Islas Canarias
Del 15 al 19 de noviembre de 2021

Eustat

XXI Jornadas de Estadística de las Comunidades Autónomas

Reingeniería del Directorio de Actividades Económicas: de los Registros administrativos al Big Data

Amparo Ruiz Gil

amparo_ruiz@eustat.eus

INDICE DE LA PRESENTACIÓN:

- Introducción
- Objetivos
- Metodología
  - Fases
  - Problematicas encontradas
- Resultados
- Conclusiones

INTRODUCCIÓN:

El Directorio de Actividades Económicas:

- Recoge información relativa a los establecimientos ubicados en la Comunidad Autónoma de Euskadi donde se ejerce cualquier actividad económica.
- Constituye la base para cualquier investigación estadística orientada a empresas
- Es necesaria una información exhaustiva y veraz.

➢ **Utilización de una gran cantidad de recursos económicos y humanos para actualizar anualmente los datos.**

➢ **EUSTAT se plantea el uso de Registros Administrativos y del Big Data**

OBJETIVOS:

- ❑ Reducir el coste económico
- ❑ Disminuir plazos y adelantar difusión
- ❑ Disminuir la carga para los encuestados

Para conseguirlo se ponen en marcha dos proyectos:

- ➢ Automatización de la actualización del Directorio de Actividades Económicas con información de ficheros administrativos.
- ➢ Obtención de datos para DIRAE mediante técnicas de Big Data y web scraping.

METODOLOGÍA:

Actualización de los datos mediante ficheros administrativos:

Algoritmo que automatice la actualización del directorio anualmente:

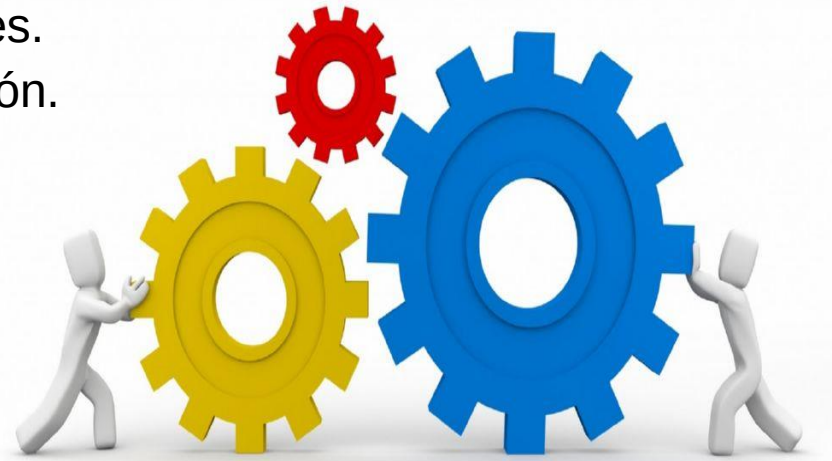
- ✓ Actualizar las altas y bajas.
- ✓ Actualizar variables básicas.

METODOLOGÍA:

Actualización de los datos mediante ficheros administrativos:

Fases del proyecto:

- ✓ Inventario de fuentes de datos.
- ✓ Definición de la parcelación o segmentación de las entidades.
- ✓ Análisis del flujo de los procesos de ejecución de la operación.
- ✓ Análisis exploratorio de los datos.
- ✓ Limpieza de datos o data cleansing.
- ✓ Transformación de los datos.
- ✓ Fusión de registros de distintas fuentes.
- ✓ Análisis del flujo de empresas.
- ✓ Validación y depuración.
- ✓ Generación de los outputs del proceso.



METODOLOGÍA:

Actualización de los datos mediante ficheros administrativos:

Problemáticas encontradas:

