Ciclo de vida de los datos: PEC 1

Francisco Alonso Sarria y Jose Luis García Bravo

1 Contexto

Vamos a trabajar con el Portal Estadístico de la Región de Murcia, sitio web del Centro de Estadística de la Región de Murcia (CREM) dependiente de la Comunidad Autónoma de la Región de Murcia (CARM). Se trata de un sitio que alberga 4012 tablas con datos socioeconómicos de interés regional organizadas en 115 grandes temas. La página:

https://econet.carm.es/web/crem/informacion-de-la-a-z (figura 1)

actúa como índice a partir del cual se puede acceder a cualquiera de estos grandes temas.

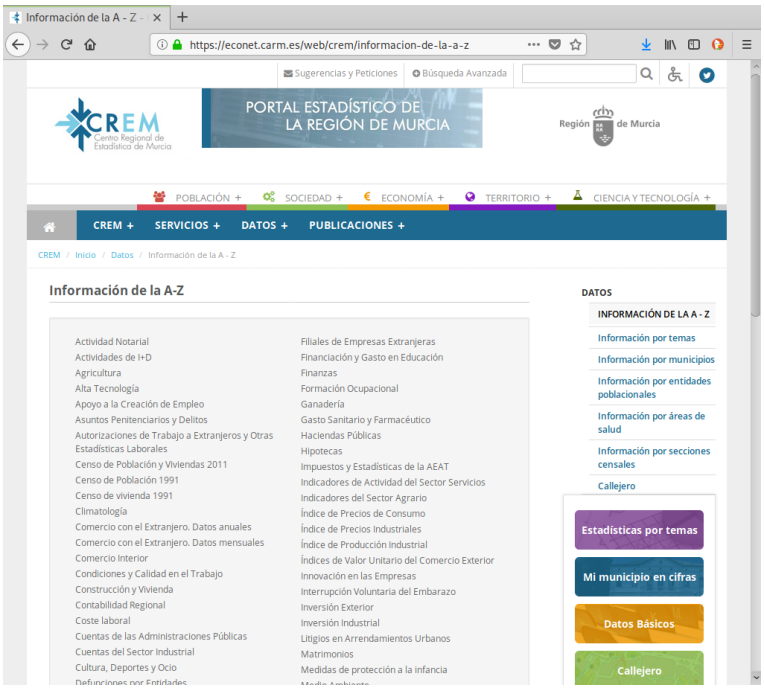


