PEC1 - Web scraping

Autores: Pablo Moreno Martínez
Macarena Palomares Pastor

1. Contexto

En el año 2007, la Agencia Andaluza del Agua junto con la Universidad de Granada publica el libro **"Manantiales de Andalucía"** con el objetivo de obtener una visión integral e integradora del rico **patrimonio andaluz de manantiales y fuentes en sus diferentes dimensiones, tanto medioambiental como socioeconómica y cultural**.

En octubre de 2007, la Agencia Andaluza del Agua de la Consejería de Medio Ambiente y la Universidad de Granada pusieron en marcha el Proyecto **"Conoce tus Fuentes" de catalogación y puesta en valor de fuentes y manantiales de Andalucía**. Estas manifestaciones de agua constituyen un valioso patrimonio ambiental, socio-económico y cultural, que urge conocer y proteger, ante nuevas amenazas, como son el cambio climático y el aumento de explotación de los recursos hídricos subterráneos. **Este catálogo permitirá identificar los manantiales y fuentes vulnerables de mayor relieve ambiental, socio-económico y cultural sobre las que centrar políticas de gestión y conservación, así como proponer el adecentamiento, la rehabilitación, la recuperación o el realce arquitectónico de las fuentes más señeras, o de aquéllas con mayores aptitudes**.

Se trata de un **proyecto pionero en España de participación ciudadana** en el que su página web está abierta a todo el que quiera colaborar en la realización de dicho catálogo mediante la cumplimentación de la encuesta estándar para incorporarlas al catálogo.

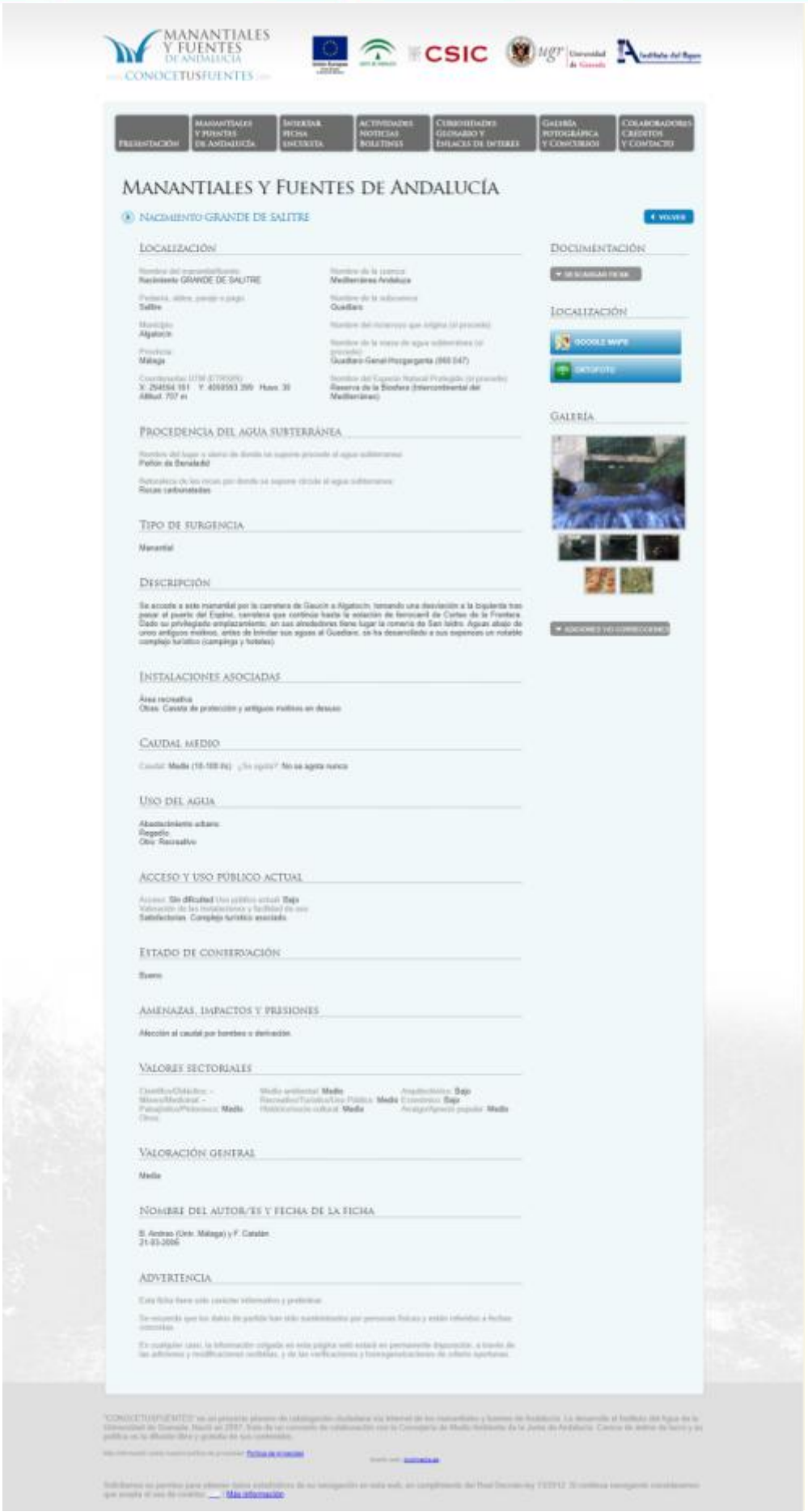
El objetivo final de dicho proyecto es además de la catalogación de dichos recursos, la conservación del agua a través de su aprecio y conocimiento.

