# Proyecto piloto sobre viabilidad de usar Internet como fuente de datos.

## Evolutivo año 2016

Junio 2019

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA

ontsi observatorio nacional de las telecomunicaciones y de la SI

red.es

# 1.

# Introducción

# Introducción

- Rastro digital tanto de las empresas como los ciudadanos en Internet.
- Mediante la recolección y explotación de dicha información es posible describir numerosos fenómenos socio-económicos casi en tiempo real.
- IaD permite identificar datos e indicadores que se pueden obtener directamente de Internet. Utilizar IaD puede proporcionar una visión rápida sobre fenómenos nuevos sobre los que las técnicas tradicionales tienen dificultad de medir.
- Pueden mejorar la calidad de las estadísticas, sobre todo cuando se combinan con las metodologías tradicionales.
- Además, puede ser una forma de reducir la carga de trabajo sobre las unidades informantes, ya sean empresas o individuos.
- De cara a la definición de políticas futuras de la Sociedad de la Información, el IaD se conforma como una alternativa posible para disponer de datos sobre los usos de Internet.

# Introducción

- La Comisión Europea publicó en octubre de 2012 un informe relativo a la viabilidad de IaD como método estadístico para recoger y analizar datos (http://ec.europa.eu/information_society/newsroom/cf/itemdetail.cfm?item_id=8701 )

- En dicho informe se describen tres posibles metodologías para utilizar IaD como un método estadístico:

  - Mediciones centradas en el usuario, que captan los cambios de comportamiento de un usuario individual analizando el uso que hace de Internet a través de sus dispositivos (PC, teléfono inteligente, tableta, etc.).
  - Mediciones centradas de la red, que se centran en la medición de las propiedades de la red subyacente.
  - Mediciones centradas en sitios web, que obtienen datos publicados en determinados servidores web mediante robots.

# Introducción

- Tanto la OCDE, como la Oficina Estadística Europea (Eurostat), como los servicios de la Comisión, animan a continuar explorando las oportunidades y beneficios que estos métodos ofrecen, solicitando a los INIs de los Estados Miembros y otros organismos productores de información estadística sobre las TIC y la SI que realicen proyectos pilotos que utilicen este técnica.

- El objetivo es compartir experiencias y mejores prácticas de forma que se puede consolidar una metodología fiable que puede ser utilizada en la investigación social como complemento a las técnicas tradicionales como las encuestas por muestreo o explotación de registros administrativos.

# Introducción

- El ONTSI desde 2014 tiene en marcha en marcha dos iniciativas que usan directamente dos de estas metodologías:
  - Por una parte, un panel de hogares on-line sobre ciberseguridad. También se está en vías de sustituir un panel de hogares sobre uso de Internet por individuos por otro on-line.
  - Por otra parte, un proyecto piloto que permita recoger datos de determinados sitios web que sirva para analizar determinados fenómenos relacionados con la Sociedad de la Información y las TIC.

