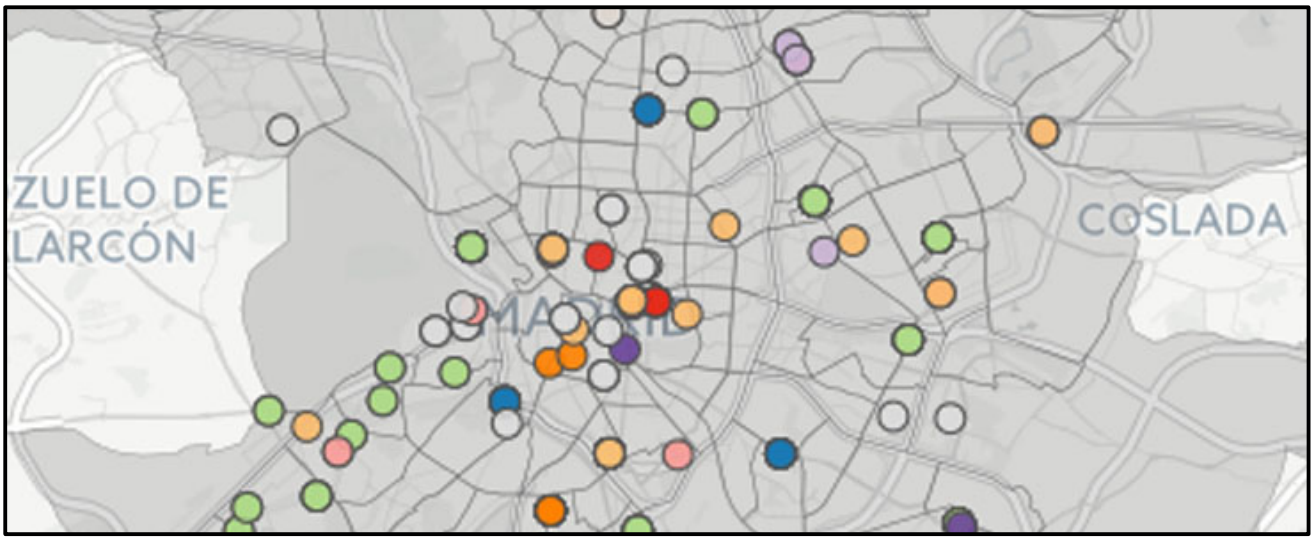

Fig. 2. Representación de los datos referidos a eventos en Madrid

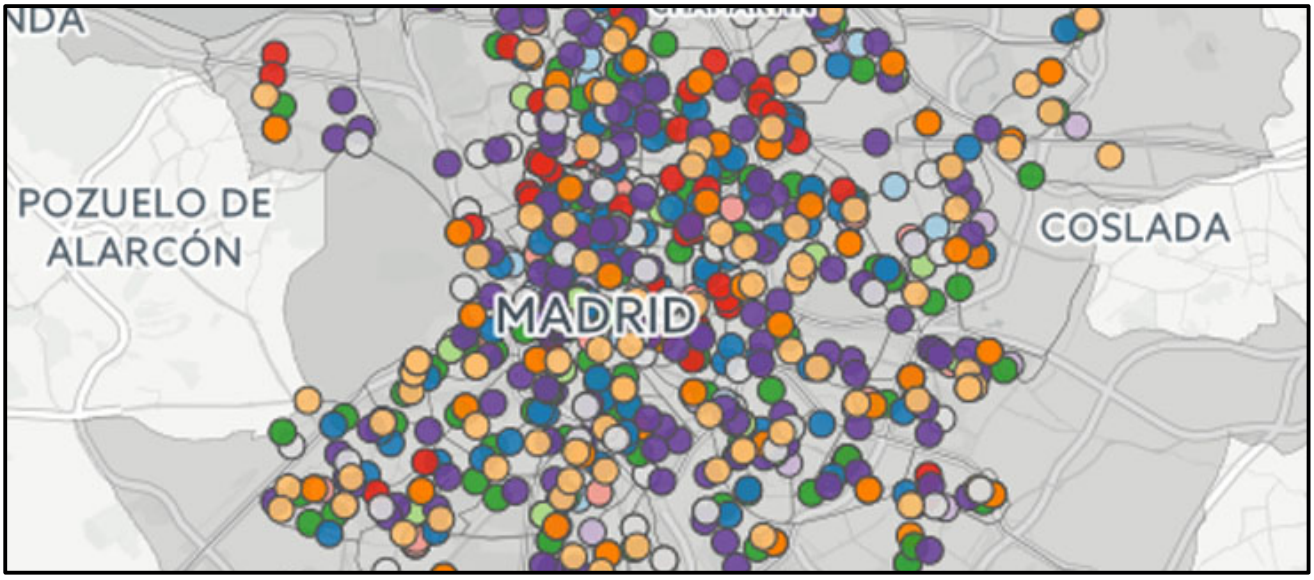

Fig. 3. Representación de los datos referidos a localizaciones en Madrid

Por último cabe destacar que cada una de las visualizaciones anteriores se han realizado a partir de datos geolocalizados, ya que como se verá en los siguientes apartados se han eliminado varios orígenes de datos por falta de tiempo o complejidad en la obtención y/o representación de los mismos.

