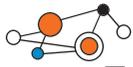


PROYECTOS EUROPEOS



Con la colaboración de la Unidad de Cultura Científica de la Universidad de Zaragoza

TECNOLOGÍAS DEL LENGUAJE

>UN ECOSISTEMA DE DATOS ENLAZADOS

Las tecnologías del lenguaje permiten crear aplicaciones informáticas para analizar y procesar el lenguaje humano, tanto hablado como escrito. Juegan un papel clave para superar las barreras idiomáticas en Europa y van allanando el camino hacia un verdadero mercado único digital. No obstante, los datos que estas aplicaciones necesitan raramente están 'listos para usar', obligando a los especialistas humanos a invertir el 80% de su tiempo recolectando, limpiando y organizando datos de diversas fuentes



ROMPER BARRERAS ¿Cómo se las arregla una tienda 'online' creada

en España para acceder a clientes de habla búlgara, estonia o portuguesa? El multilingüismo tiene un indudable valor cultural y social en Europa, pero también puede dificultar la difusión de servicios digitales en internet a través de distintos países. Para romper esas barreras del lenguaje, la Comisión Europea lanzó en octubre de 2017, en el marco del programa H2020, la convocatoria 'A multilingual Next Generation Internet'; en Prêt-à-LLOD, uno de los proyectos seleccionados, participa la Universidad de Zaragoza junto a otros nueve socios.

LISTOS PARA USAR El objetivo principal de Prêt-à-LLOD es crear un ecosistema de datos lingüísticos en la web que estén 'listos para usar' por parte de aplicaciones y servicios de procesamiento del lenguaje, así como de impulsar una nueva generación de servicios interoperables que harán uso de dichos datos. Esos datos estarán enlazados y serán accesibles mediante tecnologías de la web semántica. Se proporcionará un conjunto de aplicaciones y herramientas para cubrir las distintas necesidades en el ciclo de vida de los datos y aplicaciones: generación de datos, enlazado de los mismos, composición de servicios, etc., así como metodologías para construir cadenas de valor de los datos para una variedad de sectores y aplicaciones.

Una de las tecnologías clave en el proyecto es la de los datos lingüísticos enlazados, también referidos como LLOD por sus siglas en inglés (Linguistic Linked Open Data), que permiten, por ejemplo, convertir datos de recursos lingüísticos como diccionarios, corpus, glosarios, etc. al formato de los datos enlazados para su publicación en la web, que sean fácilmente accesibles y sea sencillo establecer relaciones (enlaces) entre ellos.

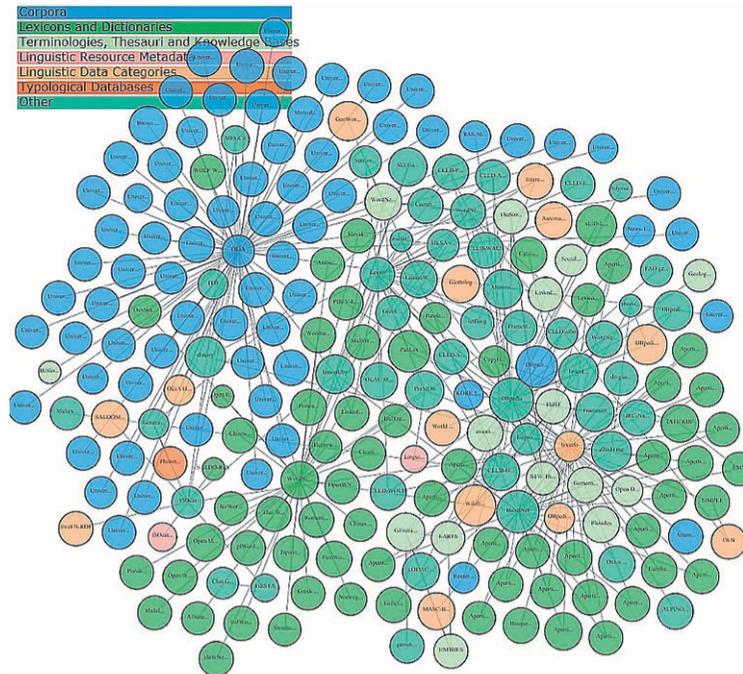
EL PROYECTO

- **NOMBRE** Prêt-à-LLOD 'Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors'.
- **SOCIOS** Dos universidades de España (Universidad de Zaragoza y Politécnica de Madrid), dos de Alemania (Universidad de Bielefeld y de Frankfurt), la National University of Ireland Galway, que coordina el proyecto, además del centro de investigación en inteligencia artificial de Alemania (DFKI) y de cuatro empresas: Oxford University Press (Reino Unido), Derilinx (Irlanda), Semalytix (Alemania) y Semantic Web Company (Austria).
- **FINANCIACIÓN** 2.997.181 euros procedentes del programa H2020.
- **DURACIÓN** Tres años: 2019-2021.
- **MÁS INFORMACIÓN** www.pret-a-llood.eu/.