Figura 1: Página principal con los enlaces a los grandes temas

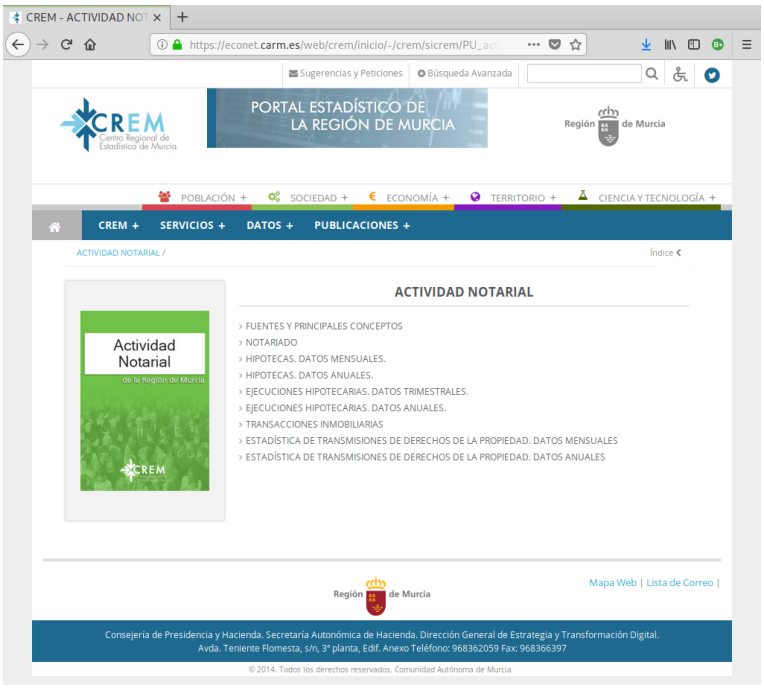


Figura 2: Página del primer gran tema con los enlaces a los subtemas

Al pinchar en alguno de los grandes temas vamos a otra página donde se desglosan los subtemas corrrespondientes (Figura 2). Al pinchar en alguno de estos enlaces, éste se desglosa en nuevos enlaces (figura 3) que nos llevan a las tablas (figura 4). Se trata por tanto, en principio, de una estructura en 3 niveles. Sin embargo, solo en algunos casos, aparece un cuarto nivel.

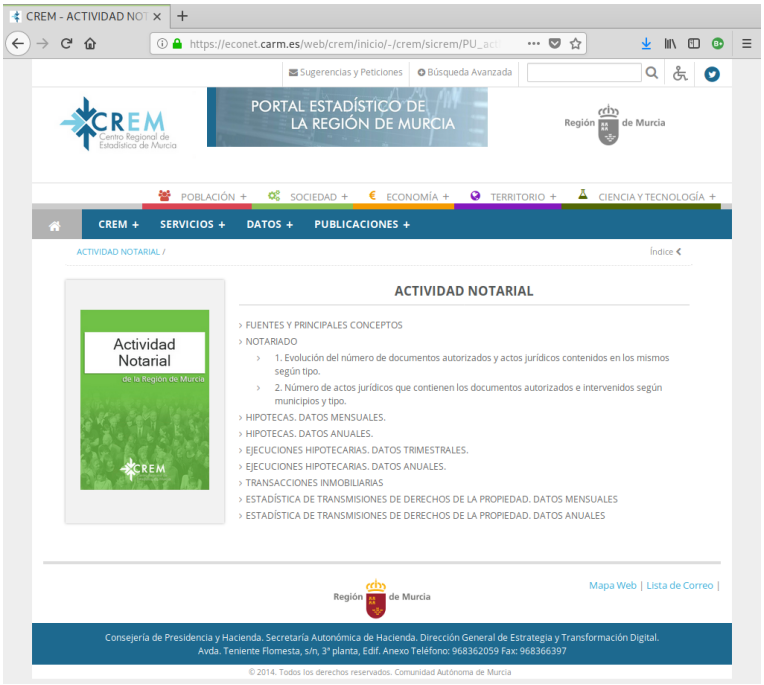
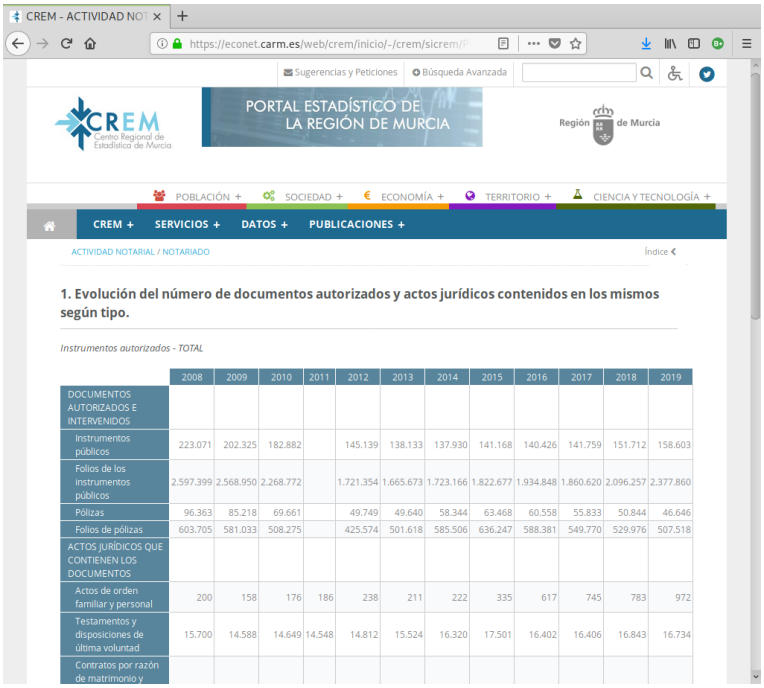