# 2.

## Análisis de comercio electrónico en empresas españolas

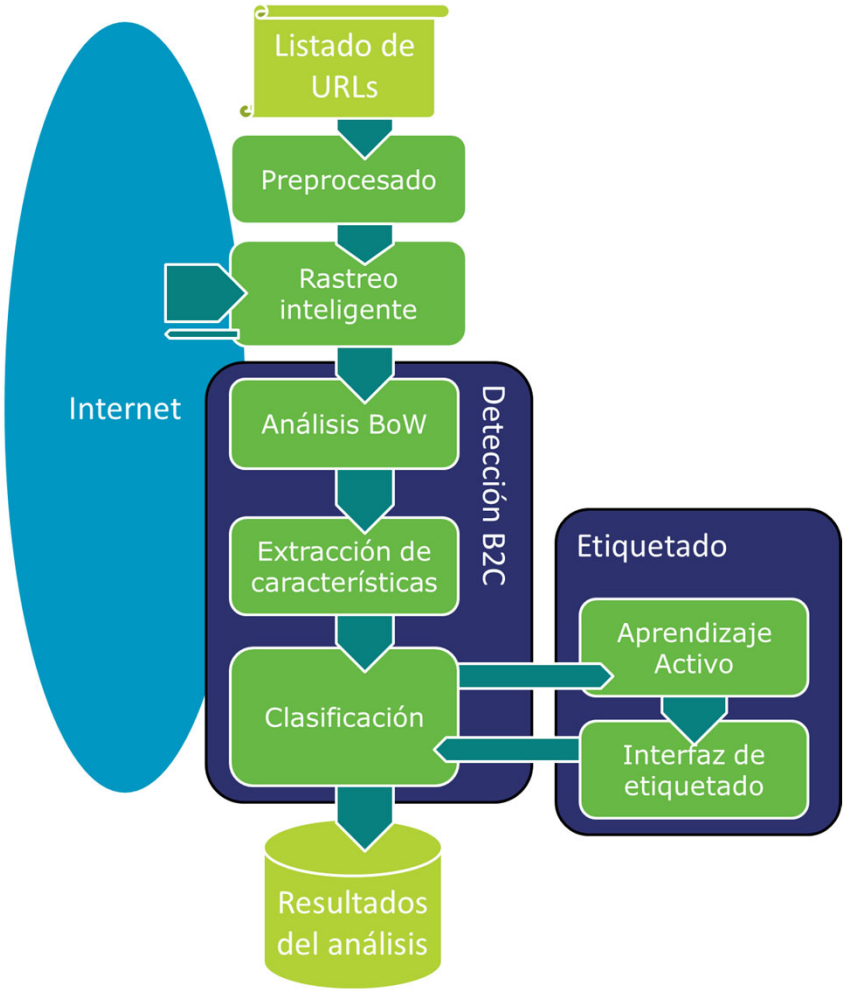
# Análisis de comercio electrónico en empresas españolas

- El objetivo principal de este estudio es cuantificar y caracterizar la presencia de B2C en las empresas españolas registradas en el CNAE.
- Para ello, se ha explorado sistemáticamente el sitio web de estas empresas, y se ha aplicado un algoritmo que, en primer lugar, detecta automáticamente la presencia de B2C y, en segundo lugar, extrae indicadores que puedan resultar útiles para caracterizar las empresas que tienen comercio electrónico y las que no.

# Metodología

- Desarrollar un SW de captura, análisis y visualización de datos.
- Aplicar el SW a las tareas de detección, perfilado y matching.
- Extraer conclusiones sobre la viabilidad del ML para IaD

# Proceso de detección automática de B2C



# Exploración web

**FUENTE DE DATOS**

**162.849**

**EMPRESAS,**

**145.920**

**DOMINIOS WEB** diferentes

**EXPLORACIÓN WEB**

El **CRAWLER** obtiene una representación de cada sitio web basada en

**8.763.024**

**TÉRMINOS,** que se reducen a

**343.780**

**EXTRACCIÓN DE CARACTERÍSTICAS**

Identifica los

**10.000**

**TÉRMINOS,** más relevantes para detección de B2C.

# Etiquetado

**2.540**

**PÁGINAS WEB ETIQUETADAS.**

## ¿Cuándo hay B2C?

**Condición 1**: Variedad de productos, añadir a cesta de la compra, completar pago, finalizar pedido.

**Condición 2**: Realizar una reserva (de habitaciones de hotel, de entradas, de billetes de viaje), pagar o señalizar la operación