Actualmente, el proyecto se lleva a cabo con financiación propia desde el Instituto del Agua de la Universidad de Granada, y apoyo técnico del Instituto Geológico y Minero de España (IGME), del Instituto de Estadística y Cartografía de Andalucía (IECA) y del Consejo Superior de Investigaciones Científicas (CSIC), habiéndose catalogado a día de hoy **12.812** recursos tomando como partida los 338 que se identificaron en el libro inicial publicado en 2007 "Manantiales de Andalucía".

El poder **disponer de esta información en formato digital, es de vital importancia para todas aquellas entidades que necesiten incorporar estos recursos en proyectos de planificación de diversa índole**, como pueden ser, la planificación de nuevos senderos, incorporación de elementos al

catálogo patrimonial, planificación de recursos turísticos, planificación de recursos hídricos, etc.

En la página web del proyecto http://www.conocetusfuentes.com/ se pone a disposición del público en general en formato HTML las fichas con toda la información recopilada para cada uno de los recursos catalogados. Las fichas tienen el siguiente formato:


MANANTIALES Y FUENTES DE ANDALUCÍA
CONOCETUSFUENTES
CSIC
PRESENTACIÓN
MANANTIALES Y FUENTES DE ANDALUCÍA
INTERIOR FICHA ENCUESTA
ACTIVIDADES NOTICIAS BOLETINES
CURIOSIDADES GLOSARIO Y ENLACES DE INTERÉS
GALERÍA FOTOGRÁFICA Y CONCURSOS
COLABORADORES CRÉDITOS Y CONTACTO
MANANTIALES Y FUENTES DE ANDALUCÍA
NACIMIENTO GRANDE DE SALITRE
VOLVER
LOCALIZACIÓN
Nombre del manantial/fuente: Nacimiento GRANDE DE SALITRE
Nombre de la cuenca: Mediterránea Andaluza
Paraje, aldea, pueblo o pago: Salitre
Nombre de la subcuenca: Guadiaro
Municipio: Algatocín
Nombre del río/arroyo que origina (si procede)
Provincia: Málaga
Nombre de la masa de agua subterránea (si procede): Guadiaro-Genal Hozgarganta (060.047)
Coordenadas UTM (ETRS89): X: 264564.161 Y: 4059593.289 Huso: 30 Altitud: 707 m
Nombre del Espacio Natural Protegido (si procede): Reserva de la Biosfera (Intercontinental del Mediterráneo)
PROCEDENCIA DEL AGUA SUBTERRÁNEA
Nombre del lugar o sierra de donde se supone procede el agua subterránea: Peñón de Benalauría
Naturaleza de las rocas por donde se supone circula el agua subterránea: Rocas carbonatadas
TIPO DE SURGENCIA
Manantial
DESCRIPCIÓN
Se accede a este manantial por la carretera de Gaucín a Algatocín, tomando una desviación a la izquierda tras pasar el puerto del Espino, carretera que continúa hasta la estación de ferrocarril de Cortes de la Frontera. Dado su privilegiado emplazamiento, en sus alrededores tiene lugar la romería de San Isidro. Aguas abajo de unos antiguos molinos, antes de brindar sus aguas al Guadiaro, se ha desarrollado a sus expensas un notable complejo turístico (camping y hoteles).
INSTALACIONES ASOCIADAS
Área recreativa
Otras: Cauda de protección y antiguos molinos en desuso
CAUDAL MEDIO
Caudal: Media (10-100 l/s) ¿Se agota? No se agota nunca
USO DEL AGUA
Abastecimiento urbano
Regadío
Otro: Recreativo
ACCESO Y USO PÚBLICO ACTUAL
Acceso: Sin dificultad Uso público actual: Bajo
Valoración de las instalaciones y facilidad de uso: Satisfactorias. Complejo turístico asociado.
ESTADO DE CONSERVACIÓN
Bueno
AMENAZAS, IMPACTOS Y PRESIONES
Afección al caudal por bombeo o detracción.
VALORES SECTORIALES
Científico/Didáctico: –
Medio ambiental: Medio
Arqueológico: Bajo
Minero/Medicinal: –
Recreativo/Turístico/Uso Público: Medio
Etnológico: Bajo
Paisajístico/Pintoresco: Medio
Histórico/cultural: Media
Paisaje/Ajeno popular: Medio
Otros:
VALORACIÓN GENERAL
Media
NOMBRE DEL AUTOR/ES Y FECHA DE LA FICHA
S. Andreo (Univ. Málaga) y F. Catalán
21-03-2006
ADVERTENCIA
DOCUMENTACIÓN
VER ALGUNAS FICHAS
LOCALIZACIÓN
GOOGLE MAPS
ORTOFOTO
GALERÍA
AÑADIMOS VD. CONOCE ALGO MÁS
"CONOCETUSFUENTES" es un proyecto pionero de catalogación divulgativa vía Internet de los manantiales y fuentes de Andalucía. Lo desarrolla el Instituto del Agua de la Universidad de Granada. Nació en 2007, fruto de un convenio de colaboración con la Consejería de Medio Ambiente de la Junta de Andalucía. Conoce de dónde beben y su política de la difusión libre y gratuita de sus contenidos.
Más información sobre nuestra política de privacidad: Política de privacidad
Diseño web: autentica.es
Solicitamos su permiso para obtener datos estadísticos de su navegación en esta web, en cumplimiento del Real Decreto-ley 13/2012. Si continua navegando consideramos que acepta el uso de cookies. OK | Más información

