E5.1 SERVICIOS DE INTEGRACIÓN

ESPACIO DE DATOS LINGÜÍSTICO (SER 15/23 OTT)

Resumen

Este documento describe y analiza las fuentes de datos que se valora integrar en el espacio de datos lingüísticos INESData. Entre los principales criterios del análisis se incluye novedad de los recursos a integrar en el ámbito de las plataformas de intercambio de recursos de tecnologías de lenguaje, cantidad de recursos a integrar, el estado tecnológico de la fuente de datos y la disponibilidad de un API para realizar la integración. Con estos criterios se analiza las plataformas TERESIA, CLARIAH-ES, CLARIN Centro de conocimiento - español, y el European Language Grid. El análisis concluye que es necesario esperar el desarrollo tecnológico de las plataformas más noveles y de interés para INESDATA para su integración en el espacio de datos.

Andrés García-Silva

28/06/2024

Expert.ai Expert.ai Language Technology Research Lab

Historia

revisión	Fecha	Descripción	Autor
0.1	20/06/2024	Primera versión del documento	Andrés García-Silva
0.2	31/06/2024	Revisión del documento	José Manuel Gómez

Contenido

1 Introducción

En este documento se analiza y evalúa diversas fuentes de datos que se consideran para su integración en el espacio de datos lingüísticos INESData. El análisis de las plataformas se realiza para garantizar que los recursos lingüísticos que se integran sean interesantes para la comunidad de las tecnologías del lenguaje en español y las lenguas cooficiales. Además, se valora posibilidad de integrar programáticamente la plataforma en el espacio de datos. La integración de recursos externos en el espacio de datos es fundamental para garantizar que espacio de datos sea útil por ejemplo para futuras investigaciones y aplicaciones en el ámbito de las tecnologías del lenguaje.

El estudio se centra en cuatro plataformas principales: TERESIA, CLARIAH-ES, el Centro de Conocimiento CLARIN - español, y el European Language Grid. Cada una de estas plataformas ofrece una serie de recursos y herramientas lingüísticas que pueden enriquecer significativamente el espacio de datos de INESData. Los criterios de análisis incluyen la novedad de los recursos ofrecidos, la cantidad y su cobertura al español y lenguas cooficiales, el estado tecnológico de las plataformas y la disponibilidad de API para facilitar su integración.

Los resultados del análisis indican que algunas plataformas de interés para INESDATA como TERESIA o CLARIAH_ES están en un estado incipiente y aún requieren un desarrollo tecnológico adicional antes de que puedan ser incorporadas de manera efectiva en INESData. Por otra parte, aunque ELG es la plataforma que está lista para integrar y ofrece mayor cobertura de recursos en español y lenguas cooficiales, se determina que su impacto en el espacio de datos lingüístico es limitado ya que ELG está integrada en el espacio de datos lingüístico europeo, lo que implicaría una redundancia de datos entre ambos espacios de datos.

En particular, se destaca la necesidad de esperar a que las plataformas más nuevas y prometedoras alcancen un nivel de madurez adecuado para asegurar una integración exitosa y beneficiosa para todos los usuarios del espacio de datos lingüísticos.

Plataformas para compartir recursos de TL

A continuación, se describe las principales plataformas y proyectos en el ámbito europeo que exponen o persiguen exponer recursos de tecnologías del lenguaje en español y lenguas cooficiales.

1.1 TERESIA

TERESIA[1] va a desarrollar un metabuscador y un portal de tecnologías de Inteligencia Artificial que se enfoca en la creación, validación y gestión de terminologías en español, con una perspectiva panhispánica, para garantizar la calidad, interoperabilidad y visibilidad de los recursos terminológicos en diversos ámbitos especializados. TERESIA se crea para abordar la fragmentación de las terminologías especializadas existentes y satisfacer la necesidad de disponer de terminologías en español que sean validadas y de alta calidad.