Figura 3: Página del primer subtema con con los enlaces a las table



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
DOCUMENTOS AUTORIZADOS E INTERVENIDOS												
Instrumentos públicos	223.071	202.325	182.882		145.139	138.133	137.930	141.168	140.426	141.759	151.712	158.603
Folios de los instrumentos públicos	2.597.399	2.568.950	2.268.772		1.721.354	1.665.673	1.723.166	1.822.677	1.934.848	1.860.620	2.096.257	2.377.860
Pólizas	96.363	85.218	69.661		49.749	49.640	58.344	63.468	60.558	55.833	50.844	46.646
Folios de pólizas	603.705	581.033	508.275		425.574	501.618	585.506	636.247	588.381	549.770	529.976	507.518
ACTOS JURÍDICOS QUE CONTIENEN LOS DOCUMENTOS												
Actos de orden familiar y personal	200	158	176	186	238	211	222	335	617	745	783	972
Testamentos y disposiciones de última voluntad	15.700	14.588	14.649	14.548	14.812	15.524	16.320	17.501	16.402	16.406	16.843	16.734
Contratos por razón de matrimonio y												

Figura 4: Página con la primera tabla del primer subtema del primer gran tema

2 Título

Recopilación de datos estadísticos de la Región de Murcia

3 Descripción del dataset

Como resultado de la práctica hemos recopilado las 4012 tablas en un fichero en formato JSON (`CREM.json`) . La figura que aparece en el apartado de representación gráfica muestra la estructura de este archivo. Lo acompaña un fichero `estructura.csv` que contiene 4012 filas (una por tabla) y 3 columnas que corresponden al tema principal, subtema y nombre de la tabla.

Adicionalmente, hemos integrado todas las tablas municipales (`tablasMunicipales.csv`) en un único fichero con 45 filas que corresponden con los 45 municipios de la Región de Murcia y 3261 columnas. La primera columna contiene los nombres de los municipios y los nombres de las restantes columnas son simplmente `columna_X`, donde `X` es un número del 2 al 3261. El fichero `clavesColumnas.csv` contiene una tabla con 3261 filas (una por cada columna en `tablasMunicipales.csv` y 2 filas. El primer campo contiene el nombre de columna que aparece en `tablasMunicipales.csv` y el segundo es una cadena de texto que incluye el tema, subtema, nombre de latabla y nombre de la columna correspondientes, separados por ";".

4 Representación gráfica

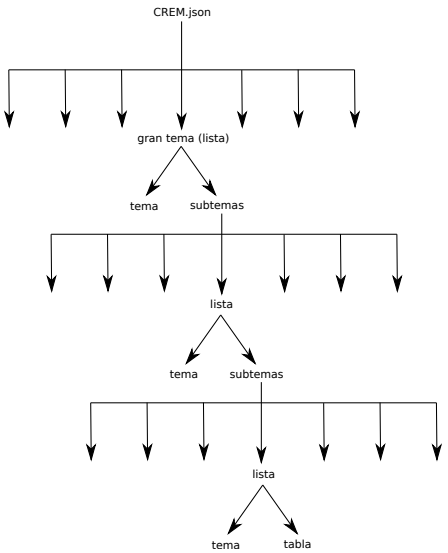


Figura 5: Representación de la lista que alberga todas las tablas extraidas

5 Contenido

En realidad son 4012 tablas por lo que no tiene sentido explicar todos los campos. En todo caso puede explorarse la estructura del fichero, sus tablas y columnas en la página web:

https://alonsarp.shinyapps.io/aplicacion

6 Agradecimientos

Los datos proceden de la página web del CREM, dependiente de la Dirección General de Estrategia y Transformación Digital de la Consejería de Presidencia y Haciendad de la Comunidad Autónoma de la Región de Murcia (CARM). Se trata del órgano responsable de la actividad estadística de interés para la CARM, a cuyo frente está José Blaya Verdú. Entre sus objetivos podemos destacar:

- Proponer, dirigir y coordinar la actividad estadística pública de interés para la CARM.
- Elaborar el anteproyecto del Plan Estadístico de la CARM y los proyectos de Programas Anuales de Estadística.
- Proponer normas, establecer requisitos y promover la coordinación metodológica con las unidades de estadística de otros organismos.
- Realizar las actividades estadísticas encomendadas por la legislación vigente.
- Promover la investigación estadística y la formación y el perfeccionamiento profesional del personal estadístico.
- Promover la difusión de las estadísticas relativas a la Región de Murcia.
- Informar todo proyecto que promueva o en el que participe la administración de la CARM para la realización de actividad estadística.
- Colaborar en el diseño y reforma de los procedimientos administrativos de la CARM que por su naturaleza puedan tener consecuencias en la producción de informaciones de posible utilización como fuente estadística.
- Promover, gestionar y centralizar la creación y mantenimiento de bancos de datos de carácter estadístico.
- Realizar investigaciones para contrastar la objetividad y corrección técnica de la metodología en las actividades estadísticas.
- Actividades de creación, elaboración, mantenimiento, actualización y tratamiento de los datos del padrón municipal de habitantes, registros de comercio exterior y contenido económico, directorios de instituciones, empresas y establecimientos que ejerzan su actividad en la Región, censos económicos y todas las estadísticas necesarias para la implantación de indicadores económicos.

- La carga de datos, actualización y mantenimiento de la base de datos estadísticos regional y municipal.

En 2010, el Centro Regional de Estadística, sustituyó el envío de publicaciones en papel por informes en su página web (https://econet.carm.es/). Estos informes se acompañan de un gran número de tablas que aparecen en diferentes páginas web y que han sido objeto de esta práctica.

7 Inspiración

Este dataset contiene todas las tablas que el CREM tiene en abierto en su dataset. Aunque las tablas son descargables, son muchas y se hace muy pesado descargarlas todas, además varían de año a año, por lo que tener un protocolo de descarga automatizado puede resultar útil.

Cualquier estudio económico, social, geográfico, etc. sobre la Región de Murcia puede beneficiarse de la explotación de estos datos. Aunque obviamente haría falta una fase de comprobación y limpieza de datos en las tablas involucradas.

Una idea interesante sería completar la aplicación shiny que hemos creado para permitir un descubrimiento y análisis exploratorio de los datos más amigable y orientado a aplicaciones concretas. La integración con información espacial para presentar como mapas de coropletas aquellas tablas que lo permitan sería otra opción a explorar.

8 Licencia

Released Under CC0: Public Domain License

En la página web del CREM no aparece ninguna nota al respecto de la propiedad intelectual de estos datos. Tampoco hay ninguna restricción en el ficherto robots.txt. Por ello consideramos estos datos como datos públicos y no consideramos adecuado añadir ninguna restrección, por tanto optamos por la licencia menos restrictiva.