Por otra parte, desde el Instituto de Estadística y Cartografía (IECA) de la Junta de Anadalucía se pone a disposición de los usuarios un servicios *WFS* de *Geoserver* dónde se puede obtener en formato *JSON* la información de los elementos catalogados: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/geoserver-ieca/conocetusfuentes/wfs?request=GetFeature&outpuFormat=json&typeName=fuentesymanantiales

El código *JSON* para la ficha mostrada en el ejemplo anterior es:


```
<wfs:member>
  <conocetusfuentes:fuentesymanantiales gml:id="fuentesymanantiales.368">
    <conocetusfuentes:gid>43770</conocetusfuentes:gid>
    <conocetusfuentes:municipio>Algatocín</conocetusfuentes:municipio>
    <conocetusfuentes:cod_ine>29006</conocetusfuentes:cod_ine>
    <conocetusfuentes:provincia>Málaga</conocetusfuentes:provincia>
    <conocetusfuentes:nombre>NACIMIENTO GRANDE DE SALITRE</conocetusfuentes:nombre>
    <conocetusfuentes:tipo>Manantial</conocetusfuentes:tipo>
    <conocetusfuentes:tipo_ent>h12</conocetusfuentes:tipo_ent>
    <conocetusfuentes:cuenca>Mediterránea Andaluza</conocetusfuentes:cuenca>
    <conocetusfuentes:coor_x>294594.000000000000000</conocetusfuentes:coor_x>
    <conocetusfuentes:coor_y>4050593.000000000000000</conocetusfuentes:coor_y>
    <conocetusfuentes:huso>30</conocetusfuentes:huso>
    <conocetusfuentes:coor_x_30>294594.000000000</conocetusfuentes:coor_x_30>
    <conocetusfuentes:coor_y_30>4050593.000000000</conocetusfuentes:coor_y_30>
<conocetusfuentes:enlace_web>http://www.conocetusfuentes.com/datos_fuente_368.html</conocetusfuentes:enlace_web>
<conocetusfuentes:enlace_fot>http://www.conocetusfuentes.com/images/fuente_1_368.jpg</conocetusfuentes:enlace_fot>
    <conocetusfuentes:geom>
      <gml:Point srsName="urn:ogc:def:crs:EPSG::25830" srsDimension="2">
        <gml:pos>294594 4050593</gml:pos>
      </gml:Point>
  </conocetusfuentes:geom>
  </conocetusfuentes:fuentesymanantiales>
</wfs:member>
```


Como podemos ver en el *JSON* sólo se obtiene la siguiente información para cada elemento catalogado: gid, municipio, cod_ine, provincia, nombre, tipo, tipo_ent, cuenca, coor_x, coor_y, enlace_web, enlace_fot.

Toda esta información se incluye en las fichas HTML, como la que se ha mostrado en el ejemplo, con la particularidad de que en las fichas HTML se incluye además mucha otra información que aquí no aparece. Además, las coordenadas que se encuentran en la ficha tienen decimales y por tanto, tienen una localización más precisa que la almacenada en el JSON, que están redondeadas y pueden dar lugar a pequeños desplazamientos en la ubicación de los recursos.

Debido a esto, se decide **obtener toda la información posible de los elementos catalogados utilizando únicamente para ello como origen las fichas HTML** y descartando así los datos provenientes del servicio WFS del IECA.

2. Título

Como título para el dataset se elige el nombre del proyecto del que se obtiene la información, junto con

el mes y el año de la extracción: "**conocetusfuentes_04_22**". Se decide poner el mes y el año de la extracción, ya que al ser un proyecto vivo de colaboración ciudadana los datos probablemente irán cambiando con el transcurso del tiempo y así queda identificado el momento de la recogida de la información.

