

Integración de Datos Medioambientales procedentes de Open Data mediante Preprocesado con R

Pavel H. Llamocca, Victoria López

Faculty of Informatics, Complutense University, Madrid, Spain
Profesor García Santesmases, s/n,
28040 Madrid, Spain

Pavel_Harold@hotmail.com, vlopez@fdi.ucm.es

Abstract—Este artículo describe el trabajo de investigación realizado en integración de datos procedentes de diversas fuentes Open Data. Concretamente se describe una fase preliminar con datos medioambientales. Se ponen de manifiesto las dificultades en la integración de los miles de datasets disponibles a pesar del interés en su análisis dado el gran valor que los resultados suponen para la ciudadanía, especialmente para el desarrollo de las smartcities. Los procesos de integración desarrollados en la actualidad son insuficientes debido a la dependencia de un analista de datos. En este trabajo se presenta una alternativa para la integración automática de diferentes datasets actualizados en tiempo real mediante preprocesado con R. Como resultado se obtienen bases de datos preparadas para su análisis y visualización en tiempo real.

Keywords—Open data, R language, data integration, environmental data

1 INTRODUCCION

A día de hoy una de las tendencias de muchos gobiernos a nivel mundial dentro del concepto de Open Government es la de hacer disponibles los datos recolectados con carácter público [1] para que el ciudadano pueda acceder a ellos de forma totalmente libre [2]. El motivo que lleva a los Gobiernos a realizar esto es una intención de ofrecer transparencia sobre sus procedimientos administrativos ante el ciudadano [3]. Incluso existe cierto compromiso por parte de los Gobiernos en hacer públicos los datos que han sido recolectados por instituciones estatales, ya que dichas instituciones subsisten gracias al aporte de los ciudadanos mediante el pago de tributos e impuestos [4]. Es así pues que el ciudadano común puede usar estos datos sin preocupación de estar transgrediendo algún derecho de Copyright. Es más, el ciudadano puede sentirse en la libertad de usar estos datos para crear servicios/aplicaciones que puedan ser de uso general para el resto de ciudadanos siempre y cuando estos servicios y aplicaciones sean también gratuitos, es decir aplicando el concepto de Open Data.

Uno de los pilares del Open Data es darle un valor y utilidad a los datos que ya nuestros Gobiernos nos proporcionan [5], sin embargo, dado el gran volumen y la cantidad de datasets que pueden estar disponibles conteniendo diferente información, nace la imperiosa necesidad de implementar soluciones con las que se pueda aprovechar esos datos, recopilarlos e integrarlos con el fin de crear aplicaciones y servicios.

En España, aunque el concepto de Open Data está en una etapa de inicio [6,7] se está extendiendo cada día más, es así que muchas Comunidades Autónomas están publicando datos que ya tienen digitalizados. La índole de la información publicada por las instituciones encargadas de publicar los datasets es demasiado vasta, pero se puede correlacionar y obtener muchos análisis y conclusiones interesantes.

Por ejemplo, podemos obtener los datos de los accidentes de tráfico y la cantidad de lluvias que hubo en los últimos años. La relación entre ambos indicadores puede fácilmente conducir a conclusiones interesantes. Sin embargo, para poder realizar este análisis, debemos tener en cuenta varios factores: si los datos proceden de fuentes distintas, el formato original de los datos, etc. Para garantizar un análisis correcto, será necesario aplicar procesos de cleaning e integración. Estos procesos pueden ser muy complejos en función del volumen y la diversidad de los datos.

A día de hoy aún no se han definido estándares internacionales para publicación de datos abiertos [8], a pesar de estar ampliamente reconocido el interés de la reutilización de los datos por parte de gobiernos e instituciones. Con estos estándares, el proceso de integración no se podría evitar, pero la complejidad de este se reduciría considerablemente, lo que es especialmente interesante cuando se trata de trabajar con datos en tiempo real ya que la mayoría de las herramientas para integración de datos requieren de la supervisión de un analista [11,12].

El primer propósito de este trabajo es mostrar un procedimiento mediante el cual se puede recopilar e integrar diferentes conjuntos de datos abiertos relativos al medio ambiente dentro de España. Un segundo propósito es crear una aplicación que dé servicio a los usuarios de este tipo de datos una vez procesados para su análisis según demanda.