Además, se espera que TERESIA lleve a cabo las siguientes actividades:

- Creación de un corpus de literatura científica en español y otras lenguas cooficiales.
- Metodología de extracción de información mediante técnicas de IA para generar, validar y sancionar nuevos términos.
- Transformación de términos a formatos compatibles con la web de datos, haciéndolos accesibles y recuperables.
- Desarrollar un enorme conjunto de datos terminológicos abiertos, siguiendo principios FAIR (Findable, Accessible, Interoperable, Reusable).
- Facilitar la interacción con comunidades de expertos para el proceso de validación.

[1] https://proyectoteresia.org/

TERESIA empezó a finales del 2023 y a la fecha de consulta el portal del proyecto no ofrece ningún resultado o información más detallada de la planificación del proyecto que permita saber que recursos se van a integrar, cuando y en qué estado serán entregados los resultados esperados del proyecto. Sin embargo, se ha incluido TERESIA en este análisis porque UPM es parte del consorcio del proyecto y cree la integración de estos recursos en el espacio de datos aportará un gran valor a la comunidad alrededor de las tecnologías del lenguaje en español y lenguas cooficiales en España.

1.2 CLARIAH-ES

CLARIAH-ES[2] es una infraestructura digital de investigación distribuida que busca promover y coordinar la participación oficial de España en las infraestructuras europeas de investigación para las humanidades y ciencias sociales: CLARIN y DARIAH.

CLARIN[3] (Infraestructura Común de Tecnología y Recursos Lingüísticos) tiene el objetivo de dar soporte al intercambio, uso y mantenimiento de datos y herramientas lingüísticas para la investigación en Ciencias Sociales y Humanidades. Además, CLARIN persigue garantizar que todos los recursos y herramientas lingüísticas integradas en la Infraestructura sean accesibles desde un portal común a todos los investigadores.

DARIAH[4] (Infraestructura de Investigación Digital de Artes y Humanidades) tiene como objetivo fortalecer la investigación y la enseñanza en las Artes y Humanidades mediante el uso de tecnología digital. Sus esfuerzos se enfocan en promover la creación, conservación e intercambio de datos, servicios y herramientas computacionales multimodales, asegurando su accesibilidad y difusión sostenida en el tiempo. Además, la infraestructura proporciona materiales educativos y oportunidades de formación para desarrollar competencias en investigación digital y fomentar la adopción de buenas prácticas a nivel transnacional y transdisciplinar.

Aunque en la página principal de la web del proyecto se describe a CLARIAH-ES como una infraestructura digital, en el apartado de proyectos se define como una red de investigación estratégica. A fecha de consulta la web del proyecto CLARIAH-ES no presenta ningún resultado en forma de una plataforma o herramienta de búsqueda de recursos de tecnologías de la lengua, ni ningún recurso relacionado con terminología. Además, tampoco se presenta una planificación que permita prever cuando se liberarán los resultados del proyecto.

1.3 CLARIN CENTRO-K-español

Los Centros de Conocimiento CLARIN (K-centres) son un componente principal de la Infraestructura de Conocimiento CLARIN (KI). La misión de la de la Infraestructura de Conocimiento CLARIN es asegurar que el conocimiento y la experiencia disponibles estén accesibles de manera organizada tanto para la comunidad CLARIN como para la comunidad de investigación en ciencias sociales y humanidades en general. El rol de Los k-centres de CLARIN es compartir conocimiento y experiencia sobre uno o más aspectos del dominio cubierto por la infraestructura CLARIN. Los K-centres son útiles para investigadores y educadores de cualquier disciplina donde el lenguaje sea un actor importante.