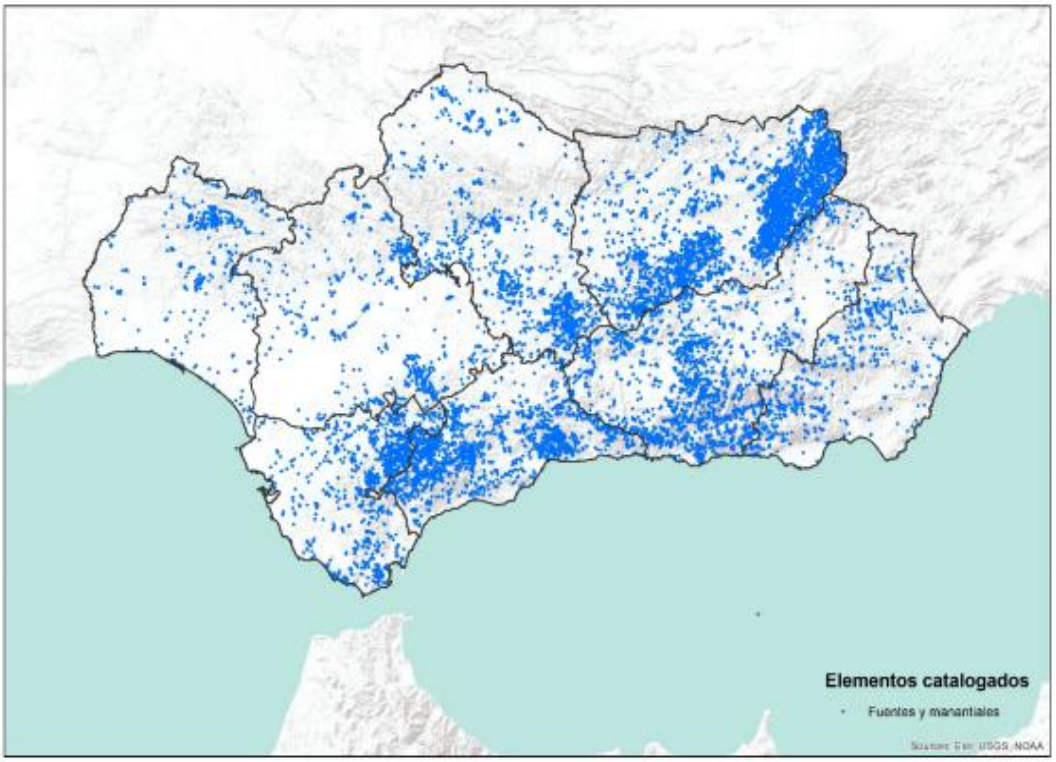
3. Descripción del dataset

Aunque en la página web aparece que hay registradas 12.812 sugerencias, se han extraído datos de 12.939 sugerencias cuyas fichas existían dentro del proyecto "Conoce tus fuentes" a día 3 de abril de 2022, proyecto pionero de participación ciudadana para la incorporación de elementos al catálogo. Para cada una de las surgencias catalogadas, se ha extraído información detallada sobre su localización, procedencia del agua subterránea, tipo de surgencia, descripción, instalaciones asociadas, caudal medio, uso del agua, acceso y uso público actual, estado de conservación, amenazas, impactos y presiones, descripción hidrogeológica, descripción arquitectónica, antecedentes históricos, aspectos culturales y etnográficos, otra información, valores sectoriales, información general, autor y fecha.

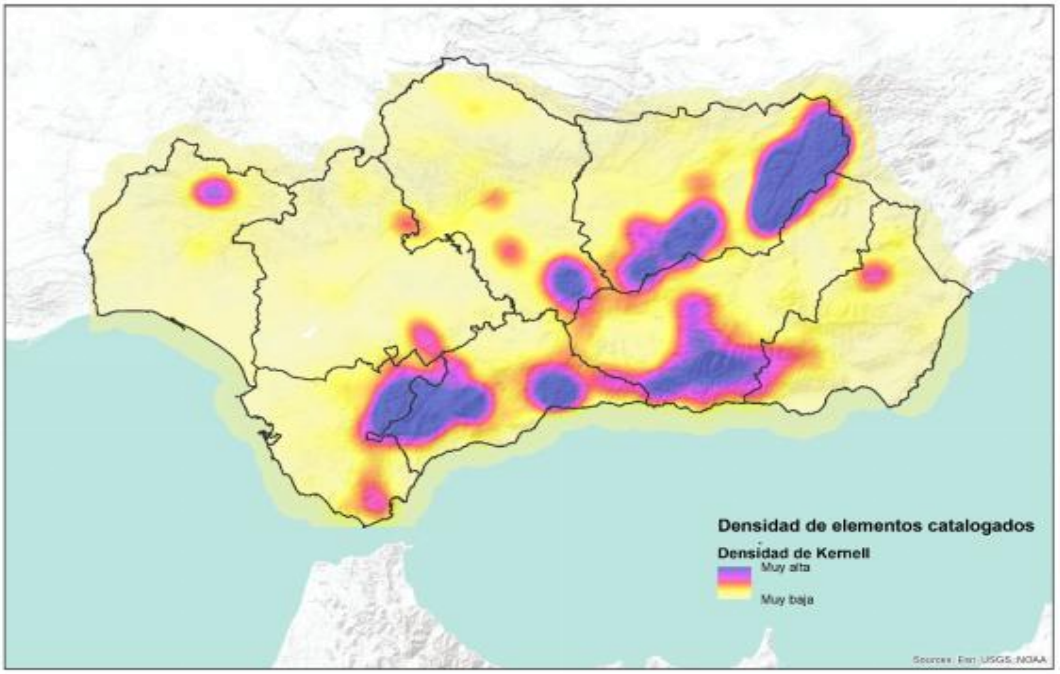
Es preciso señalar que existe un catálogo adicional categorizado como "Otros puntos de interés catalogados", de los cuales no se ha extraído información puesto que no son relevantes para el estudio.

