ESTUDIO SOBRE DATOS REUTILIZABLES COMO RECURSOS LINGÜÍSTICOS

Plan de impulso de las Tecnologías del Lenguaje

Antonio Moreno, Doroteo Torre, Ana Valverde, Leonardo Campillos

Enero 2019

Revisado en Septiembre de 2019

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital y Red.es, que no comparten necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

En el presente informe se muestra el desarrollo del contrato "Servicio para la realización de un estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del Plan de Impulso de las Tecnologías del Lenguaje", con el que se pretende identificar conjuntos de datos y recursos que las administraciones públicas ponen a disposición de los usuarios a través de sus páginas webs, y una gran parte de ellos, bajo el concepto de Datos Abiertos (DA), y que pudieran ser susceptibles de conversión en recursos lingüísticos. Todo ello con una finalidad clara: no solo censar y localizar este tipo de recursos dentro de la administración, estableciendo pautas para su reconocimiento y creación como recurso lingüístico, sino con vistas a su posterior reutilización y aprovechamiento en sistemas de Tecnologías del Lenguaje (TL).

1 INTRODUCCIÓN

Las Tecnologías del Lenguaje son una rama de la Inteligencia Artificial, que permiten, entre otras cosas, explotar automáticamente la abrumadora cantidad de información de textual y oral a la que tenemos acceso fruto del desarrollo de la digitalización y las telecomunicaciones.

Para catalizar este potencial, el Plan de Impulso de las Tecnologías del Lenguaje (en adelante, Plan TL) es una política pública impulsada por la Secretaría de Estado para el Avance Digital que tiene como objetivo, desarrollar la industria del Procesamiento del Lenguaje Natural (PLN), la Traducción Automática (TA) y los sistemas conversacionales (SSCC) en España, y especialmente, en lengua española y lenguas cooficiales, con la finalidad de impulsar el desarrollo de la Nación emprendedora en el sector de las Tecnologías del Lenguaje (y, por extensión, de la Inteligencia Artificial), y aprovechar estas novedosas tecnologías para mejorar el servicio público.

El estudio que aquí exponemos se realiza con el fin de desarrollar principalmente los ejes 1 y 3 del Plan TL:

- **Eje 1: Desarrollo de infraestructuras lingüísticas:** Dirigido a aumentar el número, la calidad y la disponibilidad de las infraestructuras lingüísticas (recursos, procesadores y campañas de evaluación) de propósito general en español y lenguas cooficiales.

- **Eje 3: La Administración Pública como impulsor de la industria del lenguaje**: Dedicado a la mejora de la calidad y capacidad del servicio público, incorporando las tecnologías del lenguaje (incluyendo el procesamiento del lenguaje natural, la traducción automática y los sistemas conversacionales), y actuando, además, como tractor de la demanda. Este aspecto apoya, también, la generación, estandarización y difusión de recursos lingüísticos creados en el

contexto de la actividad de gestión pública propia de la Administración en el marco de la **política de Reutilización de la Información del Sector Público (RISP)**.

Del mismo modo que se ha expuesto en estudios previos realizados dentro del marco del Plan TL, entendemos que es de vital importancia señalar que el "apoyo institucional al desarrollo de recursos lingüísticos a partir de los datos recensados permitirá el avance de las TL, tanto desde la investigación básica de los grupos de investigación como de la I+D activa en el tejido industrial emergente". Fruto de este avance y apoyo son los informes realizados por Bel y Rigau en 2015 para una consulta del amplio panorama de herramientas y recursos lingüísticos (RL), así como el estudio de asociaciones, grupos de investigación e iniciativas y empresas de PLN en España, también para lenguas cooficiales [1]. Un avance significativo que incluye también al informe de Soroa y otros (2017) para un análisis de herramientas PLN de aplicación industrial a gran escala [2].

Por tanto, y con el fin de continuar en esa línea de mejora y apoyo institucional al desarrollo de recursos lingüísticos, y dada la naturaleza del estudio que nos ocupa, podemos resumir los objetivos finales primordiales de nuestra investigación en:

- **Ofrecer una revisión exhaustiva de documentos publicadosrecursos por las diferentes administraciones públicas** que puedan ser **convertidos en RL.**
- **Extraer una propuesta de plan de acción que sirva para múltiples documentos de las administraciones, de forma que puedan convertirse en potenciales RL**, a partir del análisis de los recursos censados en el proyecto y de sus diferentes grados de madurez.

2 CONCEPTOS

Se definen a continuación una serie de conceptos que consideran clave para comprender el alcance de los trabajos:

- **Dato reutilizable como recurso lingüístico**

Se refiere *dato reutilizable* como toda información o parte de ella, cualquiera que sea su soporte o forma de expresión, sea esta textual, gráfica, sonora, visual o audiovisual, incluyendo los metadatos asociados y los datos contenidos con los niveles más elevados de precisión y desagregación. En línea con esta definición, se consideran también documentos los vocabularios, taxonomías, tesauros, ontologías, etc.

- **Datos abiertos**

Se consideran *datos abiertos* aquellos que cualquiera es libre de utilizar, reutilizar y redistribuir, con el único límite, en su caso, del requisito de atribución de su fuente o reconocimiento de autoría.

- **Portal de Datos abiertos**

Se entiende por *portal de datos abiertos* un espacio web dedicado en el que los órganos de la Administración informan, de manera estructurada y usable, acerca de qué documentación es susceptible de ser reutilizada, los formatos en que se encuentra disponible, las condiciones aplicables a su reutilización, los mecanismos de recuperación disponibles tales como listados, bases de datos o índices de información reutilizable, entre otros aspectos.

- **Recurso lingüístico**

Se denominan *recurso lingüístico,* tanto a las potenciales entradas y salidas de los procesadores lingüísticos, como a los textos y datos auxiliares que puedan necesitar.

- **Inteligencia competitiva**

En el ámbito del Plan TL, se entiende por *inteligencia competitiva* los aspectos que versan sobre el control y diseño de políticas públicas, que permite obtener y estructurar información para generar un conocimiento que ayude en la toma de decisiones, con el fin último de mejorar la calidad de las políticas públicas y la gestión de los servicios internos de la Administración. Los sistemas de inteligencia competitiva pueden aprovechar los beneficios de las tecnologías del lenguaje, junto con otras herramientas de análisis, para dos objetivos fundamentales:

- Dirección de Políticas públicas: Disponer de herramientas de apoyo al diseño, desarrollo, seguimiento y evaluación de políticas públicas que, mediante el análisis automático de corpus textuales de diversas fuentes, extraen conocimiento objetivo tanto actual como histórico (evolución temporal), y con capacidad de facilitar información con el grado de síntesis o detalle que se necesite.
- Evaluación y seguimiento de Ayudas: Servir de apoyo a la evaluación de la innovación y el riesgo de solicitudes de ayudas públicas, y al seguimiento y a la lucha contra el fraude.

Es oportuno precisar que en este informe, la concepción de inteligencia competitiva engloba todos los recursos de utilidad para la toma de decisiones de interés económico, político o industrial.

3 METODOLOGÍA DEL ESTUDIO

Uno de los primeros objetivos que se plantean en este estudio es **localizar conjuntos de datos o documentos de distintos organismos que resulten de interés desde el punto de vista de las Tecnologías del Lenguaje, y que sean, además, factibles para su conversión en recursos lingüísticos.** Como sabemos, este estudio se encuadra en el Plan TL antes mencionado, por lo que, para una mayor integración del análisis, se parte de metodologías ya asentadas en estudios previos dentro de este Plan, y, más concretamente, de la caracterización de recursos lingüísticos estandarizada por la Red de Tecnologías del Lenguaje (ReTeLe).[1] No obstante, tenemos que indicar que ha sido necesario adaptar en parte esa caracterización de ReTeLe, incluyendo también campos sobre el grado de proximidad de los datos a un recurso lingüístico, así como otros campos tomados de las fuentes indicadas en la Sección 3.2.2 de este estudio. Por otra parte, y como se reflejará más adelante, existen variaciones que se han ido incorporando como consecuencia de la naturaleza de los datos encontrados, así como de las nuevas consideraciones que se han ido tomando a lo largo del proyecto, generalmente, relacionadas con problemas con formatos y licencias de los documentos, así como solapamientos con las acciones de prospección desarrolladas por otros proyectos en curso dentro del mismo Plan TL.

3.1 ASPECTOS NORMATIVOS

A modo de resumen, es importante destacar aquellas normas de obligado cumplimiento para la buena ejecución de este estudio y sus tareas asociadas. Por un lado, cabe señalar la Ley sobre la Reutilización de Información del Sector Público (RISP) 18/2015, de 9 de julio de 2015, que se publicó modificando la Ley anterior 37/2007, de 16 de noviembre, para incorporar al ordenamiento jurídico español la Directiva europea 2013/37/UE, que tiene la misma finalidad: i.e. potenciar la reutilización de la información del sector público. Como se indica en su preámbulo, con esta ley *"se persigue facilitar la creación de productos y servicios de información basados en documentos del sector público, garantizar la eficacia en el uso transfronterizo de documentos del sector público por empresas privadas y ciudadanos y promover la libre circulación de información y la comunicación, garantizando el respeto a la seguridad jurídica, la protección de los datos personales, así como la propiedad intelectual e industrial"*. Respecto a esta cuestión, se debe tener en cuenta que recientemente se ha aprobado la Directiva (UE) 2019/1024, de 20 de junio de 2019, relativa a los datos abiertos y la reutilización de la

[1] http://retele.linkeddata.es

información del sector público que introduce algunos cambios significativos y que deberá ser transpuesta a nuestro ordenamiento a más tardar el 17 de julio de 2021[2].

Asimismo, como principal marco normativo de este proyecto, se contempla tanto lo previsto en la LOPDGDD (Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales) y en el nuevo Reglamento General de Protección de Datos (RGPD)[3] que entró en vigor en 2016 y cuyo plazo de adaptación concluyó el 25 de mayo de 2018. Debe destacarse que esta norma europea es directamente aplicable. También se tienen en cuenta otras normas similares que pudiesen limitar, total o parcialmente, el uso de los recursos y su tratamiento para una posterior reutilización y difusión[4]. En cualquier caso, para todo el proceso, se ha respetado también lo expresado en el Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de Ley de Propiedad Intelectual (última modificación por Ley 2/2019, de 1 de marzo).

Del mismo modo, se ha atendido a lo expuesto en el Esquema Nacional de Interoperabilidad (Real Decreto 4/2010, de 8 de enero) y en la Norma Técnica de Interoperabilidad de Reutilización de recursos de la información (Resolución de 19 de febrero de 2013 de la Secretaría de Estado de Administraciones Públicas), pues constituye un marco de referencia para, entre otros aspectos, la selección, identificación, descripción, y puesta a disposición de los documentos y recursos de información relativos al sector público, de forma que se pueda garantizar una mejor reutilización y unos formatos y estándares compatibles con documentos y recursos ya disponibles en el ámbito de la Administración.

Por tanto, a la luz de lo expuesto, se ha desarrollado un sistema de censado o elección basado en aspectos fundamentales de cada recurso. La selección se realiza en función de sus propiedades, y centrándonos en los siguientes aspectos, que se consideran fundamentales:

---

[2] Para conocer con detalle el contenido de la nueva directiva puede consultarse la nota técnica del OBSAE accesible en: https://administracionelectronica.gob.es/pae_Home/dam/jcr:7957c02d-fe20-4c25-966a-358d5e119267/2019-06-Nota-tecnica-OBSAE-nueva-directiva-RISP.pdf.

[3] https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R0679&from=ES. Al respecto, puede ser de utilidad la siguiente "Guía del RGPD para responsables del tratamiento": https://www.aepd.es/media/guias/guia-rgpd-para-responsables-de-tratamiento.pdf

[4] Para conocer el impacto de la normativa de protección de datos en las administraciones públicas puede consultarse la infografía elaborada por la Agencia Española de Protección de Datos accesible en: https://www.aepd.es/media/infografias/infografia-adaptacion-rgpd-aapp.pdf.

- Su **localización y estructura**: Aquellos que contengan una forma adecuada, estén más estructurados y expuestos o identificados de forma clara en las diferentes páginas webs.
- Su **potencial utilidad** para generar recursos lingüísticos y su posible impacto para diversas tareas de tecnologías del lenguaje en general (anotación, generación de modelos, entrenamiento de clasificadores, etc.), o para algún dominio específico de los identificados como prioritarios en el estudio (es decir, para los ámbitos de Sanidad, Justicia, Inteligencia competitiva o Cultura).
- El **volumen disponible y su calidad**, entendida como aspectos relativos a los atributos de calidad inherente a los productos de datos definidos por la norma ISO/IEC 25012 [3]. En concreto: exactitud, completitud, consistencia, credibilidad, actualidad y accesibilidad.
- El **formato, la estructura y los metadatos** del documento, como aborda la ficha técnica (sección 3.3), pues constituye uno de los criterios de identificación e interoperabilidad más consistente.
- La **licencia de uso** que pudiera limitar su aprovechamiento como recurso lingüístico abierto de algún modo y otros posibles aspectos legales (v.g.: licencias propietarias, derechos de propiedad intelectual, derechos de protección de datos, etc.). Se elegirán, por tanto, aquellos documentos o conjuntos de datos que se consideren más fáciles de reutilizar, y, por consiguiente, se dará prioridad a aquellos que contengan licencias de uso gratuito o de tipo *Creative Commons* (preferiblemente, de solo reconocimiento).
- La existencia de **tarifas** ligadas a su reutilización: Se primarán aquellos en los que no exista una tarifa, o, en todo caso, de existir, se detallará convenientemente.
- El **tratamiento necesario** para convertirlo en recurso lingüístico interoperable y reutilizable, dando claves sobre qué tipo de RL podría derivarse y qué tipo de acciones deberían de emprenderse para ello.

Esta primera caracterización del recurso y su cercanía a la norma de interoperabilidad inspiró, en parte, la **ficha de madurez** que acompaña a cada recurso estudiado, de forma que con dicha caracterización se agrupase a los recursos en diferentes tipos de madurez: baja, media, y alta, que ayudasen a discriminar los recursos más cercanos a un RL y los procesos que se tienen que seguir para convertir aquellos de menor madurez, como veremos más adelante en la sección 7.

3.2 ASPECTOS TÉCNICOS

3.2.1 Diferencia entre recursos de datos abiertos y recursos lingüísticos

Una vez establecidos los primeros pasos para la identificación de recursos, creemos necesario delimitar la diferencia entre datos abiertos susceptibles de ser convertidos en recursos lingüísticos (RL) y los que ya pueden denominarse RL.

Por **recurso lingüístico** (**RL**) se entiende **cualquier fichero electrónico que ha sido procesado para servir de fuente, entrenamiento o evaluación de un sistema de tecnologías del lenguaje**. Ejemplos de RL son corpus escritos y orales, lexicones, ontologías o listas de entidades.

Por **datos abiertos** (**DA**) se entienden aquellos que **cualquiera puede utilizar, reutilizar y redistribuir**, con el único límite, en su caso, del requisito de **atribución de su fuente o reconocimiento de autoría**.

Los datos abiertos pueden presentar diferentes formatos (texto, número e imágenes) y contener información de diversa índole (geográfica, estadística, etc.). Por tanto, es interesante indicar que, en el presente documento, se denominará **recurso** a todo conjunto de **datos abiertos** (**RDA**) que esté en formato texto, audio o video (en un formato fácilmente procesable para las tareas de PLN) y que pueda ser interesante para su conversión en recurso lingüístico.

Así, **para que los datos abiertos se conviertan en un recurso lingüístico es necesario recopilarlos y adaptarlos a formatos que sean susceptibles de ser utilizados por las aplicaciones de Tecnología Lingüística**.

Los recursos se presentan, como ya hemos mencionado anteriormente, con **distintos grados de madurez**. Lo más básico son los **datos brutos**, que no contienen ni estructura ni anotación. En muchos casos, el formato no es utilizable directamente en una herramienta PLN, como es el caso de los documentos en PDF. El siguiente grado son los **datos primarios** (p. ej. en formatos como TXT o CSV), que contienen ya cierta estructuración, pero que no están anotados lingüísticamente. Finalmente, los datos abiertos más maduros (**datos secundarios**) contienen metadatos y están anotados con información lingüística. Se pueden distinguir más grados (p. ej. recursos derivados o corpus), como se puede ver en el informe previo de Bel y Rigau (2015: 36) [1].

Como ya mencionamos anteriormente, el objetivo de este estudio es localizar y censar, ante todo, conjuntos de datos y valorar el grado de madurez para su conversión en un RL. No obstante, tras el primer análisis realizado de diferentes recursos con una naturaleza muy dispar (documentos en formatos como PDF, XML, datos provenientes de sistemas de búsqueda avanzada, etc.), se incluyeron otros recursos lingüísticos propiamente dichos (como memorias de traducción o bases terminológicas) que se encontraban en repositorios de datos abiertos, de manera que sirvieran de ejemplo de recurso con madurez más alta, como veremos más adelante en este informe.

3.2.2 Portales nacionales y extranjeros de referencia para la recogida de datos

Una vez establecida la diferencia entre RDA y RL, es necesario encontrar los datos que formarán parte del estudio, y para ello, se decidió realizar una primera búsqueda de portales que pudieran ofrecer datos de distinta naturaleza. En principio, se tomaron como punto de partida portales de datos abiertos, tanto nacionales como de diferentes comunidades autónomas y grandes ciudades (p.ej. Madrid, Barcelona, o Zaragoza). Posteriormente, con el fin de recensar el mayor número de datos posible, se amplió esta búsqueda a otras páginas web de diferentes entidades públicas de nuestro país (Institutos de Salud, Academias nacionales y Reales Academias, universidades públicas, agencias estatales, entre otros), en las que se han ido identificando aquellos recursos de mayor interés para este estudio. Esto es, fundamentalmente, aquellos que tuvieran una buena calidad, un volumen importante de datos, una estructura en diferentes lenguas cooficiales o una licencia menos restrictiva, entre otros aspectos.

A título ilustrativo, pues la totalidad de sitios consultados ha sido demasiado extensa como para citarla a continuación, se muestra una lista de los principales sitios consultados que contienen datos en diferentes administraciones de nuestro país. A partir de esta lista se ha ido implementando la consulta de datos, como ya comentamos previamente, de manera que, una vez recorridas las páginas iniciales, se ha hecho extensiva a todo tipo de páginas web de carácter oficial estatal dentro de los ámbitos considerados de prioridad para el estudio: Sanidad, Justicia, Cultura e Inteligencia competitiva.

3.2.2.1 Portales de datos abiertos de España

- Portal de datos abiertos del Gobierno: http://datos.gob.es/
- Portal de transparencia de la Administración del Estado: https://administracion.gob.es/pag_Home/espanaAdmon/Transparencia_DatosAbiertos/datos_abiertos.html
- Portal de datos abiertos de licitaciones del Ministerio de Hacienda: http://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/licitaciones_plataforma_contratacion.aspx

3.2.2.2 Portales de datos abiertos en Europa

- EU Open Data Portal: http://data.europa.eu/euodp/en/home, https://data.europa.eu/euodp/data/dataset?
- European Data Portal: https://www.europeandataportal.eu/
- Tenders Electronic Daily (TED): https://ted.europa.eu/TED/main/HomePage.do

- Europeana Linked Data: https://old.datahub.io/dataset/europeana-lod
- Francia: http://www.opendatafrance.net/
- Reino Unido: https://data.gov.uk/
- Italia: https://www.dati.gov.it/

3.2.2.3 Portales de datos abiertos en países latinoamericanos

- Argentina: http://datos.gob.ar/
- México: https://datos.gob.mx/

3.2.2.4 Portales de datos abiertos de CCAA y administraciones locales

- Junta de Andalucía: http://www.juntadeandalucia.es/datosabiertos/portal.html
- Junta de Galicia: http://abertos.xunta.gal/portada
- Castilla La Mancha: http://datosabiertos.castillalamancha.es/ , http://transparencia.castillalamancha.es/transparencia/otras-comunidades-autonomas
- País Vasco: http://opendata.euskadi.eus/inicio/
- Navarra: http://www.gobiernoabierto.navarra.es/es/open-data
- Madrid: https://datos.madrid.es/portal/site/egob/
- Ayuntamiento de Barcelona: http://opendata-ajuntament.barcelona.cat/es
- Área Metropolitana de Barcelona: www.amb.cat/es/web/area-metropolitana/dades-obertes
- Aragón: https://opendata.aragon.es/

3.2.2.5 Portales de datos abiertos de otras instituciones estatales

Sanidad

- Ministerio de Sanidad, Consumo y Bienestar Social: https://sede.mscbs.gob.es/datosabiertos/home.htm
- Agencia Española de Medicamentos y productos sanitarios: http://www.aemps.gob.es/datos_abiertos.html
- Observatorio de Salud en Cataluña: http://observatorisalut.gencat.cat/es/demanar_dades/dades_obertes/dades_obertes_salut/

Justicia

- Ministerio de Justicia:

https://sede.mjusticia.gob.es/cs/Satellite/Sede/es/servicios/reutilizacion-informacion/datos-abiertos-ministerio

- Boletín Oficial del Estado (BOE): https://www.boe.es/datosabiertos/

Inteligencia competitiva e innovación

- Ministerio de Industria, Comercio y Turismo: https://sede.minetur.gob.es/es-es/datosabiertos/catalogo-datos/Paginas/catalogo.aspx
- Oficina Española de Patentes y Marcas: https://sede.oepm.gob.es/eSede/datos/es/index.html
- Boletín Oficial del Registro Mercantil (BORM): https://www.boe.es/datosabiertos/

Cultura

- Biblioteca Nacional de España: http://datos.bne.es/inicio.html

3.2.2.6 Otros portales con datos

- Consejerías de Sanidad CC.AA.: https://www.xunta.gal/sanidade, http://www.euskadi.eus/gobierno-vasco/departamento-salud/inicio/, http://www.castillalamancha.es/gobierno/sanidad, etc.
- Justicia: http://www.poderjudicial.es/search/indexAN.jsp, https://www.tribunalconstitucional.es/es/Paginas/default.aspx
- Inteligencia competitiva e innovación: http://web.gencat.cat/es/temes/empresa/, http://www.euskadi.eus/gobierno-vasco/industria/inicio/, http://www.juntadeandalucia.es/organismos/economiayconocimiento.html
- Ministerio de Educación y Formación Profesional: http://www.educacionyfp.gob.es/portada.html
- Ministerio de Cultura y Deporte: http://www.culturaydeporte.gob.es/cultura-mecd/
- Museo del Prado: www.museodelprado.es

3.2.3 Censo de sitios y tipos de RL

Como es natural, un determinado sitio web puede contener no solo conjuntos de documentos susceptibles de ser convertidos en RL (aquellos que son objeto de nuestro estudio), sino diferentes tipos de recursos lingüísticos ya creados. Por tanto, lo primero que se realizó para cada uno de los

sitios en los que se pretendían encontrar datos fue una búsqueda de aquellos RL ya existentes, y que fueran fácilmente accesibles, siguiendo el modelo mostrado en la Tabla 1.

Sitio	Corpus textuales	Corpus multimodales	Memorias de traducción	Entidades nombradas	Recursos léxicos
A					
B					
C					

Tabla 1: Tipología de recursos lingüísticos

Para la clasificación previa de tipos de RL se utilizaron aquellos tipos más usuales dentro de las tecnologías del lenguaje, que suponen, a grandes rasgos, las herramientas o productos más importantes, relacionados con cinco áreas complejas como el habla, el léxico, los textos, y la terminología[5]. Así, esta clasificación, dependiendo de con qué áreas se relacione, se podría subdividir en los siguientes RL: **corpus textuales** (colecciones de textos), **corpus multimodales** (corpus que aúnan texto e imagen o sonido, o los tres a la vez), **memorias de traducción** (base de datos lingüística con segmentos de texto en parejas de lenguas, la de origen y la de destino -traducción-), **entidades nombradas** (listas con categorías predefinidas de personas, organizaciones, lugares, cantidades, etcétera, que se utilizan en tareas de recuperación de la información), **recursos léxicos** (categoría general que agruparía a otros recursos de índole léxico semántica, como glosarios, terminologías, tesauros y ontologías).

3.3 FICHA TÉCNICA PARA LA RECOGIDA DE INFORMACIÓN

Tras la primera localización de datos expuesta en la sección anterior, se pasó a una segunda fase en la que se fue completando el censo definitivo de recursos de datos abiertos o de conjuntos de documentos susceptibles de ser RL. Para dichos datos, se procedió a su análisis según lo identificado y caracterizado en una ficha técnica de recogida de información creada a partir de lo determinado en estudios previos relacionados con las TL y los datos abiertos en Europa.

[5] *ELRA/ELDA distingue 5 tipos de RL: speech, lexica, corpora, terminology y multimodal/multimedia.*

Así, para la configuración de los aspectos que era necesario evaluar de los recursos de datos abiertos censados, se consultaron fuentes de diferentes organizaciones e iniciativas internacionales:

- **European Language Resource Coordination** [4]: Iniciativa de la Comisión Europea para el desarrollo y la distribución de recursos en todas las lenguas oficiales europeas. En particular, se consultó el buscador de recursos que mantiene.[6]
- **European Language Resources Association (ELRA)** [5]: Constituida en 1995, es una asociación europea sin fin lucrativo para la distribución, archivado, evaluación y organización de eventos relacionados con los recursos para la tecnología lingüística. Desde 1998, organiza la conferencia internacional Language Resources and Evaluation Conference (LREC). La agencia de distribución comercial de sus recursos es ELDA (European Language Distribution Association) [6].
- **CLARIN (European Research Infrastructure for Language Resources and Technology**) [7]: Plataforma para el acceso de recursos y herramientas de tecnología lingüística, creada en 2012. Se trata de una infraestructura para la interoperabilidad de herramientas y datos compartidos (escritos, hablados y multimodales), principalmente de Ciencias Sociales y Humanidades. Como dato importante, CLARIN no define un esquema de metadatos de recursos, sino que integra los metadatos existentes en los recursos originales.
- **Dublin Core** [8]: Estándar de metadatos para describir recursos digitales con 15 elementos en su versión simple, y 18 (más los cualificadores de elementos) en su versión ampliada. De él se extrae el **Open Languages Archive Community (OLAC)**, una extensión lingüística de Dublin Core.
- **LDC (Linguistic Data Consortium, University of Pennsylvania)** [9]: Consorcio de universidades, e instituciones norteamericanas que distribuye y fomenta el desarrollo de recursos de tecnología lingüística desde 1992. Para describir los recursos, LCD propone, por ejemplo, una serie de metadatos menos atomizados que los de las iniciativas anteriormente enunciadas.
- **META-NET**: Red de excelencia formada por 60 centros de 34 países. [10] Sería el equivalente a CLARIN, pero se enfoca, sobre todo, al desarrollo industrial, reuniendo a investigadores, proveedores comerciales, usuarios privados y corporativos, e inversores. Mantiene el portal de recursos abiertos y distribuido META-SHARE, y ha redactado varios libros blancos con información relevante del sector.[7]

[6] *https://elrc-share.eu/repository/search/?q=*

[7] *www.meta-net.eu/whitepapers/overview*

Tras el estudio de las fuentes, que arrojarían luz sobre los aspectos que sería interesante valorar para establecer la selección de recursos, se identificaron **24 rasgos relevantes, agrupados en siete aspectos**. Naturalmente, no todos los recursos se pueden identificar para cada rasgo, pero supone una aproximación muy completa para la descripción del recurso, necesaria para la heterogeneidad de recursos encontrados o censados. De ellos, el punto 3.3.6 (Grado de madurez de datos conforme al modelo de la metodología) plantea **3 grados de madurez (alta, media y baja)**, que se definen considerando criterios técnicos (ej. necesidades de procesamiento) y normativos (ej. necesidad de anonimización de datos). Para cada recurso censado, se asigna un valor en la escala de madurez según el recuento de puntos identificados conforme a la Tabla 2 (incluida en la sección 3.3.8). Se trata, pues, de un modelo de madurez distinto al de otros informes (*vid*. Aguado y otros, 2016), porque fue adaptado a los atributos de los datos estudiados [11]. A continuación, se detalla el modelo de ficha técnica utilizada para el censo (los ejemplos entre paréntesis son simplemente ilustrativos, aunque se refieren, en este caso, a RL reales).