2 DATA SETS AMBIENTALES EN TIEMPO REAL EN ESPAÑA

En este trabajo estamos interesados en facilitar el análisis de datos medio ambientales. Inicialmente hemos identificado varias fuentes de datos oficiales dentro del territorio español. Actualmente hay varias comunidades que han creado portales Open Data donde publicar sus datasets[9]. Como se apuntó en la sección anterior, el objetivo de este trabajo es el desarrollo de un servicio de consulta y análisis de datos ambientales en tiempo real lo que reduce considerablemente el número de fuentes de datos, ya que muchas administraciones no actualizan sus datos en tiempo real y en cualquier caso, los formatos no coinciden entre las distintas administraciones al no aplicarse estándares de Open Data. Para este trabajo hemos optado por trabajar con datos de la Comunidad Autónoma de Madrid y con datos de la Junta de Andalucía.

2.1 Datos de la Comunidad de Madrid

Actualmente la Comunidad de Madrid está poniendo a disposición, un conjunto de datos bajo el título de “Calidad del Aire en Tiempo Real”. Estos datos, cumpliendo los Requisitos técnicos de Open Data, viene en un formato “machine readable” [10]. En este caso la información viene en un fichero “csv”.

El fichero contiene 57 columnas y se actualiza entre los minutos 20 y 30 de cada hora. El fichero está disponible en la web de datos de la Comunidad de Madrid, en la siguiente url: "<http://www.mambiente.munimadrid.es/opedata/horario.txt>".

Descripción de las columnas:

Col	Descripción	Long.
1	Comunidad Autónoma	2
2	Ciudad	3
3	Estación	3
4	Indicador	2
5	Técnica Analítica	2
6	Periodo de Análisis	2
7	Año	4
8	Mes	2
9	Día	2
10	Valor Medición a la hora 01:00	5
11	Verificación de VM 01:00 (V/N)	1
...	Valor Medición a la hora X	5
...	Verificación de la hora X (V/N)	1
56	Valor Medición a la hora 24:00	5
57	Verificación de la hora 24:00 (V/N)	1

Table 1. Estructura Fichero Comunidad de Madrid

Estos datos se publican en Tiempo Real, la actualización de los datos se realizada cada hora. En el fichero solo vendrán informados los Valores de Medición menores o iguales a la hora actual X, las columnas correspondientes a horas mayores que X se informarán con Valores de Medición a 0 y la Verificación a “N”. La url del fichero siempre será la misma: "<http://www.mambiente.munimadrid.es/opedata/horario.txt>".