4. Representación gráfica

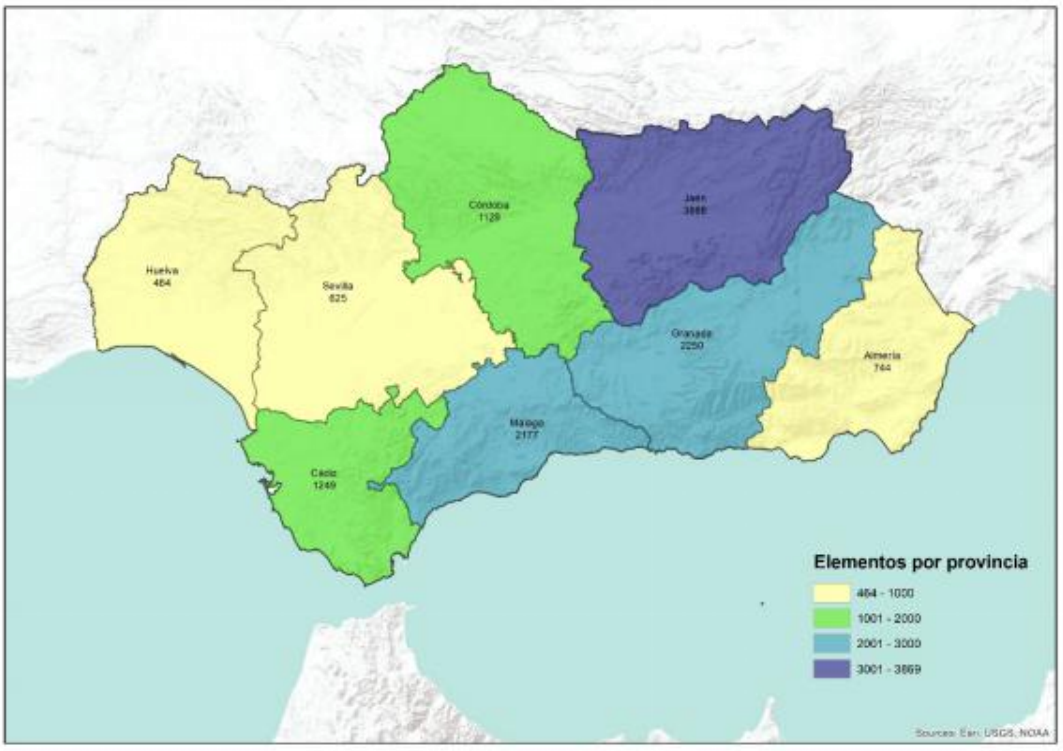
Dado que la información obtenida es geográfica puesto que tiene sus coordenadas X e Y, para representarla se seleccionan una serie de mapas dónde se va a visualizar su ubicación:



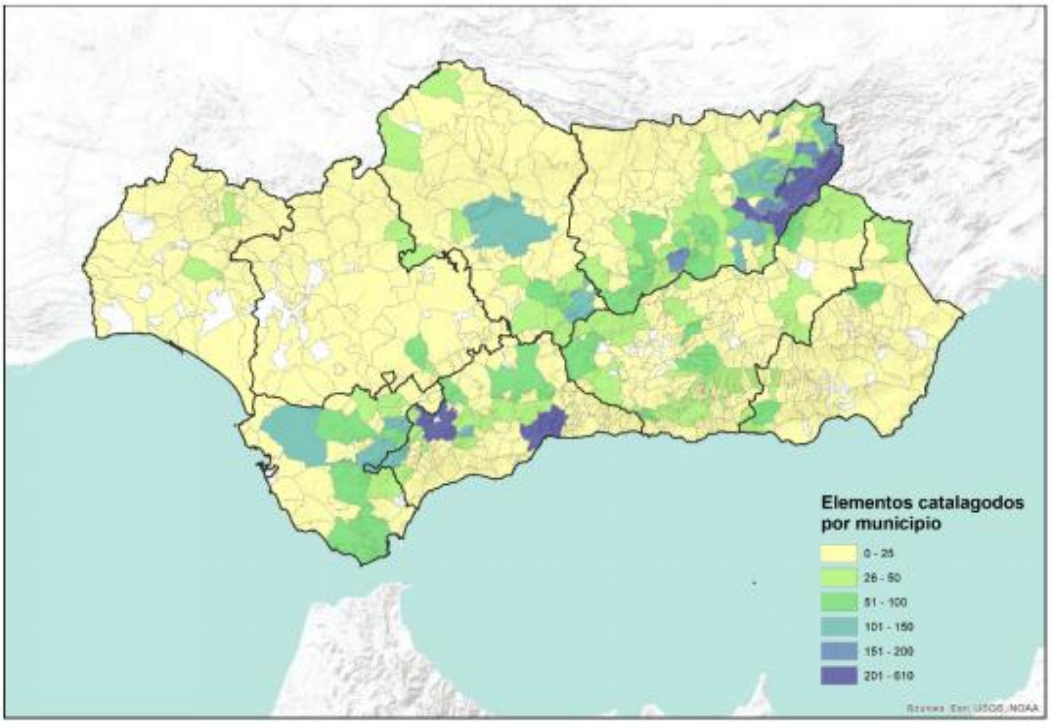
Localización de elementos catalogados



Densidad de Kernell de elementos catalogados



Número de elementos catalogados por provincia



Número de elementos catalogados por municipio

5. Contenido

Atributos que incluye el dataset:

Campo	Tipo	Descripción
Url	String	Enlace a la página HTML del catálogo
Nombre	String	Nombre del elemento
Otros_nombres	String	Otros nombres del elemento
Pedania	String	Nombre de la pedanía en la que se encuentra
Municipio	String	Nombre del municipio en el que se encuentra
Provincia	String	Provincia en la que se encuentra
Coordenada_x	Float	Coord. X en UTM (ETRS_89)
Coordenada_y	Float	Coord. Y en UTM (ETRS_89)
Cuenca	String	Nombre de la cuenca

Campo	Tipo	Descripción
Subcuenca	String	Nombre de la subcuenca
Río	String	Nombre del río/arroyo que origina (si procede)
Masa_agua	String	Nombre de la masa de agua subterránea (si procede)
ENP	String	Nombre del Espacio Natural Protegido (si procede)
Lugar	String	Nombre del lugar o sierra de dónde procede el agua subterránea
Naturaleza	String	Naturales de las rocas por dónde se supone que circula el agua subterránea
Tipo	String	Tipo de surgencia
Descripcion	String	Descripción
Instalaciones_asociadas	String	Instalaciones asociadas
Caudal	String	Caudal