3.3.1 Identificación del recurso.

- *Nombre:* (ej. C-ORAL-ROM).
- *Clasificación por tipo de recurso: (ej. corpus textual, memoria de traducción...).*
- *Clasificación por número de lenguas:* (ej. monolingüe/bilingüe).
- *Lenguas:* (ej. español, catalán, euskera... así como las posibles variantes del español)
- *Descripción del recurso:* (ej. Forma parte de un corpus multilingüe de lengua espontánea en las cuatro lenguas romance principales: francés, italiano, portugués y español. El proyecto fue financiado por la UE bajo el V Framework Programme (IST-2000-26228)).
- *Fecha de comienzo de creación (del recurso)*: (ej. desde 1998).
- *Fecha de finalización de creación (del recurso):* (ej. 2016).
- *Frecuencia de actualización*: (ej. anual).
- *Fecha de última actualización*: (ej. mayo de 2018).
- *Versión:* (ej. final (2004)).
- *Identificador del recurso (ISLRN, ISSN, ISBN, DOI u otro) y/o URL:* (ej. doi.org/10.1075/scl.15).
- *Tipo de licencia:* (ej. comercial, ELDA; libre con restricciones, Creative Commons, si aplica, DPI o IPR).
- *Descarga masiva disponible:* SI/NO.

3.3.2 Persona de contacto u organización responsable

- *Nombre y correo electrónico:*
- *Nombre organización (abreviatura, dpto., URL):*

3.3.3 Creación del recurso

- *Proveedor y/o creador:* (ej. LLI-UAM).
- *Proyecto(s) financiador(es):* (ej. C-ORAL-ROM).

3.3.4 Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro).*
- *Niveles de anotación lingüística:* (ej. POS, lematización, prosodia).
- *Conforme a los estándares (EAGLES, PAROLE, CONLL, TMX, etc.).*
- *Tamaño:* (ej. 300.000 palabras, 210 textos).
- *Unidad (términos, entradas, textos, oraciones, otro):* (ej. palabras, enunciados).
- *Formato (CSV, HTM, etc.):* (ej. texto, UTF-8).
- *Dominio (economía, legislación, etc.):* (ej. Sanidad).
- *Género (crónica, publicidad, oficial, etc.):* (ej. formal, informal, medios de comunicación).
- *Tipo de texto: (académico, blog, etc.):* (ej. habla espontánea).
- *Tipo de documento: (artículo, manual, etc.):* (ej. monólogos y conversaciones).
- *Información adicional* (URL con información relacionada, etc.).

3.3.5 Otros recursos relacionados

- *Identificación y URL de recursos relacionados.*

3.3.6 Grado de madurez de datos conforme al modelo de la metodología

- *Necesidades de procesamiento (manual o automático):* bajas/medias/altas (ej. conversión de formatos, alineamientos, anotación, transcripción, verificación, etc.).

3.3.7 Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	Valores: -, *, **, N/A (Desconocido o No aplica)	Observaciones
Aspectos técnicos (necesidad de procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…).		
2. Transcripción (ortográfica, fonológica, suprasegmental…).		
3. Alineación vídeo/sonido y texto		
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1).		
5. Anotación morfológica y/o sintáctica.		
6. Anotación de entidades nombradas.		
7. Otros tipos de anotación (semántica, pragmática, palabras clave…).		
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…).		
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran un revisor experto).		
10. Anotación conforme a estándares de la comunidad PLN.		
11. Presencia de metadatos.		
Aspectos legales		
12. Necesidad de anonimización de datos personales.		
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…).		

Tabla 2: Plantilla para la evaluación de la madurez como RL de un recurso

La escala de madurez se define según la proporción entre los aspectos presentes dentro de los aplicables. Ciertos aspectos no se evalúan si no conciernen a la tarea o al tipo de datos (ej. “N/A” para

la necesidad de alineación de sonido/vídeo y texto en datos puramente textuales). La evaluación se realiza otorgando estrellas (-, * o **) a cada punto de los analizados.

La columna de observaciones se emplea, además, para explicar cada puntuación y argumentar lo expuesto con comentarios cualitativos que justifiquen la razón de la valoración final.

3.3.8 Posibles aplicaciones del futuro recurso lingüístico

- *Ejemplo de aplicaciones posibles:* (ej. entrenamiento y evaluación de sistemas de reconocimiento de habla; modelo para el desarrollo de sistemas conversacionales).

4 CENSADO DE DOCUMENTOS

4.1 INTRODUCCIÓN

Para la identificación de los datos, se ha empleado la ficha de registro comentada en la sección precedente. En un primer momento, se analizaron en detalle un gran número de portales relacionados con las administraciones públicas de carácter local, nacional y regional, así como otras entidades públicas que pudieran ser de interés por su campo de estudio (Sanidad, Justicia, Cultura e Inteligencia competitiva). También se consultaron organismos extranjeros cuyo contenido se ajustaba a las áreas de interés definidas. De este análisis previo, se obtuvo una primera lista de 101 recursos y conjuntos de datos potenciales de ser convertidos en recursos lingüísticos. A modo de resumen, esta primera lista se dividía en:

- **Sanidad: 41** (textos, documentos multimedia, entidades nombradas, memorias bilingües, o terminología, de ámbito español o latinoamericano).
- **Justicia: 5** (textos y documentos multimedia de ámbito español).
- **Inteligencia competitiva: 7** (entidades nombradas, terminología y corpus paralelos de ámbito español o latinoamericano).
- **Cultura, Turismo y otros: 35** (textos, documentos multimedia, entidades nombradas y corpus paralelos de ámbito español).
- **Recursos excelentes sin acceso: 13** (nacionales, internacionales, textos, documentos multimedia o entidades nombradas, de ámbito español o latinoamericano).

Para analizar de forma más completa los datos, y con el objeto de obtener un plan de acción centrado en varios tipos de los mismos, se optó por describir aquellos recursos que fuesen más interesantes para su empleo en TL, y que pudieran, además, dar lugar a un mayor número de RL. Así, se eligieron 24 de ellos, de tipología variada:

- Recurso 1: Patentes, modelos de utilidad e informes técnicos digitalizados de la Oficina Española de Patentes y Marcas (OEPM).
- Recurso 2: Patentes multilingües digitalizadas en PATSTAT de European Patent Office (EPO).
- Recurso 3: Diccionarios terminológicos del Centro de Terminología (TERMCAT).
- Recurso 4: Padrón: Relación de municipios del Instituto Nacional de Estadística.

- Recurso 5: Topónimos del Instituto Geográfico Nacional (IGN).
- Recurso 6: Grabaciones de vídeo de RTVE a la carta.
- Recurso 7: Grabaciones de audio y vídeo del Archivo Audiovisual del Congreso de los Diputados de España.
- Recurso 8: Índices de clasificación de los catálogos de la BNE.
- Recurso 9: Publicaciones periódicas digitalizadas de la Hemeroteca Digital.
- Recurso 10: Documentos digitalizados de la Biblioteca Digital Hispánica.
- Recurso 11: Publicaciones en repositorio SciELO (Scientific Electronic Library Online).
- Recurso 12: Publicaciones y vídeos del Instituto de Salud Carlos III (ISCIII).
- Recurso 13: Banco de datos de enfermedades raras y medicamentos huérfanos de OrphaData.
- Recurso 14: Guías de práctica clínica (GPC) del portal Guía Salud.
- Recurso 15: Vídeos del portal web de TV del Gobierno Vasco relacionados con el tema de Salud.
- Recurso 16: Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS).
- Recurso 17: Nomenclátor de prescripción del Centro de Información de Medicamentos (CIMA).
- Recurso 18: Textos de Jurisprudencia del CENDOJ.
- Recurso 19: Textos del Boletín Oficial del Estado (BOE) Diario.
- Recurso 20: Textos de códigos electrónicos del Boletín Oficial del Estado (BOE).
- Recurso 21: Textos sobre Legislación del Boletín Oficial del Estado (BOE).
- Recurso 22: Memorias de traducción que contienen las publicaciones en el Boletín Oficial del Estado realizadas en euskera del Instituto Vasco de Administración Pública (IVAP).
- Recurso 23: Memorias públicas de traducción de la Diputación Foral de Gipuzkoa.
- Recurso 24: Grabaciones de Vistas Judiciales del Consejo General del Poder Judicial.

4.2 LISTADO DE DOCUMENTOS CENSADOS

Ante la heterogeneidad de los documentos censados y la imposibilidad de reseñarlos todos en detalle, se optó finalmente por describir los 24 recursos elegidos por su interés, englobándolos en categorías relacionadas con su dominio y con las áreas temáticas definidas como prioritarias para este estudio (Inteligencia competitiva, Sanidad, Justicia, Cultura y Otros). En este informe solo se incluyen, pues, las conclusiones generales para cada recurso elegido.

4.2.1 Inteligencia competitiva

4.2.1.1 Recurso 1: Patentes, modelos de utilidad e informes técnicos digitalizados de la Oficina Española de Patentes y Marcas (OEPM)

Identificación del recurso.

- *Nombre:* Patentes, modelos de utilidad e informes técnicos digitalizados de la Oficina Española de Patentes y Marcas (OEPM)
- *Clasificación por tipo de documento:* corpus textual.
- *Clasificación por número de lenguas:* monolingüe.
- *Lenguas:* español.
- *Descripción del recurso:* catálogo de patentes, modelos de utilidad y expedientes digitalizados.
- *Fecha de comienzo de creación*: variable (véase apdo. Tamaño).
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*: la actualización de los textos completos suele ser diaria, pero varía de unos meses a otros. *Fecha de última actualización*: diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Conjunto completo de datos abiertos de la OEPM: https://sede.oepm.gob.es/eSede/datos/es/catalogo/
- *Tipo de licencia:* se permiten la reutilización para fines comerciales y no comerciales[8].
- *Descarga masiva disponible:* Sí. Se dispone de descarga masiva para Catálogos de patentes y modelos de utilidad (información bibliográfica, Clasificación Cooperativa de Patentes, citaciones y European Classification System, folletos y documentación legal), Boletín Oficial de la Propiedad Industrial (BOPI) y Clasificación Internacional de Patentes (CIP).

Persona de contacto u organización responsable:

- *Nombre y correo electrónico:* Portal de Open Data: opendata@oepm.es
  Base de Datos Expedientes Digitalizados: archivoenlinea@oepm.es
- *Nombre organización:* Oficina Española de Patentes y Marcas (OEPM). www.oepm.es

Creación del recurso

[8] *www.oepm.es/es/avisoLegal.html*

- *Proveedor y/o creador:* Oficina Española de Patentes y Marcas (OEPM).

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal; los folletos para usuarios presentan un estilo divulgativo.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:*

  Catálogos de patentes y modelos de utilidad: la información bibliográfica y la documentación legal está disponible para la descarga desde 2010 hasta la actualidad; la Clasificación Cooperativa de Patentes (CPC), desde el año 2013; las citaciones y ECLA (European Classification System), desde 1994; los folletos (PDF), desde el año 1987; y los textos completos, están disponibles desde 2004 hasta la actualidad.
- *Unidad (términos, entradas, textos, oraciones, otro):* textos.
- *Formato*: Los documentos se presentan mayoritariamente en formato PDF y XML. (Los textos completos de CPC).
- *Dominio:* propiedad intelectual, legislación industrial, inteligencia competitiva, innovación tecnológica, información divulgativa industrial para usuarios.
- *Género:* patentes, folletos informativos, documentación técnica.
- *Tipo de texto: textos científico-técnicos y administrativos.*
- *Otros recursos relacionados:*
  - Catálogos de patentes y modelos de utilidad: información bibliográfica, Clasificación Cooperativa de Patentes (CPC), citaciones y European Classification System (ECLA) en formato SGML, folletos (PDF) y documentación legal. Los archivos (XML y PDF) comprimidos (ZIP) se encuentran en: https://sede.oepm.gob.es/eSede/datos/es/catalogo/datos.html?catalogo=invenciones.
    Los textos completos de CPC (XML) ofrecen mayores posibilidades de utilidad para crear corpus y extraer terminología
  - Boletín Oficial de la Propiedad Industrial (BOPI): https://sede.oepm.gob.es/bopiweb/descargaPublicaciones/formBusqueda.action, de actualización diaria, recoge publicaciones en 3 tomos: Tomo 1, Marcas y otros signos distintivos; Tomo 2, Invenciones; y Tomo 3, Diseños industriales. Documentos disponibles desde 1987 hasta la fecha actual. El BOPI está disponible en XML, PDF y HTML. Se pueden

descargar los tomos mediante plantillas XSD: https://sede.oepm.gob.es/eSede/datos/es/catalogo/catalogo.html?catalogo=otros

- Base de datos de diseños: http://consultas2.oepm.es/DisenosWeb/faces/busquedaInternet.jsp No es posible la descarga masiva de datos desde este buscador, pero sí desde la dirección: https://sede.oepm.gob.es/eSede/datos/es/catalogo/catalogo.html?catalogo=disenos
  Contiene 376.969 referencias bibliográficas y datos de modelos y diseños industriales desde 1966. Incluye imágenes publicadas en el BOPI desde 1998. Actualización de periodicidad variable.

- Base de datos de invenciones (INVENES): buscador disponible en: http://invenes.oepm.es/InvenesWeb/faces/contenidoBases.jsp (no es posible la descarga masiva de datos). Permite acceder a 1.245.082 archivos PDF (base de datos INTERPAT) y 332.063 archivos PDF (base de datos LATIPAT). INVENES permite consultar (pero no descargar) datos bibliográficos de Privilegios Reales desde 1826 hasta 1878 y de Patentes de la Restauración desde 1878 hasta 1929; documentos de Patentes y Modelos de Utilidad tramitados por el Estatuto de la Propiedad Industrial desde 1929, por la Ley de Patentes de 20 de marzo de 1986 y por la Ley de Patentes de 24 de julio de 2015. También incluye las Patentes Europeas y solicitadas que designan a España (textos en español). Actualización diaria.

- Clasificación Internacional de Patentes (CIP): taxonomía bilingüe (inglés y español) que contiene más de 76.000 entradas de conceptos organizados jerárquicamente en formato PDF y XML; disponible desde 2006 hasta el año actual. Es una clasificación para patentes, certificados de inventor, modelos de utilidad y certificados de utilidad. Como recurso lingüístico, se trata de un corpus paralelo español e inglés. Disponible en: http://cip.oepm.es/descargas. Y más detalladamente en WIPO (XML y DTD): https://www.wipo.int/classifications/ipc/ipcpub

- Expedientes digitalizados: Buscador que permite consultar expedientes por código identificador de solicitud, publicación o fecha: http://archivoenlinea.oepm.es/register/regviewer (no es posible la descarga masiva de datos). Se pueden consultar las referencias de Patentes Nacionales (fecha de solicitud entre 1940 y 1972), Modelos de Utilidad (solicitud entre 1940 y 1968), Patentes Nacionales

y Modelos de Utilidad (desde 1986 hasta la fecha), Patentes Europeas validadas en España, Marcas Nacionales desde 1878 (actualización progresiva hasta la fecha actual).

- Folletos divulgativos para consumidores y usuarios: www.oepm.es/es/propiedad_industrial/publicaciones/folletos_informativos/index.html Aproximadamente 50 recursos (tamaño variable). Actualización de periodicidad variable. En formato PDF, y algunos también en formato HTML.

- Estudios, artículos, guías sectoriales, monografías y libros: aproximadamente 40 (tamaño y frecuencia de actualización variable): http://www.oepm.es/es/propiedad_industrial/publicaciones

- Normativa: www.oepm.es/es/propiedad_industrial/Normativa/

- Catálogo de marcas y nombres comerciales (signos distintivos): https://sede.oepm.gob.es/eSede/datos/es/catalogo/catalogo.html?catalogo=marcas http://www.oepm.es/es/signos_distintivos/index.html
  Con relación a recursos de interés de las marcas, cabe señalar la clasificación internacional de Viena, que incluye una clasificación descriptiva de los elementos figurativos de los logos de las marcas que puede tener interés para entrenar relaciones entre imágenes y texto. En el siguiente enlace, se puede consultar la BBDD de marcas en el que se recoge dicha información para las imágenes en el campo (531): https://www.wipo.int/branddb/es/

- Catálogo de estadísticas: https://sede.oepm.gob.es/eSede/datos/es/catalogo/catalogo.html?catalogo=estadisticas

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):*
  La Oficina Europea de Patentes y Marcas (OEPM) recoge documentos adecuados para crear un corpus de patentes y modelos de utilidad. Dichos textos permitirían extraer terminología o entrenar sistemas de OCR, dado que algunas patentes escaneadas (PDF) se ofrecen también en XML.

  La Clasificación Internacional de Patentes (CPC, CIP), además, contiene textos bilingües en inglés y español, lo que posibilita crear un corpus paralelo.

Por otra parte, una gran cantidad de patentes registradas en la OEPM son traducción de patentes internacionales (70-90% dependiendo del año). Dichas patentes son identificadas en el estandar ST36 como patentes tipo T3. Sería muy interesante, para el ámbito de la traducción automática, que se dispusiera de un "corpus" bilingüe de segmentos paralelos de texto en inglés y español, análogamente al corpus creado por WIPO llamado COPPA [9](Corpus de Solicitudes de Patentes Paralelas de Patentscope), que ofrece un "corpus" bilingüe de más de 8 millones de segmentos paralelos de texto en inglés y francés que abarcan más de 170 millones de palabras.

Los textos completos de la Clasificación Internacional de Patentes y el Boletín Oficial de la Propiedad Industrial (BOPI) se presentan en XML, formato que requiere menor necesidad de tratamiento. Los folletos divulgativos o técnicos, o publicaciones e informes se encuentran en PDF; la extracción de textos a partir de los mismos añade dificultades de conversión de formato y revisión manual posterior.

Por estos motivos, se valora como recurso con potencial y de **madurez media-alta** en su estado actual.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	**	Solo catálogo de patentes, la CIP, el BOPI y los textos de CPC
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	XML de la CIP y textos CPC
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	

[9] https://www.wipo.int/export/sites/www/patentscope/en/data/pdf/wipo-coppa-technicalDocumentation.pdf

7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	**	En los textos completos (XML), los tipos de etiquetas XML definen las categorías semánticas.
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas de revisión por experto)	*	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	**	Solo en los textos en XML hay metadatos como Inventor, Fecha, País, etc.
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media-alta

Tabla 3: Madurez del recurso 1: Oficina Española de Patentes y Marcas (OEPM).

Posibles aplicaciones del futuro recurso lingüístico

- Recogida de corpus de textos de dominio industrial.
- Extracción de terminología científico-técnica.
- La Clasificación Internacional de Patentes (CIP) permite construir un corpus paralelo inglés/español de terminología industrial de utilidad para traducción automática.

Recomendaciones: Construir un corpus paralelo inglés/español de terminología industrial, de modo análogo al corpus COPPA existente para inglés/francés.

4.2.1.2 Recurso 2: Patentes multilingües digitalizadas de European Patent Office (EPO)

Identificación del recurso.

- *Nombre: PATSTAT de European Patent Office (EPO).*
- *Clasificación por tipo de documento:* corpus textual.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* alemán, francés e inglés; en menor cantidad, español, portugués e italiano.
- *Descripción del recurso:* solicitudes de patentes de nivel europeo (en alemán, francés e inglés, y ocasionalmente en italiano); si la patente tiene un equivalente nacional, se ofrece el documento correspondiente en español. El portal de patentes europeas EPO, con ámbito de aplicación en más de 38 países, permite también la consulta de patentes relacionadas de ámbito extra-europeo. Aloja una base de datos a la que se accede con los buscadores Espacenet (ámbito mundial) y Latipat-Espacenet (destinado al ámbito iberoamericano).
- *Fecha de comienzo de creación*: los buscadores Espacenet y Latipat-Espacenet permiten acceder a datos de patentes concedidas desde 1836 hasta la actualidad (la fecha de inicio varía de un país a otro). Véanse las fechas por país en: www.epo.org/searching-for-patents/technical/full-text-additions.html.
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
    - *Frecuencia de actualización*: variable (depende de cada oficina nacional de patentes).
- *Fecha de última actualización*: diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Patstat: https://www.epo.org/searching-for-patents/business/patstat.html
    - Conjuntos de datos disponible para la descarga (requieren licencia):
        - Datos bibliográficos de patentes europeas (EP bibliographic data, EBD): www.epo.org/searching-for-patents/data/bulk-data-sets/ebd.html
        - Datos del registro europeo de patentes (European Patent Register): www.epo.org/searching-for-patents/data/bulk-data-sets/register-data.html
        - Textos completos de patentes (EP full-text data): www.epo.org/searching-for-patents/data/bulk-data-sets/data.html

- Base de datos bibliográfica mundial (EPO worldwide bibliographic database, DOCDB): www.epo.org/searching-for-patents/data/bulk-data-sets/docdb.html
- Base de datos bibliográfica mundial de estatus legal de patentes (EPO worldwide legal status database, INPADOC): www.epo.org/searching-for-patents/data/bulk-data-sets/inpadoc.html
- Imágenes de la primera página (formato TIFF): www.epo.org/searching-for-patents/data/bulk-data-sets/first-page-images.html
- Listados de secuencias de nucleótidos y aminoácidos de publicaciones de la EPO (formato TXT): www.epo.org/searching-for-patents/data/bulk-data-sets/sequence-listing.html
- Archivos de patentes por países (Francia, España, Suiza y Reino Unido): http://www.epo.org/searching-for-patents/data/bulk-data-sets/national-full-text-data.html
- Decisiones de los comités de solicitud de la EPO (Decisions of the EPO boards of appeal): http://www.epo.org/searching-for-patents/data/bulk-data-sets/boards-of-appeal-decisions.html

- *Tipo de licencia:* la información disponible en el portal está sometida a leyes de propiedad intelectual. Se permite la distribución, o traducción si se menciona la fuente de los datos[10]. El acceso a la base de datos para profesionales requiere una suscripción de pago.
- *Descarga masiva disponible:* El conjunto de datos disponible para la descarga masiva (archivos comprimidos en formato ZIP) requiere comprar una licencia: www.epo.org/searching-for-patents/data/bulk-data-sets.html. Sí pueden descargarse muestras de cada recurso en: https://publication.epo.org/raw-data/product?productId=94

Persona de contacto u organización responsable:

- *Nombre y correo electrónico:* European Patent Office (EPO). President of the European Patent Office, Mr António Campinos. Contacto: https://forms.epo.org/service-support/contact-us/contact0-form.html.
- *Nombre organización:* European Patent Office (EPO).

[10] *www.epo.org/footer/terms.html*

Creación del recurso

- *Proveedor y/o creador:* la Oficina Europea de Patentes (EPO) reúne los datos de las correspondientes oficinas nacionales de patentes, alberga la base de datos y mantiene el servicio de consulta Espacenet. El banco de datos de la EPO también aporta datos al buscador Latipat, dirigido al ámbito iberoamericano.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:* los buscadores Espacenet y Latipat-Espacenet permiten acceder gratuitamente a más de 80 millones de patentes concedidas por más de 90 autoridades. La cobertura de patentes por país está disponible en: www.epo.org/searching-for-patents/technical/full-text-additions.html
  Están disponibles las bases de datos de patentes completas desde 1978-1980.
- *Unidad (términos, entradas, textos, oraciones, otro):* textos e imágenes.
- *Formato*: PDF y XML (documentos completos; bases de datos bibliográficas solo en XML), TIFF (imágenes de la primera página), TXT (listas de secuencias de nucleótidos y aminoácidos) y CSV/XLS (resultados de búsquedas).
- *Dominio:* propiedad intelectual, legislación industrial, inteligencia competitiva, innovación tecnológica.
- *Género:* patentes documentación técnica.
- *Tipo de texto:* textos científico-técnicos y administrativos.

Otros recursos relacionados:

- Espacenet: https://worldwide.espacenet.com
- Latipat-Espacenet: https://es.espacenet.com/
- Buscador para expertos: servicio destinado a profesionales; la consulta requiere una suscripción de pago. https://data.epo.org/expert-services/index-2-4-3.html
- Patentes, modelos industriales y de utilidad digitalizados de ámbito iberoamericano (véase:
  INAPI (Chile): https://ion.inapi.cl/Patente/ConsultaAvanzadaPatentes.aspx,
  IMPI (México): http://siga.impi.gob.mx/newSIGA/content/common/principal.jsf,

https://eservicios.impi.gob.mx/seimpi/action/invencionesenlinea ,

SIC (Colombia): http://www.sic.gov.co/base-de-datos,

PROSUR: http://prosur.org/tramitacion/buscador-de-patentes/).

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* El portal European Patent Office recoge un gran volumen de documentos internacionales de patentes y modelos de utilidad, que permitirían obtener un corpus de dominio industrial y extraer terminología. Los documentos se encuentran en formato XML y PDF, que necesitaría su conversión a texto, previa a su revisión manual. Desde el buscador Espacenet no es posible la descarga masiva de documentos bilingües, sino la consulta manual a partir de código de solicitud de patente, código de publicación, título o palabras clave, entre otros datos. Aunque pueden descargarse muestras de cada recurso, el conjunto de datos disponible para la descarga masiva (archivos comprimidos en formato ZIP) requiere comprar una licencia: www.epo.org/searching-for-patents/data/bulk-data-sets.html. Por estas razones, se valora como recurso con potencial de aplicación a las tareas de PLN, pero de **baja madurez** en su estado actual.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Solo gratuito parte de los archivos XML.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	-	XML en UTF8, pero otros TXT en ISO-8859-1.
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	

8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	*	Solo en ciertos documentos.
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	*	
		Madurez baja

Tabla 4: Madurez del recurso 2: Patentes, modelos de utilidad e informes técnicos digitalizados de la Oficina Española de Patentes y Marcas (OEPM).

Posibles aplicaciones del futuro recurso lingüístico

- Recogida de corpus de textos de dominio industrial.
- Extracción de terminología científico-técnica.
- Construcción de un corpus paralelo multilingüe de terminología industrial.
- Obtención de metadatos y extracción de relaciones entre documentos de dominio industrial.
- Entrenamiento de clasificadores.
- Creación de recursos lingüísticos de variedades no peninsulares a partir de las patentes iberoamericanas existentes (véase Otros recursos relacionados)

Recomendaciones: facilitar acceso a investigadores con cuotas más accesibles; ampliar la cobertura hispanoamericana de patentes; crear un corpus paralelo del español, con variantes iberoamericanas del español y los pares de lenguas traducidos al español, semejante al Corpus of Parallel Patent Applications (COPPA)[11]; entrenamiento de clasificadores de citas para ver relaciones entre documentos.

[11] https://pdfs.semanticscholar.org/68c1/c201e98abcb757547e26602f6187d77fae22.pdf#page=20

4.2.1.3 Recurso 3: Diccionarios terminológicos del Centro de Terminología (TERMCAT)[12]

Identificación del recurso.

- *Nombre:* Centro de Terminología (TERMCAT).
- *Clasificación por tipo de documento:* repositorio pluridisciplinar de diccionarios de términos, en catalán y español (algunos recursos también en inglés, francés o alemán).
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* catalán, español (algunos diccionarios también en inglés, francés o alemán).
- *Descripción del recurso:* 112 diccionarios terminológicos de cinco grupos temáticos: Ciencias humanas, Ciencias de la salud y la vida, Ciencias jurídicas y económicas, Deportes e Industria y tecnología.
- *Fecha de comienzo de creación*: variable, según cada diccionario.
- *Fecha de finalización:* N/A.
- *Frecuencia de actualización*: N/A.
- *Fecha de última actualización*: diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:* Diccionarios (descargables anteriormente, no a fecha de revisión de la última versión del estudio): www.termcat.cat/es/TerminologiaOberta/
- *Tipo de licencia:* Creative Commons 3.0 Reconocimiento No adaptada y Creative Commons Reconocimiento Sin obras derivadas 3.0 No adaptada.
- *Descarga masiva disponible:* Sí anteriormente; no a fecha de la última versión del estudio,.

Persona de contacto u organización responsable:

- *Contacto:* TERMCAT. www.termcat.cat/es/Contacte/
- *Nombre organización:* Centro de Terminología (TERMCAT).

Creación del recurso

- *Proveedor y/o creador:* Centro de Terminología (TERMCAT).

Descripción del recurso

[12] Este recurso no se englobaría en Inteligencia competitiva de acuerdo con la definición del Plan de TL, pero se ajusta al criterio de los autores del informe, explicado en el apartado 2, “Conceptos”.

- *Variedad de la lengua (estándar, dialecto, argot, otro):* dependiente de los contenidos de cada diccionario; la mayoría, en variedad estándar y formal.
- *Niveles de anotación lingüística:* palabra: forma, lema, traducción, categoría e información morfológica (género y número), tipo de palabra (como abreviatura o sigla) o área temática. La información morfológica no está disponible para todas las lenguas.
- *Conforme a los estándares*: sí respecto a los estándares de tipo de contenido lingüístico, pero no respecto a estándares como TEI o CoNLL.
- *Tamaño y cobertura:* El tamaño y la cobertura temporal o temática de cada diccionario es variable.
- *Unidad (términos, entradas, textos, oraciones, otro):* entradas (formas de palabra, lemas, información categorial y morfológica, y traducciones).
- *Formato*: HTML (no todos los diccionarios están en todos los formatos); algunos también en XML y PDF (actualmente se requiere registro de usuario, y algunos contenidos están vacíos).
- *Dominio:* los contenidos de los diccionarios se agrupan en 5 áreas temáticas: Ciencias humanas, Ciencias de la salud y la vida, Ciencias jurídicas y económicas, Deportes e Industria y tecnología.
- *Género:* diccionarios terminológicos.
- *Tipo de texto:* terminología especializada.