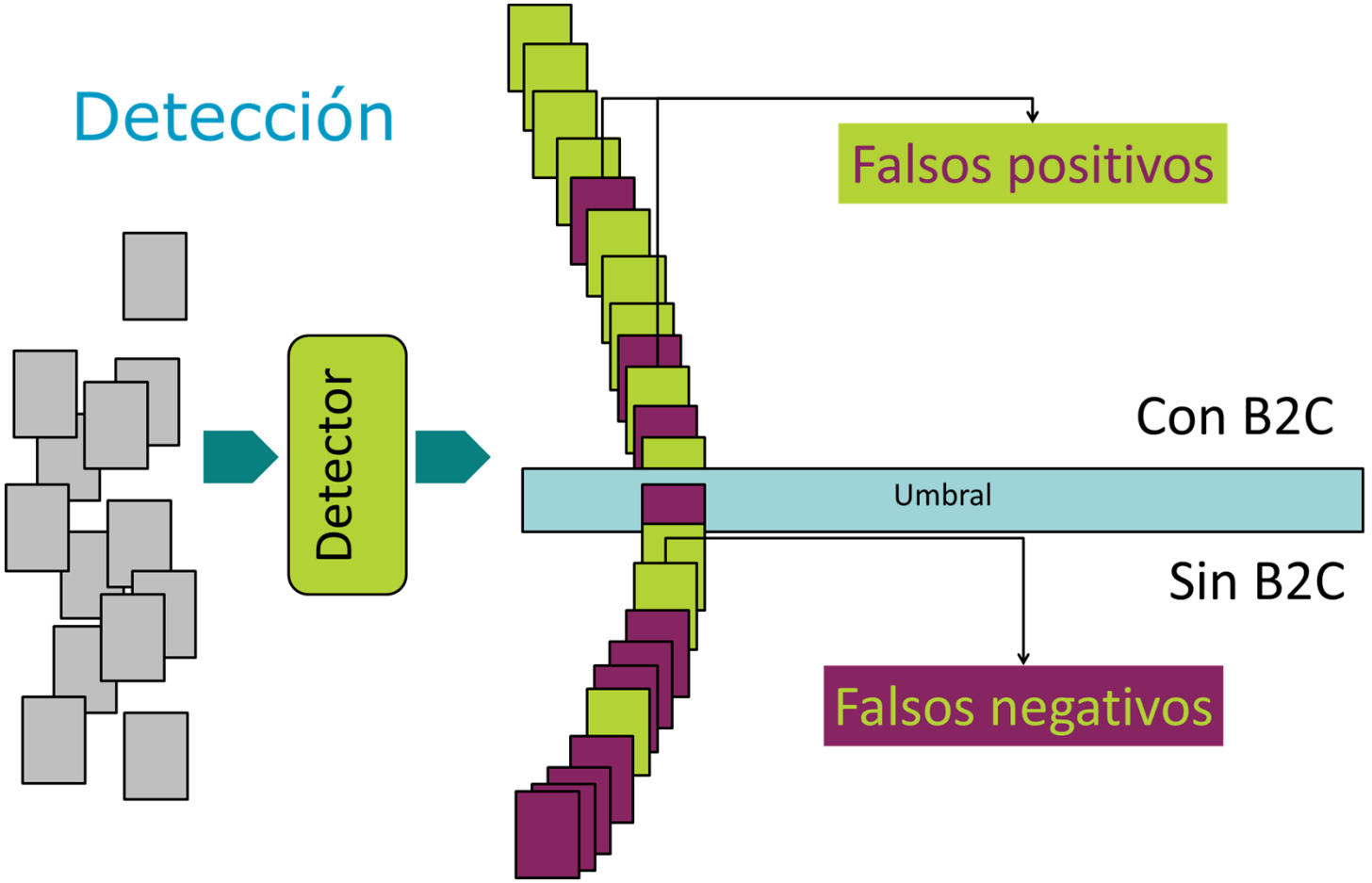

**1.606**

webs **SIN B2C**

**776**

webs **CON B2C**

# Detección


Detección
Detector
Falsos positivos
Con B2C
Umbral
Sin B2C
Falsos negativos

# Concusiones análisis B2C

El Machine Learning para IaD permite procesar más de

145.000 webs de empresas ….

…etiquetando solo un 0,8% de los sitios web ….