9 Código

```
# Crea un data.frame con los nombres y los links presentes en obj vía el xpath
extraeLinks <- function(obj, xpath, urlBase=NULL) {
    nodos    = rvest::html_nodes(obj, xpath=xpath)
    nombres  = rvest::html_text(nodos)
    enlaces  = rvest::html_attr(nodos, "href")
    if (!is.null(urlBase)) enlaces = paste0(urlBase,"/",enlaces)
    return(data.frame(nombres=nombres, enlaces=enlaces))
}
```

```
# Reemplaza NA por val en un data.frame.
replace_na <- function(tabla,val) {
     for (c in 1:ncol(tabla)) {
        tabla[which(is.na(tabla[,c])),c] = val
     }
     return(tabla)
}

espera = 2 # Tiempo en segundos antes de cada llamada

# Compruebo en primer lugar que puedo proceder:

robotstxt::get_robotstxt(domain= "https://econet.carm.es")

#Leo la página base que es la que contiene los enlaces a los grandes temas.

url0 <- "https://econet.carm.es/web/crem/informacion-de-la-a-z"
html0 <- xml2::read_html(url0)

# Los enlaces a grandes temas aparecen en dos columnas y a cada columna se accede
# con un xpath diferente. Así que accedo a cada columna por separado y luego las
# integro

xpath1 = "//div/div/div[1]/div/div/div/div/div/div/div/div[1]/ul/li/a"
xpath2 = "//div[2]/ul/li/a"
nivel1 = rbind(extraeLinks(html0, xpath1),
               extraeLinks(html0, xpath2))

# Para cada uno de los grandes temas leo el segundo nivel. Guardo los enlaces
# junto con los nombres en nivel2. El vector temas1 contiene el número del tema
# principal de cada entrada.

for (n1 in 1:nrow(nivel1)) {
    cat(n1, "/", nrow(nivel1), "\n")
    url1 = nivel1[n1,2]
    html1 <- xml2::read_html(url1)

    urlBase=url1
    if(grepl("html",url1)) {
      urlBase=paste0(head(strsplit(url1, "/", fixed=TRUE)[[1]],-1), collapse="/")
    }

    xpath = "//td/table/tbody/tr/td/a"
    niv2 = extraeLinks(html1, xpath=xpath, urlBase=urlBase)
    niv2 = niv2[grep("Indice",niv2[,2]),]  # Me quedo con enlaces a datos
    if (n1==1) {
        nivel2 = niv2
        temas1 = rep(n1, nrow(niv2))
```

```
    } else {
        nivel2=rbind(nivel2, niv2)
        temas1 = c(temas1, rep(n1, nrow(niv2)))
    }
    Sys.sleep(espera)
}

#Acceso a las páginas de subtemas temas

# El siguiente paso es repetir el proceso anterior, pero ahora para descubrir y
# extraer las entradas de nivel 3 que hay en las de nivel 2. En la mayoría de
# los casos solo se llega al nivel 3, pero en algunos se alcanza el nivel 4.
# Cuando el nivel 3 contiene tablas, los nombres de los ficheros html contienen
# las palabras clave sec o pagina. Cuando contiene las entradas de nivel 4,
# la palabra clave es indice.

# El problema es que en el nivel 4 los xpath no son sistemáticos como en los
# niveles anteriores, por lo que ha habido que analizar cada caso por separado.
# El conocimiento adquirido se integra en una función llamada getnt que en función
# de la entrada de nivel 2 en la que estemos devuelve un número con el que formar
# la cadena correcta para el xpath.

#Para simplificar la estructura de datos final, todas las tablas de nivel 4 se
# almacenan como tablas de nivel 3 ya que los captions correspondientes son
# distintos:

source("getnc.R")
for (n2 in 1:nrow(nivel2)) {

    cat(n2, "/", nrow(nivel2), "\n")
    if (n2==88) next # excepción
    url2 = nivel2[n2,2]

    urlBase=url2
    if(grepl("html",url2)) {
      urlBase=paste0(head(strsplit(url2, "/", fixed=TRUE)[[1]],-1), collapse="/")
    }

    html2 <- xml2::read_html(url2)
    xpath = "//td[2]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td/a"
    niv3 = extraeLinks(html2, xpath=xpath, urlBase=urlBase)

    # Pero que pasa cuando hay un enlace al cuarto nivel????
    # Cuales tienen 4º nivel pero no graficos?