Otros recursos relacionados:

- Consulta en línea: www.termcat.cat/ca/Diccionaris_En_Linia/.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático)*: TERMCAT recoge valiosos diccionarios multidisciplinares anotados en formato XML. Su procesamiento permite extraer equivalentes de traducción en diferentes lenguas, y obtener información categorial y morfológica acerca de lemas no recogidos en diccionarios electrónicos de dominio general. El acceso y descarga de los mismos era inmediata, pero el estado de algunos ha cambiado a fecha de la última versión de este documento. Recomendamos que la institución facilite su descarga (como se realizaba anteriormente) para que mejore su grado de madurez, y que dichos recursos permitan enriquecer recursos léxicos de PLN.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	
2. Transcripción (ortográfica, fonológica, suprasegmental…)	-	No incluyen la transcripción fonológica
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	Recursos en UTF-8
5. Anotación morfológica y/o sintáctica	**	
6. Anotación de entidades nombradas	0	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	**	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	**	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media

Tabla 5: Madurez del recurso 3: Diccionarios terminológicos del Centro de Terminología (TERMCAT).

Posibles aplicaciones del futuro recurso lingüístico

- Extracción de terminología multilingüe de dominio especializado.

- Extracción de información categorial y morfológica para enriquecer diccionarios y recursos léxicos de dominio especializado.
- Modelo para desarrollar un recurso semejante en español ("TERMESP"). Existe la iniciativa TERMINESP, [13] aunque su grado de desarrollo todavía no es comparable.

Recomendaciones: Crear un análogo de TERMCAT para el español ("TERMESP").

4.2.1.4 Recurso 4: Padrón: Relación de nombres de personas del Instituto Nacional de Estadística[14].

Identificación del recurso.

- *Nombre:* Padrón. Población por municipios: Apellidos y nombres más frecuentes.
- *Clasificación por tipo de documento:* Lista de entidades nombradas.
- *Clasificación por número de lenguas:* monolingüe.
- *Lenguas:* español.
- *Descripción del recurso:* Relación de nombres y apellidos más frecuentes de los residentes en España según su provincia de nacimiento y de residencia, así su década de nacimiento, sexo y nacionalidad; nombres más frecuentes de recién nacidos por comunidad autónoma y sexo.
- *Fecha de comienzo de creación*: 01/01/2002.
- *Fecha de finalización:* 01/01/2018.
- *Frecuencia de actualización*: Anual.
- *Fecha de última actualización*: 01/01/2018
- *Versión:* N/A.
- *Identificador del* recurso:
  - En el portal del INE: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177009&menu=resultados&idp=1254734710990

[13] *http://www.wikilengua.org/index.php/Wikilengua:Terminesp*

[14] Este recurso no se englobaría en Inteligencia competitiva de acuerdo con la definición del Plan de TL, pero se ajusta al criterio de los autores del informe, explicado en el apartado 2, "Conceptos".

- *Tipo de licencia:* Libre, con restricciones: Debe citarse la fuente de la información objeto de reutilización y mencionarse la fecha de la última actualización de la información objeto de reutilización, siempre y cuando estuviera incluida en el original.[15]
- *Descarga masiva disponible:* No, existe un fichero xls para cada una de las listas de nombres y apellidos.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*: http://www.ine.es/infoine/
- *Nombre organización:* Instituto Nacional de Estadística.

Creación del recurso

- *Proveedor y/o creador:* Instituto Nacional de Estadística.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* no anotado.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:* Ficheros individuales conteniendo distintas listas con nombres y apellidos según frecuencias de aparición, y comparadas con el total nacional y la provincia de residencia.
- *Unidad (términos, entradas, textos, oraciones, otro):* Términos, entidades nombradas.
- *Formato*: XLS.
- *Dominio:* Geografía, demografía.
- *Género:* Antropónimos.
- *Tipo de texto:* N/A.

Otros recursos relacionados:

- Nomenclátor: Población del Padrón Continuo por unidad poblacional: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177010&menu=resultados&secc=1254736195532&idp=1254734710990

Grado de madurez de los datos conforme al modelo

[15] *Aviso legal disponible en : www.ine.es/ss/Satellite?L=0&c=Page&cid=1254735849170&p=1254735849170&pagename=Ayuda%2FINELayout#*

- *Necesidades de procesamiento (manual o automático):* Este recurso posee una **madurez media**, ya que su contenido está bastante estandarizado, y se entrega en un formato fácilmente adaptable a su uso para RL, como es el **XLS**. Su interés es alto, porque sería un buen punto de partida de diferentes RL relacionados con las entidades nombradas. Sería primordial una conversión a formatos estandarizados en PLN como JSON, TXT o XML, más fácilmente reutilizables e interoperables. Se necesitaría un procesamiento para aislar aquellas cifras que señalan las frecuencias de aparición, dejando exclusivamente los caracteres (es decir, únicamente los antropónimos). Sería importante que se reseñara fácilmente como dato abierto, puesto que, a priori, el recurso (diferentes listas en formato XLS) no se ofrece en la sección dedicada a los datos abiertos del propio portal del INE, aunque sí aparece el buscador en HTML.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Posibilidad de convertirlo a otros formatos más interoperables.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	N/A	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	N/A	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	

9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	*	
10. Anotación conforme a estándares de la comunidad PLN	N/A	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media.

Tabla 6: Madurez del recurso 4: Padrón: Relación de municipios del Instituto Nacional de Estadística.

Posibles aplicaciones del futuro recurso lingüístico

- Creación de recursos léxicos bilingües, en su relación con la antroponimia de otras comunidades autónomas con lengua cooficial.
- Uso en etiquetadores automáticos (p. ej., morfológicos o semánticos) o en sistemas de reconocimiento y anotación de entidades nombradas, así como para recursos de anonimización o estudios que pretendan trabajar con este tipo de proceso.

Recomendaciones: Sería primordial una conversión a formatos estandarizados en PLN como JSON, TXT o XML, más fácilmente reutilizables e interoperables. Los topónimos de nombres propios de personas son de gran interés para reconocedores de entidades nombradas (NER por sus siglas en inglés).

4.2.1.5 Recurso 5: Topónimos del Instituto Geográfico Nacional (IGN)[16].

Identificación del recurso.

- *Nombre: Nomenclátor Geográfico de Municipios y Entidades de Población.*
- *Clasificación por tipo de documento:* Lista de entidades nombradas.
- *Clasificación por número de lenguas:* monolingüe.

[16] Este recurso no se englobaría en Inteligencia competitiva de acuerdo con la definición del Plan de TL, pero se ajusta al criterio de los autores del informe, explicado en el apartado 2, "Conceptos".

- *Lenguas:* español, con indicaciones de equivalentes en catalán, gallego y euskera.
- *Descripción del recurso:* Nomenclátor Geográfico de Municipios y Entidades de Población, que contiene las denominaciones, coordenadas, altitud y población, entre otros atributos, correspondientes a los municipios y entidades de población españolas.
- *Fecha de comienzo de creación*: 01/01/2010.
- *Fecha de finalización:* Mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*: Anual.
- *Fecha de última actualización*: 01/01/2018 (existen modificaciones posteriores, en versión HTML con fecha de 23/06/18).
- *Versión:* N/A.
- *Identificador del* recurso:
    - En datos.gob.es: https://datos.gob.es/es/catalogo/e00125901-spaignnomenclatorgeograficomunicipiosentpob201503240000
    - Acceso directo no disponible como recurso único, sólo desde portal de catálogo de datos: https://centrodedescargas.cnig.es/CentroDescargas/catalogo.do?Serie=NGMEN
- *Tipo de licencia:* Creative Commons BY 4.0.
    - *Descarga masiva disponible:* Sí, fichero comprimido que contiene varias bases de datos en diferente formato.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*: Instituto Goegráfico Nacional, ign@fomento.es
- *Nombre organización:* Centro Nacional de Información Geográfica (MINISTERIO DE FOMENTO).

Creación del recurso

- *Proveedor y/o creador:* Centro Nacional de Información Geográfica (Ministerio de Fomento).

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* no anotado.
- *Conforme a los estándares*: no.

- *Tamaño y cobertura:* Dos ficheros con base de datos odb y .mdb. Relación de todos los municipios.
- *Unidad (términos, entradas, textos, oraciones, otro):* Términos, entidades nombradas.
- *Formato*: ZIP (ODB, MDB).
- *Dominio:* Geografía, ordenación del territorio.
- *Género:* Toponimia.
- *Tipo de texto:* N/A.

Otros recursos relacionados:

- Nomenclátor Geográfico Básico de España: https://datos.gob.es/es/catalogo/e00125901-spaignngbe
- Relación de municipios y sus códigos por provincias. Instituto Nacional de Estadística: https://www.ine.es/daco/daco42/codmun/codmun19/19codmun.xlsx

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* Este recurso posee una **madurez media**, ya que su contenido está bastante estandarizado, y se entrega en un formato habitual para bases de datos. Su interés es alto, porque sería un buen punto de partida de diferentes RL relacionados con las entidades nombradas, así como para la realización de recursos léxicos bilingües (dada la existencia de toponimias en otras lenguas como el gallego, catalán o euskera). Sería primordial una conversión a formatos más estandarizados en PLN como JSON, TXT o XML, más fácilmente reutilizables e interoperables. Se necesitaría un procesamiento para aislar aquellas cifras que muestran las coordenadas de longitud y latitud, y priorizar los topónimos y otras denominaciones que pudieran ser de interés.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		

1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Posibilidad de convertirlo a otros formatos más interoperables.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	N/A	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	N/A	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	*	
10. Anotación conforme a estándares de la comunidad PLN	N/A	
11. Presencia de metadatos	*	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media.

Tabla 7: Madurez del recurso 5: Topónimos del Instituto Geográfico Nacional (IGN)

Posibles aplicaciones del futuro recurso lingüístico

- Creación de recursos léxicos bilingües, en su relación con la toponimia de otras comunidades autónomas con lengua cooficial.
- Uso en etiquetadores automáticos (p. ej., morfológicos o semánticos) o en sistemas de reconocimiento y anotación de entidades nombradas.

Recomendaciones: Sería primordial una conversión a formatos más estandarizados en PLN como JSON, TXT o XML, más fácilmente reutilizables e interoperables.

4.2.1.6 Recurso 6: Grabaciones de vídeo de RTVE a la carta[17].

Identificación del recurso.

- *Nombre:* Grabaciones de vídeo de RTVE a la carta.
- *Clasificación por tipo de documento:* Documentos multimedia producidos o emitidos por RTVE incluyendo informativos, series, programas de entretenimiento, etc.
- *Clasificación por número de lenguas:* Multilingüe.
- *Lenguas:* Principalmente castellano, aunque también catalán (RTVE Catalunya).
- *Descripción del recurso:* Programas producidos por RTVE incluyendo informativos, series, programas de entretenimiento, etc.
- *Fecha de comienzo de creación*: La mayor parte son contenidos de los últimos 10 años.
- *Fecha de finalización:* Mantenimiento y actualización continua de los datos.
    - *Frecuencia de actualización*: varias veces al día: se añaden cada día los programas que se han emitido.
- *Fecha de última actualización*: Diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - RTVE a la carta: http://www.rtve.es/alacarta/
- *Tipo de licencia*: RTVE no comercializa sus contenidos a particulares, pero su consulta sí está disponible para trabajos de investigación y docencia.[18] Televisión Española también facilita ciertos contenidos a instituciones con fines docentes. Estas peticiones se gestionan en atencion.instituciones@rtve.es. Por otro lado, los usuarios de RTVE.es, tienen a su disposición en TVE a la Carta los últimos contenidos emitidos de producción propia.
- *Descarga masiva disponible:* NO.

Persona de contacto u organización responsable:

[17] Este recurso no se englobaría en Inteligencia competitiva de acuerdo con la definición del Plan de TL, pero se ajusta al criterio de los autores del informe, explicado en el apartado 2, "Conceptos".

[18] *http://www.rtve.es/comunes/aviso_legal.html*

- *Nombre y correo electrónico:* Corporación de RTVE. Contacto: atencion.instituciones@rtve.es o formulario web: http://www.rtve.es/participacion/consultas/
- *Nombre organización:* Radio Televisión Española (RTVE).

Creación del recurso

- *Proveedor y/o creador:* Radio Televisión Española (RTVE).

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* De todo tipo, dependiendo del programa, pero con más énfasis en lengua estándar y formal.
- *Niveles de anotación lingüística:* Datos no anotados, salvo algunos programas (por ejemplo, los telediarios) que incluyen subtítulos y, algunos, traducción a lengua de signos. Los subtítulos aparecen en la web de RTVE como "transcripción completa", pero no es una transcripción exacta, y aunque están alineados temporalmente con el audio, tampoco es una alineación fiable. Si RTVE dispusiera de transcripciones exactas sería un recurso de gran utilidad para las TL.
- *Conforme a los estándares*: no.
    - Tamaño y cobertura: RTVE A la Carta incluye cerca de 100.000 horas de televisión en alta calidad, con los vídeos de las series, documentales, informativos (incluye cerca de 10.000 telediarios) y otros programas.
- *Unidad (términos, entradas, textos, oraciones, otro):* Vídeos.
- *Formato*: N/A (solo se permite la reproducción en el reproductor web propio).
- *Dominio:* Dominios de todo tipo, dependiendo del programa y del apartado del programa.
- *Género:* Series, documentales, informativos, etc.
- *Tipo de texto:* De todo tipo, dependiendo del tipo de programa.

Otros recursos relacionados:

- Archivo multimedia de RTVE: http://www.rtve.es/television/archivo/. Incluye algunas series y programas antiguos que se van digitalizando. Contiene programas desde la década de 1960.
- Archivo sonoro de RNE: http://www.rtve.es/alacarta/audios/archivo-sonoro/. Incluye programas de RNE, algunos con transcripciones.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* El acceso a los documentos está pensado para la visualización de vídeos individuales. No es posible descargar los contenidos, solo reproducirlos en un reproductor web propio. RTVE no comercializa sus contenidos a particulares, pero sí están disponibles para su consulta para trabajos de investigación y docencia. Televisión Española también facilita ciertos contenidos a instituciones con fines docentes. Con esta salvedad, la licencia es muy restrictiva: tienen licencia propia para reproducir contenidos en otras websites. No se permite la distribución, y no parece que se puedan descargar (ni siquiera individualmente). La información de los subtítulos alineados temporalmente existe, pero no es posible descargarla. Es posible copiar los contenido de los subtítulos pero ha de hacerse frase a frase o, como mucho, en pequeños grupos de frases. Tal y como está el recurso actualmente resultaría imposible convertirlo en un recurso lingüístico. Sin embargo, la información que atesora es extremadamente rica y valiosa. Los datos no están anotados, salvo algunos programas (por ejemplo, los telediarios) que incluyen subtítulos y, algunos, traducción a lengua de signos. Los subtítulos aparecen en la web de RTVE como "transcripción completa", pero no es una transcripción exacta, y aunque están alineados temporalmente con el audio, tampoco es una alineación fiable. Si se permitiese el acceso al menos a algunas secciones, de forma que se pudiesen descargar los vídeos y los subtítulos, se podría convertir en un recurso lingüístico de gran valor. Para ello, sería necesario volver a transcribir y alinear el audio, pero dado el pobre alineamiento y la imprecisión de los subtítulos disponibles, estos apenas resultarían una ayuda en este proceso. De modo que prácticamente resultaría necesario hacer la transcripción y el alineamiento de frases desde cero. También resultaría muy interesante etiquetar los locutores que aparecen en los vídeos para sistemas de reconocimiento y segmentación de locutores. Por todo ello, la **madurez** del recurso es **baja**, pero con un potencial muy alto.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	N/A	

2. Transcripción (ortográfica, fonológica, suprasegmental…)	*	Algunos contenidos transcritos, no literalmente (subtítulos)
3. Alineación vídeo/sonido y texto	*	Vídeo y texto alineado sin excesiva precisión (subtítulos) en algunos contenidos.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	-	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	N/A	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	**	No requiere.
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	-	No se pueden usar más que para reproducción privada actualmente.
		Madurez baja[19]

Tabla 8: Madurez del recurso 6: Grabaciones de vídeo de RTVE a la Carta.

[19] Este conjunto de documentos incluye un número variado de recursos, cada uno de los cuales tiene un grado de madurez propio. Por tanto, se ofrece una valoración global del conjunto.

Posibles aplicaciones del futuro recurso lingüístico

- Entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-To-Text) para subtitulado automático.
- Entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-To-Text) para facilitar la búsqueda y recuperación de información en formato vídeo (por ejemplo, en el archivo de RTVE y RTVE A la Carta).
- Entrenamiento y evaluación de sistemas de reconocimiento de locutores y segmentación de locutores para facilitar la búsqueda y recuperación de información en formato vídeo (por ejemplo en el archivo de RTVE y RTVE A la Carta).
- Entrenamiento y evaluación de sistemas de traducción de voz a lengua de signos (empleando para ello los programas con traducción a lengua de signos).
- Análisis de la evolución del lenguaje castellano estándar a lo largo del tiempo (empleando para ello el archivo histórico de RTVE).

Recomendaciones: Que se permita el acceso al menos a algunas secciones, de forma que se pudiesen descargar los vídeos y los subtítulos, así como los archivos sonoros y sus transcripciones. Cabe diferenciar entre lo que es subtitulado, que no supone transcripción exacta, sino resumida para facilitar la lectura síncrona y lo que es transcripción exacta del audio. Sería de gran utilidad para las tecnologías del lenguaje disponer de los script con las transcripciones exactas. Estos vídeos y audios se podrían convertir en recursos lingüísticos de gran valor. El recurso sería útil incluso si se publicaran fragmentos desordenados. También resultaría muy interesante etiquetar los locutores que aparecen en los vídeos para sistemas de reconocimiento y segmentación de locutores. Asimismo, permitir la descarga de secciones de los territoriales sería de gran utilidad para el desarrollo de recursos lingüísticos en lenguas cooficiales.

4.2.1.7 Recurso 7: Grabaciones de audio y vídeo del Archivo Audiovisual del Congreso de los Diputados de España[20].

Identificación del recurso.

[20] Este recurso no se englobaría en Inteligencia competitiva de acuerdo con la definición del Plan de TL, pero se ajusta al criterio de los autores del informe, explicado en el apartado 2, "Conceptos".

- *Nombre:* Grabaciones de audio y vídeo del Archivo Audiovisual del Congreso de los Diputados de España.
- *Clasificación por tipo de documento:* Grabaciones de audio o vídeo de intervenciones, sesiones plenarias y comisiones del Congreso de los Diputados, muchos de ellos incluyendo transcripción, y algunos segmentados por intervenciones.
- *Clasificación por número de lenguas:* Mayoritariamente monolingüe.
- *Lenguas:* Principalmente castellano.
- *Descripción del recurso:* Grabaciones de audio de las intervenciones desde 1977 y hasta 1996. Transcritos de forma no completamente literal (diario de sesiones). Grabaciones de vídeo desde 2004 hasta la actualidad. Algunos contenidos no están transcritos. Sí están transcritas las sesiones parlamentarias. Los vídeos aparecen segmentados por intervenciones. Las transcripciones de las sesiones, al parecer, están disponibles traducidas a otros idiomas oficiales del Estado a través de un servicio denominado cortesi@ proporcionado por Presidencia del Gobierno[21].
- *Fecha de comienzo de creación*: 1977.
- *Fecha de finalización:* Se actualiza continuamente.
- *Frecuencia de actualización*: Diaria, siempre que haya sesiones o comisiones.
- *Fecha de última actualización*: Diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Consulta en línea (Archivo Audiovisual): www.congreso.es/portal/page/portal/Congreso/Congreso/CongresoTV/HistEmisionFecha
    - Consulta en línea (Archivo de Audio): http://www.congreso.es/portal/page/portal/Congreso/Congreso/Intervenciones/ArchivoAudio

- *Tipo de licencia:* Se permite explícitamente su reutilización sin condiciones.[22]
- *Descarga masiva disponible:* No.

Persona de contacto u organización responsable:

[21] *https://administracionelectronica.gob.es/ctt/cortesia*

[22] *http://www.congreso.es/portal/page/portal/Congreso/Congreso/Publicaciones/Aviso_Legal*

- *Contacto:* https://administracionelectronica.gob.es/ctt/cortesia
- *Nombre organización:* Congreso de los Diputados de España.

Creación del recurso

- *Proveedor y/o creador:* Congreso de los Diputados de España.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* Estándar formal.
- *Niveles de anotación lingüística:* Datos no anotados, pero los vídeos y audios tienen transcripciones.
- *Conforme a los estándares*: No.
- *Tamaño y cobertura:* El tamaño es muy elevado, cubre un rango temporal muy extenso, si bien no se ha podido realizar una estimación fiable del tamaño del archivo. Dado que el tipo de temáticas tratadas es muy diverso, la cobertura temática es también muy amplia.
- *Unidad (términos, entradas, textos, oraciones, otro):* audios y vídeos, muchas veces emparejados con sus transcripciones no completamente literales.
- *Formato*: MP3 y MP4.
- *Dominio:* Dado que el tipo de temáticas tratadas es muy variado, la cobertura temática es también muy amplia, si bien todas ellas tratadas en el contexto de la vida política del parlamento.
- *Género:* Sesiones y comisiones parlamentarias.
- *Tipo de texto:* Audios y vídeos de sesiones y comisiones parlamentarias.

Otros recursos relacionados:

- Archivo, biblioteca y documentación del Congreso de los Diputados (accesibles desde http://www.congreso.es/portal/page/portal/Congreso/Congreso/SDocum).

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* Los audios y vídeos en general no están transcritos, aunque las sesiones parlamentarias en particular sí están transcritas. Las transcripciones no son completamente literales (al menos las sesiones parlamentarias en el diario de sesiones), pero quizás podrían servir como base para facilitar una transcripción manual más detallada. Los vídeos (al menos los últimos) están segmentados por

intervenciones identificando a la persona que interviene (no única, pero sí mayoritariamente). Algunos contenidos no están transcritos. Las transcripciones de las sesiones, al parecer, están disponibles traducidas a otros idiomas oficiales del Estado a través del servicio denominado cortesi@. Se permite explícitamente su reutilización sin condiciones, si bien no está disponible la descarga por bloques. Se trata de un recurso con **madurez media**, dentro de los recursos multimedia. Una forma de mejorar el grado de madurez del recurso de forma sencilla sería permitir su descarga de forma masiva.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	N/A	
2. Transcripción (ortográfica, fonológica, suprasegmental…)	*	Algunos contenidos transcritos, otros no. Los transcritos no lo están de forma totalmente literal.
3. Alineación vídeo/sonido y texto	*	Alineados vídeo y locutores, pero no vídeo y texto.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	*	Anotado por intervinientes
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	-	

9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	N/A	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	**	No requiere.
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	No requiere autorización previa.
		Madurez media

Tabla 9: Madurez del recurso 7: Grabaciones de audio y vídeo del Archivo Audiovisual del Congreso de los Diputados de España.

Posibles aplicaciones del futuro recurso lingüístico

- Como base de datos de entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-to-Text) que posteriormente permitan facilitar la transcripción automática o la búsqueda en este tipo de contenidos.
- Como base de datos de entrenamiento y evaluación de sistemas de identificación del hablante.

Recomendaciones: Sería de gran utilidad que se habilitara la posibilidad de descarga masiva de secciones de los audios y vídeos del Archivo Audiovisual del Congreso de los Diputados de España.

4.2.2 Cultura

4.2.2.1 Recurso 8: Índices de clasificación de los catálogos de la BNE

Identificación del recurso.

- *Nombre: Índices de clasificación de los catálogos en DATOS.BNE.ES.*
- *Clasificación por tipo de documento:* Publicación de datos como Linked Open Data, basado en tecnologías y estándares de la Web. La Biblioteca ofrece la versión íntegra de su catálogo automatizado en ficheros comprimidos creados a partir de los distintos formatos bibliográficos y registros de autoridad. De cada uno de los formatos bibliográficos y de autoridad se incluyen

dos ficheros comprimidos: uno con los registros bibliográficos o de autoridad en XML, y otro, con el mismo contenido, en MRC (ISO 2709: formato de intercambio e integración de información bibliográfica en entornos automatizados).

- *Clasificación por número de lenguas:* Multilingüe.
- *Lenguas:* Los datos de autoridades y bibliográficos están en múltiples lenguas, destacando el español, inglés y francés.
- *Descripción del recurso:* Contiene todos los índices de clasificación del catálogo, incluyendo persona, persona-título, entidad, congreso, título, materia, geográfico, subencabezamientos de materia y género.
- *Fecha de comienzo de creación*: 2008.
- *Fecha de finalización:* N/A.
- *Frecuencia de actualización*: El contenido de todos los ficheros se actualiza a primeros del mes en curso.
- *Fecha de última actualización*: Diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Descarga masiva: http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/SuministroRegistro/Ficheros/
- *Tipo de licencia:* Los datos del Catálogo tienen licencia Creative Commons, su uso es gratuito y no requiere autorización; pero la BNE agradecerá la mención del origen de los registros.[23]
- *Descarga masiva disponible:* SÍ.

Persona de contacto u organización responsable:

- *Contacto:* http://www.bne.es/es/Servicios/ReproduccionDocumentos/index.html.
- *Nombre organización:* Biblioteca Nacional de España.

Creación del recurso

- *Proveedor y/o creador:* Biblioteca Nacional de España.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* Estándar, formal.

[23] *http://www.bne.es/es/NavegacionRecursiva/Pie/avisoLegal/*

- *Niveles de anotación lingüística:* Metadatos.
- *Conforme a los estándares*: Sí respecto a los estándares de bibliotecas digitales internacionales XML y MRC (ISO 2709).
- *Tamaño y cobertura:* El tamaño y la cobertura temporal o temática de cada fichero es variable, pero son muy completos.
- *Unidad (términos, entradas, textos, oraciones, otro):* Registros bibliográficos: cada registro corresponde a una entidad (autor, lugar, etc.), título o materia.
- *Formato*: XML y MRC.
- *Dominio:* Todos los dominios y cualquier tipo de documento
- *Género:* General y variado.
- *Tipo de texto:* Registro de base de datos.

Otros recursos relacionados:

- RDF obras, autores y materias: http://datos.bne.es/inicio.html
- Este recurso está relacionado con el censado como número 10 (Biblioteca Digital Hispánica), que incluye un subconjunto de este catálogo.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* Los registros no están anotados lingüísticamente, pero contienen información estructurada. Es una colección muy completa, y los datos aparecen como Linked Open Data (LOD). Lo más destacable es que la BNE ofrece íntegramente todo el catálogo en XML, lo que lo hace directamente utilizable como RL.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad de procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	**	Directamente en XML para su procesamiento
2. Transcripción (ortográfica, fonológica, suprasegmental…)	-	

3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	Ficheros en UTF-8
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	*	No directamente, pero se puede inferir de los ficheros
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	Se generan automáticamente de catálogo de la BNE
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran un revisor experto)	**	Se generan automáticamente de catálogo de la BNE
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	*	Información publicada como LOD
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	no requiere autorización previa
		Madurez alta

Tabla 10: Madurez del recurso 8: Índices de clasificación de los catálogos de la BNE.

Posibles aplicaciones del futuro recurso lingüístico

- Sus datos de materias y submaterias se pueden reutilizar para reconocedores de entidades nombradas (NER por sus siglas en inglés) y clasificadores de entidades, materias, títulos, etc.

Recomendaciones: Enriquecer con anotación lingüística y emplear para entrenamiento de clasificadores.

4.2.2.2 Recurso 9: Publicaciones periódicas digitalizadas de la Hemeroteca Digital

Identificación del recurso.

- *Nombre:* Publicaciones periódicas digitalizadas de la Hemeroteca Digital.
- *Clasificación por tipo de documento:* Contiene más de 2.066 publicaciones periódicas digitalizadas y un total de 55.395.726 páginas que abarcan desde 1683 [fecha de consulta: 2 de julio de 2018]. La oferta de títulos va a ir ampliándose hasta cubrir la evolución histórica de la prensa española, desde sus inicios hasta principios del siglo XX, respetando siempre las limitaciones que marca nuestra legislación en temas de propiedad intelectual. El criterio que ha guiado la composición de esta colección ha sido seleccionar periódicos y revistas representativos de su época, que reflejaran la riqueza temática de la edición hemerográfica hispana y de los que se conservaran colecciones completas. En 2012 se ha puesto en producción una nueva versión de la aplicación de Hemeroteca Digital, que cumple con los estándares internacionales OAI (Open Archives Initiative) y EUROPEANA, y con algunas mejoras en la interfaz de búsqueda para facilitar su consulta. Los principales editores de prensa actuales han firmado una colaboración para que digitalicen diariamente sus PDF en la aplicación. Esto ha provocado que el porcentaje de páginas con derechos de propiedad intelectual sea mucho mayor que el de libre acceso. Hay un filtro que permite buscar solamente en los títulos de libre acceso (prensa histórica). Con la actualización de enero de 2019 la Hemeroteca Digital alcanza 60.120.500 páginas y 2.179 títulos:
  - 1.804 títulos (6.826.989 páginas) del fondo histórico de prensa y revistas.
  - 374 títulos (53.293.511 páginas) del fondo moderno, accesibles únicamente desde las sedes de la Biblioteca Nacional de España al estar sujetos a derechos de autor.

