*FAIR*ización 2.0 del INE: Automatización de cubos de datos semánticos

Vilches-Blázquez, Luis M.; Conde-Herreros, Diego; Noya, Isaac; Lorente Morata, Ana Cecilia; Corcho, O.

Tema: Nº. 4 - Inteligencia artificial

Resumen

En el contexto de interoperabilidad actual, la publicación de datos conforme a los principios FAIR (*Findable*, *Accessible*, *Interoperable*, and *Reusable*) (Wilkinson *et al.*, 2016) es prioritaria en el ámbito de la información geográfica y estadística. Para superar las limitaciones de los enfoques tradicionales, este trabajo presenta una arquitectura avanzada (versión 2.0) para la FAIRización y publicación sostenible de datos estadísticos del Instituto Nacional de Estadística (INE) mediante la generación automatizada de cubos de datos semánticos.

La principal innovación radica en el desarrollo de un pipeline de automatización que explota dinámicamente las bases de datos y la API JSON del INE, rompiendo la dependencia de transformaciones tabulares estáticas (ver Figura 1). El núcleo de la transformación RDF se apoya en el motor Morph-KGC, bajo el cual se han adoptado vocabularios estándar como RDF Data Cube Vocabulary, SKOS y XKOS. Asimismo, se construyó una ontología de dimensiones y medidas estadísticas a partir del estándar SDMX a través de un proceso asistido por grandes modelos de lenguaje (LLM, por sus siglas en inglés). En esta fase, mediante ingeniería de *prompts*, se evaluaron y compararon diversos aspectos del proceso de construcción de la ontología para refinar y consolidar la semántica inicial del modelo.

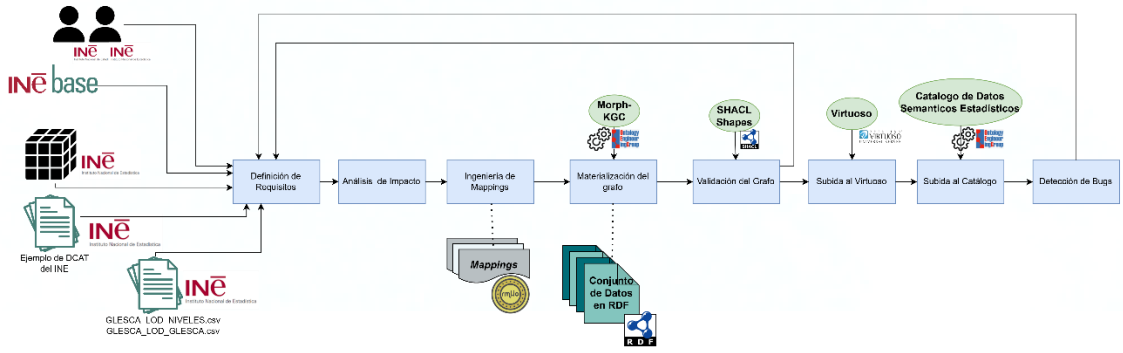

Figura 1. Pipeline para la construcción automática de cubos de datos semánticos

El despliegue de este pipeline parametrizado ha permitido la publicación de más de 40 cubos de datos semánticos en diversos temas, así como la identificación sistemática de problemas latentes de calidad en los datos de origen, tales como inconsistencias estructurales en las respuestas de la API JSON y heterogeneidades en las codificaciones de las clasificaciones estadísticas.

Por otra parte, para resolver la brecha entre datos estadísticos y geográficos, nuestro trabajo se integró de forma nativa con interfaces OGC API *Features*, permitiendo cruzar dinámicamente las geometrías de secciones censales con las observaciones estadísticas. Para la validación de esta integración se desarrolló un caso de uso sobre la gentrificación en la ciudad de Madrid. Este escenario demostró la viabilidad de realizar consultas SPARQL para integrar múltiples cubos de datos que previamente operaban como silos. En definitiva, este trabajo demuestra cómo la combinación de tecnologías semánticas, la orquestación de APIs, la curación asistida por *LLM* y la utilización de estándares para datos estadísticos pueden transformar la publicación de datos del INE, sirviendo como modelo replicable para avanzar hacia una gestión abierta, interoperable y sostenible de las estadísticas oficiales.

Palabras clave

Datos estadísticos, Principios FAIR, cubos de datos semánticos, ontología, OGC API *Features*, API JSON del INE, Modelos de Lenguaje (LLM), integración estadística-espacial.

Referencias

Wilkinson, M. D., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18.

Autores

Luis M. Vilches-Blázquez
luis.vilches@upm.es
Ontology Engineering Group
Universidad Politécnica de Madrid

Diego Conde-Herreros
diego.conde.herreros@upm.es
Ontology Engineering Group
Universidad Politécnica de Madrid

Isaac Noya
isaac.noya@alumnos.upm.es
Ontology Engineering Group
Universidad Politécnica de Madrid

Ana Cecilia Lorente Morata
ana.lorente.morata@ine.es
Instituto Nacional de Estadística

Oscar Corcho
correo@gmail.es
Ontology Engineering Group
Universidad Politécnica de Madrid