La siguiente imagen muestra un ejemplo del fichero de Datos Ambientales en Tiempo Real de la Comunidad de Madrid a fecha-hora: 25-09-2015 06:30:00.

```

28,079,099,01,38,02,2015,09,25,00006,V,00006,V,00006,V,00006,V,00005,V,00007,V,00000,N,00000,N
28,079,099,06,48,02,2015,09,25,00031,V,00031,V,00031,V,00031,V,00032,V,00032,V,00032,V,00000,N,00000,N
28,079,099,07,08,02,2015,09,25,00007,V,00007,V,00003,V,00002,V,00003,V,00005,V,00000,N,00000,N
28,079,099,08,08,02,2015,09,25,00034,V,00035,V,00028,V,00018,V,00016,V,00032,V,00000,N,00000,N
28,079,099,09,47,02,2015,09,25,00008,V,00008,V,00007,V,00005,V,00006,V,00000,N,00000,N,00000,N
28,079,099,10,47,02,2015,09,25,00013,V,00013,V,00011,V,00009,V,00009,V,00000,N,00000,N,00000,N
28,079,099,12,08,02,2015,09,25,00045,V,00045,V,00033,V,00021,V,00020,V,00037,V,00000,N,00000,N
28,079,099,14,06,02,2015,09,25,00043,V,00036,V,00043,V,00055,V,00056,V,00000,N,00000,N,00000,N
28,079,099,20,59,02,2015,09,25,0020,V,0021,V,0020,V,0018,V,0016,V,0016,V,0016,V,00000,N,00000,N,00000,N
28,079,099,30,59,02,2015,09,25,00033,V,00033,V,00033,V,00033,V,00033,V,00033,V,00000,N,00000,N,00000,N
28,079,099,35,59,02,2015,09,25,00002,V,00002,V,00002,V,00002,V,00002,V,00002,V,00000,N,00000,N,00000,N
28,079,099,42,02,02,2015,09,25,0150,V,0153,V,0148,V,0147,V,0153,V,00000,N,00000,N,00000,N
28,079,099,43,02,02,2015,09,25,0131,V,0133,V,0131,V,0131,V,0131,V,0136,V,00000,N,00000,N,00000,N
28,079,099,44,02,02,2015,09,25,0019,V,0020,V,0017,V,0016,V,0016,V,0018,V,00000,N,00000,N,00000,N
28,079,099,80,98,02,2015,09,25,00001,V,00001,V,00001,V,00001,V,00001,V,00000,N,00000,N,00000,N
28,079,099,81,98,02,2015,09,25,0107,V,0107,V,0105,V,0107,V,0105,V,0033,V,00000,N,00000,N,00000,N
28,079,099,82,98,02,2015,09,25,00134,V,00149,V,00171,V,00102,V,00201,V,00182,V,00000,N,00000,N,00000,N
28,079,099,83,98,02,2015,09,25,0205,V,0197,V,0195,V,0194,V,0181,V,0165,V,00000,N,00000,N,00000,N
28,079,099,85,98,02,2015,09,25,00020,V,00020,V,00020,V,00020,V,00020,V,00019,V,00000,N,00000,N,00000,N
28,079,099,86,98,02,2015,09,25,00033,V,00035,V,00035,V,00035,V,00038,V,00064,V,00000,N,00000,N,00000,N
28,079,099,87,98,02,2015,09,25,00936,V,00937,V,00937,V,00937,V,00937,V,00000,N,00000,N,00000,N
28,079,099,88,98,02,2015,09,25,00011,V,00011,V,00011,V,00011,V,00011,V,00000,N,00000,N,00000,N
28,079,099,89,98,02,2015,09,25,00000,V,00000,V,00000,V,00000,V,00000,V,00000,N,00000,N,00000,N
28,079,004,01,38,02,2015,09,25,00006,V,00007,V,00006,V,00006,V,00006,V,00006,V,00000,N,00000,N,00000,N
28,079,004,06,48,02,2015,09,25,00033,V,00033,V,00033,V,00033,V,00033,V,00033,V,00000,N,00000,N,00000,N
28,079,004,07,08,02,2015,09,25,00006,V,00008,V,00004,V,00002,V,00002,V,00005,V,00000,N,00000,N,00000,N
28,079,004,08,08,02,2015,09,25,00046,V,00054,V,00037,V,00024,V,00020,V,00032,V,00000,N,00000,N,00000,N
28,079,004,12,08,02,2015,09,25,00055,V,00066,V,00043,V,00027,V,00023,V,00037,V,00000,N,00000,N,00000,N
28,079,004,81,98,02,2015,09,25,00070,V,0042,V,0121,V,0084,V,0065,V,0033,V,00000,N,00000,N,00000,N
28,079,004,82,98,02,2015,09,25,00048,V,00042,V,00067,V,00078,V,00082,V,00182,V,00000,N,00000,N,00000,N
28,079,004,83,98,02,2015,09,25,02108,V,0207,V,0209,V,0207,V,0195,V,0165,V,00000,N,00000,N,00000,N
28,079,004,85,98,02,2015,09,25,00020,V,00020,V,00020,V,00020,V,00020,V,00019,V,00000,N,00000,N,00000,N
28,079,004,86,98,02,2015,09,25,00045,V,00047,V,00048,V,00048,V,00051,V,00064,V,00000,N,00000,N,00000,N
28,079,004,89,98,02,2015,09,25,00000,V,00000,V,00000,V,00000,V,00000,V,00000,N,00000,N,00000,N
28,079,008,01,38,02,2015,09,25,00011,V,00012,V,00011,V,00011,V,00011,V,00011,V,00000,N,00000,N,00000,N
28,079,008,06,48,02,2015,09,25,00004,V,00004,V,00003,V,00003,V,00003,V,00003,N,00000,N,00000,N,00000,N
28,079,008,07,08,02,2015,09,25,00006,V,00018,V,00004,V,00003,V,00001,V,00002,N,00000,N,00000,N,00000,N
28,079,008,08,08,02,2015,09,25,00044,V,00053,V,00043,V,00029,V,00015,V,00021,N,00000,N,00000,N,00000,N

```

Fig. 1. Imagen A: Fichero de la Comunidad de Madrid

2.2 Datos de Andalucía

La Junta de Andalucía, fue una de las primeras instituciones en España en hacer públicos en Tiempo Real los datos ambientales de su región. Dentro de su portal: <http://www.redhidrosurmedioambiente.es>, los datos los podemos ver en formato de tabla, pero también permite obtener los datos en formato “machine-readable”.