  Los títulos de fondo moderno incluyen las publicaciones que, gracias a la colaboración de los editores de prensa, ingresan a diario en formato electrónico y se actualizan periódicamente, así como las que se agregan procedentes de la colección de prensa actual digitalizada por la BNE en años anteriores y que hasta ahora se consultaba en CD-ROM. Está previsto continuar migrando imágenes de este soporte a la Hemeroteca Digital.
- *Clasificación por número de lenguas:* Multilingüe.
- *Lenguas:* Español, catalán, francés, alemán, inglés.
- *Descripción del recurso:* Contiene un variado catálogo de publicaciones periódicas históricas de diversa naturaleza y de diferentes ámbitos geográficos, no solo España sino

Hispanoamérica, Gran Bretaña, Francia, Estados Unidos, Alemania, Italia o Marruecos. El fondo moderno no se puede consultar en línea.

- *Fecha de comienzo de creación*: 2007.
- *Fecha de finalización:* N/A.
- *Frecuencia de actualización*: se incorporan nuevas revistas del fondo histórico de manera periódica y todos los días se actualiza el fondo moderno.
- *Fecha de última actualización*: Enero 2019.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Consulta en línea: http://www.bne.es/es/Catalogos/HemerotecaDigital/
- *Tipo de licencia:* Salvo que se especifique expresamente lo contrario, las imágenes en dominio público que se encuentren en el dominio bne.es están bajo una licencia de Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional de Creative Commons. Solo se puede consultar el fondo histórico, ya que al fondo moderno (con derechos de propiedad intelectual) solo se tiene acceso en la sede de la BNE.[24]
- *Descarga masiva disponible:* Sí.

Persona de contacto u organización responsable:

- *Contacto:* http://www.bne.es/es/Servicios/ReproduccionDocumentos/index.html.
- *Nombre organización:* Biblioteca Nacional de España.

Creación del recurso

- *Proveedor y/o creador:* Biblioteca Nacional de España.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* Estándar.
- *Niveles de anotación lingüística:* Metadatos de catalogación.
- *Conforme a los estándares*: Sí respecto a los estándares de bibliotecas digitales internacionales OAI (Open Archives Iniciative) y Europeana.

[24] *http://www.bne.es/es/NavegacionRecursiva/Pie/avisoLegal/*

- *Tamaño y cobertura:* El tamaño supera los 60 millones de páginas y más de 2000 publicaciones. El espacio temporal va desde 1683 hasta nuestros días, actualizándose periódicamente, tanto el fondo histórico como el actual.
- *Unidad (términos, entradas, textos, otro):* Registros bibliográficos (cada registro corresponde a un número de la revista).
- *Formato*: PDF.
- *Dominio:* Todos los dominios y cualquier tipo de documento.
- *Género:* General y variado.
- *Tipo de texto:* Registro de base de datos y visualización de la página escaneada en PDF.

Otros recursos relacionados:

- Índices de clasificación de los catálogos de la BNE (recurso 8).
- Documentos digitalizados de la Biblioteca Digital Hispánica (recurso 10).

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* Los registros no están anotados lingüísticamente; solo contienen metadatos de catalogación. El buscador permite hacer búsquedas en el texto de las publicaciones. Cumple los estándares internacionales de datos abiertos y bibliotecas digitales. Su punto débil es su escasa madurez para ser reutilizado como RL: al estar en formato PDF, se necesitaría, en primer lugar, procesarlo con OCR, revisar la transcripción, y luego, comenzar con su anotación. Presenta **madurez baja**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad de procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	-	En ocasiones, el PDF es fácilmente convertible en TXT; otras veces se requiere OCR.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	

3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	-	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	-	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran un revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	*	Información de los registros del catálogo.
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	*	No requiere autorización previa para el fondo histórico. El fondo moderno es de libre acceso.
		Madurez baja.

Tabla 11: Madurez del recurso 9: Publicaciones periódicas digitalizadas de la Hemeroteca Digital.

Posibles aplicaciones del futuro recurso lingüístico

- Creación de un recurso histórico general tipo Google Books para estudios lingüísticos, históricos y culturales de la evolución de la prensa en España y otros países relacionados históricamente con nuestro país.
- Generación de lexicones de especialidad, de materia, por ámbito geográfico, por tipo de publicación, etc.
- Generación de modelos de lenguaje por tipos de publicaciones, ámbitos geográficos, lingüísticos, etc.
- Entrenamiento de clasificadores.

- Extracción de entidades nombradas en la prensa histórica.

Recomendaciones: Convertir de formato PDF a texto mediante OCR y revisión de la transcripción. También sería de utilidad enriquecerlo con anotación lingüística.

4.2.2.3 Recurso 10: Documentos digitalizados de la Biblioteca Digital Hispánica

Identificación del recurso.

- *Nombre:* Documentos digitalizados de la Biblioteca Digital Hispánica.
- *Clasificación por tipo de documento:* La Biblioteca Digital Hispánica es la biblioteca digital de la Biblioteca Nacional de España. Proporciona acceso libre y gratuito a miles de documentos digitalizados, entre los que se cuentan libros impresos entre los siglos XV y XIX, manuscritos, dibujos, grabados, folletos, carteles, fotografías, mapas, atlas, partituras, prensa histórica y grabaciones sonoras.
- *Clasificación por número de lenguas:* Multilingüe.
- *Lenguas:* Destaca las obras en español, pero las hay de numerosas lenguas, incluidas lenguas minoritarias o de familias lingüísticas muy diferentes (japonés, chino, árabe, etc.)
- *Descripción del recurso:* Más de 221. 000 documentos (enero 2019).
- *Fecha de comienzo de creación*: 2008.
- *Fecha de finalización:* N/A.
- *Frecuencia de actualización*: Mensual
- *Fecha de última actualización*: Enero 2019.
- *Versión:* N/A.
- *Identificador del recurso:*
  - Set completo en Datos.gob.es: https://datos.gob.es/es/catalogo/e00123904-biblioteca-digital-hispanica-bdh-set-completo
  - Consulta en línea en BNE: http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/index.html
- *Tipo de licencia:* Portal libre y gratuito de documentos digitalizados de la BNE. El usuario queda autorizado por la BNE a utilizar las imágenes accesibles en cualquiera de las páginas pertenecientes al dominio bne.es, siempre y cuando dicha utilización no tenga fines comerciales o lucrativos. Por tanto, y salvo que se especifique expresamente lo contrario, las

imágenes en dominio público que se encuentren en el dominio bne.es están bajo una licencia de Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional de Creative Commons (CC-BY-NC-SA): El uso público no comercial de las imágenes de la Biblioteca Digital Hispánica (con contenido en dominio público) es gratuito y no requiere autorización previa. El uso de dichas imágenes no conlleva una cesión en exclusiva e implicará citar la procedencia de la obra reproducida como perteneciente a los fondos de la Biblioteca Nacional de España.

- *Descarga masiva disponible:* SÍ.

Persona de contacto u organización responsable:

- *Contacto:* http://www.bne.es/es/Servicios/ReproduccionDocumentos/index.html.
- *Nombre organización:* Biblioteca Nacional de España.

Creación del recurso

- *Proveedor y/o creador:* Biblioteca Nacional de España.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* Dependiente de los contenidos de cada obra; la mayoría, en variedad estándar y formal.
- *Niveles de anotación lingüística:* Metadatos, texto y audio.
- *Conforme a los estándares*: Sí respecto a los estándares de bibliotecas digitales internacionales (MARC XML, METS-PREMIS, protocolo OAI-PMH de exportación de metadatos).
- *Tamaño y cobertura:* El tamaño y la cobertura temporal o temática de cada obra es variable.
- *Unidad (términos, entradas, textos, oraciones, otro):* Texto, separado por páginas, al tratarse de obras digitalizadas que reproducen el original. Se puede descargar cada página en formato PDF y JPEG. En la descarga masiva de todo el conjunto se pueden encontrar en los formatos CSV, JSON, ODS, TXT y XML, pero no están los textos individuales. En la descarga masiva hay enlaces a cada documento en su versión digital, y en texto OCR sin revisar.
- *Formato*: PDF, JPEG (textos individuales); CSV, JSON, ODS, TXT y XML (catálogo completo).
- *Dominio:* Los contenidos de las obras son de múltiples temáticas: Ciencia y cultura en general, Filosofía, Psicología, Religión, Teología, Ciencias sociales, Ciencias puras y Ciencias naturales, Ciencias aplicadas, Medicina, Tecnologías, Bellas artes, Espectáculos, Deportes, Lingüística. Literatura, Geografía. Biografías, Historia, etc.
- *Género:* General y variado.

- *Tipo de texto:* Texto, registros sonoros musicales y no musicales, dibujos, cartografía, grabados, fotografías.

Otros recursos relacionados:

- Textos digitales en formato EPUB de la BNElab (ver consulta en línea: www.bne.es/bnelab, libros interactivos: www.bne.es/es/Colecciones/LibrosInteractivos/index.html, Biblioteca Digital Hispánica: http://bdh.bne.es/bnesearch/Search.do?&destacadas1=Epub&home=true&languageView=es). Este recurso es ideal para la creación de modelos de lenguaje diacrónicos y de variantes del español, por ejemplo, como mecanismo para mejorar los resultados de sistemas basados en OCR para digitalización de documentos históricos.

- Relacionado con el recurso 8, Índices de clasificación de los catálogos de la BNE, donde está todo el catálogo completo, del que la BDH es un subconjunto.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* Las obras no están anotadas lingüísticamente, sino en formato imagen o audio. Un aspecto interesante es que la mayoría de los textos están procesados por OCR y el texto transcrito está disponible. Desgraciadamente, estos textos no están revisados y su calidad es baja. Los documentos son de acceso libre y sin el requerimiento de solicitar permiso para su uso. Recientemente, se han habilitado en diferentes formatos, lo que permite su descarga masiva. Su grado de **madurez** para convertirse en RL es **bajo,** porque necesita ser revisado completamente. Su interés para el PLN es enorme por su variedad de dominios, épocas y lenguas.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad de procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Están accesibles en PDF o JPEG, además de la conversión a TXT, pero el fichero resultante no está revisado y contiene

		muchos errores para ser usado directamente como RL. Los metadatos están en varios formatos (XML, JSON,CSV).
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	Se ofrece una versión en formato TXT, sin revisión.
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	-	Es necesaria.
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran un revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	**	Los datos de cada texto son descargables en varios formatos (JSON, CSV, XML).
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	No requiere autorización previa.
		Madurez baja

Tabla 12: Madurez del recurso 10: Documentos digitalizados de la Biblioteca Digital Hispánica.

Posibles aplicaciones del futuro recurso lingüístico

- *Ejemplo de aplicaciones posibles:*

- Base para crear un repositorio como Google Books, Google n-gram viewer, y generar modelos word2vec o similares.
- Entrenamiento de clasificadores.
- Creación de modelos de lenguaje diacrónicos y de variantes del español, por ejemplo, como mecanismo para mejorar los resultados de sistemas basados en OCR para la digitalización de documentos históricos.

Recomendaciones: Revisión de los textos procesados con OCR para mejorar su calidad. Serían de gran utilidad para la mejora de sistemas OCR para textos históricos.

4.2.3 Sanidad

4.2.3.1 Recurso 11: Publicaciones en repositorio SciELO (Scientific Electronic Library Online)

Identificación del recurso.

- *Nombre:* Publicaciones en repositorio SciELO (Scientific Electronic Library Online).
- *Clasificación por tipo de documento:* Artículos de revistas científicas.
- *Clasificación por número de lenguas:* Multilingüe.
- *Lenguas:* Predominantemente, español e inglés; también otras como portugués o francés.
- *Descripción del recurso:*
  - Repositorio que reúne más de 1.280 colecciones de revistas científicas de ámbito mayoritariamente iberoamericano; cada país mantiene un portal para sus revistas:
    - Argentina: www.scielo.org.ar
    - Bolivia: www.scielo.org.bo
    - Brasil: www.scielo.br
    - Chile: www.scielo.cl
    - Colombia: www.scielo.org.co
    - Costa Rica: www.scielo.sa.cr
    - Cuba: http://scielo.sld.cu
    - Ecuador: http://scielo.senescyt.gob.ec
    - España: http://scielo.isciii.es
    - México: www.scielo.org.mx
    - Paraguay: http://scielo.iics.una.py
    - Perú: www.scielo.org.pe
    - Portugal: www.scielo.mec.pt
    - Sudáfrica: www.scielo.org.za
    - Uruguay: www.scielo.edu.uy
    - Venezuela: www.scielo.org.ve

- *Fecha de comienzo de creación*: variable (dependiente de cada revista).
- *Fecha de finalización:* N/A.

- *Frecuencia de actualización*: variable (dependiente de cada revista).
- *Fecha de última actualización*: el portal general indica que septiembre de 2017; el portal de cada país mantiene una actualización diferente.
- *Versión:* N/A.
- *Identificador del recurso:* www.scielo.org
- *Tipo de licencia*: Variable de una publicación a otra, y según el repositorio SciELO de cada país; el portal SciELO España tiene una licencia Creative Commons Atribución BY-NC-SA 4.0.

- *Descarga masiva disponible:* SI, de cada artículo individual, aunque no en todas las revistas, pues algunas solo presentan los artículos en HTML.

Persona de contacto u organización responsable:

- *Contacto:* Email: scielo@scielo.org Tel.: (55 11) 5083-3639. Dirección de correo: Av. Onze de Junho, 269 - Vila Clementino 04041-050 São Paulo SP Brasil.
- *Nombre organización:* SciELO - Scientific Electronic Library Online.

Creación del recurso

- *Proveedor y/o creador:* Repositorio general (www.scielo.org): FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) - CAPES - CNPq - BIREME (Centro Latinoamericano y del Caribe de información en Ciencias de la Salud) – FapUNIFESP.
  El portal de revistas de cada país es gestionado por una institución nacional; por ejemplo, la Biblioteca Nacional de Ciencias de la Salud del Instituto de Salud Carlos III (ISCIII) en el caso de España. Cada revista incluida en el repositorio es producida de manera independiente por asociaciones científicas o culturales.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* Estándar, formal.
- *Niveles de anotación lingüística:* Datos no anotados.
- *Conforme a los estándares*: No.
  - *Tamaño y cobertura:*
    - A fecha de la última actualización del portal web (septiembre de 2017), se recogían 1.285 revistas activas, 52.356 números, 745.182 artículos y 16.943.454 citas.

Consúltese el portal de cada país y de cada revista incluida para detalles más concretos.

- *Unidad (términos, entradas, textos, oraciones, otro):* Textos.
- *Formato*:
    - PDF y/o XML y/o HTML (depende del formato disponible en cada revista)
- *Dominio:* Ciencias Agrícolas, Ciencias Biológicas, Ciencias de la Salud, Ciencias Exactas y de la Tierra, Ciencias Sociales Aplicadas, Humanidades, Ingenierías, Lingüística, Letras y Artes.
- *Género:* Artículos e informes científicos.
- *Tipo de texto:* Publicaciones científicas y culturales.

Otros recursos relacionados:

- El Instituto de Salud Carlos III (ISCIII) publica varias revistas e informes en formatos semejantes, pero pocas se incluyen en el repositorio SciELO España, de modo que se analizan en la ficha del recurso número 12.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento:* las publicaciones del repositorio SciELO constituyen un amplio corpus de textos principalmente de dominio sanitario, pero también del área de inteligencia competitiva o cultura. La calidad y representatividad de cada área temática depende de cada revista. Sucede lo mismo con los formatos de archivo, puesto que muchos se encuentran solo en PDF (que requiere la conversión a formatos más adecuados para PLN), y otras publicaciones también están en HTML y XML. Los artículos en formato XML suelen presentar una estructura del documento poco variable de unas revistas a otras, lo que facilita la extracción de información. La descarga masiva de contenidos varía de unas revistas a otras. Con un alto potencial como fuente de corpus, valoramos que posee **madurez media**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Ciertas publicaciones solo en PDF.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	*	La codificación de caracteres suele estar en ISO-8859-1, pero puede variar de unas revistas a otras (UTF-8).
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	*	Los artículos en XML suelen tener la misma estructura y etiquetas.
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	**	Los textos están redactados por investigadores y profesionales.
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	*	Cada artículo se clasifica por palabras clave; en el formato XML, se recoge en las etiquetas “<kwd>”.
Aspectos legales		

12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	En general, la licencia es Creative Commons, pero recomendamos confirmar su uso en cada revista.
		Madurez media

Tabla 13: Madurez del recurso 11: Publicaciones en repositorio SciELO (Scientific Electronic Library Online).

Posibles aplicaciones del futuro recurso lingüístico

- Extracción de terminología y corpus de dominio especializado, sobre todo de ámbito sanitario.

Constitución de corpus paralelos a partir de los resúmenes (*abstracts*) de artículos, que suelen presentarse en inglés además de en la lengua de cada publicación.

4.2.3.2 Recurso 12: Publicaciones y vídeos del Instituto de Salud Carlos III (ISCIII)

Identificación del recurso.

- *Nombre:* Publicaciones y vídeos del Instituto de Salud Carlos III (ISCIII).
- *Clasificación por tipo de documento:* boletines e informes científicos; vídeos.
- *Clasificación por número de lenguas:* monolingüe.
- *Lenguas:* español (ciertas monografías solo en inglés).
- *Descripción del recurso:*
  - *Monografías ISCIII:* publicaciones editadas por diferentes centros: Dirección General del ISCII, Agencia de Evaluación de Tecnologías Sanitarias (AETS), Telemedicina, Escuela Nacional de Sanidad, Centro Nacional de Epidemiología o Centro Nacional de Microbiología.
  - *Boletín Epidemiológico Semanal*: el Centro Nacional de Epidemiología edita esta publicación con trabajos acerca de salud pública.
  - *Medicina y Seguridad del Trabajo:* la Escuela Nacional de Medicina del Trabajo (ENMT) publica esta revista con artículos acerca de la salud laboral.

- *Boletín del ECEMC (Estudio Colaborativo Español de Malformaciones Congénitas)*: publica trabajos del grupo de investigación ECEMC en el Centro de Investigación sobre Anomalías Congénitas (CIAC).
- Portal Vídeos: contiene grabaciones en torno a 4 temáticas: Institucional, Formación, Científico-técnicos, y Otros.

- *Fecha de comienzo de creación*: variable (dependiente de cada publicación).
- *Fecha de finalización:* N/A.
- *Frecuencia de actualización*:
  - *Boletín Epidemiológico Semanal:* semanal.
  - *Medicina y Seguridad del Trabajo*: trimestral.
  - *Boletín del ECEMC*: anual.
  - Portal Vídeos: variable.
- *Fecha de última actualización*: septiembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
  - Monografías ISCIII: https://publicaciones.isciii.es/
  - *Boletín Epidemiológico Semanal* (eISSN:2173-9277): http://revista.isciii.es/index.php/bes/issue/archive
  - *Medicina y Seguridad del Trabajo* (ISSN 1989-7790 de la versión en línea): http://scielo.isciii.es/scielo.php?script=sci_serial&pid=0465-546X&lng=es&nrm=iso.
  - *Boletín del ECEMC* (ISSN: 0210-3893): http://revista.isciii.es/index.php/ecemc/issue/archive
  - Portal Vídeos: http://portal-videos.isciii.es/

- *Tipo de licencia:*
  - Monografías del ISCIII: derechos de propiedad intelectual del ISCIII, uso restringido o permitido si cita a autores (consúltese cada publicación).
  - Revistas del ISCIII: licencia Creative Commons Atribución BY-NC-SA 4.0.
  - Portal Vídeos: Derechos de propiedad intelectual del ISCII (contactar).
- *Descarga masiva disponible:* Sí, de cada revista, independientemente.

Persona de contacto u organización responsable:

- *Contacto:* programaeditorial@isciii.es / http://publicaciones.isciii.es/contact.jsp

- *Nombre organización:* Publicaciones Online del Instituto de Salud Carlos III.

Creación del recurso

- *Proveedor y/o creador:* Instituto de Salud Carlos III.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
    - *Tamaño y cobertura:*
        - *Boletín Epidemiológico Semanal:* más de 2.500 boletines desde 1997.
        - *Medicina y Seguridad del Trabajo*: 62 revistas desde 2004.
        - *Boletín del ECEMC:* 12 boletines entre 2002 y 2014.
        - Portal Vídeos: 47 grabaciones desde 2014.
- *Unidad (términos, entradas, textos, oraciones, otro):* textos y vídeos.
- *Formato*:
    - Monografías del ISCII: PDF y/o EPUB.
    - *Boletín Epidemiológico Semanal* y *Medicina y seguridad del trabajo*: HTML / PDF / EPUB (no todos los números, depende de la antigüedad).
    - *Boletín del ECEMC:* casi todos los números en PDF.
    - Portal Vídeos: N/A (solo se permite la visualización en el reproductor propio)
- *Dominio:* medicina, investigación clínica, epidemiología.
- *Género:* boletines, artículos e informes científicos. Vídeos: jornadas de formación, conferencias, grabaciones destinadas a la formación continua.
- *Tipo de texto:* terminología sanitaria. Vídeos: especializados y divulgativos.

Otros recursos relacionados:

- Conjunto de revistas del repositorio SciELO analizado como recurso número 11.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento:* El portal del Instituto de Salud Carlos III recoge publicaciones e informes de interés para tareas de PLN en dominio sanitario (p. ej. creación de corpus y extracción de terminología). Asimismo, el Portal Vídeos incluye grabaciones con finalidad científica, formativa

y divulgadora, cuyas transcripciones son de interés para entrenar sistemas de reconocimiento de voz. No obstante, la mayoría de datos textuales se encuentran en PDF, lo que requiere la conversión de contenidos previa al procesamiento; pocas publicaciones están en HTML y permitirían ser procesados de inmediato. Por otra parte, los vídeos no están disponibles para la descarga, es preciso contactar para obtenerlos. No es posible la descarga masiva de contenidos. Pese a su potencial como fuente de creación de recursos, estimamos que posee **madurez media** (textos) y **baja** (vídeos).

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	* (textos) N/A (vídeos)	Solo ciertas publicaciones en HTML.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A (textos) * (vídeos)	Los vídeos tienen subtítulos generados automática-mente.
3. Alineación vídeo/sonido y texto	N/A (textos) * (vídeos)	Subtítulos generados automática-mente.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	** (textos) N/A (vídeos)	Las revistas en HTML (UTF-8).
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	*	Los subtítulos de los vídeos requerirían revisión.
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de	**	Los textos están redactados por

subconjunto de datos, o tareas que requieran revisor experto)		profesionales sanitarios.
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	Las mono-grafías poseen una clasificación temática, pero no codificada en los archivos.
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	** (textos) - (vídeos)	Es preciso contactar acerca del estatuto de uso de los vídeos.
TOTAL	Textos: ** Vídeos: *	Textos: madurez media Vídeos: madurez baja

Tabla 14: Madurez del recurso 12: Publicaciones y vídeos del Instituto de Salud Carlos III (ISCIII).

Posibles aplicaciones del futuro recurso lingüístico

- *Ejemplo de aplicaciones posibles:*
    - Textos: extracción de terminología de dominio sanitario; recogida de corpus de textos de dominio médico.
    - Vídeos: entrenamiento de sistemas de reconocimiento de voz (Speech-To-Text) en el dominio sanitario, para aplicaciones de detección de palabras clave (Key-word Spotting), subtitulado automático, búsqueda y recuperación de información, o reconocimiento y segmentación de locutores.

Recomendaciones: Habilitar la descarga masiva.

4.2.3.3 Recurso 13: Banco de datos de enfermedades raras y medicamentos huérfanos de OrphaData.

Identificación del recurso.

- *Nombre:* OrphaData.
- *Clasificación por tipo de documento:* banco de datos anotado de enfermedades raras y medicamentos huérfanos.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* alemán, español, francés, inglés, italiano, neerlandés y portugués; el inventario de enfermedades raras también disponible en checo y polaco. Solo disponibles en inglés la clasificación jerárquica de enfermedades raras, la lista de enfermedades raras asociadas con genes y la linearización de enfermedades.
- *Descripción del recurso:* subconjunto de datos de uso gratuito, extraído de Orphanet (www.orpha.net):
    - Orphanet Rare Diseases Ontology (ORDO): vocabulario estructurado y ontología de enfermedades raras, con relaciones anotadas entre síndromes, genes y datos epidemiológicos, que incluye códigos en terminologías de referencia (Medical Subject Headings, MeSH; Medical Dictionary for Regulatory Activities, MedDRA; Unified Medical Language System, UMLS), bancos de datos genéticos (HUGO Gene Nomenclature Committee, HGNC; Online Mendelian In Man Database, OMIM; GenAtlas, UniProt Knowledgebase, UniProtKB; Ensembl; International Union of Basic and Clinical Pharmacology database, IUPHAR-DB; Reactome) y clasificaciones (International Classification of Diseases vs. 10, ICD-10; Genetic and Rare Diseases; GARD).
    - Lista de enfermedades raras y referencias (Rare diseases and cross-referencing): nomenclatura y alineamiento con terminologías o clasificaciones de referencia (MeSH, UMLS, MedDRA, OMIM, ICD-10, GARD).
    - Fenotipos asociados con enfermedades raras (*Phenotypes associated with rare disorders*): contiene información asociadas a las enfermedades raras: frecuencia, criterios diagnósticos o signos patognomónicos.
    - Los siguientes conjuntos de datos requieren un Acuerdo de Transferencia de Datos (*Data Transfer Agreement*) con fines académicos: Información textual sobre enfermedades raras, Organizaciones de pacientes, Centros expertos, Test diagnósticos y laboratorios clínicos, y Medicamentos huérfanos.
- *Fecha de comienzo de creación*: variable (véase apdo. Tamaño).
- *Fecha de finalización:* N/A.
- *Frecuencia de actualización*: mensual.

- *Fecha de última actualización*: diciembre de 2018.
- *Versión:* Orphadata V. 0.9.8. Orphanet Rare Disease Ontology (ORDO) vs. 2.5.

- *Identificador del recurso:* www.orphadata.org/cgi-bin/index.php/
- *Tipo de licencia:* Creative Commons Reconocimiento-SinObraDerivada. Citar de este modo: *Orphadata: Free access data from Orphanet. © INSERM 1997. Available on http://www.orphadata.org. Data version (XML data version).*[25] Otros datos creados por Orphanet requieren un Acuerdo de Transferencia de Datos (*Data Transfer Agreement*) con fines académicos: Información textual sobre enfermedades raras, Organizaciones de pacientes, Centros expertos, Test diagnósticos y laboratorios clínicos, y Medicamentos huérfanos.
- *Descarga masiva disponible:* SI.

Persona de contacto u organización responsable:

- *Contacto:* OrphaData. www.orphadata.org/cgi-bin/contact.php
- *Nombre organización:* Orphanet INSERM (Institut national de la santé et de la recherche médicale).

Creación del recurso

- *Proveedor y/o creador:* Orphanet INSERM.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* datos anotados con códigos en terminologías médicas de referencia (MeSH, MedDRA, UMLS), bancos de datos genéticos (HGNC, OMIM, GenAtlas, UniProtKB, Ensembl, IUPHAR-DB, Reactome) y clasificaciones (ICD-10, GARD).
- *Conforme a los estándares*: sí, respecto a las nomenclaturas de dominio médico.
- *Tamaño y cobertura:*
    - Lista de enfermedades raras y referencias: más de 9500 entradas descriptivas de enfermedades raras.

[25] http://www.orphadata.org/cgi-bin/inc/legal.inc.php

    - Fenotipos asociados con enfermedades raras: más de 67000 entradas de fenotipos asociados a códigos de Human Phenotype Ontology (HPO), respecto a más de 3100 síndromes.
- *Unidad (términos, entradas, textos, oraciones, otro):* entradas (banco de datos).
- *Formato*:
    - Orphanet Rare Disease ontology (ORDO): OWL. Incluye un punto de acceso para consultas SPARQL: www.orpha.net/sparql
    - Inventario de enfermedades raras y referencias (Rare diseases and cross-referencing): disponible en XML y JSON.
    - Clasificaciones jerárquicas de enfermedades raras, Fenotipos asociados con enfermedades raras, Banco datos de enfermedades y genes relacionados, y Linearización de enfermedades: XML.
- *Dominio:* medicina, genética, investigación clínica, epidemiología, farmacología.
- *Género:* banco de datos.
- *Tipo de texto:* terminología científico-técnica.