Campo	Tipo	Descripción
Se_agota	String	¿Se agota?
Uso_agua	String	Uso del agua
Acceso	String	Acceso
Uso_publico	String	Uso público actual
Valoración_acceso	String	Valoración de las instalaciones y facilidad de uso
Conservacion	String	Estado de conservación
Amenazas	String	Amenazas, impactos y presiones
Descripcion_hidrogeologica	String	Descripción hidrogeológica
Descripcion_arquitectonica	String	Descripción arquitectónica
Antecedentes_historicos	String	Antecedentes históricos
Aspectos_culturales	String	Aspectos culturales y etnográficos

Campo	Tipo	Descripción
Otra_informacion	String	Otra información adicional
Científico	String	Valor sectorial científico/didáctico
Minero	String	Valor sectorial minero/medicinal
Paisajistico	String	Valor sectorial paisajístico/pintoresco
Otros	String	Valor sectorial otros
Medioambiental	String	Valor sectorial medio-ambiental
Recreativo	String	Valor sectorial recreativo/turístico/uso público
Historico	String	Valor sectorial histórico/socio-cultural
Arquitectonico	String	Valor sectorial arquitectónico
Economico	String	Valor sectorial económico
Arraigo	String	Valor sectorial arraigo/aprecio popular

Campo	Tipo	Descripción
Valoracion	String	Valoración general
Autor	String	Nombre del autor
Fecha	Date	Fecha de la ficha
Imagenes	String	Enlaces de las imágenes del elemento

Los datos que se recogen son desde octubre del 2007, que se pone en marcha el proyecto, hasta abril de 2022 que se descargan.

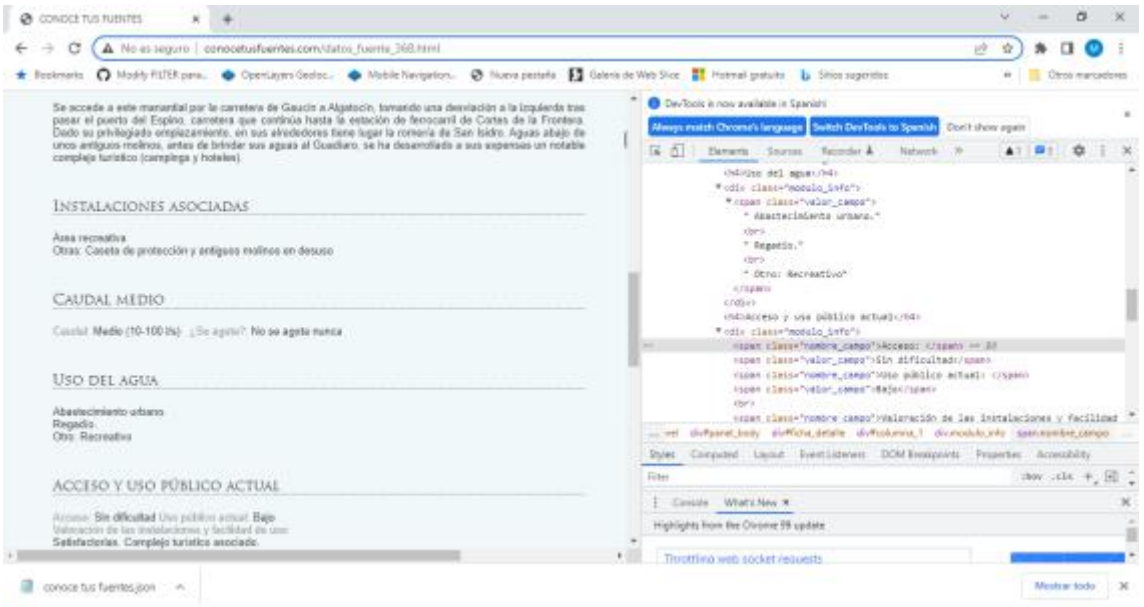
La recopilación de dicha información por parte de los promotores del proyecto, se ha realizado mediante participación ciudadana a través de unos formularios/encuestas que se encuentran disponibles en su web.

Como paso previo a la recopilación, se ha realizado una evaluación inicial con la que se han obtenido las siguientes conclusiones:

- El sitio web no dispone de archivo robots.txt ni de Sitemap.
- La estimación del tamaño web con google es de 22.700 resultados, lo cual es coherente con la cantidad de páginas extraídas.
- Las tecnologías identificadas en el sitio web a través de la librería *WebTech* de *Python* son: Apache, PHP y Youtube.
- El propietario del sitio web, según la función *whois* es: Agencia Andaluza del Agua - Consejería de Medio Ambiente.

La recopilación de dicha información por nuestra parte se ha realizado mediante técnica de web scraping utilizando la librería *Beautiful Soup* de *Python* sobre las fichas *HTML* de los elementos catalogados que tienen publicada, ya que como se comentó anteriormente descartamos el utilizar el fichero *JSON* del servidor *geoserver* del *IECA* por disponer de menos información y más imprecisa que la que hay en estas páginas.

En primer lugar, se ha analizado la estructura de varios HTML de fichas de los elementos, observando que los valores que teníamos que recuperar se encontraban dentro de etiquetas * /* cuya propiedad *class='valor_campo'* y que los títulos de los distintos apartados se encuentran en etiquetas *<h4>*, y los nombres de los campos en etiquetas * /* con el atributo *class='nombre_campo'*.



Por otra parte, también se detectó de que hay algunas fichas que tienen más campos de información que el resto, como, por ejemplo, http://www.conocetusfuentes.com/datos_fuente_406.html, que incorpora en su ficha un nuevo apartado denominado "Otra información".