¿QUÉ SON LOS DATOS LINGÜÍSTICOS ENLAZADOS?

La idea de los datos enlazados es disponer de identificadores únicos a escala de la web para representar cualquier entidad (por ejemplo, 'Zaragoza') y poder relacionarla con otras entidades (que pueden estar descritas en cualquier otra parte de la web) para aportar información sobre ellas ('Zaragoza es capital de Aragón'). Cuando esa idea se aplica a información de tipo lingüístico, se pueden representar aspectos léxicos y gramaticales, como por ejemplo que la palabra 'Zaragoza' es un nombre propio o que 'Saragosse' es la traducción al francés de 'Zaragoza'. El resultado de convertir recursos lingüísticos al formato de los datos enlazados es que dichos recursos dejan de estar aislados y representados en formatos propietarios y pasan a formar parte de un grafo que conecta y relaciona toda esa información lingüística para un mejor acceso y reutilización a escala de la web.

JORGE GRACIA DEL RÍO INVESTIGADOR PRINCIPAL DEL PROYECTO PRÊT-À-LLOD POR PARTE DE LA UNIVERSIDAD DE ZARAGOZA



Nube de los datos lingüísticos enlazados en la web. Los círculos representan distintos recursos lingüísticos, clasificados por colores, y las líneas indican enlaces entre ellos. UNIZAR

¿CUÁLES SON LAS APLICACIONES PRÁCTICAS?

Las herramientas desarrolladas en Prêt-à-LLOD se aplicarán en diversos sectores como farmacia, compañías tecnológicas o gobierno electrónico, por parte de las empresas que participan en el consorcio. En farmacia, por ejemplo, está cobrando una creciente importancia la búsqueda de 'evidencias del mundo real' (o RWE por sus siglas en inglés). Consiste en reunir pruebas de la efectividad y seguridad de un producto farmacológico, extraídas de datos que no sean estudios clínicos controlados para revelar información sobre el tratamiento en condiciones reales, fuera de un laboratorio. La extracción de RWE requiere analizar grandes volúmenes de datos textuales donde médicos y pacientes valoran subjetivamente el producto, y es ahí donde el proyecto puede ayudar. Se trabajará con ese tipo de contenido, que será anotado semánticamente y enlazado a otras fuentes de datos para mejorar la extracción de RWE, así como su adaptación a otros idiomas.

Otra de las aplicaciones del proyecto dará soporte al desarrollo de servicios abiertos de gobierno electrónico a través de las fronteras. Se aplicarán tecnologías del lenguaje desarrolladas en Prêt-à-LLOD para mejorar las relaciones con el ciudadano (por ejemplo mejorando las capacidades multilingües de los portales de las Administraciones públicas) en proyectos de ciudades inteligentes como Smart Dublin y Smart Cork Gateway.

¿QUÉ VENTAJAS TIENE ENLAZAR RECURSOS LINGÜÍSTICOS?

Imaginemos varios diccionarios bilingües en formato electrónico: inglés-español, español-portugués, portugués-italiano. Si enlazamos las entradas de dichos diccionarios, por ejemplo para representar que la palabra 'casa' en el primer diccionario es la misma que 'casa' en el segundo, seríamos capaces de obtener traducciones que no están disponibles en los diccionarios originales pero sí en su versión 'enlazada', por ejemplo del inglés al portugués o al italiano, sin más que recorrer los enlaces y aplicar algoritmos que calculan el grado de confianza que tiene la nueva traducción. Adicionalmente, las entradas de estos diccionarios pueden enlazarse a otros recursos de la web como DBpedia (la versión en datos enlazados de la Wikipedia) o Babelnet, para enriquecer la información inicial con información enciclopédica, imágenes, descripciones textuales, traducciones a otros idiomas, etc., sin tener que modificar los datos originales (solo añadiendo los enlaces correspondientes).