Otros recursos relacionados:

- No se han identificado recursos relacionados.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* OrphaData recoge recursos con alto potencial para tareas de PLN en dominio sanitario (extracción de terminología multilingüe, extracción de listas de entidades nombradas, confección de tesauros anotados o extracción de relaciones y minería de textos). Los datos están ricamente anotados conforme a terminologías de referencia, se distribuyen en XML, o permiten el acceso dinámico mediante consultas SPARQL. La descarga de datos es inmediata y su distribución y actualización es continua. Por todos estos factores, se trata de un recurso de **madurez alta** y de alto interés.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	**	
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	Recursos en español en ISO-8859-1.
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	**	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	**	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	*	Solo ciertos archivos.
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez alta.

Tabla 15: Madurez del recurso 13: Banco de datos de enfermedades raras y medicamentos huérfanos de OrphaData.

Posibles aplicaciones del futuro recurso lingüístico

- Extracción de terminología multilingüe de dominio médico.
- Extracción de listas de entidades nombradas de tipos semánticos de dominio médico: p. ej. enfermedades, genes, fenotipos y nombres de medicamentos.
- Confección de tesauros anotados con códigos en terminologías de referencia.
- Extracción de conocimiento ontológico (relaciones entre enfermedades, fenotipos y genes) para su aplicación en minería de textos.

Recomendaciones: No se han identificado; en nuestra opinión, dispone de madurez alta.

4.2.3.4 Recurso 14: Guías de práctica clínica (GPC) del portal Guía Salud

Identificación del recurso.

- *Nombre:* Guía Salud.
- *Clasificación por tipo de documento:* corpus de textos.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* predominantemente, español; algunas guías en inglés, catalán, euskera o gallego.
- *Descripción del recurso:* guías de práctica clínica (GPC) del Sistema Nacional de Salud.
- *Fecha de comienzo de creación*: 2002.
- *Fecha de finalización:* N/A.
- *Frecuencia de actualización*: variable.
- *Fecha de última actualización*: 2018.
- *Versión:* N/A.
- *Identificador del recurso:* http://portal.guiasalud.es
- *Tipo de licencia:* Creative Commons BY-NoComercial-SinObraDerivada 3.0.[26]
- *Descarga masiva disponible:* Sí, aunque cada guía se descarga independientemente.

Persona de contacto u organización responsable:

- *Contacto:* Secretaría de Guía Salud. Instituto Aragonés de Ciencias de Salud (IACS), Zaragoza. iacs@guiasalud.es
- *Nombre organización:* Sistema Nacional de Salud.

Creación del recurso

[26] *www.guiasalud.es/web/guest/aviso-legal*

- *Proveedor y/o creador:* los contenidos son elaborados por organismos (consejerías y departamentos de salud) de las 17 comunidades autónomas.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:* más de 170 guías de práctica clínica que cubren los siguientes tipos de trastornos (clasificados según CIE): anomalías congénitas, causas externas, enfermedades con origen en el periodo perinatal, complicaciones del embarazo, parto y puerperio, enfermedades de la piel y del tejido subcutáneo, enfermedades de la sangre y de los órganos hematopoyéticos, enfermedades del aparato digestivo, enfermedades del aparato genitourinario, enfermedades del aparato respiratorio, enfermedades del sistema circulatorio, enfermedades del sistema nervioso y de los órganos de los sentidos, enfermedades del sistema osteo-mioarticular y tejido conectivo, enfermedades infecciosas y parasitarias, enfermedades endocrinas, de la nutrición y metabólicas y trastornos de la inmunidad, factores que influyen en la salud, lesiones y envenenamientos, neoplasias, síntomas, signos y estados mal definidos, y trastornos mentales.
- *Unidad (términos, entradas, textos, oraciones, otro):* textos.
- *Formato*: PDF y HTML.
- *Dominio:* medicina, atención sanitaria.
- *Género:* guías de práctica clínica, manuales metodológicos, información para pacientes.
- *Tipo de texto:* textos científico-técnicos.

Otros recursos relacionados:

- Existen guías de práctica clínica en el Ministerio de Salud de Chile.[27]

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* los documentos más adecuados para crear un corpus son los textos completos en HTML. Los textos en PDF añaden dificultades de conversión a texto y revisión manual, y no presentan anotaciones ni informaciones adicionales. Tiene una

[27] http://diprece.minsal.cl/le-informamos/auge/acceso-guias-clinicas/

licencia muy flexible (Creative Commons Reconocimiento, CC-BY). Consideramos que se trata de un recurso de **madurez media**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	**	
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	HTML general-mente en ISO-8859-1.
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	*	

10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media.

Tabla 16: Madurez del recurso 14: Guías de práctica clínica (GPC) del portal Guía Salud.

Posibles aplicaciones del futuro recurso lingüístico

- Creación de corpus paralelo monolingüe/multilingüe de dominio médico.
- Recogida de textos de dominio médico.
- Extracción de terminología de dominio médico.
- Análisis de texto prescriptivo.

Recomendaciones: Conversión de los PDF a texto y, deseablemente, revisión de los mismos.

4.2.3.5 Recurso 15: Vídeos del portal web TV del Gobierno Vasco relacionados con el tema de Salud

Identificación del recurso.

- *Nombre:* Vídeos del portal web TV del Gobierno Vasco relacionados con el tema de Salud.
- *Clasificación por tipo de documento:* documentos multimedia sobre temas de salud.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* euskera, castellano.
- *Descripción del recurso:* El conjunto forma parte del portal de datos abiertos del Gobierno Vasco, IREKIA-Gobierno Abierto (https://www.irekia.euskadi.eus/es). El recurso seleccionado consiste en los vídeos de esta plataforma relacionados con el dominio de la salud (https://www.irekia.euskadi.eus/es/departments/85/videos#middle). En la fecha de realización de esta ficha el conjunto contenía 809 vídeos, en formato MP4, descargables uno a uno, no como colección. También están disponibles solo los audios formato MP3. En algunos

casos, junto con los datos multimedia se incluyen textos a modo de noticias relacionadas con el vídeo. Algunos de ellos están subtitulados automáticamente.

- *Fecha de comienzo de creación*: diciembre de 2009.
- *Fecha de finalización:* Se añaden vídeos de forma continuada.
- *Frecuencia de actualización*: variable (durante los últimos días se añade aproximadamente un vídeo nuevo por día).
- *Fecha de última actualización*: Diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - IREKIA: https://www.irekia.euskadi.eus/es
- *Tipo de licencia:* Creative Commons BY 4.0.[28] Esta licencia es la elegida para dar cobertura legal al principio del gobierno abierto por el que la ciudadanía podrá disponer de toda la información, material audiovisual y multimedia que se genere en este espacio. Así, siempre y cuando se cite la fuente y el autor, se permite su uso libre para cualquier fin.
    - *Descarga masiva disponible:* NO.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico:* Secretaría General de Presidencia del Gobierno Vasco. Contacto: https://www.irekia.euskadi.eus/es/departments/85/proposals#middle
- *Nombre organización:* Secretaría General de Presidencia del Gobierno Vasco.

Creación del recurso

- *Proveedor y/o creador:* Secretaría General de Presidencia del Gobierno Vasco.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:* 809 vídeos en la actualidad. Cubre noticias, declaraciones y actos relacionados con el área de salud del Gobierno Vasco.

[28] https://www.irekia.euskadi.eus/es/site/page/tos

- *Unidad (términos, entradas, textos, oraciones, otro):* vídeos y audios (en algunos casos acompañados de textos y en algunos otros de transcripciones automáticas).
- *Formato*: MP4 (vídeos), MP3 (audios), HTML/PDF (textos), HTML (transcripción).
- *Dominio:* salud.
- *Género:* declaraciones, entrevistas, intervenciones parlamentarias, actos.
- *Tipo de texto:* vídeos y audios.

Otros recursos relacionados:

- No se han identificado recursos relacionados.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* los vídeos se encuentran en formato MP4; los audios en MP3; los textos (cuando están disponibles) en HTML o PDF. Los documentos necesitan su conversión a texto, previa a su revisión manual. No es posible la descarga masiva de documentos, sino la descarga uno a uno, no como colección. Los vídeos/audios no están clasificados por idioma, así que sería necesario también determinar el idioma. Para convertirse en un recurso lingüístico útil sería primordial descargar todo el contenido, clasificarlo por idiomas, y transcribirlo manualmente, tomando quizás como base la transcripción automática (cuando está disponible). En algunos casos, junto con los datos multimedia se incluyen textos relacionados a modo de noticias que tienen que ver con el vídeo. Lo más reseñable de este conjunto es su licencia abierta Creative Commons Reconocimiento (CC-BY), que permite su reutilización sin apenas restricciones y la posibilidad de descargar los vídeos/audios, aunque no en bloque. Por otro lado, la falta de anotaciones fiables hace que la **madurez** del recurso sea **media**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	N/A	

2. Transcripción (ortográfica, fonológica, suprasegmental…)	*	Algunos vídeos están subtitulados automáticamente.
3. Alineación vídeo/sonido y texto	*	Algunos vídeos están subtitulados. automáticamente.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	-	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	N/A	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	**	No requiere.
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	No requiere autorización previa. Distribuido mediante licencia CC-BY.
		Madurez media.

Tabla 17: Madurez del recurso 15: Vídeos del portal web TV del Gobierno Vasco relacionados con el tema de salud.

Posibles aplicaciones del futuro recurso lingüístico

- Evaluación de tecnología de reconocimiento de voz (Speech-To-Text) para subtitulado/transcripción automática.
- El conjunto resulta demasiado pequeño y limitado en cuanto al número de locutores para ser empleado por sí solo para el entrenamiento de tecnología de reconocimiento de voz (Speech-To-Text), aunque podría utilizarse en combinación con otros conjuntos

(por ejemplo, con vídeos de otros temas de la misma fuente, que parece contener más de 10.000 vídeos).

Recomendaciones: Habilitar la descarga masiva. En relación a los documentos en PDF, sería deseable su conversión a texto y revisión de los mismos.

4.2.3.6 Recurso 16: Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS).

Identificación del recurso.

- *Nombre:* Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), Fichas técnicas de medicamentos y Prospectos para pacientes, publicados por el Centro de Información de Medicamentos (CIMA) dependiente de la AEMPS.
- *Clasificación por tipo de documento:* corpus textual.
- *Clasificación por número de lenguas:* monolingüe.
- *Lenguas:* español.
- *Descripción del recurso:* CIMA proporciona un buscador de medicamentos para consultar y descargar Fichas técnicas y prospectos de medicamentos. La AEMPS publica estudios de investigación clínico-farmacológica, información sobre fármacos, circulares, boletines informativos y trípticos divulgativos.
- *Fecha de comienzo de creación*: variable (véase apartado de Tamaño).
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*:
    - Fichas técnicas de medicamentos y prospectos: continua y variable.
    - Publicaciones, trípticos divulgativos, artículos y boletines: variable. El Boletín sobre fármacos de uso humanos se publica mensualmente; el Boletín sobre medicamentos veterinarios y el Boletín de productos cosméticos son trimestrales; el Boletín de Farmacovigilancia Veterinaria es anual.
- *Fecha de última actualización*: diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Fichas técnicas de medicamentos (dirigidas a profesionales) y prospectos (dirigidos a pacientes) accesibles en PDF a partir del buscador: https://cima.aemps.es/cima/publico/buscadoravanzado.html

    - Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS): www.aemps.gob.es/publicaciones/portada/home.htm

- *Tipo de licencia:* se autoriza la reproducción total o parcial de los contenidos de la web, siempre que se cite expresamente su origen. [29]
- *Descarga masiva disponible:* No, solo algunas publicaciones se pueden descargar una a una.

Persona de contacto u organización responsable:

- *Contacto:* Formulario disponible en: https://enviotelematico.aemps.es/enviotelematico/informacion.jsp
- *Nombre organización:* Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) (www.aemps.gob.es/home.htm).

Creación del recurso

- *Proveedor y/o creador:* Agencia Española de Medicamentos y Productos Sanitarios (AEMPS).

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal; los trípticos para usuarios presentan un estilo divulgativo.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
    - *Tamaño y cobertura:*
        - Fichas técnicas de medicamentos y prospectos: a fecha de la última versión del informe, contiene información sobre 14.478 medicamentos, 2.482 principios activos y 31.948 presentaciones (además de 311 biosimilares y 193 huérfanos).
        - Publicaciones, artículos, boletines y trípticos divulgativos: variable. Los Boletines sobre medicamentos de uso humano y farmacovigilancia veterinaria están disponibles desde 2007; los Boletines de productos cosméticos, desde 2018; otros informes y memorias de actividades, desde 2012 hasta 2018.
- *Unidad (términos, entradas, textos, oraciones, otro):* textos.
- *Formato*: las fichas técnicas y los prospectos, en PDF y HTML; las publicaciones, artículos, boletines y trípticos están en mayoritariamente en PDF; solo algunos también en HTML.

[29] *https://www.aemps.gob.es/avisoLegal/home.htm#derechos*

- *Dominio:* farmacología, productos sanitarios, epidemiología, medicina.
- *Género:* fichas técnicas, prospectos, boletines, informes científicos, folletos divulgativos.
- *Tipo de texto:* textos científico-técnicos.

Otros recursos relacionados:

  - Alertas farmacológicas (ver: www.aemps.gob.es/informa/alertas/home.htm): Alertas farmacéuticas y sanitarias sobre medicamentos de uso humano, de uso veterinario, productos sanitarios, y productos cosméticos y de higiene. Suelen estar en HTML.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* los documentos más adecuados para crear un corpus son los textos completos en HTML (fichas técnicas, prospectos, alertas sanitarias desde 2008, además de PDF). No resultaría difícil extraer textos utilizables para aplicaciones en farmacovigilancia. Los textos en PDF (folletos divulgativos o técnicos, boletines e informes) añaden dificultades de conversión de formato PDF a texto y revisión manual (véase tabla de madurez). A partir de estos aspectos, consideramos que globalmente gozan de **madurez media**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Las fichas técnicas, prospectos, alertas sanitarias y algunos informes están en HTML.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	*	Los archivos HTML suelen estar en ISO-8859-1.
5. Anotación morfológica y/o sintáctica	-	

6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	*	Las alertas sanitarias requerirían revisión.
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media.

Tabla 18: Madurez del recurso 16: Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS).

Posibles aplicaciones del futuro recurso lingüístico

- Recogida de corpus de textos de dominio médico-farmacológico. Los textos de las alertas sanitarias y medicamentosas son especialmente útiles para minería de textos y extracción de relaciones con aplicación a la farmacovigilancia.
- Extracción de terminología en el área de sanidad.

Recomendaciones: Habilitar la descarga masiva. Sería deseable la conversión a texto y revisión de los documentos en PDF.

4.2.3.7 Recurso 17: Nomenclátor de prescripción del Centro de Información de Medicamentos (CIMA) de la AEMPS

Identificación del recurso.

- *Nombre:* Nomenclátor de prescripción del Centro de Información de Medicamentos (CIMA).
- *Clasificación por tipo de documento:* banco de datos sobre productos sanitarios.
- *Clasificación por número de lenguas:* monolingüe.
- *Lenguas:* español.
- *Descripción del recurso:*
    - Banco de datos con información de productos sanitarios, anotada en XML: entre otros, nombre comercial, denominación común internacional (DCI), código de clasificación según la Clasificación Anatómica terapéutica y química (ATC), excipiente, presentación farmacéutica, unidades de dosis o vía de administración, e interacciones conocidas con otros fármacos.
- *Fecha de comienzo de creación*: variable (véase apdo. Tamaño).
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*:
    - Diaria y progresiva desde 2014.
- *Fecha de última actualización*: 27 de diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Nomenclátor de prescripción (CIMA): www.aemps.gob.es/cima/publico/nomenclator.html, http://listadomedicamentos.aemps.gob.es/prescripcion.zip
    - API (servicio REST): https://cima.aemps.es/cima/rest/[METODO] (resultados en formato JSON).
- *Tipo de licencia:* libre y gratuita citando la fuente: "cualquier utilización posible de todos o parte de los datos contenidos en el nomenclátor, deberá contener (...) la siguiente mención expresa a la autoría de la AEMPS sobre los mismos: Fuente de la información: Agencia Española de Medicamentos y Productos Sanitarios". [30]
- *Descarga masiva disponible:* Sí.

Persona de contacto u organización responsable:

[30] *http://listadomedicamentos.aemps.gob.es/Aviso_Legal_Nomenclator.pdf*

- *Contacto:* Formulario disponible en: https://enviotelematico.aemps.es/enviotelematico/informacion.jsp
- *Nombre organización:* Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) (www.aemps.gob.es/home.htm).

Creación del recurso

- *Proveedor y/o creador:* Centro de Información de Medicamentos (CIMA) de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS).

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal.
- *Niveles de anotación lingüística:* datos anotados.
- *Conforme a los estándares*: no respecto a los estándares de la comunidad PLN, pero conforme a las nomenclaturas de dominio (p. ej., códigos de la ATC, DCI, etc.).
    - *Tamaño y cobertura:*
        - Datos de más de 26000 medicamentos, incluyendo los comercializados desde 2014 hasta la actualidad, y fármacos suspendidos, revocados o fuera de comercialización desde mayo de 2013.
- *Unidad (términos, entradas, textos, oraciones, otro):* recoge más de 26000 entradas de información sobre fármacos.
- *Formato*:
    - Disponible en XML y XLS. La API (servicio REST) de CIMA presenta los datos en formato JSON.
- *Dominio:* farmacología, productos sanitarios, epidemiología, medicina.
- *Género:* banco de datos anotados.
- *Tipo de texto:* banco de datos anotados.

Otros recursos relacionados:

- Árbol de medicamentos: es la versión simplificada del Nomenclátor en formato XLS: http://listadomedicamentos.aemps.gob.es/Arbol_Medicamentos.zip
- Buscador de medicamentos del CIMA, con acceso a Fichas técnicas y Prospectos (recursos referenciados en la ficha anterior): https://cima.aemps.es/cima/publico/buscadoravanzado.html

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* El Nomenclátor de prescripción es un recurso completísimo, etiquetado en formato XML, que incluye anotada la información clave (p. ej., nombres comerciales de fármacos, principios activos o DCI, códigos de la clasificación Anatómica, Terapéutica y Química, ATC, dosis, excipientes, o interacciones conocidas con otros medicamentos), de la que es posible extraer información farmacológica. El CIMA ofrece una API y servicio REST para consultarlo. El Árbol de medicamentos se presenta en formato Excel (XLS), lo que permite obtener datos en formato tabular para su procesamiento posterior. Se trata de datos de **gran valor y potencial** para tareas de PLN en dominio sanitario. A partir de estos puntos, consideramos que presenta una madurez **alta**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	**	Hay una API y servicio REST para el Nomenclátor.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	El Nomenclátor en XML y UTF-8. El Árbol de medicamentos en formato XLS.
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	**	Permite extraer listas de entidades según las etiquetas XML.
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	**	Anotación de información farmacológica: códigos ATC, DCI, dosis, composición, excipientes e interacciones.

8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	**	El personal de la ANMPS mantiene y actualiza los datos del Nomenclátor.
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez alta.

Tabla 19: Madurez del recurso 17: Nomenclátor de prescripción del Centro de Información de Medicamentos (CIMA).

Posibles aplicaciones del futuro recurso lingüístico

- Listas de entidades nombradas (nombres de medicamentos, principios activos, clases farmacológicas, compañías farmacéuticas, etc.).
- Extracción de terminología farmacológica.

Recomendaciones: según nuestra opinión, es recurso de madurez alta que se puede tomar como modelo de diseño y distribución para otros conjuntos de datos analizados en el presente informe.

4.2.4 Justicia

4.2.4.1 Recurso 18: Textos de Jurisprudencia del CENDOJ

Identificación del recurso.

- *Nombre:* Textos de Jurisprudencia del CENDOJ (Centro de Documentación Judicial): Jurisprudencia del Tribunal Supremo (incluye Sala de lo Civil, de lo Penal, de lo Contencioso Administrativo, de lo Social, de lo Militar), Audiencia Nacional, Tribunal Superior de Justicia, Audiencias provinciales y Tribunales militares y unipersonales.

- *Clasificación por tipo de documento:* corpus textual.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* español, catalán, gallego y euskera.
- *Descripción del recurso:* autos (auto aclaratorio, recurso de casación, auto de admisión, auto de inadmisión), sentencias, sentencias de casación y acuerdos.
- *Fecha de comienzo de creación*: Histórico de sentencias desde 1979.
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*: diaria.
- *Fecha de última actualización*: diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Buscador de resoluciones/sentencias en formato PDF: http://www.poderjudicial.es/search/indexAN.jsp
    - Cada sentencia individual en formato PDF cuenta con su propio identificador. Ejemplo: http://www.poderjudicial.es/search/openDocument/acaaa913fbf710d7/20180709
- *Tipo de licencia:* Reutilización*:* Privada*.* Todos los derechos reservados*.* Consulta: Gratuita y libre.[31]
- *Descarga masiva disponible:* No, solo es posible obtener archivos separados por cada sentencia.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*: cendoj@poderjudicial.es
- *Nombre organización:* Centro de Documentación Judicial, Consejo General del Poder Judicial.

Creación del recurso

- *Proveedor y/o creador:* Centro de Documentación Judicial, Consejo General del Poder Judicial.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal, lenguaje jurídico.
- *Niveles de anotación lingüística:* datos no anotados.

31

www.poderjudicial.es/portal/site/cgpj/menuitem.65d2c4456b6ddb628e635fc1dc432ea0/?vgnextoid=47e7d17dc7a99210VgnVCM100000cb34e20aRCRD&vgnextfmt=default&vgnextlocale=es_ES

- *Conforme a los estándares*: no.
- *Tamaño y cobertura:*
    - El histórico de sentencias está disponible (en español) para la descarga desde 1979 hasta la actualidad. (> 6.468.208).
    - El histórico de sentencias en catalán aparece desde 1997 hasta la actualidad (>75.000).
    - El histórico de sentencias en gallego aparece desde 1996 hasta la actualidad (>12.500).
    - El histórico de sentencias en euskera aparece desde 2000 hasta mayo de 2018 (159).
- *Unidad (términos, entradas, textos, oraciones, otro):* textos.
- *Formato*: Los documentos de todas las sentencias se presentan en formato PDF.
- *Dominio:* legislación, jurisprudencia.
- *Género:* autos judiciales (auto aclaratorio, recurso de casación, auto de admisión, auto de inadmisión), sentencias, sentencias de casación y acuerdos.
- *Tipo de texto:* textos de ámbito jurídico.

Otros recursos relacionados:

- Resúmenes ("Abstracts"): hay como medio millón de resúmenes de sentencias.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* costaría bastante esfuerzo convertir este recurso en un RL, primero, porque su recuperación a través del buscador no es fácil, y segundo, porque su formato en PDF añade dificultades de conversión a texto y la necesidad de revisión manual posterior. Sería deseable su conversión a formatos fácilmente reutilizables como XML. Además, su reutilización tiene derechos restringidos, por lo que sería obligatoria la petición de un permiso para ello. Aun así, su interés es alto porque podría servirnos para crear un corpus textual de amplitud, con el que realizar otros RL. Posteriormente, podría también ser de interés para la creación de glosarios bilingües español-euskera o de otros recursos léxicos bilingües, pues contiene sentencias en otras lenguas como el catalán, gallego o euskera, que son fácilmente localizables a través del buscador. Sería recomendable alinear los textos para su reutilización en traducción automática. Consideramos, por todo ello, que posee una **madurez baja.**

Para facilitar su utilización sería deseable, además, que se permitiese la descarga masiva de los datos. También sería mucho más fácil su utilización si estuviese disponible la descarga de datos anonimizados

que posibilitaran su uso sin problemas relativos a la ley de protección de datos personales vigente.*Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico*

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	-	Necesidad de conversión de PFD a formato procesable.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	*	Recurso normalizado con el identificador europeo de jurisprudencia (ECLI)
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	Metadatos asociados en el buscador, que podrían recuperarse.

Aspectos legales		
12. Necesidad de anonimización de datos personales	-	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	-	Todos los derechos reservados para el CGPJ.
		Madurez media-alta.

Tabla 20: Madurez del recurso 18: Textos de Jurisprudencia del CENDOJ.

Posibles aplicaciones del futuro recurso lingüístico

- Recogida de corpus de textos de ámbito jurídico.
- Extracción de terminología jurídica.
- Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales).
- Creación de un corpus paralelo de sentencias que estén disponibles en español y otra lengua cooficial (observado con euskera).

Recomendaciones: Habilitar la descarga masiva, preferiblemente en formatos fácilmente reutilizables como XML. También sería de gran utilidad la publicación de conjuntos con marcado de anonimización (recurso anonimizado y su correspondiente sin anonimizar).

4.2.4.2 Recurso 19: Textos del Boletín Oficial del Estado (BOE) Diario.

Identificación del recurso.

- *Nombre:* Textos del Boletín Oficial del Estado (BOE) Diario
- *Clasificación por tipo de documento:* corpus textual.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* español; catalán, valenciano, gallego y euskera (suplementos específicos).
- *Descripción del recurso:*
  - Disposiciones generales de los órganos del Estado y los tratados o convenios internacionales.
  - Disposiciones generales de las Comunidades Autónomas.

  - Resoluciones y actos de los órganos constitucionales del Estado.
  - Disposiciones, resoluciones y actos de los departamentos ministeriales y otros órganos del estado y administraciones públicas.
  - Convocatorias, citaciones, requisitorias y anuncios dispuestos por ley o real decreto.
- *Fecha de comienzo de creación*:
  - *Colecciones históricas (Gazeta)*: De 1661 a 1959.
  - *Boletín Oficial del Estado*: desde 1 de enero de 1960 a la actualidad.
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*: diaria.
- *Fecha de última actualización*: enero de 2019.
- *Versión:* N/A.
- *Identificador del recurso:*
  - En portal de datos abiertos: http://datos.gob.es/es/catalogo/e04761001-boletin-oficial-del-estado-boe.
- *Tipo de licencia:* Consulta: Gratuita y libre. Reutilización: Permitida bajo condiciones: citación de autoría, fecha de actualización y conservación de metadatos. [32]
- *Descarga masiva disponible:* No. Archivos separados para cada disposición. No se puede descargar en conjunto ni siquiera el diario de forma completa.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*: Formulario de contacto: https://www.boe.es/legislacion/informacion/formulario_web.php. También en el 060.
- *Nombre organización:* Agencia Estatal Boletín Oficial del Estado.

Creación del recurso

- *Proveedor y/o creador:* Agencia Estatal Boletín Oficial del Estado.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal, lenguaje jurídico.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.

[32] *http://www.boe.es/sede_electronica/informacion/aviso_legal.php*

- *Tamaño y cobertura:*
    - El histórico de diarios oficiales está disponible (en español) para la descarga desde 1 de enero de 1960 hasta la actualidad.
- *Unidad (términos, entradas, textos, oraciones, otro):* textos.
- *Formato*: Los documentos se presentan en varios formatos: HTML, PDF, XML. Además, los PDF tienen varios tamaños, coincidiendo con la totalidad del diario, o bien, por cada una de sus secciones. El formato EPUB solo está disponible para la legislación consolidada.
- *Dominio:* legislación, jurisprudencia.
- *Género:* Legislación española: leyes, reales decretos, disposiciones generales, sentencias, actos, resoluciones.
- *Tipo de texto:* textos de ámbito jurídico.

Otros recursos relacionados:

- Buscador de diarios por calendario: https://www.boe.es/diario_boe/
- Cada diario tiene su propio identificador asociado a la fecha: https://www.boe.es/boe/dias/2018/07/10/
- Cada suplemento en diferentes idiomas, posee, además, su propio buscador: https://www.boe.es/diario_boe/calendarios.php?c=g; https://www.boe.es/diario_boe/calendarios.php?c=v; https://www.boe.es/diario_boe/calendarios.php?c=e; https://www.boe.es/buscar/suplementos_ca.php .

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* Este recurso presenta contenido bastante estandarizado, y se entrega bajo diferentes formatos. Su interés es alto, porque sería un buen punto de partida de diferentes RL. Por ejemplo, destacaríamos la creación de un corpus textual de amplitud, aunque su formato en PDF (individuales por cada diario, e incluso, secciones) añade dificultades de conversión a texto y una posterior revisión manual. Escollo suplido, en parte, por los formatos XML de los diarios, que también necesitaría una reconversión. Por otro lado, sería igualmente útil para la creación de corpus paralelos y memorias de traducción a partir de disposiciones en castellano con suplementos en otras lenguas cooficiales. Para ello, es preciso identificar aquellos suplementos en lenguas cooficiales que tengan correspondencia con disposiciones en castellano, y revisar

manualmente (los documentos no se encuentran hiperenlazados entre sí, o bajo alguna indización determinada). También sería necesario alinear los textos para permitir su utilización en aplicaciones de traducción automática. Habría que identificar cuáles de ellos son realmente auténticos y coincidentes para establecer esa alineación del corpus. Consideramos, por todo lo expuesto, que posee una **madurez media**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Necesidad de conversión de PFD a formato procesable en algunos de ellos. Según el tipo de recursos lingüísticos, sería necesaria un mayor esfuerzo.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	Para corpus paralelos sería necesaria una alineación entre textos de lenguas.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base	-	

de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)		
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	Metadatos asociados en el buscador, que podrían recuperarse.
Aspectos legales		
12. Necesidad de anonimización de datos personales	*	Posiblemente sería necesaria en casos donde aparezcan datos sensibles como nombres, apellidos o DNI.
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media.