El centro de conocimiento español de CLARIN[5] tiene como objetivo ofrecer los conocimientos y la experiencia de los grupos que inicialmente lo componen en la utilización de tecnología para la investigación en humanidades y ciencias sociales. El k-centre español ofrece un punto de acceso unificado y actúa como un agente de distribución de solicitudes de servicios: https://ixa2.si.ehu.eus/clarin-es/en/recursos

[2] https://www.clariah.es/es

[3] http://www.clarin.eu/

[4] http://www.dariah.eu/

[5] https://ixa2.si.ehu.eus/clarin-es/en/node/41

El portal que lista los recursos y servicios disponibles contiene 33 recursos y servicios en total que ofrecen soporte al catalán, euskera, gallego, español, inglés y otros lenguajes. Las categorías usadas para clasificar los recursos y servicios son: Corpora, Gramáticas y modelos de lenguaje, Herramientas y servicios, y Recursos Léxicos. Del total de recursos y servicios solo 8 tienen un enlace de descarga directa. Sin embargo, no se ha podido encontrar un API que permita la consulta y el acceso a los recursos y servicios de forma programática.

1.4 European Language Grid

El European Language Grid (ELG) tiene como objetivo proporcionar una plataforma tecnológica y comercial para la industria y la investigación en el ámbito de las tecnologías lingüísticas en Europa. La plataforma ofrece acceso a una amplia variedad de recursos lingüísticos, herramientas y servicios relacionados con el procesamiento del lenguaje natural y la inteligencia artificial aplicada a los idiomas. Un usuario puede compartir sus recursos en ELG de diferentes desde una infraestructura externa a ELG o por medio de servicios de almacenamiento y ejecución en la nube de ELG. ELG contiene un gran conjunto de recursos y herramientas de las tecnologías del lenguaje para el español y las lenguas cooficiales (ver Tabla 1 y Figura 1). ELG se ha nutrido de recursos integrados de importante plataformas europeas como el catálogo de ELRA[6], el repositorio META SHARE[7], recursos alojados en Huggingface[8] o Zenodo[9], y recursos subidos por los usuarios de la plataforma.

Language	***Basque***	***Catalan***	***Galician***	***Spanish***
Text Processing	52	83	53	264
Speech Processing	18	38	11	93
Information Extraction and Retrieval	26	48	27	171
Image / Video Processing	1	2	2	17
Translation Technologies	30	57	23	179
Natural Language Generation	14	19	14	51
Support operation	25	91	24	212
Human Computer Interaction	1	7	1	7
Other functions of software	23	45	20	97
Corpus	158	222	107	1031
Model	10	28	10	141
Grammar	1	4	0	7
Lexical / Conceptual resource	40	130	34	655
Uncategorized Language Description	1	1	0	2
Total	*400*	*775*	*326*	*2927*