Así, como campo detectado que no aparece en todas las fichas tenemos "*Otros nombres conocidos*" en

el apartado "*Localización*". Y como apartados que pueden aparecer o no en las fichas tenemos, en este orden de aparición: "*Descripción hidrogeológica*", "*Descripción arquitectónica*", "*Antecedentes históricos*", "*Aspectos culturales y etnográficos*" y "*Otra información*". Aunque el resto de apartados aparecen en todas las fichas, se ha diseñado el web scraping con idea de que en alguna ficha nueva se pueda omitir alguno, aunque siempre tendrán que mantener el mismo orden de aparición.

Ante estos cambios de estructura entre unas y otras fichas, nos hemos visto obligados a definir como estructura base la que contiene todos los apartados con toda la información y a recorrer la estructura completa de la página HTML de cada uno de los elementos catalogados para poder insertar la información de cada atributo en su lugar correspondiente y dejar vacíos aquellos valores de los que no tenga información.

Ha sido necesario crear un pequeño diccionario de datos para parsear los valores de los campos y eliminar aquellos caracteres especiales que podrían causarnos problemas en la extracción.

Se ha añadido un pequeño retraso de 1 segundo entre las peticiones a los HTML de los elementos catalogados para evitar posibles bloqueos por parte del servidor de origen. Se ha descargado la información completa sin que se hayan producido bloqueos por su parte.

Toda esta información se ha almacenado en un CSV de salida que ha sido publicado en Zenodo, como se indicará más adelante.

6. Agradecimientos

En el documento denominado "El proyecto Conoce Tus Fuentes: Cuatro años dando a conocer los manantiales y fuentes de Andalucía" cuyos autores son "Luis Sánchez-Díaz, Virginia María Robles-Arenas, Antonio Castillo y José María Fernández-Palacios" que podemos encontrar en la siguiente URL: http://www.conocetusfuentes.com/documentos/doc_89.pdf, en la página 5 del documento se incluye la siguiente imagen:

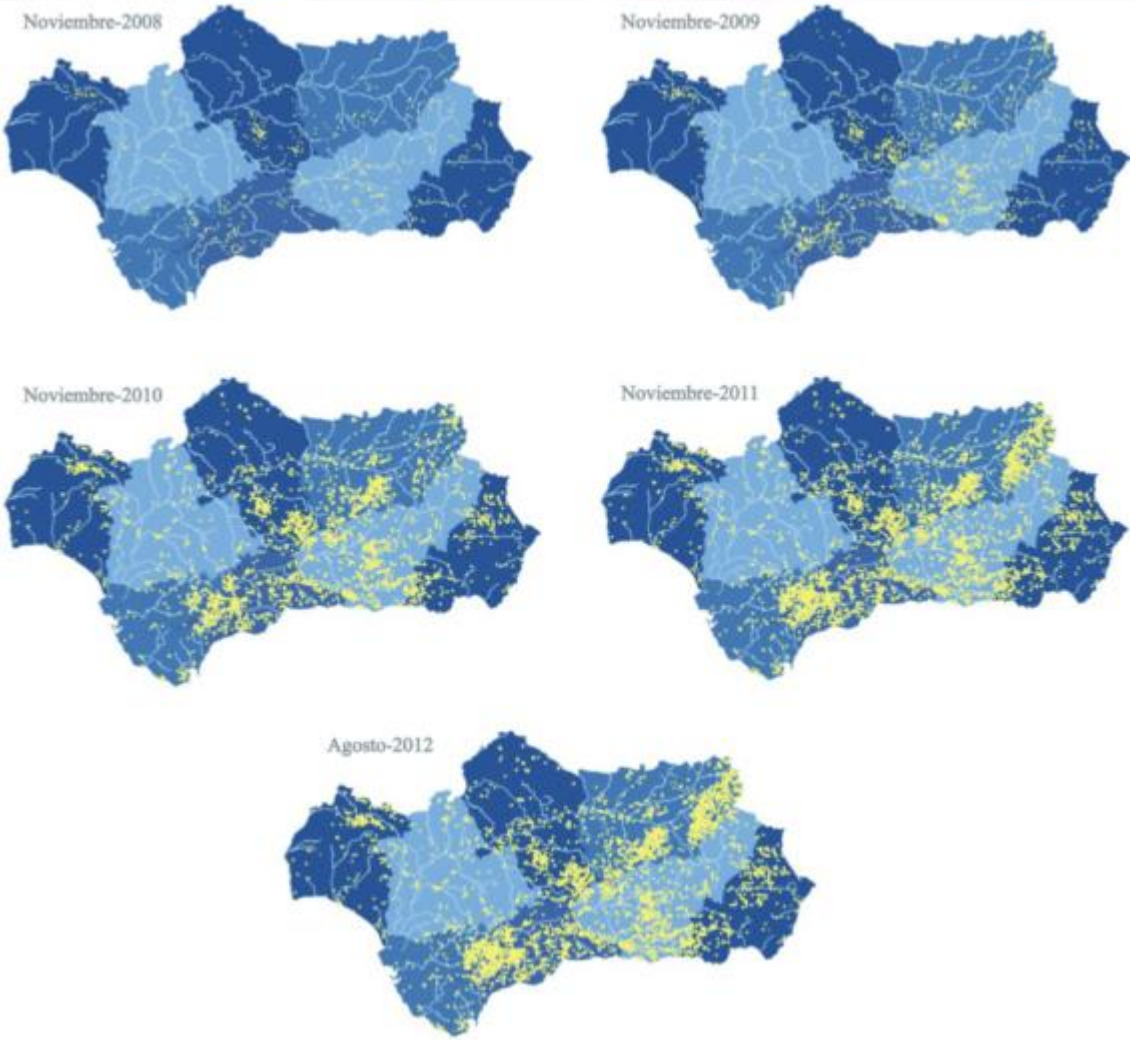


Figura 2. Evolución de la distribución espacial de las surgencias inventariados en Andalucía desde noviembre de 2008 a agosto de 2012

Dónde se muestra cómo fue evolucionando el número de sugerencias catalogadas desde noviembre de 2008 a agosto de 2012.