Mediante este portal, se puede consultar datos como la Temperatura o la Pluviometría rellenando un simple formulario en el que es necesario indicar una Fecha Inicial, Fecha Final y la Estación. Esta consulta devolverá una tabla pero también existe la posibilidad de pasarlo a un fichero plano.

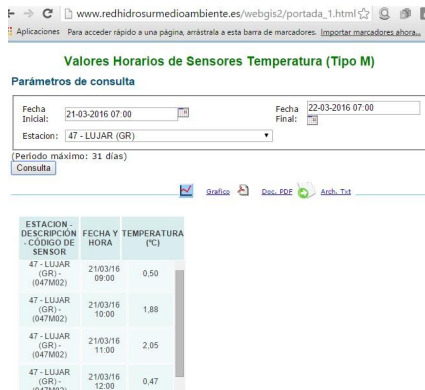


Fig. 2. Ejemplo de datos devueltos de la Junta de Andalucía

Al hacer click sobre el enlace “Arch. Txt” se generará un fichero de texto con los datos consultados. En dicho fichero se usará el “,” (punto y coma) como separador de campos.

Este fichero contiene las siguientes columnas:

Col.	Descripción	Longitud
1	Cod. Estación – Desc.de Estación – Cod. de sensor	Delimitado por separador
2	Fecha y Hora	Delimitado por separador
3	Valor de Medición	Delimitado por separador

Table 2. Estructura Fichero Junta de Andalucía

El primer campo contiene información del “Código de Sensor” que para el alcance de este proyecto, no es necesario. Por tanto solo mostraremos en esta tabla el Código y Descripción de cada estación para esta región. Estos datos se publican en Tiempo Real. Los datos se actualizan cada hora.

Un detalle que habrá que solucionar es la automatización de la descarga debido a que, primero es necesario informar los campos de “Fecha Inicio”, “Fecha Fin”, “Estación” y luego, al pulsar sobre “Arch. Txt”, se genera finalmente el fichero plano para descargarlo.

Una alternativa bastante sencilla, es, aprovechando que la consulta al servidor de datos se realiza con el método “GET”, indicarle los parámetros en la url. Esto se verá con mayor detalle más adelante.

La siguiente imagen muestra un ejemplo del fichero de Datos Ambientales en Tiempo Real de la Junta de Andalucía entre el 07-Mar-2016 y el 10-Mar-2016 para la estación de: EMBALSE DE CHARCO REDONDO.

```

Estación - Descripción - Código de sensor; Fecha Hora; Temperatura °C;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 00:00; 10,590;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 01:00; 10,200;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 02:00; 10,070;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 03:00; 9,560;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 04:00; 9,110;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 05:00; 8,630;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 06:00; 8,300;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 07:00; 8,220;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 08:00; 8,590;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 09:00; 9,090;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 10:00; 10,120;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 11:00; 11,250;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 12:00; 12,770;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 13:00; 12,400;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 14:00; 10,860;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 15:00; 11,090;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 16:00; 13,980;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 17:00; 15,610;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 18:00; 16,340;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 19:00; 16,760;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 20:00; 13,350;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 21:00; 12,490;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 22:00; 11,840;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 07/03/16 23:00; 11,390;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 00:00; 10,710;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 01:00; 10,550;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 02:00; 9,930;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 03:00; 9,490;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 04:00; 9,140;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 05:00; 8,740;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 06:00; 8,480;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 07:00; 8,490;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 08:00; 8,000;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 09:00; 7,870;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 10:00; 9,920;
3 - EMBALSE DE CHARCO REDONDO (CA) - (003M02); 08/03/16 11:00; 12,110;

```

Fig. 3. Fichero de la Junta de Andalucía

3 FORMATO DE INTEGRACION

El objetivo de esta sección es mostrar una Estructura Final de datos que será el “output” del proceso de Integración de ambos Data Sets: Comunidad de Madrid y Junta de Andalucía.