… para detectar la presencia de B2C con menos de un 8% de errores

# 3.

## Análisis de perfiles profesionales TICC y programas formativos

# Oferta y demanda de perfiles profesionales

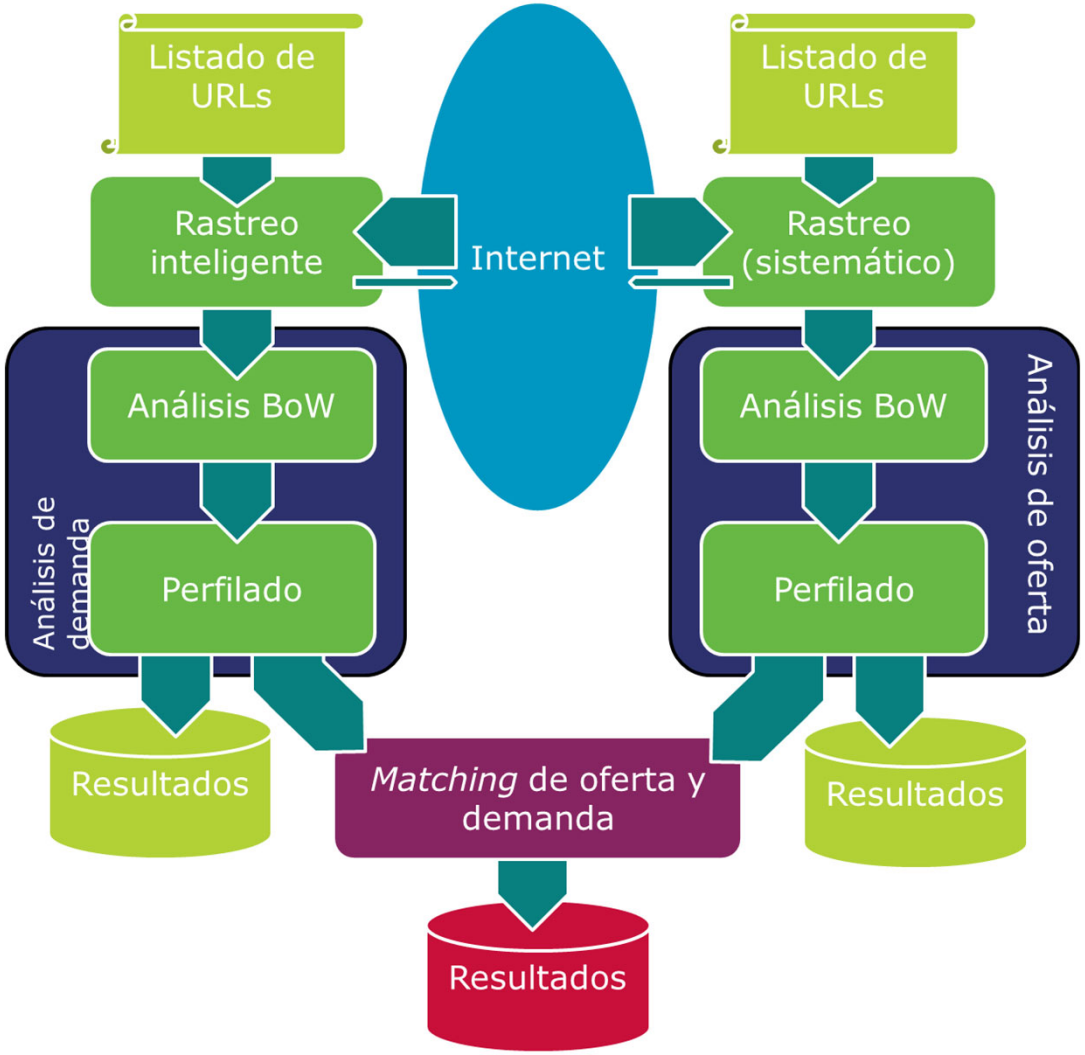
- Oferta y demanda de perfiles profesionales de las Tecnologías de la Información, las Comunicaciones y los Contenidos (TICC) se sustenta en la detección automática de los perfiles profesionales más demandados y ofertados por empresas y centros educativos en:
  - Portales de empleo
  - Sitios webs de empresas del sector TICC
  - Webs oficiales relativas a títulos universitarios, de formación profesional y cursos de formación en empresas

# Objetivos

- Objetivo 1:
  - Utilizar fuentes de datos disponibles en Internet para estimar y modelar la demanda de profesionales del Sector TICC
- Objetivo 2:
  - Ídem para la oferta formativa disponible
- Objetivo 3 (“matching”):
  - Analizar la adecuación de la oferta formativa a la demanda de profesionales

- Se han aplicado técnicas de ML que evitan la necesidad de etiquetado manual y que permiten detectar grupos de términos “coherentes” en los documentos

# Estructura global



# Rastreo de oferta formativa

Para cada plan formativo en TICC el crawler se encarga de obtener la siguiente información:

- **Plan de estudios**: donde se recogen las competencias generales y específicas de cada titulación universitaria o el listado de módulos y competencias asociados a una cualificación profesional.
- Un conjunto de **meta-datos** que se almacenan en una serie de ficheros csv que serán de utilidad para el procesado posterior y, principalmente, para la visualización.
    - En titulaciones universitarias se ha almacenado el nombre del título, universidad, rama del conocimiento, nivel y enlace al plan de estudios.
    - En cualificaciones profesionales se obtiene el nombre de la cualificación, la familia profesional a la que pertenece y el nivel de cualificación que tiene asignado.