Tabla 1. Recursos y servicios en ELG para el español y lenguas cooficiales

[6] https://catalog.elra.info/en-us/

[7] http://www.meta-share.org/

[8] https://huggingface.co/

[9] https://zenodo.org/

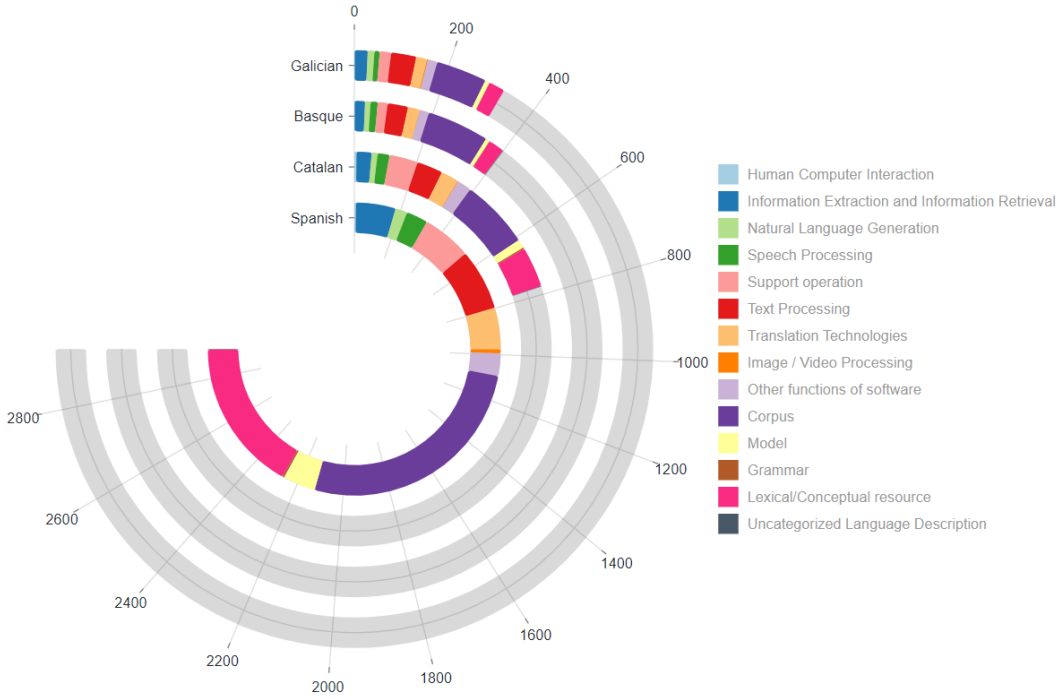

Figura 1. Comparación de los recursos y servicios en ELG para el español y las lenguas cooficiales.

Además de la gran cobertura de recursos y servicios para los lenguajes considerados principales en el espacio de datos lingüístico INESDATA, ELG ofrece acceso a su catálogo por medio de una completa API in Python.

2 Análisis de las plataformas para compartir recursos de TL

En la Tabla 2 se presenta las diferentes plataformas y proyectos para la compartición de recursos de tecnologías del lenguaje. Para las plataformas en desarrollo se indica si hay planificación disponible que permita identificar que recursos y cuando estarán disponibles. Además, se indica si hay una plataforma tecnológica operativa que soporte el intercambio de recursos y un API para el acceso programático. Finalmente, en la última columna para las plataformas operativas se reporta el número de recursos disponibles para el español y las lenguas cooficiales.

Plataforma	**Planificación**	**Operativa**	**API**	**Recursos/Servicios**
TERESIA	No	No	No	-
CLARIAH-ES	No	No	No	-
CLARIN K-centro de conocimiento	No	Si	No	33 de los que 8 son descargables.
ELG	-	Si	Si	2927 español, 775 catalán, 400 euskera, 326 gallego

Tabla 2. Comparación de los proyectos y plataformas actuales para compartir recursos.

Se puede observar que TERESIA y CLARIAH-ES están en un estado incipiente de desarrollo. Además, la no disponibilidad de información de la planificación del proyecto dificulta la evaluación de si será posible o no integrar recursos de estas plataformas en el espacio de datos lingüístico

antes de la finalización de este proyecto. Las dos únicas plataformas operativas son CLARIN k-centro de conocimiento y ELG.

Sin embargo, CLARIN k-centro de conocimiento actualmente solo expone 8 recursos descargables y no ofrece un API lo que dificulta la integración automática de los recursos en el espacio de datos. Dado que actualmente son solo 8 recursos es posible integrarlos de forma manual. Sin embargo, el impacto de estos recursos es limitado en la comunidad de las tecnologías del lenguaje.

ELG es la plataforma que está operativa, tiene el mayor número de recursos disponibles para el español y lenguas cooficiales, y ofrece un API lo que facilita el proceso de integración en el espacio de datos lingüísticos. Sin embargo, existe un conector de ELG al espacio de datos lingüístico europeo[10] lo que limita el impacto, principalmente en términos de novedad, de integrar ELG en el espacio de datos lingüístico de INESData.

3 Conclusiones

Dado que las plataformas de interés para la dirección del proyecto INESData como TERESIA están en un estado incipiente, y que el impacto de integrar plataformas como ELG es limitado, se decide postergar la integración con otras plataformas a la próxima versión de este entregable, tiempo en el que se espera que la dirección del proyecto de UPM informe de la planificación de TERESIA y/o tome una decisión en cuanto a la integración de ELG en el espacio de datos lingüístico.

[10] https://language-data-space.ec.europa.eu/index_en