Tabla 21: Madurez del recurso 19: Textos del Boletín Oficial del Estado (BOE) Diario.

Posibles aplicaciones del futuro recurso lingüístico

- Recogida de corpus de textos de ámbito jurídico.
- Extracción de terminología jurídica.
- Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales).
- Creación de un corpus paralelo de suplementos de disposiciones o sentencias que estén disponibles en español y otra lengua cooficial.
- Entrenamiento para recuperación y extracción de la información.

Recomendaciones: Habilitar descarga masiva e identificar pares traducidos (español-lengua cooficial).

4.2.4.3 Recurso 20: Textos de Códigos electrónicos del Boletín Oficial del Estado (BOE).

Identificación del recurso.

- *Nombre:* Textos de Códigos electrónicos del Boletín Oficial del Estado (BOE).
- *Clasificación por tipo de documento:* corpus textual.

- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* español; francés, inglés, italiano y alemán (solo Constitución Española); catalán, valenciano, gallego y euskera para versiones de la Constitución Española.
- *Descripción del recurso:* Compilaciones electrónicas de normas vigentes del ordenamiento jurídico, agrupadas bajo ramas del Derecho y otros ámbitos de competencia:
    - Constitución Española.
    - Derecho Constitucional.
    - Derecho Administrativo General.
    - Organización Administrativa.
    - Función Pública.
    - Seguridad Vial, Transporte y Telecomunicaciones.
    - Defensa y Seguridad.
    - Derecho Tributario.
    - Derecho Financiero.
    - Derecho Civil.
    - Derecho Penal.
    - Derecho Mercantil.
    - Sociedades Mercantiles.
    - Mercados, Entidades y Operaciones Financieros.
    - Auditoría y Contabilidad.
    - Legislación Social.
    - Derecho procesal.
    - Educación.
    - Sanidad y Farmacia.
    - Deporte.
    - Cultura.
    - Energía.
    - Agricultura y Alimentación.
    - Comunidades Autónomas.
    - Derecho Urbanístico.
    - Vivienda.
    - Medio Ambiente.
    - Otros: Código de aguas de la parte española de la demarcación hidrográfica del Miño Sil; Código de Aguas de la Cuenca del Duero; Código de Patronos.
- *Fecha de comienzo de creación*: Desde 1960 a la actualidad.
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*: Siempre que se aprueba un cambio en ellas.
- *Fecha de última actualización*: Diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - En portal de datos abiertos: http://datos.gob.es/es/catalogo/e04761001-codigos-y-otros-libros-electronicos.
    - Acceso general a los Códigos electrónicos: https://www.boe.es/legislacion/codigos/

- *Tipo de licencia:* Consulta: Gratuita y libre. Reutilización: Permitida bajo condiciones: citación de autoría, fecha de actualización y conservación de metadatos. [33]
- *Descarga masiva disponible:* No. Archivos separados para cada código electrónico consolidado. Ficheros individuales para cada lengua (si existe la traducción de ese código consolidado), y un fichero distinto para cada formato (PDF, EPUB).

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*: Formulario de contacto: https://www.boe.es/legislacion/informacion/formulario_web.php. También en el 060.
- *Nombre organización:* Agencia Estatal Boletín Oficial del Estado.

Creación del recurso

- *Proveedor y/o creador:* Agencia Estatal Boletín Oficial del Estado.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal, lenguaje jurídico.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:* En total existen unos 201 códigos hasta enero de 2019.
- *Unidad (términos, entradas, textos, oraciones, otro):* textos.
- *Formato*: Los documentos se presentan en varios formatos: HTML, PDF, EPUB y XML.
- *Dominio:* legislación, jurisprudencia.
- *Género:* Legislación española: leyes, normas y códigos legislativos.
- *Tipo de texto:* textos de ámbito jurídico.

Otros recursos relacionados:

- Relacionado con el recurso 21 (Textos sobre Legislación del Boletín Oficial del Estado, BOE).

Grado de madurez de los datos conforme al modelo

Necesidades de procesamiento (manual o automático): este recurso posee una **madurez media**, ya que su contenido está bastante estandarizado, y se muestra bajo diferentes formatos (y principalmente, en EPUB), y algunos de ellos en distintas lenguas, lo que haría mucho más fácil su conversión a RL. El

[33] http://www.boe.es/sede_electronica/informacion/aviso_legal.php

formato en PDF añade dificultades de conversión a texto y la necesidad de una posterior revisión manual. Su interés es alto, porque sería un buen punto de partida para la creación de un corpus paralelo de legislación que esté disponible en español y otra lengua cooficial, así como la creación de memorias de traducción (formato TMX) o de recursos léxicos para diferentes lenguas. No obstante, habría primero que identificar qué recursos están traducidos, puesto que solo la Constitución Española está plenamente traducida a varios idiomas, y realizar algunas conversiones en los formatos de los ficheros para una mejor reutilización. Los códigos traducidos en varias lenguas requieren revisión manual. Habría que identificar cuáles son realmente auténticos y coincidentes. En esos casos sería recomendable realizar la alineación de los textos para facilitar su utilización en traducción automática. Igualmente, es necesaria la conversión a formatos compatibles y el uso de programas para la creación de memorias de traducción.*Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico*

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Los documentos se encuentran en formato EPUB.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	N/A	Para corpus paralelos sería necesaria una alineación entre textos de lenguas.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	

9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media.

Tabla 22: Madurez del recurso 20: Textos de códigos electrónicos del Boletín Oficial del Estado (BOE).

Posibles aplicaciones del futuro recurso lingüístico

- Recogida de corpus de textos de ámbito jurídico.
- Extracción de terminología jurídica.
- Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales de España y de Europa).
- Creación de memorias de traducción entre lenguas cooficiales y castellano, o bien español y otras lenguas de la Unión Europea (francés, inglés, alemán e italiano).
- Creación de un corpus paralelo de suplementos de disposiciones o sentencias que estén disponibles en español y otra lengua cooficial.
- Entrenamiento para recuperación y extracción de la información.

Recomendaciones: Identificar pares de documentos traducidos (archivo original – traducción) y sería recomendable realizar alineación de los textos para facilitar su utilización en traducción automática, así como la conversión a formatos reutilizables.

4.2.4.4 Recurso 21: Textos sobre Legislación del Boletín Oficial del Estado (BOE)

Identificación del recurso.

- *Nombre:* Textos sobre Legislación del Boletín Oficial del Estado (BOE).

- *Clasificación por tipo de documento:* corpus textual.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* español; otras lenguas cooficiales si existe la disposición traducida en alguno de los suplementos creados para cada lengua.
- *Descripción del recurso:*
    - Normativa estatal publicada en el Boletín Oficial del Estado y disposiciones anteriores a 1960 que siguen vigentes.
    - Normas con rango de ley de las Comunidades Autónomas.
    - Sentencias del Tribunal Constitucional sobre procedimientos de inconstitucionalidad y conflictos de competencia.
    - Normativa europea vigente: reglamentos, directivas, decisiones y recomendaciones que afectan a España publicadas en el DOUE.
    - Sentencias del Tribunal de Justicia de la Unión Europea y del Tribunal General sobre normativa comunitaria.
- *Fecha de comienzo de creación*:
    - Normativa estatal publicada en el Boletín Oficial del Estado desde 1960 hasta la actualidad.
    - Normas con rango de ley de las Comunidades Autónomas desde 1980 a la actualidad.
    - Normativa europea vigente desde 1982 a la actualidad.
- *Fecha de finalización:* mantenimiento y actualización progresiva de datos.
- *Frecuencia de actualización*: Diaria.
- *Fecha de última actualización*: enero de 2019.
- *Versión:* N/A.
- *Identificador del recurso:*
    - En portal de datos abiertos: http://datos.gob.es/es/catalogo/e04761001-busqueda-de-legislacion.
    - Buscador de legislación: https://www.boe.es/legislacion/legislacion.php
- *Tipo de licencia:* Consulta: Gratuita y libre. Reutilización: Permitida bajo condiciones: citación de autoría, fecha de actualización y conservación de metadatos.[34]
- *Descarga masiva disponible:* No, se accede mediante un buscador (HTML) a los distintos datos. Ficheros individuales en distintos formatos.

[34] *http://www.boe.es/sede_electronica/informacion/aviso_legal.php*

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*: Formulario de contacto: https://www.boe.es/legislacion/informacion/formulario_web.php. También en el 060.
- *Nombre organización:* Agencia Estatal Boletín Oficial del Estado.

Creación del recurso

- *Proveedor y/o creador:* Agencia Estatal Boletín Oficial del Estado.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal, lenguaje jurídico.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:* En total existen unos 210.953 registros hasta enero de 2019. Se incluyen en este resultado:
    - Normativa estatal publicada en el Boletín Oficial del Estado desde 1960 hasta la actualidad (133.773).
    - Normas con rango de ley de las Comunidades Autónomas desde 1980 a la actualidad (8.782).
    - Normativa europea vigente desde 1982 a la actualidad. (68.394)
- *Unidad (términos, entradas, textos, oraciones, otro):* textos.
- *Formato*: Los documentos se presentan en varios formatos: HTML, PDF, XML y EPUB (aunque el formato EPUB no está disponible para todas las disposiciones).
- *Dominio:* legislación, jurisprudencia.
- *Género:* Legislación española: leyes, reales decretos, disposiciones generales, sentencias, actos, resoluciones.
- *Tipo de texto:* textos de ámbito jurídico.

Otros recursos relacionados:

- Memorias de traducción de directivas y reglamentos europeos del Instituto Vasco de Administración Pública (IVAP) (ver: http://opendata.euskadi.eus/catalogo/-/memorias-de-traduccion-de-directivas-y-reglamentos-europeos/): Memoria de traducción de directivas y reglamentos europeos.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* el contenido está muy estandarizado, y se muestra bajo diferentes formatos (XML, y principalmente, en EPUB), lo que haría mucho más fácil su conversión a RL. Aunque necesitaría, ya desde el principio, de procesamiento de formatos para algunos de sus archivos en otras lenguas cooficiales (solo disponibles en PDF), una revisión manual posterior, así como la inclusión de metadatos. Su interés es alto, porque sería un buen punto de partida para la creación de un corpus textual de amplitud, de corpus paralelos de legislación consolidada que esté disponible en español y otra lengua cooficial, así como el desarrollo de memorias de traducción en formato TMX o de recursos léxicos (jerarquizados o no, como ontologías) para diferentes lenguas. Para ello, es preciso identificar aquellas disposiciones en lenguas cooficiales que tengan correspondencia con disposiciones en castellano y establecer una revisión manual del proceso. Habría que identificar cuáles son auténticos y coincidentes para alinear los textos, así como hacer una conversión de PDF a texto. Para crear memorias de traducción, sería necesaria su conversión a formatos compatibles y su procesamiento con ayuda de programas específicos para el formato TMX. Hay que indicar que no toda la legislación se encuentra en formatos compatibles como EPUB, principalmente, las disposiciones en otras lenguas, por lo que sería necesaria su conversión y revisión, así como anotación según lo dispuesto por las convenciones de PLN para cada tipo de RL final. Este recurso posee una **madurez media.**

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	*	Necesidad de conversión de PFD a formato procesable en algunos de ellos.
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	

3. Alineación vídeo/sonido y texto	N/A	Para corpus paralelos sería necesaria una alineación entre textos.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	**	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	-	
11. Presencia de metadatos	-	Metadatos asociados en el buscador, que podrían recuperarse.
Aspectos legales		
12. Necesidad de anonimización de datos personales	*	Posiblemente sería necesaria en casos donde aparezcan datos sensibles como nombres, apellidos o DNI.
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez media.

Tabla 23: Madurez del recurso 21: Textos de Legislación del Boletín Oficial del Estado (BOE).

Posibles aplicaciones del futuro recurso lingüístico

- Recogida de corpus de textos de ámbito jurídico.

- Extracción de terminología jurídica.
- Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales).
- Creación de memorias de traducción entre lenguas cooficiales y castellano.
- Creación de un corpus paralelo de legislación en español y otra lengua cooficial.
- Entrenamiento para recuperación y extracción de información, y de modelos de traducción automática.

Recomendaciones: Conversión de aquellos archivos en lenguas cooficiales que solo están disponibles en PDF a formato texto, así como identificación de los pares de documentos traducidos y alineación de textos.

4.2.4.5 Recurso 22: Memorias de traducción que contienen las publicaciones en el Boletín Oficial del Estado realizadas en euskera del Instituto Vasco de Administración Pública (IVAP)

Identificación del recurso.

- *Nombre:* Memorias de traducción que contienen las publicaciones en el Boletín Oficial del Estado realizadas en euskera del Instituto Vasco de Administración Pública (IVAP).
- *Clasificación por tipo de documento:* Memoria de traducción.
- *Clasificación por número de lenguas:* bilingüe.
- *Lenguas:* español y euskera.
- *Descripción del recurso:* Memoria de traducción de lo publicado en euskera en el Boletín Oficial del Estado con sumarios y legislación estatal traducidos por el Servicio Oficial de traductores.
- *Fecha de comienzo de creación*: 20/09/2016.
- *Fecha de finalización:* 31/12/2017, hasta el momento.
- *Frecuencia de actualización*: Anual.
- *Fecha de última actualización*: 02/01/2018.
- *Versión:* N/A.
- *Identificador del recurso:*
  - En el portal de datos de Euskadi: http://opendata.euskadi.eus/contenidos/ds_recursos_linguisticos/memo_boe/opendata/boe.tmx

  - En el portal datos.gob.es: http://datos.gob.es/es/catalogo/a16003011-memorias-de-traduccion-que-contienen-las-publicaciones-en-el-boletin-oficial-del-estado-realizadas-en-euskera

- *Tipo de licencia:* Creative Commons 4.0 Reconocimiento (CC BY 4.0).
- *Descarga masiva disponible:* Sí.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*: Buzón de consultas sede electrónica Euskadi: https://www.euskadi.eus/y22-izapide/es/x43kToolkitWar/form/fdp?procedureId=1013901&tipoPresentacion=1&language=es&formDataPreload=%7b%22idIniciativa%22:%22105%22%7d
- *Nombre organización:* IVAP - Herri Arduralaritzaren Euskal Erakundea, IZO (Servicio Oficial de Traductores).

Creación del recurso

- *Proveedor y/o creador:* Comunidad Autónoma del País Vasco.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal, lenguaje jurídico.
- *Niveles de anotación lingüística:* no anotado.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura:* Fichero de 115.59 MB, que contiene datos desde 2011 a 2017.
- *Unidad (términos, entradas, textos, oraciones, otro):* unidades de traducción con segmentos equivalentes entre idiomas.
- *Formato*: TMX (Translation Memory eXchange).
- *Dominio:* legislación, jurisprudencia.
- *Género:* Legislación española: leyes, normas, códigos legislativos.
- *Tipo de texto:* textos de ámbito jurídico.

Otros recursos relacionados:

- Relacionado con el recurso 20 (Textos de códigos electrónicos del Boletín Oficial del Estado) .

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* se trata de un recurso lingüístico ya creado, por lo que no se necesitaría ningún tipo de procesamiento, salvo para la creación de otros recursos dependientes de él, o nuevas memorias de traducción para otros pares de lenguas. Su interés es alto, pues es un recurso libre de derechos. Sin embargo, quizá sería necesaria la anonimización de datos personales, la verificación del correcto alineamiento de textos o de traducción, o un tratamiento determinado para su conversión en otro tipo de RL (etiquetado, anotación, creación de ontología con extracción de parte de la terminología, léxicos bilingües, entre otros). Posee una **madurez alta**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	**	
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	**	Alineación de texto.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	N/A	
6. Anotación de entidades nombradas	N/A	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	N/A	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	*	

10. Anotación conforme a estándares de la comunidad PLN	**	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	N/A	
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez alta.

Tabla 24: Madurez del recurso 22: Memorias de traducción que contienen las publicaciones en el Boletín Oficial del Estado realizadas en euskera del Instituto Vasco de Administración Pública (IVAP).

Posibles aplicaciones del futuro recurso lingüístico

- Extracción de terminología jurídica.
- Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales de España).
- Creación de memorias de traducción entre otros pares de lenguas.
- Entrenamiento de modelos de traducción automática.

Recomendaciones: Anonimización de datos personales.

4.2.4.6 Recurso 23: Memorias públicas de traducción de la Diputación Foral de Gipuzkoa.

Identificación del recurso.

- *Nombre:* Memorias públicas de traducción de la Diputación Foral de Gipuzkoa.
- *Clasificación por tipo de documento:* Memoria de traducción.
- *Clasificación por número de lenguas:* bilingüe.
- *Lenguas:* español y euskera.
- *Descripción del recurso:* Memorias de traducción extraídas de la base de datos de traducciones de la Diputación Foral de Gipuzkoa, ordenados por diferentes materias: administración; economía y empresa; asuntos sociales; medio ambiente; deportes; cultura; ordenación del territorio; agricultura; política; aguas.

- *Fecha de comienzo de creación*: 03/02/2015.
- *Fecha de finalización:* 01/07/2018.
- *Frecuencia de actualización*: Bimensual.
- *Fecha de última actualización*: Diciembre de 2018.
- *Versión:* 1.7.
- *Identificador del recurso:* a8ea037f-7835-4299-9482-b42d1c53a062
    - En el portal de datos.gob.es: https://datos.gob.es/es/catalogo/l02000020-memorias-publicas-de-traduccion-de-la-diputacion-foral-de-gipuzkoa4
    - En portal de datos abiertos de la Diputación Foral: http://www.gipuzkoairekia.eus/es/datu-irekien-katalogoa/-/openDataSearcher/detail/detailView/a8ea037f-7835-4299-9482-b42d1c53a062
- *Tipo de licencia:* Creative Commons 4.0 Reconocimiento (CC BY 4.0) Aviso legal complementario en PDF (http://api.gipuzkoairekia.eus/dataset/recurso/cce3c0c3-ac63-4616-aefa-efc2148e1e08/descargar)
- *Descarga masiva disponible:* Sí.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico*:
    - Formulario de contacto para datos abiertos: http://www.gipuzkoairekia.eus/es/datu-irekien-eskaera
    - Email soporte: euskara@gipuzkoa.eus
- *Nombre organización:* Diputación Foral de Gipuzkoa.

Creación del recurso

- *Proveedor y/o creador:* Gipuzkoako Foru Aldundiaren Euskara Zuzendaritza Nagusia / Dirección General de Euskera de la Diputación Foral de Gipuzkoa.

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* estándar, formal, lenguaje específico o técnico según el campo de conocimiento.
- *Niveles de anotación lingüística:* no anotado.
- *Conforme a los estándares*: no.

- *Tamaño y cobertura:* 10 ficheros TMX de distinto tamaño, uno para cada área de trabajo (administración; economía y empresa; asuntos sociales; medio ambiente; deportes; cultura; ordenación del territorio; agricultura; política; aguas), con datos de 2015 hasta la actualidad.
- *Unidad (términos, entradas, textos, oraciones, otro):* unidades de traducción con segmentos equivalentes entre idiomas.
- *Formato*: TMX (Translation Memory eXchange).
- *Dominio:* sector público, legislación, administración; economía y empresa; asuntos sociales; medio ambiente; deportes; cultura; ordenación del territorio; agricultura; política; aguas.
- *Género:* Legislación, textos explicativos, textos técnicos.
- *Tipo de texto:* textos de ámbito técnico y jurídico.

Otros recursos relacionados:

- No se han identificado recursos relacionados.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* recurso lingüístico ya creado, por lo que no se necesitaría ningún tipo de procesamiento, salvo para la creación de otros recursos dependientes de él, o nuevas memorias de traducción para otros pares de lenguas. Su interés es alto, pues posee diferentes campos de estudio, es un recurso libre de derechos, y está constantemente siendo actualizado. Sin embargo, quizá sería necesaria la verificación del correcto alineamiento de textos o de traducción, o un tratamiento determinado para su conversión en otro tipo de RL (etiquetado, anotación, creación de ontología con extracción de parte de la terminología, entre otros). Este recurso posee una **madurez alta**; y podríamos afirmar que se trata ya de un RL puro.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		

1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	**	
2. Transcripción (ortográfica, fonológica, suprasegmental…)	N/A	
3. Alineación vídeo/sonido y texto	**	Alineación de texto en dos lenguas ya realizada.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	**	
5. Anotación morfológica y/o sintáctica	N/A	
6. Anotación de entidades nombradas	N/A	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	N/A	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	**	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	N/A	
10. Anotación conforme a estándares de la comunidad PLN	N/A	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	**	Aunque nos comentan que están anonimizados, sería necesaria una revisión.
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	**	
		Madurez alta.

Tabla 25: Madurez del recurso 23: Memorias públicas de traducción de la Diputación Foral de Gipuzkoa.

Posibles aplicaciones del futuro recurso lingüístico

- Extracción de terminología precisa sobre alguno de los campos de estudio.
- Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales de España, u otras lenguas).
- Creación de memorias de traducción entre otros pares de lenguas.
- Entrenamiento de modelos de traducción automática.

Recomendaciones: No se han identificado recomendaciones. Ya dispone de madurez alta.

4.2.4.7 Recurso 24: Grabaciones de Vistas Judiciales del Consejo General del Poder Judicial

Identificación del recurso.

- *Nombre:* Grabaciones de Vistas Judiciales del Consejo General del Poder Judicial.
- *Clasificación por tipo de documento:* documentos multimedia (audios y vídeos) incluyendo grabaciones de las vistas y comparecencias orales públicas y otras actuaciones orales públicas del poder judicial.
- *Clasificación por número de lenguas:* multilingüe.
- *Lenguas:* castellano y lenguas cooficiales (catalán, gallego, euskera).
- *Descripción del recurso:* Conjunto de documentos multimedia (audios y vídeos) de grabaciones de las vistas orales públicas y otras actuaciones orales públicas del poder judicial. Desde el año 2000 la Ley 1/2000, de 7 de enero, de Enjuiciamiento Civil establece que *la documentación de las actuaciones podrá llevarse a cabo, no solo mediante actas, notas y diligencias, sino también con los medios técnicos que reúnan las garantías de integridad y autenticidad. Y las* vistas *y comparecencias orales habrán de registrarse o grabarse en soportes aptos para la reproducción*. También la Ley Orgánica del Poder Judicial de 1994 establecía que las Administraciones *Promoverán el empleo de los medios técnicos, audiovisuales e informáticos de documentación con que cuente la unidad donde prestan sus servicios*, y atribuye funciones de auxilio en esta materia al Cuerpo de Auxilio Judicial. Por tanto, desde hace años, los juzgados de toda España están grabando vistas y comparecencias orales, que deben archivarse. Aunque es difícil estimar el tamaño de los datos, para hacernos una idea, consultamos el portal de estadísticas judiciales[35] y, a modo de ejemplo, en 2017 y solo en la

[35] www.poderjudicial.es/cgpj/es/Temas/Estadistica-Judicial

Comunidad de Madrid se dictaron más de 200.000 sentencias y más de 400.000 autos.[36] Muy posiblemente, muchas no requiriesen ninguna vista ni comparecencia oral, pero solo con que un pequeño porcentaje lo requiriesen el recurso resulta muy amplio.

- *Fecha de comienzo de creación*: Indeterminada, aunque dada la legislación podría empezar a haber grabaciones desde el año 1995 aproximadamente.
- *Fecha de finalización:* Se siguen realizando grabaciones de vistas y comparecencias orales probablemente todos los días.
- *Frecuencia de actualización*: probablemente se realicen grabaciones todos los días.
- *Fecha de última actualización*: diciembre de 2018.
- *Versión:* N/A.
- *Identificador del recurso:*
    - Los datos no están accesibles *públicamente*. La Junta de Andalucía tiene ya un sistema de descarga de las grabaciones de las vistas judiciales, pero solo para profesionales de la justicia autorizados (https://sede.justicia.juntadeandalucia.es/portal/adriano/es/tramites-y-servicios/Descarga-de-Vistas-Judiciales/).
- *Tipo de licencia*: Los datos ni siquiera están accesibles públicamente.
- *Descarga masiva disponible:* No.

Persona de contacto u organización responsable:

- *Nombre y correo electrónico:* Consejo General del Poder Judicial (CGPJ), Audiencia Nacional o Audiencias Provinciales. Contacto mediante formulario web: www.poderjudicial.es/portal/site/cgpj/menuitem.178d5a0b03ac7b23737c9ad3dc432ea0/?vgnextoid=d6c0e0f71b83a210VgnVCM100000cb34e20aRCRD&vgnextfmt=default&vgnextlocale=es_ES
- *Nombre organización:* Consejo General del Poder Judicial (CGPJ), Audiencia Nacional o Audiencias Provinciales.

Creación del recurso

- *Proveedor y/o creador:* Consejo General del Poder Judicial (CGPJ) Audiencia Nacional o Audiencias Provinciales.

[36] *www.poderjudicial.es/stfls/CGPJ/ESTADÍSTICA/RESÚMENES%20ESTADÍSTICOS/FICHERO/CA-2017-13-00-MADRID-T1T4_1.0.0.pdf*

Descripción del recurso

- *Variedad de la lengua (estándar, dialecto, argot, otro):* lenguaje jurídico, pero también dialectos y lenguaje más informal contenido en las declaraciones.
- *Niveles de anotación lingüística:* datos no anotados.
- *Conforme a los estándares*: no.
- *Tamaño y cobertura*: Es difícil hacer una estimación del tamaño del recurso. Para hacernos una idea, en un año en Madrid se dictan más de 200.000 sentencias y 400.000 autos, aunque no todas ellas requieran vistas ni comparecencias orales. En cualquier caso, la cobertura es de todas las comunidades autónomas y provincias. También incluye casos de derecho civil, penal, contencioso-administrativo y social.
- *Unidad (términos, entradas, textos, oraciones, otro):* audios y vídeos.
- *Formato*: N/A (los datos no son públicos).
- *Dominio:* Todas las ramas del derecho (civil, penal, contencioso-administrativo y social).
- *Género:* vistas y comparecencias orales en tribunales de justicia.
- *Tipo de texto:* vistas y comparecencias orales en tribunales dejusticia.

Otros recursos relacionados:

- No se han identificado recursos relacionados.

Grado de madurez de los datos conforme al modelo

- *Necesidades de procesamiento (manual o automático):* actualmente no es posible acceder a estos datos, por lo que no es fácil evaluar su madurez. Lo más probable es que los datos no estén etiquetados de ninguna forma (más allá de la documentación del caso, fecha e intervinientes). Por ello, sería necesario, en primer lugar, establecer un modo de acceso a los datos. Probablemente se pueda conseguir acceso solicitándolo al Consejo General del Poder Judicial (CGPJ)[37] o al Centro de Documentación Judicial (CENDOJ), o a la Audiencia Nacional. Algunas audiencias provinciales tienen ya sistemas de descarga de las grabaciones de las vistas judiciales, pero solo para profesionales de la justicia autorizados, por ejemplo, la de Andalucía. Posteriormente, sería necesario crear de cero la transcripción y el alineamiento de frases. También resultaría muy interesante etiquetar los locutores que aparecen en los vídeos para

[37] www.poderjudicial.es

sistemas de reconocimiento y segmentación de locutores. El grado de **madurez** del recurso es muy **bajo**.

Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico

Puntos que considerar para evaluar el grado de madurez	**Valores: -, *, **, N/A (Desconocido o No aplica)**	**Observaciones**
Aspectos técnicos (necesidad procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…)	N/A	
2. Transcripción (ortográfica, fonológica, suprasegmental…)	-	
3. Alineación vídeo/sonido y texto	-	
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1)	-	
5. Anotación morfológica y/o sintáctica	-	
6. Anotación de entidades nombradas	-	
7. Otros tipos de anotación (semántica, pragmática, palabras clave…)	-	
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…)	-	
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran revisor experto)	-	
10. Anotación conforme a estándares de la comunidad PLN	N/A	
11. Presencia de metadatos	-	
Aspectos legales		
12. Necesidad de anonimización de datos personales	-	Seguramente sí.
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…)	-	Actualmente no están disponibles en abierto.
		Madurez muy baja

Tabla 26: Madurez del recurso 24: Grabaciones de Vistas Judiciales del Consejo General del Poder Judicial.