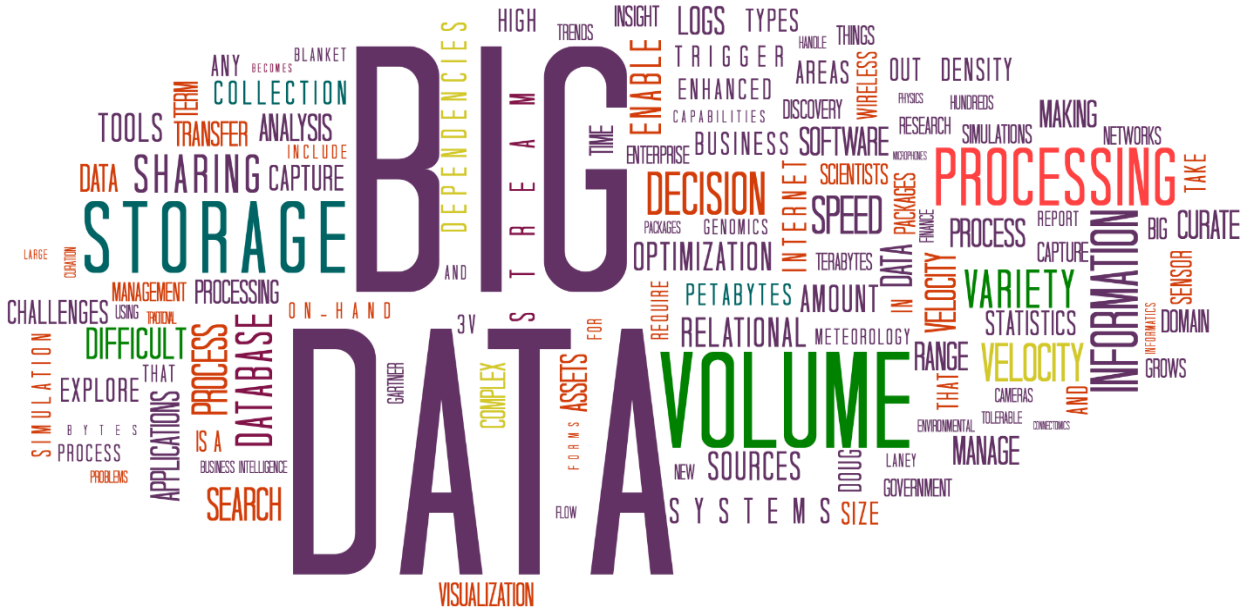
- ✓ Armonización y normalización de las variables.
- ✓ Identificación de las unidades (tanto unidades jurídicas como unidades locales).

METODOLOGÍA:

Actualización de los datos mediante Big Data:

Big Data:

- Datos masivos.
- Gran variedad.
- Gran velocidad.
- Datos no estructurados.
- Fuentes muy diversas.



METODOLOGÍA:

Actualización de los datos mediante Big Data:

Limitaciones del Big Data:

- ✓ Difícil acceso a los datos.
- ✓ Usabilidad de los datos: estabilidad en el tiempo.
- ✓ Dificultades metodológicas: selección, representatividad y validez.
- ✓ Problemas con la privacidad y la confidencialidad de los datos.
- ✓ Infraestructura y seguridad.
- ✓ Necesidad de nuevas destrezas.
- ✓ Dificultad en el paso de la investigación a la producción.

METODOLOGÍA:

Actualización de los datos mediante Big Data:

Fases del proyecto:

- Identificación inequívoca de los sitios webs oficiales de las empresas con actividad económica en Euskadi.
- Utilizar la información de esas páginas web para actualizar diferentes variables del directorio.

METODOLOGÍA:

Actualización de los datos mediante Big data:

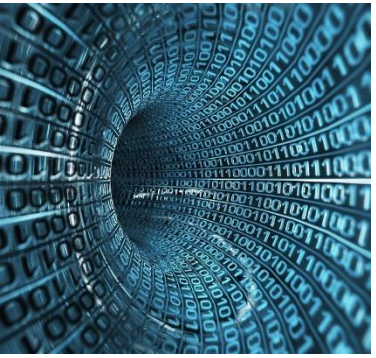
Identificación de los sitios webs oficiales :

- Se fundamenta en el proyecto ESSnet Big Data de Eurostat (SummaIstat Scraping Tools).
- Se utilizan técnicas de web scraping/crawling y modelos deterministas para calcular la URL oficial de cada Unidad Jurídica.
- El proceso incluye la utilización de un motor de búsqueda como principal fuente de información.

METODOLOGÍA:

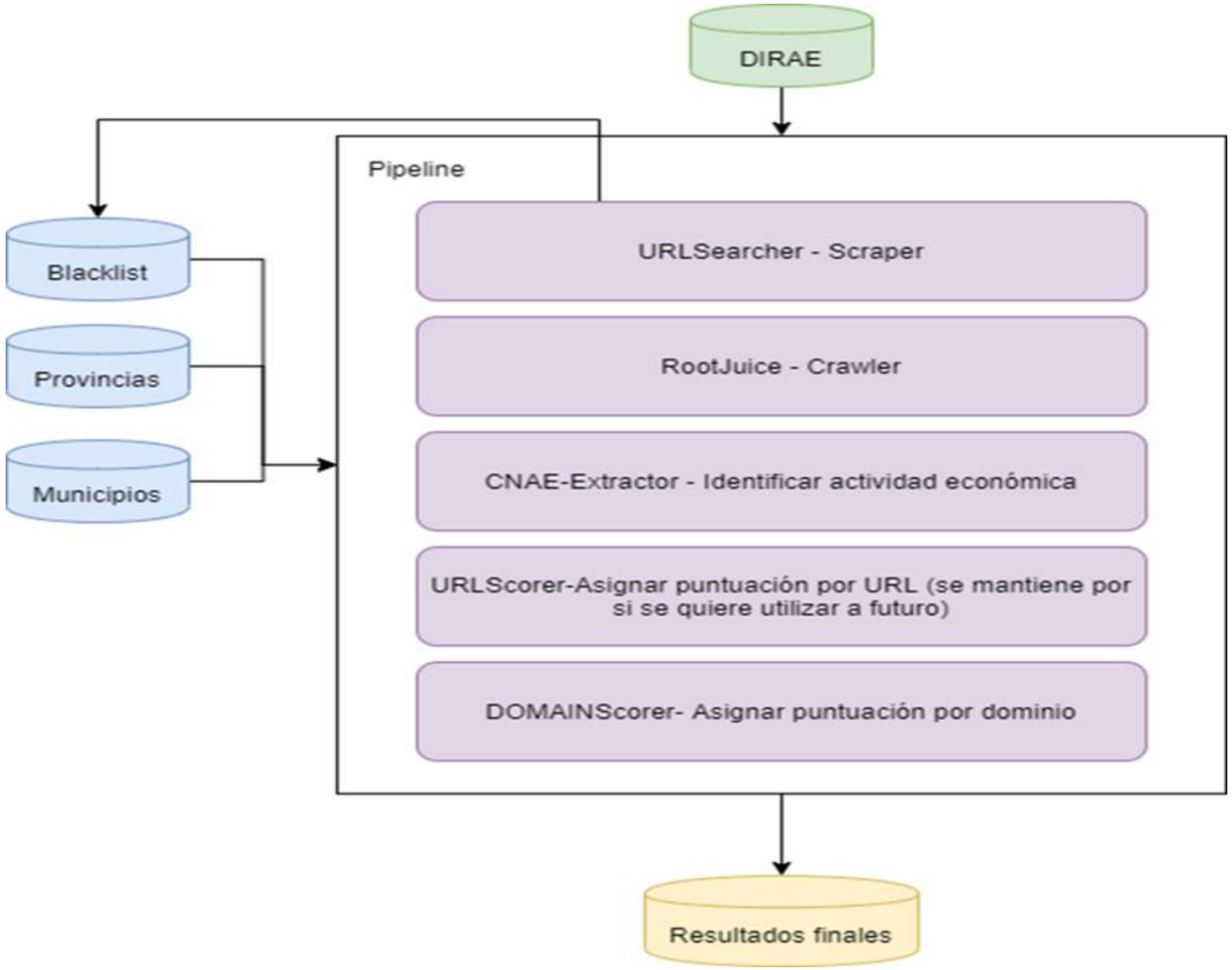
Actualización de los datos mediante Big Data:

Herramientas utilizadas:



- URLSEARCHER: permite realizar búsquedas mediante BING y GOOGLE.
- ROOTJUICE: recopila y almacena datos de la web.
- URLSCORER: asigna una puntuación de las páginas según su contenido.
- URLMATCHTABLEGENERATOR: genera un conjunto de entrenamiento para que el modelo sea capaz de reconocer la URL correcta para cada empresa.
- PIPELINE: configura todo el proceso y parámetros.
- CNAE-EXTRACTOR: herramienta que identifica la actividad económica (basado en diccionarios de A38) en cada URL obtenida.
- DOMAIN SCORER: agrupa las URL descargadas por dominio.
- DICCIONARIOS DE ACTIVIDADES ECONÓMICAS: a partir del procesamiento del lenguaje natural y la lematización de sustantivos.

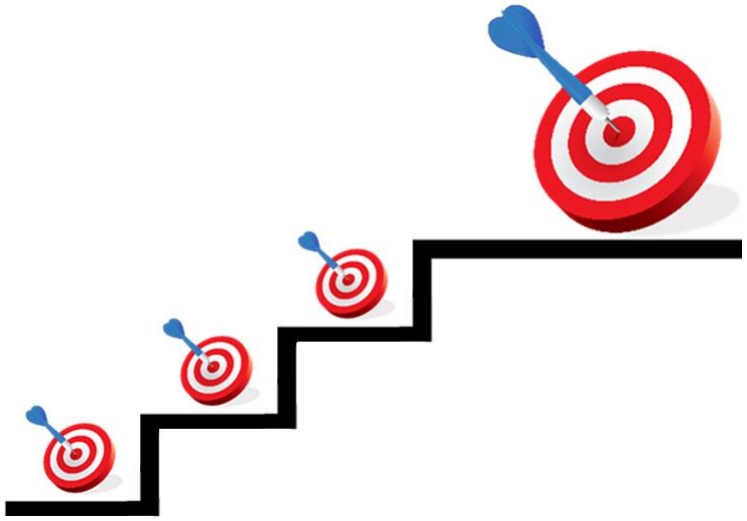
METODOLOGÍA:



METODOLOGÍA:

Actualización de los datos mediante Big data:

Desarrollos en marcha:



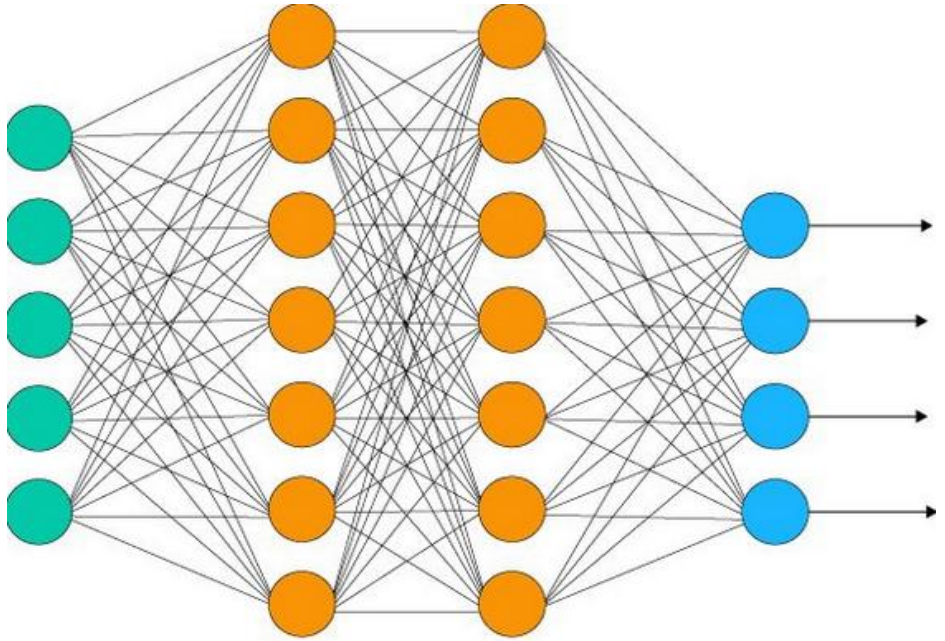
- ✓ Detección de idiomas de las páginas (incluido euskera).
- ✓ Detección de ofertas de trabajo en las webs de las empresas.
- ✓ Mortalidad de las empresas: hacer uso de Google para obtener listados de empresas que probablemente estén de baja.
- ✓ Detección de URL que tenemos en DIRAE y son obsoletas.
- ✓ Análisis de procedimientos para conseguir información de las empresas pequeñas (las de 1, 2 y 3 de empleo). Estas no suelen tener una página web propia, sino que utilizan Facebook, Instagram…

METODOLOGÍA:

Actualización de los datos mediante Big data:

Desarrollos futuros:

- ✓ Aplicar métodos de NLP para la extracción de características de las empresas.
- ✓ Aplicar redes neuronales para determinar la clasificación CNAE.



RESULTADOS:

Actualización de los datos mediante ficheros administrativos:

Datos 2019	→	Modificaciones y bajas
Modificaciones	→	36.708 (21,4%)
Bajas	→	4.772 (2,8%)

TOTAL → 41.480 (24,2%)

RESULTADOS:

Actualización de los datos mediante Big Data:

Testeo con 2.400 web → 85% de las web buenas

3.000 empresas sin web en DIRAE → recuperadas 50% de webs correctas

CONCLUSIÓN:

Tanto los registros administrativos como el Big Data son herramientas indispensables para la estádística y en concreto para la actualización del Directorio de actividades económicas.

Aportaciones importantes:

- Reducción de la carga de trabajo de los informantes
- Reducción de los plazos de publicación
- Reducción de los costes económicos
- Mantener la hegemonía en la difusión de datos frente a otras organizaciones y empresas

Retos:

- Cambios en la formación, metodología, infraestructura y tecnología.

XXI Jornadas de Estadística de las Comunidades Autónomas

Reingeniería del Directorio de Actividades Económicas: de los Registros administrativos al Big Data

Amparo Ruiz Gil

GRACIAS POR VUESTRA ATENCIÓN

MILA ESKER

amparo_ruiz@eustat.eus