Desarrollo del proyecto

En un primer momento se recogieron más de 60 datasets (68) con el fin de poblar la aplicación con el mayor número de datos interesantes en relación a un barrio determinado. Por determinadas causas (formato de los datos, tiempo de realización del proyecto y algunos datos no tenían geolocalización o la tenían en un determinado sistema de referencia espacial) se ha reducido a un total de 11 repartidos en dos grandes grupos: eventos y localizaciones.

Se ha utilizado una herramienta para limpieza y curación de datos masivos llamada Open Refine (anteriormente, Google Refine) que permite ejecutar de forma automatizada estas operaciones en un momento determinado gracias a su API embedida via REST. Es por ello que el proyecto tiene diferentes ficheros .json con cada una de las operaciones ejecutadas sobre los datos para su posterior extracción de forma normalizada y lo más limpia posible.

Además el script de descarga de datasets está desarrollado de forma genérica pudiendo meter tantos datasets como requiramos y sus correspondientes operaciones y/o configuraciones de creación de proyecto (si es un fichero de texto, csv, tsv, codificaciones, uso de comillas, etc)

En resumen, la parte backend del proyecto consta de una serie de configuraciones y ficheros que permiten crear proyectos en Refine para su posterior limpieza y curación de datos, obteniendo así una actualización de dichos datos (si utilizásemos un gestor de procesos en background) y una agregación en dos únicos datasets con propiedades comúnes.

Es importante comentar que a medida que se han analizado cada uno de los ficheros descargados del portal, se han encontrado diferentes errores (menores o graves) que han determinado la inclusión en esta versión del proyecto.

Pulse aquí para visualizar la carpeta con los errores obtenidos

Por otra parte, se han realizado tres datasets con sus mapas correspondientes en la infraestructura de mapas y cartografía denominada Carto.

Los dos primeros se pueden realizar de forma sencilla con la importación de los datasets generados por los scripts y su posterior personalización con las herramientas que ofrece la plataforma, pero el tercero es la intersección de los diferentes puntos encontrados y la geometría de cada uno de los barrios, dando así el número de instancias recogidas, que junto a alguna query en lenguaje SQL permite obtener las puntuaciones visualizadas en el mapa (Fig. 1).

Pulse aquí para visualizar el mapa en la plataforma Carto

Los dos primeros se pueden realizar de forma sencilla con la importación de los datasets generados por los scripts y su posterior personalización con las herramientas que ofrece la plataforma, pero el tercero es la intersección de los diferentes puntos encontrados y la geometría de cada uno de los barrios, dando así el número de instancias recogidas, que junto a alguna query en lenguaje SQL permite obtener las puntuaciones visualizadas en el mapa (Fig. 1).

NOTA: todo el código fuente del proyecto así como los scripts, configuraciones comentadas, website y en definitiva todo lo necesario para reconstruir estas visualizaciones está publicado en github de forma pública y con licencia MIT.

Datasets utilizados (11 de 68)

- Distritos y barrios de Madrid (dataset base)
- Agenda y eventos culturales
- Agenda de actividades deportivas
- Listado de polideportivos
- Piscinas municipales
- Parques y jardines
- Bibliotecas universitarias o especializadas
- Centros culturares de interés
- Parques de bomberos
- Centros de salud
- Centros de atención médica especializados
- Unidades por distrito de la policía municipal

Viabilidad y siguientes pasos

Debido a la gran magnitud y la modularización del proyecto, su viabilidad se basa en la utilización de las diferentes técnicas provistas en el mismo.

Se han desarrollado partes de código que podrían valer tanto para scrapping de datos de forma automatizada como la generación de históricos mediante validación SHA, limpieza de datos y creación para nuevos usos.

La unificación de servicios y optimización de las técnicas comentadas pueden definirse como los pasos a seguir en posteriores versiones.

Un ejemplo claro es el posible uso del dataset de calidad de aire o contaminación acústica que podría ser archivado históricamente de forma fácil con los scripts planteados y al estar geolocalizada su estación de medida, podría relacionarse con un barrio o distrito de Madrid.

En pocas palabras los siguientes pasos se podrían definir en:

- Uso de herramientas para normalizar datos geo-espaciales y así usarlos fácilmente. Ejemplos de ello: GeoKettle u ogr2ogr.
- Normalización y curación de datos de todos los datasets del portal.
- Unificación de los diferentes recursos para una mayor trazabilidad (ejemplo dado de la contaminación del aire y su barrio)
- Generación de datos unificados a gran escala cómo el proyecto dado.
- Técnicas de Geocoding para saltar la barrera de convertir una dirección a una latitud y longitud específica.
- Generación de scripts de automatización y actualización de datos para crear históricos y datasets nuevos en el portal de forma fácil y sencilla.
- Poblar las visualizaciones con diferentes perspectivas para que el usuario pueda por ejemplo buscar recursos o instancias en el mapa indicado.

TuBarrioAlDetalle es un ejemplo del potencial de este tipo de portales y una posible primera piedra para fomentar el uso del mismo y la visualización de los datos que en gran medida, el usuario no usa a veces por desconocer su existencia.