# Extracción de perfiles

- Preprocesado del corpus de datos
- Construcción del vocabulario y extracción de bolsas de palabras
- Aprendizaje de perfiles
  - Un perfil queda caracterizado por una colección de palabras que son susceptibles de ser observadas con alta probabilidad para los documentos de dicho perfil
  - Se aprenden de forma automática un conjunto de perfiles para el corpus de datos analizado, junto con el grado de pertenencia de cada documento a cada uno de los perfiles identificados

# Matching

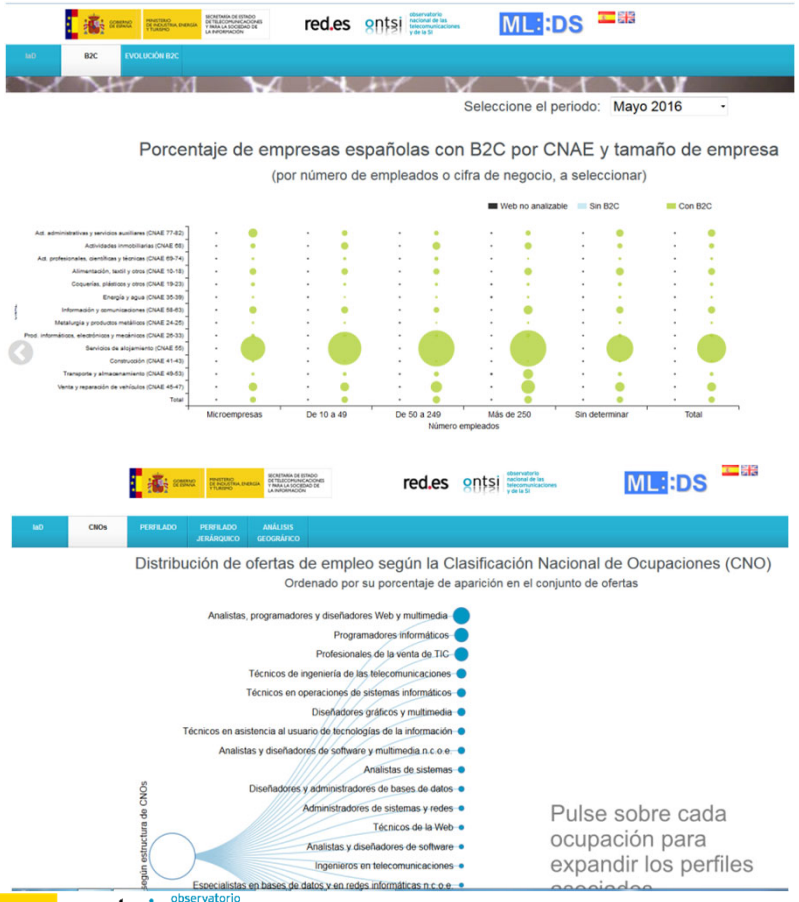
- En esta fase se analiza conjuntamente la demanda de empleo y la oferta formativa, con objeto de obtener diferentes medidas de “ajuste” de la oferta curricular (del sistema universitario y de formación profesional español) a la demanda de profesionales por parte de las empresas.
- La aplicación estimará las siguientes similitudes:
  - Similitud entre cada oferta de trabajo y cada oferta curricular.
  - Similitud entre cada oferta de trabajo y uno de los perfiles identificados sobre el conjunto de la oferta curricular.
  - Similitud entre cada perfil de ofertas de trabajo y cada titulación (universitaria o de formación profesional)
  - Similitud entre cada par de perfiles de ofertas de trabajo y de ofertas curriculares.
- Estrategia para el aprendizaje de similitudes:
  - Búsqueda de la similitud entre documentos al listado de términos de mayor relevancia para los perfiles de las ofertas de empleo

# 4.

## Herramienta de visualización

# Herramienta de visualización

- Se ha desarrollado una herramienta de visualización que permite hacer una completa explotación de los resultados.




## Enlaces

- Comercio electrónico: http://iad.ontsi.es/B2C/
- Perfiles profesionales: http://iad.ontsi.es/perfilado/
- Perfiles profesionales según la clasificación nacional de ocupaciones (CNO): http://iad.ontsi.es/perfiles_CNO/

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA
ontsi observatorio nacional de las telecomunicaciones y de la SI
red.es

# Muchas gracias