También se incorpora una tabla resumen dónde se indican el número de sugerencias por provincia y los municipios de cada provincia donde más se han catalogado:

Provincia	nº de fichas	Municipio destacado	nº de fichas
Almería	467	Oria	61
Cádiz	349	Ubrique	58
Córdoba	588	Priego de Córdoba	75
Granada	1.203	Otívar	70
Huelva	298	Aracena	22
Jaén	1.608	Santiago-Pontones	175
Málaga	1.301	Ronda	235
Sevilla	236	Cazalla de la Sierra	18

Tabla 1. Numero de fichas catalogadas por provincias, junto al municipio de cada uno de ellas con mayor representación

De acuerdo con los términos y condiciones del sitio web, los contenidos son de difusión libre y gratuita. Además, como no se dispone de protocolo de exclusión de robots, hay libertad de acceso a las páginas web. Aún con todo esto, se han seguido los consejos de mejores prácticas en web scraping, como el espaciado de peticiones al servidor y la modificación del *user agent.*

7. Inspiración

Los manantiales tienen una especial importancia en la hidrología subterránea, son fundamentales en la interrelación que presentan las aguas subterráneas y superficiales, abastecen de agua a usos urbanos y agrícolas, crean zonas de importante interés medioambiental, paisajístico y recreativo, dan lugar a nacimientos fluviales, etc., por lo que, es importante llevar a cabo una buena caracterización y estudio de dichos puntos de agua.

Por otro lado, las fuentes de agua constituyen un recurso público con el que abastecer a la población de agua potable. A medida del crecimiento de la población y las ciudades, la demanda de sistemas públicos de agua y las nuevas tecnología de tratamiento y suministro de agua hicieron que aumentara el uso de las fuentes públicas. Sin embargo, en las últimas décadas, han ido desapareciendo de los espacios públicos por varias razones, como la llegada de agua embotellada, los riesgos sanitarios de las fuentes y la disminución de la inversión pública en infraestructuras urbanas.

El conjunto de datos recogido pretende actuar como una fuente de datos actualizada de los elementos catalogados en el sitio web, de forma que se puede evaluar la evolución y distribución por provincias y municipios. Asimismo, disponer de los datos en forma de texto sin formato facilita su organización y manipulación para todas aquellas entidades que no tengan acceso a otro medio de exportación de los datos del sitio web.

De los datos obtenidos, podemos mostrar el número de recursos que se han ido catalogando por cada año, quedando la distribución de la siguiente forma:

A continuación, se muestra la evolución de la catalogación de recursos por series anuales por provincias:

Como se observa claramente, Jaén es la provincia en la que más recursos se han catalogado durante este periodo de tiempo, siendo además la que cuenta con más números de recursos catalogados.

La tabla resumen de los municipios por provincia dónde más surgencias se han catalogado a fecha actual quedaría de la siguiente forma:

Provincia	**Nº de recursos**	**Municipio**	**Nº de recursos**	**Porcentaje**
Almería	764	Orja	85	11,13%
Cádiz	1.303	Ubrique	129	9,90%

Provincia	Nº de recursos	Municipio	Nº de recursos	Porcentaje
Córdoba	1.161	Priego de Córdoba	129	11,11%
Granada	2.338	Loja	76	3,25%
Huelva	486	Valverde del Camino	38	7,82%
Jaén	3.980	Santiago-Pontones	634	15,93%
Málaga	2.217	Málaga	304	13,71%
Sevilla	690	Alcalá de Guadaira	44	6,38%

Con respecto a los datos de 2012 presentados en el documento anteriormente mencionado, encontramos, que el número de recursos por provincia ha aumentado por provincia de la siguiente forma:

Provincia	Nº recursos 2016	Nº recursos 2022	Variación
Almería	467	764	297
Cádiz	349	1.303	954
Córdoba	588	1.161	573
Granada	1.203	2.338	1.135
Huelva	298	486	188
Jaén	1.608	3.980	2.372
Málaga	1.301	2.217	916
Sevilla	236	690	454

8. Licencia

Se ha escogido como licencia **CC BY-NC-SA 4.0 License**. Esto permite compartir y adaptar el material de forma libre, siempre y cuando se reconozca adecuadamente la autoría, y no tenga fines comerciales.

9. Código

Se ha creado un repositorio público en GitHUB con la documentación del proyecto, cuya URL de acceso es: https://github.com/mpalomarespastor/Tipologia_y_ciclo_de_vida_PRA1

10. Database

Se ha publicado en Zenodo el Dataset obtenido:

DOI: 10.5281/zenodo6425641

11. Vídeo

El enlace al vídeo es el siguiente:

https://drive.google.com/file/d/12Q37fd1OqgNho_y9UEL0cNcXhk_sce0I/view?usp=sharing

12. Contribuciones

Contribuciones	Firma
Investigación previa	PMM, MPP
Redacción de las respuestas	PMM, MPP
Desarrollo del código	PMM, MPP