Posibles aplicaciones del futuro recurso lingüístico

- Entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-To-Text) para facilitar la búsqueda y la recuperación de información en audio y vídeo de vistas y comparecencias orales en juicios. Dado que cada vez es mayor la cantidad de estos contenidos multimedia, y que el análisis y acceso a los mismos es muy costoso en tiempo, es necesario acelerar el acceso a estos datos por parte de abogados y jueces.
- Entrenamiento y evaluación de sistemas de reconocimiento y segmentación de locutores para facilitar la búsqueda y recuperación de información en audio y vídeo de vistas y comparecencias orales en juicios, por los mismos motivos que el punto anterior.

Recomendaciones: Habilitar el acceso y descarga masiva.

5 CONCLUSIONES PRELIMINARES SOBRE LOS CONJUNTOS DE DATOS ANALIZADOS

La **mayoría de los conjuntos de datos analizados** en este informe se quedan en los estadios de **madurez baja** o **media**. Esto es comprensible y esperable, dado que los requisitos para ser considerado un recurso *maduro* son muy estrictos, en el sentido de que solo los recursos ya procesados y en formatos directamente usables por los investigadores de PLN (por ejemplo, XML o TMX) pueden ser considerados propiamente RL.

Posibles aplicaciones TL	Recomendaciones al Organismo
1. Patentes de la Oficina Española de Patentes y Marcas (OEPM)	
- Recogida de corpus de textos de dominio industrial. - Extracción de terminología científico-técnica. - La Clasificación Internacional de Patentes (CIP) permite construir un corpus paralelo inglés/español de terminología industrial de utilidad para traducción automática.	- Construir un corpus paralelo inglés/español de terminología industrial, de modo análogo al Corpus of Parallel Patent Applications (COPPA) existente para inglés/francés.
2. Patentes multilingües digitalizadas en PATSTAT de European Patent Office (EPO)	
- Recogida de corpus de textos de dominio industrial.	- Facilitar acceso a investigadores con cuotas más accesibles.

- Extracción de terminología científico-técnica. - Construcción de un corpus paralelo multilingüe de terminología industrial. - Obtención de metadatos y extracción de relaciones entre documentos de dominio industrial. - Entrenamiento de clasificadores. - Creación de recursos lingüísticos de variedades no peninsulares a partir de las patentes iberoamericanas existentes (véase Otros recursos relacionados)	- Ampliar la cobertura hispanoamericana de patentes. - Crear un corpus paralelo del español, con variantes iberoamericanas del español y los pares de lenguas traducidos al español, semejante al Corpus of Parallel Patent Applications (COPPA). - Entrenamiento de clasificadores de citas para ver relaciones entre documentos.
3. Diccionarios terminológicos del TERMCAT	
- Extracción de terminología multilingüe de dominio especializado. - Extracción de información categorial y morfológica para enriquecer diccionarios y recursos léxicos de dominio especializado. - Modelo para desarrollar un recurso semejante en español ("TERMESP"). Existe la iniciativa TERMINESP, [38] aunque su grado de desarrollo todavía no es comparable.	- Crear un análogo de TERMCAT para el español ("TERMESP").
4. Padrón: Relación de nombres de personas del Instituto Nacional de Estadística	
- Creación de recursos léxicos bilingües, en su relación con la antroponimia de otras comunidades autónomas con lengua cooficial. - _Uso en etiquetadores automáticos (p. ej., morfológicos o semánticos) o sistemas de reconocimiento y anotación de entidades.	- Sería primordial una conversión a formatos estandarizados en PLN como JSON, TXT o XML, más fácilmente reutilizables e interoperables. Los topónimos de nombres propios de personas son de gran interés para reconocedores de entidades nombradas (NER por sus siglas en inglés).
5. Topónimos del Instituto Geográfico Nacional (IGN)	
- Creación de recursos léxicos bilingües, en su relación con la toponimia de otras comunidades autónomas con lengua cooficial. - Uso en etiquetadores automáticos (p. ej., morfológicos o semánticos) o en sistemas de reconocimiento y anotación de entidades nombradas.	- Sería primordial una conversión a formatos más estandarizados en PLN como JSON, TXT o XML, más fácilmente reutilizables e interoperables.
6. Grabaciones de vídeo de RTVE a la carta	

[38] *http://www.wikilengua.org/index.php/Wikilengua:Terminesp*

- Entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-To-Text) para subtitulado automático. - Entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-To-Text) para facilitar la búsqueda y recuperación de información en formato vídeo (por ejemplo, en el archivo de RTVE y RTVE A la Carta). - Entrenamiento y evaluación de sistemas de reconocimiento de locutores y segmentación de locutores para facilitar la búsqueda y recuperación de información en formato vídeo (por ejemplo en el archivo de RTVE y RTVE A la Carta). - Entrenamiento y evaluación de sistemas de traducción de voz a lengua de signos (empleando para ello los programas con traducción a lengua de signos). - Análisis de la evolución del lenguaje castellano estándar a lo largo del tiempo (empleando para ello el archivo histórico de RTVE).	- Permitir el acceso y descarga a algunos vídeos y los subtítulos, así como los archivos sonoros y sus transcripciones.. Sería de gran utilidad para las tecnologías del lenguaje disponer de los script con las transcripciones exactas. También resultaría muy interesante etiquetar los locutores que aparecen en los vídeos para sistemas de reconocimiento y segmentación de locutores. Asimismo, permitir descargar secciones de los territoriales, para desarrollar recursos en lenguas cooficiales.
7. Audios y vídeos del Archivo Audiovisual del Congreso de los Diputados de España	
- Base de datos de entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-to-Text) que posteriormente permitan facilitar la transcripción automática o la búsqueda en este tipo de contenidos. - Base de datos de entrenamiento y evaluación de sistemas de identificación del hablante.	- Sería de gran utilidad que se habilitara la posibilidad de descarga masiva de secciones de los audios y vídeos del Archivo Audiovisual del Congreso de los Diputados de España.
8. Índices de clasificación de los catálogos de la BNE	
- Sus datos de materias y submaterias se pueden reutilizar para reconocedores de entidades nombradas y clasificadores de entidades, materias, títulos, etc.	- Enriquecer con anotación lingüística y emplear para entrenamiento de clasificadores.
9. Publicaciones periódicas digitalizadas de la Hemeroteca Digital	
- Creación de un recurso histórico general tipo Google Books para estudios lingüísticos, históricos y culturales de la evolución de la prensa en España y otros países relacionados históricamente con nuestro país.	- Convertir de formato PDF a texto mediante OCR y revisión de la transcripción. También sería de utilidad enriquecerlo con anotación lingüística.

- Generación de lexicones de especialidad, de materia, por ámbito geográfico, por tipo de publicación, etc. - Generación de modelos de lenguaje por tipos de publicaciones, ámbitos geográficos, lingüísticos, etc. - Entrenamiento de clasificadores. - Extracción de entidades nombradas en la prensa histórica.	
10. Documentos digitalizados de la Biblioteca Digital Hispánica	
- Base para crear un repositorio como Google Books, Google n-gram viewer, y generar modelos word2vec o similares. - Entrenamiento de clasificadores. - Creación de modelos de lenguaje diacrónicos y de variantes del español, por ejemplo, como mecanismo para mejorar los resultados de sistemas basados en OCR para la digitalización de documentos históricos.	- Revisión de los textos procesados con OCR para mejorar su calidad. Serían de gran utilidad para la mejora de sistemas OCR para textos históricos.
11. Publicaciones en repositorio SciELO	
- Textos: extracción de terminología de dominio sanitario; recogida de corpus de textos de dominio médico. - Vídeos: entrenamiento de sistemas de reconocimiento de voz (Speech-To-Text) en el dominio sanitario, para aplicaciones de detección de palabras clave (Key-word Spotting), subtitulado automático, búsqueda y recuperación de información, o reconocimiento y segmentación de locutores.	- Habilitar la descarga masiva.
12. Publicaciones y vídeos del Instituto de Salud Carlos III	
- Textos: extracción de terminología de dominio sanitario; recogida de corpus de textos de dominio médico. - Vídeos: entrenamiento de sistemas de reconocimiento de voz en dominio sanitario, de detección de palabras clave, subtitulado automático, búsqueda y recuperación de información, o recono-cimiento y segmentación de locutores.	- Habilitar la descarga masiva.
13. OrphaData	

- Extracción de terminología multilingüe de dominio médico. - Extracción de listas de entidades nombradas de tipos semánticos de dominio médico: p. ej. enfermedades, genes, fenotipos y nombres de medicamentos. - Confección de tesauros anotados con códigos en terminologías de referencia. - Extracción de conocimiento ontológico (relaciones entre enfermedades, fenotipos y genes) para su aplicación en minería de textos.	- No se han identificado; en nuestra opinión, dispone de madurez alta.
14. Guías de práctica clínica (GPC) del portal Guía Salud	
- Creación de corpus paralelo monolingüe/multilingüe de dominio médico. - Recogida de textos de dominio médico. - Extracción de terminología de dominio médico. - Análisis de texto prescriptivo.	- Conversión de los PDF a texto y, deseablemente, revisión de los mismos.
15. Vídeos del portal web TV del Gobierno Vasco relacionados con Salud	
- Evaluación de tecnología de reconocimiento de voz (Speech-To-Text) para subtitulado/transcripción automática. - El conjunto resulta demasiado pequeño y limitado en cuanto al número de locutores para ser empleado por sí solo para el entrenamiento de tecnología de reconocimiento de voz (Speech-To-Text), aunque podría utilizarse en combinación con otros conjuntos (por ejemplo, con vídeos de otros temas de la misma fuente).	- Habilitar la descarga masiva. En relación a los documentos en PDF, sería deseable su conversión a texto y revisión de los mismos.
16. Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios	
- Recogida de corpus de textos de dominio médico-farmacológico. Los textos de las alertas sanitarias y medicamentosas son especialmente útiles para minería de textos y extracción de relaciones con aplicación a la farmacovigilancia. - Extracción de terminología en el área de sanidad.	- Habilitar la descarga masiva. - Sería deseable la conversión a texto y revisión de los documentos en PDF.
17. Nomenclátor de prescripción del Centro de Información de Medicamentos	

- Listas de entidades (nombres de medicamentos, principios activos, clases farmacológicas, compañías farmacéuticas, etc.). - Extracción de terminología farmacológica.	- Según nuestra opinión, es recurso de madurez alta que se puede tomar como modelo de diseño y distribución para otros conjuntos de datos analizados en el presente informe.
18. Textos de Jurisprudencia del CENDOJ	
- Recogida de corpus de textos de ámbito jurídico. - Extracción de terminología jurídica. - Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales). - Creación de un corpus paralelo de sentencias que estén disponibles en español y otra lengua cooficial (observado con euskera).	- Habilitar la descarga masiva, preferiblemente en formatos fácilmente reutilizables como XML. - También sería de gran utilidad la publicación de conjuntos con marcado de anonimización (recurso anonimizado y su correspondiente sin anonimizar).
19. Textos del Boletín Oficial del Estado (BOE) Diario.	
- Recogida de corpus de textos de ámbito jurídico. - Extracción de terminología jurídica. - Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales). - Creación de un corpus paralelo de suplementos de disposiciones o sentencias que estén disponibles en español y otra lengua cooficial. - Entrenamiento para recuperación y extracción de la información.	- Habilitar descarga masiva - Identificar pares traducidos (español-lengua cooficial).
20. Textos de Códigos electrónicos del Boletín Oficial del Estado (BOE)	
- Recogida de corpus de textos de ámbito jurídico. - Extracción de terminología jurídica. - Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales de España y de Europa). - Creación de memorias de traducción entre lenguas cooficiales y castellano, o bien	- Identificar pares de documentos traducidos (archivo original - traducción) - Sería recomendable realizar alineación de los textos para facilitar su utilización en traducción automática, así como la conversión a formatos reutilizables.

español y otras lenguas de la Unión Europea (francés, inglés, alemán e italiano). - Creación de un corpus paralelo de suplementos de disposiciones o sentencias que estén disponibles en español y otra lengua cooficial. - Entrenamiento para recuperación y extracción de la información.	
21. Textos sobre Legislación del Boletín Oficial del Estado (BOE)	
- Recogida de corpus de textos de ámbito jurídico. - Extracción de terminología jurídica. - Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales). - Creación de memorias de traducción entre lenguas cooficiales y castellano. - Creación de un corpus paralelo de legislación en español y otra lengua cooficial. - Entrenamiento para recuperación y extracción de información, y de modelos de traducción automática.	- Conversión de aquellos archivos en lenguas cooficiales que solo están disponibles en PDF a formato texto. - Identificación de los pares de documentos traducidos y alineación de textos.
22. Memorias de traducción con publicaciones en el BOE en euskera del IVAP	
- Extracción de terminología jurídica. - Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales de España). - Creación de memorias de traducción entre otros pares de lenguas. - Entrenamiento de modelos de traducción automática.	- Anonimización de datos personales.
23. Memorias públicas de traducción de la Diputación Foral de Gipuzkoa	
- Extracción de terminología precisa sobre alguno de los campos de estudio. - Creación de ontologías jurídicas y otros recursos léxicos (posibilidad de crear un léxico bilingüe entre pares de lenguas cooficiales de España, u otras lenguas).	-

- Creación de memorias de traducción entre otros pares de lenguas. - Entrenamiento de modelos de traducción automática.	
24. Grabaciones de Vistas Judiciales del Consejo General del Poder Judicial	
- Entrenamiento y evaluación de sistemas de reconocimiento de voz (Speech-To-Text) para facilitar la búsqueda y la recuperación de información en audio y vídeo de vistas y comparecencias orales en juicios. - Entrenamiento y evaluación de sistemas de reconocimiento y segmentación de locutores para facilitar la búsqueda y recuperación de información en audio y vídeo de vistas y comparecencias orales.	- Habilitar el acceso y descarga masiva.

Tabla 27: Posibles aplicaciones y recomendaciones de los recursos analizados en el estudio

Otro requisito imprescindible para todos los conjuntos de datos analizados en este informe es que estén en dominio abierto y sus condiciones de uso permitan fácilmente la reutilización por parte de la comunidad investigadora. En ese sentido, **se han excluido colecciones** que están **muy maduras** y son **muy apreciables** en términos de interés y de tamaño, pero que no están accesibles para los investigadores. Por ejemplo, podríamos señalar aquí los corpus anotados de la RAE, el Diccionario de Términos Médicos de la Real Academia Nacional de Medicina, la Biblioteca Cochrane Plus de Evidencia Sanitaria, el Diccionario Español de Ingeniería, o el Archivo de la Web Española de la BNE.

A pesar de esta primera observación, el estado actual de los datos abiertos que pueden convertirse en RL es muy prometedor. En primer lugar, porque existen colecciones muy amplias de documentos digitalizados. Queremos destacar en este punto las colecciones de la BNE (Hemeroteca Digital, Biblioteca Digital Hispánica, EPUBs de la BNE) o los informativos de RTVE a la carta. La importancia y la cobertura temática y temporal de algunas de estas colecciones les hace candidatos a ser propuestos en la lista de proyectos de conversión a RL. Efectivamente, una de las características de los fondos digitalizados es que están en formato PDF y necesitan una conversión por OCR para acceder a su contenido en formato texto UTF-8. A partir de esta primera conversión, y tras una revisión de errores, los textos estarían a disposición de la comunidad investigadora para su procesamiento por etiquetadores PLN.

En los materiales multimedia existe también mucho interés en su conversión a formatos reutilizables por la comunidad de procesamiento de habla. En concreto, los archivos sonoros de RTVE y las vistas

orales de juicios son recursos muy demandados por esos investigadores, dado que contendrían un gran número de horas de audio (en el caso de RTVE, de gran calidad, y posiblemente, con transcripciones fieles en formato texto).

Entre los recursos maduros y fácilmente reutilizables nos encontramos con la colección de índices de materias, autores y títulos del catálogo de la BNE, ya en formato XML y siguiendo la metodología de Linked Open Data, que se actualizan mensualmente, y cuentan con una licencia de uso que no requiere autorización previa. Sin duda, es el recurso analizado en este informe más avanzado para su uso directo en PLN (por ejemplo, para elaborar reconocedores de entidades y clasificadores por temas).

También destacan por su madurez los conjuntos de memorias de traducción del Instituto Vasco de Administración Pública, tanto de la legislación española como europea al euskera, así como las de la Diputación Provincial de Gipuzkoa. Los recursos terminológicos del Centro de Terminología (TERMCAT) proporcionaban excelentes datos listos para ser utilizados como RL, pero el estado de disponibilidad de algunos diccionarios ha cambiado a fecha de última versión del informe.

Por último, destacamos el Nomenclátor de prescripción (CIMA) y el portal OrphaData, que recogen recursos con alto potencial para tareas de PLN en dominio sanitario (extracción de terminología multilingüe, extracción de listas de entidades nombradas, confección de tesauros anotados o extracción de relaciones y minería de textos). Los datos están ricamente anotados conforme a terminologías de referencia, se distribuyen en XML, o permiten el acceso dinámico mediante consultas SPARQL. La descarga de datos es inmediata y su distribución y su actualización es continua.

6 DATOS ABIERTOS Y SU USO COMO RL EN OTROS PAÍSES

Para comprender mejor en qué lugar se sitúa España en referencia al uso de datos abiertos como RL, es primordial conocer cómo se han desarrollado este tipo de iniciativas en otros lugares del mundo. En la Sección 6.1, abordamos estudios e iniciativas de impulso de recursos y tecnologías lingüísticas en otros países, y en la Sección 6.2, ofrecemos un panorama comparativo por países de datos abiertos o con potencial de convertirse en RL.

6.1 ESTUDIOS SIMILARES E INICIATIVAS DE IMPULSO DE RECURSOS Y TECNOLOGÍAS LINGÜÍSTICAS EN OTROS PAÍSES

A lo largo de esta sección, vamos a comparar las iniciativas de impulso de RL y TL realizadas en España con otros países de ámbito hispano-americano, europeo, y otras grandes potencias como Estados Unidos o Canadá. Cabe destacar que las prioridades de programas de investigación actuales (p. ej. H2020) integran las tecnologías de PLN junto a la investigación en *Big Data*. El apoyo al desarrollo de recursos lingüísticos ha cambiado respecto a programas de financiación anteriores como FP7 (vid. informe de Zabala Innovation Consulting 2018) [12]. El desarrollo ha sido impulsado, pues, por iniciativas nacionales (p. ej., Francia y Estonia) o en proyectos tecnológicos de investigación genérica (no propiamente de creación de recursos).

6.1.1 Hispanoamérica

En **México,** el Consejo Nacional de Ciencia y Tecnología (CONACYT) mantiene una **Red Temática de Tecnologías del Lenguaje[39].** Impulsa la investigación y la sinergia de iniciativas empresariales y académicas, con especial atención al español de México, en cuatro líneas de acción: Difusión, Formación, Creación de recursos y Colaboración. Organiza periódicamente jornadas científicas y divulgativas, mantiene un repositorio de publicaciones, y ha coordinado informes o eventos acerca del estado del arte.[40] Destaca la promoción de estancias breves para investigadores y la organización de un premio para la creación de recursos lingüísticos.

No hemos documentado redes temáticas similares en **Argentina, Chile** o **Perú**. En estos países, las iniciativas parten de grupos de investigación independientes y de grado heterogéneo de madurez.

6.1.2 Europa

En Europa, hay diferentes instituciones a nivel internacional que aglutinan la recopilación de RL. En primer lugar, hay que destacar ELRA (European Language Resources Association). [4] Su creación en la década de los 90 fue favorecida por la Comisión Europea para ser el repositorio de los RL que se estaban creando con numerosos proyectos financiados públicamente, pero dando también cabida a RL de financiación privada. En sus estatutos, se proclama:

> *The mission of the Association is to promote language resources (henceforth LRs) and evaluation for the Human Language Technology (HLT) sector in all their forms and all their uses,*

[39] www.redttl.mx
[40] http://goo.gl/Nyw275

> *in a European and international context. Consequently, the goals are: to coordinate and carry out identification, production, validation, distribution, standardisation of LRs, as well as support for evaluation of systems, products, tools, etc. - related to language resources. Other resources will be considered as well if developments of the field make this desirable: e.g. multimedia resources both with and without language.*

Por tanto, ELRA ha sido la institución a nivel europeo que se lleva encargando 20 años de *identificar*, *producir*, *validar*, *estandarizar*, *distribuir* y *evaluar* RL. Si bien, todas estas actividades están directamente implicadas con el tema de este estudio, debemos insistir en que ELRA básicamente se centra en RL ya desarrollados. O, dicho de otra manera, su misión principal no es identificar conjuntos de datos públicos (abiertos o no) para su posible conversión en RL utilizables por la comunidad PLN.

Por otra parte, encontramos el portal LT-INNOVATE[41], que pertenece a la Asociación Internacional de la Industria de las Tecnologías del Lenguaje, con sede en Bruselas. Entre los servicios ofrecidos está el LT-Observatory, un portal que contiene información sobre las políticas lingüísticas por países y fuentes de financiación. Además, se puede consultar un catálogo de RL, principalmente formado por aportaciones de los socios industriales.

Del mismo modo, en la Comisión Europea, la DG CONNECT (G3 Learning, Multilingualism & Accessibility) es la unidad responsable para las Tecnologías Lingüísticas. El documento más reciente de posicionamiento sobre políticas en este campo es de enero de 2017, y lleva por título: *Assesment of the State of the EU Language Technology sector and EU policy recommendations* [42]. Una de sus conclusiones es que se necesitan RL para el desarrollo de soluciones de traducción automática en dominios específicos, dado que *existing LR repositories are not usable operationally, particularly not for comercial purposes*. Dicho de otra manera, la Comisión Europea es consciente de que hay que invertir dinero en preparar RL aptos para ser reutilizados en diferentes entornos. El informe de DG CONNECT también destaca el desfase entre la industria americana de TL y la europea, y la necesidad de desarrollar un Mercado Único Digital basado en el tratamiento multilingüe, para que las empresas europeas puedan competir con las americanas o las chinas. Entre los aspectos positivos, hay que reseñar que hay una institución europea que ha conseguido este objetivo (es decir, un mercado único digital multilingüe): la European Union Intellectual Property Office (EUIPO).

[41] www.lt-innovate.org

[42] http://www.lt-innovate.org/sites/default/files/Assessment of the state of Language Technologies and EU policy recommendations.pdf

Entre las recomendaciones del informe DG CONNECT, queremos destacar las siguientes:

1. *Language interoperability will depend on widespread sharing of EU "big language data." US firms have been able to capitalise on access massive data resources through their platforms to build algorithms. The EC has initiated a European Language Resources Coordination action to gather language data in support of the European public services developed in the framework of the CEF Programme. The EC should broaden this effort and make all multilingual data resources created and gathered with public funding available for public AND commercial purposes as a major social and business asset.*
2. *It is vital to adjust European copyright law to enable the use of very large data banks in multiple languages for technical purposes such as the training of machine translation engines, by for example "anonymising" data sources.*
3. *The one-sided policy emphasis on open source software should be toned down as it has negative consequences on commercial software product development in Europe.*

La recomendación más importante de este informe es que **Europa necesita una infraestructura para PLN**, que denominan European Language Infrastructure (ELI). Todo desarrollo de PLN depende de esta infraestructura, que es costosa de desarrollar y mantener y que, además, es necesaria para cualquier lengua. Esta ELI debería proporcionar las funcionalidades básicas a través de APIs (Application Programming Interfaces) como la tokenización, el etiquetado morfosintáctico, el reconocimiento de entidades, etc., para todas las lenguas comunitarias, con un calidad básica y universal para todos los usuarios.

Por tanto, con los objetivos de compilar y preparar esos *big language data* para el desarrollo de aplicaciones, buscamos aquellas estrategias nacionales que pudieran entroncar con esta iniciativa.

Así, no se encontraron estudios que hablasen de la reutilización de datos públicos o de contenido de webs de diferentes administraciones públicas para la creación o el mantenimiento de RL en los países francófonos (p. ej., **Francia** y **Bélgica**). La única iniciativa destacable en Francia (aunque no se enfoca a la reutilización de datos públicos abiertos) es el buscador MultiTal,[43] desarrollado por el Instituto Nacional de Lenguas y Civilizaciones Orientales (INALCO) de París. Este recurso recopila 142 recursos lingüísticos multilingües para la ingeniería lingüística a fecha de julio de 2018. Un modelo que sería interesante adaptar para recensar recursos de ámbito iberoamericano.

[43] http://multital.inalco.fr/

En relación a este campo de estudio podemos señalar, sin embargo, el portal europeo de recursos lingüísticos de la European Language Resource Coordination, con una lista de recursos existentes en distintos países de Europa y que tienen carácter abierto[44]. Del mismo modo, encontramos un inventario de recursos lingüísticos de las lenguas de Francia, elaborado por la Delegación de la lengua francesa y de las lenguas de Francia en colaboración con ELDA,[45] aunque con un carácter informativo y más centrado en lenguas o usos dialectales, y, en ningún caso, como punto de partida para la elaboración de recursos de TL o que incidan en el PLN.

Francia también se ocupa de su lengua a través de la Dirección General de la lengua francesa y de las lenguas de Francia (DGLFLF), que mantiene igualmente un portal dedicado en exclusiva a las lenguas. En dicho portal podemos encontrar los decretos sobre la terminología o el uso de la lengua en la administración, recursos jurídicos u otras publicaciones de interés.[46]

En los países germanoparlantes, no hay ninguna estrategia estatal sobre temas de RL. Por ejemplo, en Alemania,[47] toda iniciativa sobre lenguas se gestiona a nivel de regiones o *Länders*, lo que contrasta con la importancia del PLN en universidades y pequeñas y medianas empresas (PYMEs) alemanas. En Austria ocurre algo similar: no hay ningún plan estratégico específico para TL, aunque existe un apoyo institucional fuerte al consorcio CLARIN-AT.

En otros países europeos con lenguas con menos hablantes, pero con una gran tradición en PLN, como **Holanda**, **Suecia**, **Estonia** o **Finlandia**, existen estrategias estatales para financiar RL.

En los países del Sur de Europa, nos encontramos con diferentes situaciones. **Portugal** tiene programas específicos asociados a la agenda Portugal 2020 dentro de su Estrategia Nacional de Especialización Inteligente (ENEI). En concreto, se establece como línea prioritaria dentro de la TIC el desarrollo de tecnologías de la lengua portuguesa, por cuanto:

O português é uma língua com grande implantação mundial e falada em países com grande crescimento, tornandose portanto decisiva e crítica a aposta no cruzamento do potencial TIC identificado com a relevância socioeconómica e histórico-cultural da Língua Portuguesa[48].

[44] https://elrc-share.eu/repository/search/?q=

[45] http://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Les-technologies-de-la-langue-et-la-normalisation/Inventaire-des-ressources-linguistiques-des-langues-de-France

[46] http://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France

[47] http://www.lt-innovate.org/lt-observe/germany

[48] https://www.fct.pt/esp_inteligente/docs/ENEI_Anexo%20B_%20PrioridadesEstrategicas_05junho2014.pdf

En **Italia**, la estrategia en TL se coordina desde los programas genéricos de I+D+i. Constitucionalmente, el estado no tiene competencia legislativa en materia lingüística y corresponde a las administraciones regionales. Desde 2013 y dentro de la Strategia per la crescita digitale (2014-2020), se ha dado impulso a la agenda digital, pero no se menciona específicamente ninguna medida para la TL[49].

6.1.3 Reino Unido, Estados Unidos y Canadá

En el **Reino Unido**, la situación es mucho mejor. Desde mediados de los 90 a través del Ministerio de Comercio e Industria se han estado financiando proyectos en TL. La principal agencia financiadora de proyectos sobre PLN es el Engineering and Physical Sciences Research Council (EPSRC).[50] Además, lenguas cooficiales como el galés y el gaélico escocés también han recibido apoyo para creación de RL.

En **Canadá**, es interesante nombrar los recursos para la redacción, léxicos o diccionarios o documentos de distintas organizaciones sobre alfabetización, bilingüismo o Francofonía del Portal Lingüístico de Canadá.[51] Dicho portal, creado por el gobierno, ofrece regularmente una actualización sobre recursos lingüísticos de la lengua del país.

Finalmente, en relación con los datos lingüísticos, en EE.UU. existe el Linguistic Data Consortium (LDC) en la Universidad de Pensilvania [9]. El LDC centraliza la distribución de recursos lingüísticos, desde textos a material audiovisual transcrito, pasando por léxicos, *treebanks*, etc. Actualmente, atesora más de 700 recursos lingüísticos.

6.1.4 Comparación por países de datos con potencial de convertirse en RLs

En lo que a datos abiertos se refiere, encontramos varios organismos que mantienen información actualizada comparable por países.