Antes de empezar a desarrollar el proceso de Integración, es fundamental conocer el formato de la estructura de datos, para lo cual debemos tener en cuenta dos puntos:

- Alcance del Trabajo. El proceso de Integración puede llegar a ser muy complejo considerando la cantidad de información que existe en cada dataset y los distintos formatos en que podemos encontrar los datos. Es importante saber qué datos son los que realmente se quieren integrar y qué datos no se quieren integrar para disminuir la amplitud de nuestros datos.
- Información Común. Los datasets que queremos integrar deben tener información en común sin importar el formato. La integración significa aunar los mismos indicadores con distintos formatos en un formato único. Es recomendable analizar ambos datasets para encontrar en ambos solo la información que realmente nos interesa integrar.

Técnicamente, lo que se busca es crear una estructura de datos, sobre la cual se vayan añadiendo datos sin importar sin vienen de la Comunidad de Madrid o de la Junta de Andalucía. En este caso, la estructura final de integración es un fichero plano

(csv) tras aplicar un proceso de formateo. La estructura final del fichero de datos integrados tiene las siguientes columnas:

Col.	Descripción	Longitud
1	C. Autónoma (28 : Madrid, 2 : Andalucía)	Delimitado por separador
2	Ciudad (79 : Madrid, 2 : Andalucía)	Delimitado por separador
3	Estación	Delimitado por separador
4	Indicador	Delimitado por separador
5	Fecha-Hora	Delimitado por separador
6	Valor de Medición	Delimitado por separador

Table 3. Estructura final de datos

La apariencia del fichero Final después del proceso de integración es la siguiente:

```

2,2,66,83,2016-03-17 21:00:00,5,73
2,2,67,83,2016-03-17 21:00:00,7,96
2,2,68,83,2016-03-17 21:00:00,2,37
2,2,75,83,2016-03-17 21:00:00,8,53
2,2,76,83,2016-03-17 21:00:00,5,6
2,2,80,83,2016-03-17 21:00:00,3,93
2,2,81,83,2016-03-17 21:00:00,9,96
2,2,82,83,2016-03-17 21:00:00,8,86
2,2,88,83,2016-03-17 21:00:00,12,88
2,2,89,83,2016-03-17 21:00:00,14,98
2,2,91,83,2016-03-17 21:00:00,7,07
2,2,92,83,2016-03-17 21:00:00,11,75
2,2,97,83,2016-03-17 21:00:00,14,98
2,2,103,83,2016-03-17 21:00:00,14,39
2,2,104,83,2016-03-17 21:00:00,16,83
2,2,130,83,2016-03-17 21:00:00,14,24
28,79,99,83,2016-03-17 01:00:00,9,1
28,79,4,83,2016-03-17 01:00:00,9,9
28,79,18,83,2016-03-17 01:00:00,10
28,79,24,83,2016-03-17 01:00:00,6,4
28,79,38,83,2016-03-17 01:00:00,12,9
28,79,54,83,2016-03-17 01:00:00,9,4
28,79,56,83,2016-03-17 01:00:00,8,4
28,79,57,83,2016-03-17 01:00:00,8,8
28,79,59,83,2016-03-17 01:00:00,7,5
28,79,99,83,2016-03-17 02:00:00,8,3
28,79,4,83,2016-03-17 02:00:00,9,9
28,79,18,83,2016-03-17 02:00:00,9,3
28,79,24,83,2016-03-17 02:00:00,4,8
28,79,38,83,2016-03-17 02:00:00,13,8
28,79,54,83,2016-03-17 02:00:00,8,7
28,79,56,83,2016-03-17 02:00:00,7,9
28,79,57,83,2016-03-17 02:00:00,7,7
28,79,59,83,2016-03-17 02:00:00,6,4
28,79,99,83,2016-03-17 03:00:00,7,9
28,79,4,83,2016-03-17 03:00:00,8,7
28,79,18,83,2016-03-17 03:00:00,9
28,79,24,83,2016-03-17 03:00:00,4,3
    
```

Fig. 4. Fichero de la estructura final de datos

4 PROCESO DE INTEGRACIÓN

El fichero final se obtiene tras la ejecución de un proceso automático en tiempo real y cuyo objetivo es transformar los datos [10] de los distintos datasets con formatos variables en un único formato (ver Tabla 4). Esta tarea se realiza utilizando procesos Hadoop para cada uno de los datasets que se desee incorporar, por tanto el proceso de integración para un dataset en concreto realiza tareas muy distintas al proceso de

integración de otro dataset. Técnicamente, el proceso de integración se compone de 4 subprocesos según se muestran en la Tabla 4.