    w = which(grepl("Indice",niv3[,2]) & !(grepl("GRÁFICOS",niv3[,1])))
    # para todos ellos, ábrelo y extrae las tablas.

    for (ww in w) {
```

```
        html3 <- xml2::read_html(niv3[ww,2])

        nt=getnt(n2,ww)

        xpath=paste0("//tr[",nt,"]/td[2]/table/tbody/tr/td/table/tbody/tr/td/a")
        if (n2==35 & ww==1) {
           xpath="//td/table/tbody/tr[2]/td[2]/table/tbody/tr/td/table/tbody/tr/td/a"
        }

        niv4 = extraeLinks(html3, xpath=xpath, urlBase=urlBase)
        niv3 = rbind(niv3, niv4)
    }
    niv3 = niv3[grep("sec|pagina",niv3[,2]),]  # Me quedo con enlaces a tablas

    if (n2==1) {
        nivel3 = niv3
        temas2 = rep(n2, nrow(niv3))
    } else {
        nivel3=rbind(nivel3, niv3)
        temas2 = c(temas2, rep(n2, nrow(niv3)))
    }
    Sys.sleep(espera)
}

# Acceso a las tablas
# El último paso del scraping es leer las tablas. Este proceso es algo más sencillo
# ya que, aunque todas las páginas que contienen una tabla tienen en realidad
# varias, siempre es la número 4 o la número 6 la que contiene la información,
# con lo que es fácil extraerla combinando las funciones html_table y html_nodes
# del paquete rvest:

tablas = list()
nt = 0
for (n3 in 1:nrow(nivel3)) {
    cat(n3, "/", nrow(nivel3), "\n")
    url3 = nivel3[n3,2]
    html3 <- xml2::read_html(url3)
    nt = nt+1
    tbs = rvest::html_nodes(html3, "table")
    if(length(tbs)==0) next
    if (length(tbs)>=6) {
        tablas[[nt]] <- rvest::html_table(tbs[6])[[1]]
    } else {
        tablas[[nt]] <- rvest::html_table(tbs[4],fill = TRUE)[[1]]
    }
    Sys.sleep(espera)
}
```

```
# Un poco de data cleaning en las tablas
# En este momento hacemos un poco de data cleaning en las tablas. Consiste en
# convertir los “.” en "" y luego las “,” en “.”. El mismo script detecta tablas
# nulas que hemos comprobado que son enlaces rotos.

for (tb in 1:length(tablas)) {
    if (is.null(tablas[[tb]])) {
       cat(tb, "/", length(tablas), "\n")
    } else {
       # Convierto en numérico las celdillas que cumplen esta regExp
       rexpReal = "^(\\+|-)?((\\d*).?)*\\,?(\\d*)$"
       for (r in 1:nrow(tablas[[tb]])) { for (c in 1:ncol(tablas[[tb]])) {
           if (grepl(rexpReal, tablas[[tb]][r,c])) {
               tablas[[tb]][r,c] = gsub(",", ".",
                                    gsub(".","",tablas[[tb]][r,c], fixed=TRUE),
                                    fixed=TRUE)
           }
       }}
    }
}



#  Meterlo todo en una lista. La figura incluida en el punto 4 representa
# la estructura de esta lista.

lista_temas1 = list()
nt = 0
for (n1 in 1:nrow(nivel1)) {
    lista_temas1[[n1]] = list(tema = nivel1[n1,1], subtemas=list())

    w = which(temas1==n1)
    for (n2 in 1:length(w)) {
        ll = list(tema=nivel2[w[n2],1], subtemas=list())
        lista_temas1[[n1]]$subtemas[[n2]] = ll

        w2 = which(temas2==w[n2])
        if (length(w2)>0) {
           for (n3 in 1:length(w2)) {
              nt = nt+1
              if (nt %in% c(1321,2021,2293)) {
                 ll = list(tema=nivel3[w2[n3],1], tabla=NULL)
              } else {
                 ll = list(tema=nivel3[w2[n3],1], tabla=tablas[[nt]])
                 ll$tabla = replace_na(ll$tabla,-9999)
              }
              lista_temas1[[n1]]$subtemas[[n2]]$subtemas[[n3]] = ll
```

```
            }
          }
      }
}

# Y guardarla en un fichero JSON
jsonlite::write_json(lista_temas1, "CREM.json")
```

10 Dataset

En la página https://alonsarp.shinyapps.io/aplicacion se tiene acceso a todo el documento JSON mediante una aplicación shiny.

Los 4 archivos CSV previamente introducidos pueden accederse en Zenodo...