Nº	Sub-Proceso	Descripción
1	Descarga	Se descarga el Data Set publicado desde la Web Institucional
2	Carga a Estructura Temporal	Se carga los datos del Data Set en una estructura Temporal para posteriormente procesarlo
3	Verticalización o Formateo	Realiza las comprobaciones, validaciones y/o formateos necesarios antes de proceder a almacenar los nuevos datos
4	Inserción en Estructura Final	Finalmente se almacenan los datos en el formato especificado en la Tabla B

Table 4. Subprocesos del Proceso de Integración

Este proceso de Integración, actualmente se ha implementado para los dos datasets descritos previamente, consiguiendo al final el formato de integración descrito en la Sección 3. Existen herramientas genéricas que permiten la integración de distintos datasets [11], sin embargo, estas herramientas siempre necesitan la intervención humana como último paso para integrar los datos. Para este trabajo, se pretende realizar uno totalmente automático a medida de los datasets que se desean integrar.

5 VISUALIZACIÓN DE RESULTADOS

Una vez integrados los datos es posible crear servicios comunes a todos los datasets que hayan pasado por el proceso de integración.

Esta interfaz gráfica se ha implementado totalmente en Open-Source, mediante el lenguaje R y la librería Shiny que permite realizar interfaces gráficas sencillas y muy eficientes. Shiny permite implementar gráficos con muy poco código pero es necesario seguir una estructura de código propia. Esta estructura consiste de únicamente elaborar dos ficheros para toda la interfaz según se muestran en la Tabla 5. La apariencia se muestra en la Figura 5. También hemos utilizado la librería dplyr para optimizar el rendimiento de las consultas a la estructura de datos.

Nº	Script	Descripción
1	server.r	Script de Servidor. Contiene instrucciones que el ordenador necesita para construir la aplicación.
2	ui.r	Script de Interface de Usuario. Controla el layout y apariencia de la aplicación.

Table 5. Módulos de la Interface Gráfica



Fig. 5. Interfaz gráfica y ejemplo de uso

6 CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se ha mostrado una implementación de una aplicación que permite aprovechar los datos abiertos por las instituciones en tiempo real para realizar análisis estadístico y visualización. La aplicación se ha implementado usando la herramienta Open-Source R. Se ha demostrado que con esta herramienta es posible elaborar un proceso de integración así como también una interfaz gráfica de manera bastante sencilla y con un buen rendimiento. Como resultado se obtiene una aplicación donde se pueden realizar análisis comparativos en tiempo real de las distintas bases de datos tras su integración.

Como trabajo futuro está la incorporación de nuevos datos en territorio español y europeo mediante la implementación de procesos de ejecución paralela con RHadoop.

REFERENCIAS

- [1]. A. Aggarwal, *Managing Big Data Integration in the Public Sector*, Information Science Reference, IGI Global, 2016. USA
- [2]. Open Government Partnership (2015). The open overnment guide special Edition: Implementing the 2030 sustainable development agenda. *Open Government Guide*, 5-6.
- [3]. M. Harrison, T., Guerrero, Santiago (2011). *Open Government and E-Government: Democratic Challenges from a Public Value Perspective*, 2-5.
- [4]. Westmore Nicola (2011). *Transparency Opening up Government*, 3-3.
- [5]. European Commission (2015). *Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources*, 28-37.
- [6]. Nicandro Cruz-Rubio, C. (2014). *Gobierno Abierto y Open Data*. *Actualidad Administrativa*, 816-817
- [7]. Darbshire, H. (2015). *Access Info presenta una queja formal sobre la falta de compromiso de España con el OGP*. *Access Info*.

- [8]. Say M. (2015). Open data: barriers to be broken. UK Authority. Recuperado de : <http://www.ukauthority.com/>
- [9]. Meijueiro, L. (2014). Mapa actual de las iniciativas Open Data en España. Centro Tecnológico de la Información y la Comunicación (CTIC). Recuperado de : <http://datos.fundacionctic.org/2014/03/mapa-actual-de-las-iniciativas-open-dataen-espana/>
- [10]. European Commission (2016). Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, 5-5. Recuperado de : http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
- [11]. M. Stonbraker et al. Data Curation at Scale: The Data Tamer System, 6th Biennial Conference on Innovative Data Systems Research (CIDR '13) January 6-9, 2013, Asilomar, California, USA
- [12]. S. Kandel et al. Wrangler: Interactive Visual Specification of Data Transformation Scripts, ACM Human Factors in Computing Systems (CHI), 2011