El Portal Europeo de Datos (Open Data in Europe),[52] financiado por la Comisión Europea, recoge una serie de indicadores sobre el grado de madurez del desarrollo de políticas nacionales de promoción de los Datos Abiertos. En concreto, en su versión de 2017, se recoge la siguiente información[53]: *"España está a la cabeza de Europa, detrás de Irlanda, y es un modelo de tendencias (trendsetter) en cuanto a madurez de datos abiertos*" (Figura 1).

[49] http://www.lt-innovate.org/lt-observe/italy

[50] https://epsrc.ukri.org/

[51] https://www.noslangues-ourlanguages.qc.ca/fr/index

[52] www.europeandataportal.eu/es

[53] https://www.europeandataportal.eu/es/dashboard#2017.

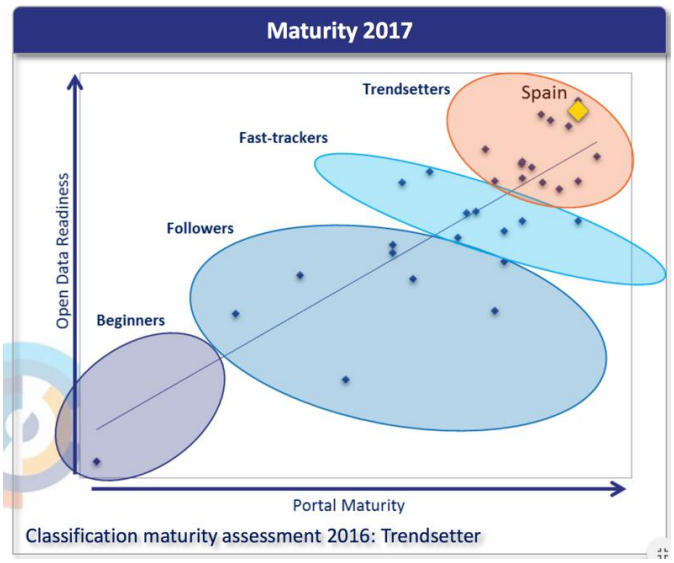


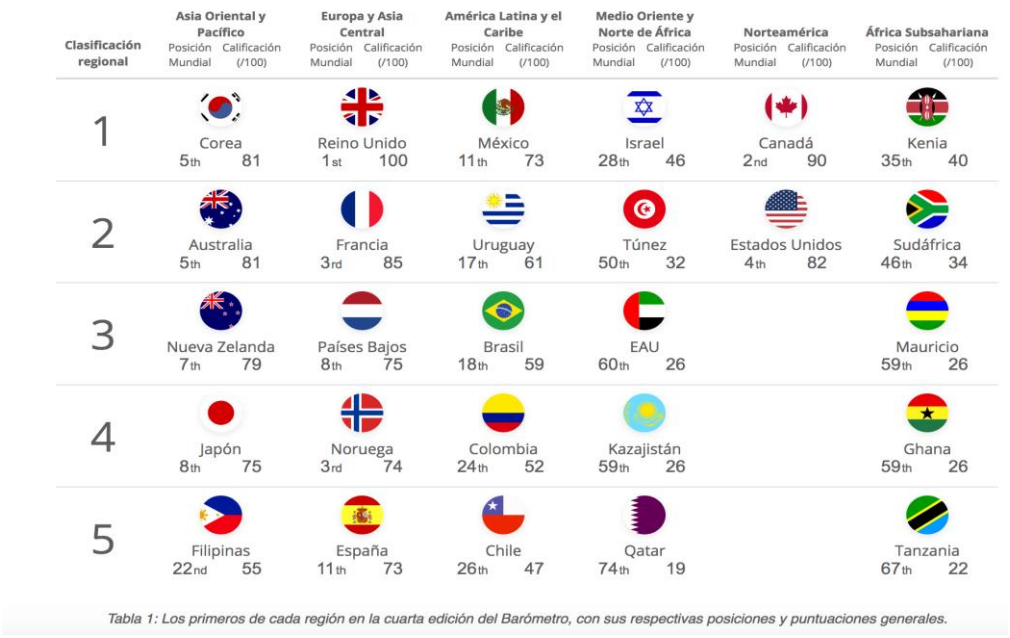
Figura 1: Posición de España en Europa en cuanto a madurez de datos abiertos

Además, en los tres subindicadores (Disponibilidad, Usabilidad e Impacto), España también ha alcanzado la plena madurez, incluso aumentando ligeramente el uso de los datos abiertos con respecto a 2016. Sin embargo, estos datos contrastan con los que aparecen en otros índices internacionales, como el Open Data Barometer o el Open Data Index Rank, que sitúan a España entre el grupo de cabeza, pero en posiciones por debajo del puesto 10.

Otra de las iniciativas internacionales para estadística de datos abiertos es el portal Open Data Barometer[54] (OPD), que realiza una clasificación por países a fin de analizar la posición de estos en términos de la disponibilidad de datos gubernamentales.[55] En nuestra consulta de julio de 2018, según los últimos censos disponibles (2016), España se encuentra en el rango 11, junto a México, por debajo de otros países como Reino Unido, Canadá, Francia o Estados Unidos, y por encima de Dinamarca, Austria, Alemania, Suecia, Colombia, Chile o Uruguay. La Figura 2 muestra la clasificación de los 5 primeros de cada región, tomada de la página del Open Data Barometer. Hay que destacar que los tres países más avanzados en datos abiertos son Reino Unido, Canadá y Francia, con una valoración muy superior al resto.

[54] www.opendatabarometer.org

[55] *La clasificación realizada por el World Wide Consortium para los años 2013-2015 se discute en el artículo de Sandoval y Ospina Mercy (2016): "Open data: realidades sobre apertura de datos en Venezuela". Enl@ce, 13(2), 56-70. Disponible en: (www.redalyc.org/jatsRepo/823/82349540004/html/index.html)*



Clasificación regional	Asia Oriental y Pacífico Posición Mundial / Calificación (/100)	Europa y Asia Central Posición Mundial / Calificación (/100)	América Latina y el Caribe Posición Mundial / Calificación (/100)	Medio Oriente y Norte de África Posición Mundial / Calificación (/100)	Norteamérica Posición Mundial / Calificación (/100)	África Subsahariana Posición Mundial / Calificación (/100)
1	Corea 5th 81	Reino Unido 1st 100	México 11th 73	Israel 28th 46	Canadá 2nd 90	Kenia 35th 40
2	Australia 5th 81	Francia 3rd 85	Uruguay 17th 61	Túnez 50th 32	Estados Unidos 4th 82	Sudáfrica 46th 34
3	Nueva Zelanda 7th 79	Países Bajos 8th 75	Brasil 18th 59	EAU 60th 26		Mauricio 59th 26
4	Japón 8th 75	Noruega 3rd 74	Colombia 24th 52	Kazajistán 59th 26		Ghana 59th 26
5	Filipinas 22nd 55	España 11th 73	Chile 26th 47	Qatar 74th 19		Tanzania 67th 22

Tabla 1: Los primeros de cada región en la cuarta edición del Barómetro, con sus respectivas posiciones y puntuaciones generales.

Figura 2: Posición de España en Europa en cuanto a madurez de datos abiertos. La cifra debajo a la izquierda de cada bandera indica la posición en el ranquin mundial, y la cifra a la derecha, la puntuación asignada.

En los **países latinoamericanos,** Argentina, México y Uruguay tienen portales de datos abiertos comparables a los del resto de países del mundo. En el otro extremo nos encontramos con Venezuela, Guatemala, Honduras, Cuba o Belice, que carecen de portales de datos abiertos.[56] En Venezuela, ciertos organismos recogen datos abiertos (p. ej., el portal del Centro Nacional de Tecnologías de la Información, CNTI),[57] pero su grado de disponibilidad y apertura no es comparable al del resto de países iberoamericanos. Con todo, existen iniciativas ciudadanas, como el portal Transparencia Venezuela[58] o el Instituto de Prensa y Sociedad de Venezuela.[59] Estas iniciativas se han aglutinado en el portal Vendata,[60] que ofrece datos de la Gaceta Oficial, las Memorias y las Cuentas.

[56] *Véase el artículo de M. Steinberg y D. Castro, "The State of Open Data Portals in Latin America" (2/07/2017, https://www.datainnovation.org/2017/07/the-state-of-open-data-portals-in-latin-america/), traducido al español como "El estado de los portales de datos abiertos de América Latina": https://blogs.iadb.org/abierto-al-publico/2017/07/18/el-estado-de-los-portales-de-datos-abiertos-en-latinoamerica/*

[57] *www.cnti.gob.ve*

[58] *https://transparencia.org.ve*

[59] *http://ipysvenezuela.org*

[60] *https://vendata.org*

Honduras, igualmente, no ofrece un portal de datos abiertos, de modo que la información se distribuye mediante iniciativas independientes. El Salvador solo cuenta con una web mantenida por ciudadanos voluntarios.[61]

Costa Rica, pese a no contar con un conjunto de datos gubernamentales, cuenta con la iniciativa ciudadana Abriendo Datos.[62] Con todo, existen consorcios de ámbito latinoamericano, como la Iniciativa Latinoamericana por los Datos Abiertos (ILDA),[63] que aglutina investigaciones e informes en torno a las temáticas de Justicia, Educación o Enfermedades, entre otros.

La Tabla 28 muestra el volumen de datos abiertos por países iberoamericanos (salvo los países sin portales gubernamentales o sin iniciativa privada), a fecha de julio de 2018.

Sin ánimo de exhaustividad, podemos comparar los datos abiertos de países latinoamericanos con los españoles en cuanto a:

- **Formatos**: los más comunes suelen ser CSV o Excel (XSL/XSLX), y en menor medida, PDF. XML está bastante extendido como estándar solo en determinados países (España y Chile), pero el volumen de datos en este formato suele ser menor. Argentina, Colombia, España, Guatemala, México, Paraguay, y Uruguay también distribuyen en formato JSON. El tipo de contenido de los datos determina, en mayor medida, el formato de distribución.

- **Contenido y categorías**: el número de clases de datos oscila entre 10 y 29, propia de clasificaciones más finas como las de Chile, Colombia y España. El volumen de datos en cada categoría no es homogéneo y varía de un país a otro.
- **Volumen**: los países que ofrecen mayor cantidad de datos son, por este orden, México, España, Colombia y Chile (marcados en negrita en la tabla).

	Argentina	Bolivia	Chile	Colombia
# Conjunto datos	640	25	**3500**	**8695**
# Organismos	20	6	521	1119
# Categorías	13	10	23	29
URL	https://datos.gob.ar	https://datos.gob.bo	https://datos.gob.cl	www.datos.gov.co

[61] *www.datoselsalvador.org*
[62] *https://abriendodatoscostarica.org*
[63] *https://idatosabiertos.org/*

	Costa Rica	República Dominicana	Ecuador	El Salvador
# Conjunto datos	Desconocido	495	119	70
# Organismos	Desconocido	132	16	Desconocido
# Categorías	13	11	15	12
URL	http://datosabiertos.presidencia.go.cr	http://datos.gob.do	http://catalogo.datosabiertos.gob.ec	www.datoselsalvador.org
	España	**Guatemala**	**México**	**Paraguay**
# Conjunto datos	**18990**	25	**37248**	220
# Organismos	115	19	265	117
# Categorías	22	12	11	25
URL	http://datos.gob.es	http://ckan.concyt.gob.gt	https://datos.gob.mx	www.datos.gov.py
	Panamá	**Perú**	**Puerto Rico**	**Uruguay**
# Conjunto datos	282	1047	Desconocido	106
# Organismos	41	1300	Desconocido	36
# Categorías	16	12	13	16
URL	www.datosabiertos.gob.pa	www.datosabiertos.gob.pe	https://data.pr.gov	http://datos.gub.uy

Tabla 28. Volumen de datos abiertos en los países hispanoamericanos

Volviendo a los países considerados francófonos, en **Francia**, **Canadá** y **Bélgica**, nos encontramos con el siguiente panorama:

- **Formatos**: los más comunes en Francia suelen ser JSON, SHP o ZIP, seguido de CSV y XLS, y en menor medida, PDF. El tipo de contenido de los datos determina, generalmente, el formato de distribución. Así, en Canadá, destaca el uso de HTML, SHP, PDF, y XML. Es curioso, por ejemplo, que no se obtengan demasiados resultados en JSON para dicho país, dado que es un formato que actualmente está siendo bastante reconocido en este ámbito, y en el científico en general. Bélgica se inclina más hacia el uso de WMS, WFS, CSV, JSON y XLS. En menor medida, los datos se ofrecen en PDF o XML, que, curiosamente, es un formato bastante aprovechable para el campo de las tecnologías del lenguaje.

- **Contenido y categorías**: En Francia se agrupa en torno a 9 categorías claras, aunque su buscador es bastante completo y nos ofrece múltiples formas de encontrar los datos (organización, palabras clave, licencias, fecha o rango de fechas, cobertura espacial, formato,

granularidad territorial, o reutilizaciones conocidas: no reutilizado, poco reutilizado, bastante reutilizado, frecuentemente reutilizado). En Canadá existe una segmentación mayor, con 19 categorías, que permite refinar todavía más las búsquedas. Bélgica apuesta, a su vez, por una segmentación en 14 categorías generales, parecidas a las del resto de países, en la que el número de datos mayor se encuentra en medio ambiente y en lo que denominan sector público (Public Sector).

- **Volumen**: El volumen de datos ofrecidos por Francia es, junto a Canadá, de los mayores del mundo.

En los países germanoparlantes (**Alemania** y **Austria**) la reutilización de datos abiertos no está muy avanzada, aunque existen portales con múltiples recursos.[64]

En **Reino Unido**, existe un portal de datos abiertos (UK Data Service),[65] que contiene una gran colección de datos abiertos, incluyendo, sobre todo, datos de estudios sociales y económicos. También existe otro portal de datos abiertos,[66] que engloba todo tipo de datos, aunque muchos de los conjuntos que aparecen no han sido todavía liberados y, por tanto, no son accesibles.

En relación con los corpus lingüísticos, la mayor parte de ellos aparecen facilitados y mantenidos por universidades como la de York[67] y la de Oxford.[68]

Estados Unidos, sin embargo, tiene un portal de datos abiertos muy desarrollado,[69] que incluye más de 285.000 conjuntos de datos. Junto a este, encontramos otro sitio web que publica sistemas para acceder y procesar datos abiertos.[70] También existen portales específicos de datos abiertos por temáticas; por ejemplo, el del Departamento de Justicia.[71]

En general, y de manera universal, pese a la disponibilidad de datos y su volumen, este tipo de **datos aún no presenta una madurez adecuada para su explotación avanzada** como recurso lingüístico. En su estado, podrían utilizarse para la extracción de listas de entidades nombradas, la recogida de corpus o tareas de clasificación de textos por categoría. Sabemos que los datos no suelen estar anotados

[64] *Datos abiertos de Alemania: www.govdata.de ; Open Knowledge Foundation (https://okfn.de/en/)*
[65] *https://www.ukdataservice.ac.uk/*
[66] *https://data.gov.uk/*
[67] *https://www.york.ac.uk/language/current/resources/corpora/*
[68] *http://ota.ox.ac.uk/about/oxford.xml*
[69] *https://www.data.gov/*
[70] *https://usopendata.org/*
[71] *https://www.justice.gov/open/open-data*

según los estándares de la comunidad PLN, de modo que **requerirían procesamiento previo a su uso** en tareas de tipo supervisado o no supervisado. La mayoría de los datos, por otra parte, y en lo que se refiere a formatos y reutilización, hay que destacar que no se distribuyen en formato texto (TXT), aunque los contenidos que se ofrecen en CSV sí que pueden procesarse de manera más directa.

Otro punto negativo que hay que destacar a nivel internacional es que los conjuntos de datos **no están disponibles para una descarga masiva**, lo que constituye *uno de los principales desafíos para las administraciones públicas si quieren mejorar la calidad de sus servicios* (European Data Portal).

6.2 COMPARACIÓN POR TEMÁTICAS DE INTERÉS.

A continuación, comparamos someramente los recursos censados en este informe con respecto a los existentes en el ámbito internacional, atendiendo a sus diferentes temáticas de interés.

6.2.1 Inteligencia competitiva

Algunos de los recursos considerados como prioritarios en este estudio poseen equivalentes en organismos **latinoamericanos.** Por ejemplo:

- Existen datos relativos a patentes y modelos de invención (véanse fichas de recursos), que se suelen presentar en formato PDF.
- Igualmente, encontramos publicaciones de las Cámaras de Comercio de diferentes países latinoamericanos (aunque algunas detentan derechos reservados de uso):
    - Estudios e informes en PDF[72] y normativa en DOC[73] de la Cámara de Comercio de Argentina.[74]
    - Publicaciones e informes[75] en PDF de la Cámara de Comercio de Lima.[76]
    - Publicaciones, estudios y legislación[77] en PDF de la Cámara de Comercio de Santiago (Chile);[78] que permite usar la información citando la fuente.

[72] www.cac.com.ar/institucional/informes_uepe_1975
[73] http://comercioexterior.cac.com.ar/institucional/Normativa_441
[74] www.cac.com.ar
[75] www.camaralima.org.pe/principal/categoria/informacion-empresarial/17/c-17
[76] www.camaralima.org.pe
[77] https://www.ccs.cl/html/promocion_negocios/estudios.html, https://www.ccs.cl/estudios/estudios.html
[78] www.ccs.cl

- Publicaciones[79] en PDF de la Cámara de Comercio y Servicios de Uruguay.[80]

En **Europa**, destacan los datos ofrecidos por la Oficina de Publicaciones de la Comisión, que se distribuyen en diferentes portales. Para inteligencia competitiva, son relevantes los documentos del Suplemento al Diario Oficial de la UE con centenares de miles de noticias al año, del portal TED (Tenders Electronic Daily).[81] Son descargables en XML para usuarios registrados[82] y cuentan también con una versión más reducida en formato CSV.[83]

Para entidades nombradas, son especialmente relevantes los portales EU Whoiswho[84] y el European Media Monitor, desarrollado por el Joint Research Center, y que ofrece sus catálogos de datos en acceso abierto[85]. Se trata de un repositorio muy completo y extenso (2009 conjuntos de datos, 68 colecciones y 10 áreas temáticas a finales de julio de 2018) y está muy bien organizado, incluyendo hasta una clasificación por tipo de formato (XML, HTML, etc.).

Otros conjuntos de datos significativos que se pueden descargar en diferentes formatos del Portal de Datos Abiertos Europeo[86] son el corpus CORDIS (con todos los proyectos de investigación de la UE) y los tesauros multilingües EuroVoc (en XML y RDF).

A continuación, analizaremos, con algo más de detalle, la situación por algunos países europeos.

Respecto a documentos de patentes y modelos de invención, en **Francia** se pueden obtener a partir del Institut National de la Propriété Industrielle (INPI)[87], encargado de gestionar las patentes del país. Los documentos de patentes pueden recuperarse en su portal, bajo la rúbrica Base Brevets[88]. Este buscador de patentes da acceso a unos 8 millones de datos, entre otros: peticiones de patentes de Francia publicadas a partir de 1902 y peticiones europeas y mundiales desde 1978. En este último caso, el acceso se suele dar a través de Espacenet Worldwide, que contiene, a su vez, las patentes de la European Patent Office. Hay que destacar que la información o datos que se obtienen aparece solo en

[79] *http://www.cncs.com.uy/articulosinformes/*, *http://www.cncs.com.uy/informacion-economica/*
[80] *www.cncs.com.uy*
[81] *https://ted.europa.eu/TED/main/HomePage.do*
[82] *http://ted.europa.eu/TED/misc/xmlPackagesDownload.do*
[83] *https://data.europa.eu/euodp/en/data/dataset/ted-csv*
[84] *http://europa.eu/whoiswho/public/index.cfm*
[85] *https://data.jrc.ec.europa.eu/*
[86] *http://data.europa.eu/euodp/en/data*
[87] *https://www.inpi.fr/fr*
[88] *https://bases-brevets.inpi.fr/fr/accueil.html*

formato HTML (a menudo, un resumen), y no es posible la descarga de archivos en PDF. Sin embargo, es posible descargar muchos datos abiertos en su portal específico[89], donde podemos encontrar:

- Resúmenes de patentes europeas con su traducción en francés (en caso de estar en inglés o alemán) de 2004 a 2008 y resúmenes de patentes europeas desde 2010. Un archivo que se actualiza de forma quincenal y que se ofrece, entre otros formatos, en XML.[90]
- Patentes francesas, tanto contemporáneas como el histórico del siglo XIX (de 1791 a 1891, unos 180.000 documentos).[91]
- Base de datos de marcas francesas desde 1976, con actualización semanal (unos 2,5 millones de documentos).[92]
- Jurisprudencia y decisiones o procedimientos de oposición de ámbito nacional sobre marcas, patentes, diseños y modelos, con una actualización mensual (unos 45.000 documentos).[93]
- Base de datos de diseños y modelos franceses desde 1910 del INPI, con más de 700.000 documentos, y actualizado bimensualmente.[94]

Es importante destacar que estos recursos se tratan como datos abiertos, aunque para la reutilización de los datos se necesita aceptar una determinada licencia que hay que cumplimentar y mandar al INPI. Como aspectos positivos, destacamos la claridad en la exposición de los recursos abiertos, ya que se nos aporta con detalle su contenido (tipo de datos: documentos, resúmenes, imágenes, etc.; fechas o rango de fechas que cubre, tipo de actualización, etc.), qué volumen de datos se estima, y hasta la posibilidad de descarga de un fichero de ejemplo, de manera que no es necesario descargarse todo el recurso para saber lo que contiene o para comprobar cómo se visualiza. Esto sería interesante de cara a diferentes conversiones futuras o al tratamiento para recursos de TL, pues permite probar primero con datos de menor tamaño.

En **Bélgica** no encontramos un acceso a las patentes del país, ni desde su portal de datos abiertos nacional (pese a que se indica su existencia) ni desde la web del Ministerio de Economía, aunque sí nos muestran el acceso a la base de datos de patentes de la EPO, Esp@cenet. [95]

[89] *https://www.inpi.fr/fr/services-et-prestations-domaine/open-data*
[90] *https://www.data.gouv.fr/fr/datasets/brevets-europeens/*
[91] *https://www.data.gouv.fr/fr/datasets/brevets-francais/* o *https://www.inpi.fr/fr/open-data-brevets-francais*
[92] *https://www.inpi.fr/fr/open-data-marques-francaises*
[93] *https://www.inpi.fr/fr/open-data-jurisprudence-et-decisions-d-opposition*
[94] *www.inpi.fr/fr/open-data-dessins-et-modeles*
[95] *https://worldwide.espacenet.com/*

En el ámbito de las patentes, sin embargo, nos parece importante señalar otro de los recursos que contienen, el EPATRAS (European Patent Translation System)[96] o buscador de traducciones de patentes de inventos europeos depositados en la Oficina de propiedad intelectual belga y otros documentos asociados al procedimiento administrativo. Este sistema, pese a su mal funcionamiento a priori, nos permite encontrar traducciones para diversas lenguas: inglés, francés, alemán y holandés. Este tipo de bases de datos podría ser un ejemplo de RL que puede ser realizado con la ayuda de otros datos abiertos considerados como susceptibles de convertirse en RL.

En **Canadá**, es posible acceder a 2.330.000 documentos de patentes a través de un buscador de la OPIC[97] (Office de la propriété intellectuelle du Canada).[98] Los documentos están accesibles de forma individual en PDF, junto a un resumen en HTML. La descarga de la base de datos de forma completa (es decir, de todo el recurso) solo es accesible mediante pago.

6.2.2 Sanidad

En el ámbito de la sanidad, las correspondientes Academias de Medicina de ciertos **países latinoamericanos** ofrecen datos que se asemejan a los censados para España contenidos en este estudio:

- Boletines[99] (en PDF), así como la Clasificación Internacional de Enfermedades vs. 10 (ICD-10)[100] en HTML de la Academia de Medicina de Argentina.[101]
- Boletines[102] y publicaciones[103] (en PDF) de la Academia Chilena de Medicina.[104]
- Boletines[105] y libros en PDF,[106] y contenido multimedia[107] (1381 vídeos) de sesiones y simposios (aunque no disponibles para su descarga) de la Academia Nacional de Medicina de México.[108]

[96] https://epatras.economie.fgov.be/rech.jsp?l=fr

[97] http://www.ic.gc.ca/opic-cipo/cpd/fra/recherche/simple.html

[98] https://www.canada.ca/fr/office-propriete-intellectuelle.html

[99] https://www.acamedbai.org.ar/boletin.php

[100] http://www.biblioteca.anm.edu.ar/icd.htm

[101] www.acamedbai.org.ar

[102] www.academiachilenademedicina.cl/cont.php?id=112

[103] www.academiachilenademedicina.cl/cont.php?id=146

[104] www.academiachilenademedicina.cl

[105] https://www.anmm.org.mx/publicaciones/publicaciones-periodicas/boletin-en-la-academia

[106] https://www.anmm.org.mx/publicaciones/libros

[107] https://www.anmm.org.mx/multimedia/videoteca

[108] www.anmm.org.mx

- Publicaciones (en PDF)[109] y conferencias en vídeo[110] (no descargables) de la Academia Nacional de Medicina de Perú.[111]
- Vídeos (no descargables)[112] y un Diccionario Académico de la Medicina[113] de la Academia Nacional de Medicina de Colombia.

Datos destinados a ciudadanos (folletos o divulgación) o a profesionales (alertas) se pueden recuperar en portales de algunos Ministerio de Salud; entre ellos, destacamos:

- Archivos sobre temas de salud,[114] y guías de práctica clínica[115] (ambos en PDF y HTML, y con licencia Creative Commons Reconocimiento, CC-BY) del Ministerio de Salud de Chile.
- Contenidos textuales, de imágenes y audio (en PDF, JPG y MP3)[116] y boletines e informes en PDF,[117] con licencia Creative Commons BY, del Ministerio de Salud de Argentina.
- Datos abiertos acerca de productos de salud, en JSON y CSV (de uso permitido si se cita la fuente) del Ministerio de Salud Pública de Paraguay.[118]
- Informes y boletines en PDF,[119] que se pueden usar citando su fuente y contactando por correo electrónico, del Ministerio de Salud de Panamá.

Otros estudios y folletos de información general del ámbito de la salud se pueden obtener también en el portal del ministerio competente en salud en **Francia** (Ministerio des Solidarités y de la Santé). Algunos recursos podrían ser convertidos, no sin un proceso exhaustivo de revisión y conversión de PDF a otros formatos más fácilmente reutilizables en PLN, en recursos lingüísticos.[120]

Los datos abiertos franceses relativos al ámbito de la sanidad y el bienestar social en el portal general del gobierno incluyen todo tipo de datos heterogéneos, similares a los de nuestro país. Muchos de ellos tienen solo un carácter estadístico o muestran listados de acciones, agentes, edificios, etcétera.

[109] http://www.acadnacmedicina.org.pe/publicaciones.html
[110] http://www.acadnacmedicina.org.pe/videoconferencia.html
[111] www.acadnacmedicina.org.pe
[112] http://anmdecolombia.net
[113] http://dic.idiomamedico.net/P%C3%A1gina_principal
[114] http://diprece.minsal.cl/temas-de-salud/temas-de-salud/
[115] http://diprece.minsal.cl/le-informamos/auge/acceso-guias-clinicas/
[116] www.msal.gob.ar/index.php?option=com_ryc_contenidos
[117] www.msal.gob.ar/index.php?option=com_bes_contenidos
[118] http://datos.mspbs.gov.py/data
[119] www.minsa.gob.pa/informacion-salud/epidemiologia, www.minsa.gob.pa/informacion-salud, www.minsa.gob.pa/informacion-salud/informes-publicaciones-y-boletines
[120] http://solidarites-sante.gouv.fr/grands-dossiers/

Lo mismo ocurre con los datos encontrados en el portal de datos abiertos específico del propio Ministerio.[121] Algunos que podrían ser destacables desde el punto de vista lingüístico serían:

- La Base de Datos Pública de Medicamentos,[122] constantemente actualizada, y que se puede descargar de forma masiva, junto con otros ficheros con las especialidades, presentaciones, composiciones o grupos genéricos.[123] Esta página contiene, además, un pequeño glosario en HTML,[124] que podría ser de utilidad también para la generación de recursos lingüísticos como, por ejemplo, léxicos bilingües.
- La lista de medicamentos y de sustancias dopantes, para su uso como entidades nombradas o en recursos terminológicos, disponible en XLSX, junto con su descripción en formato DOC.[125] Del mismo modo, existe un PDF con las sustancias actualizadas, proveniente del diario oficial de la República.[126]

En **Bélgica**, los conjuntos de datos que encontramos en su portal general no contienen excesivos datos de interés, ya que, en su mayoría, son listados de farmacias, hospitales, y otros de tipo estadístico. Algunos datos que destacan del conjunto, aunque no están listados en dicho portal, podrían ser:

- Publicaciones (trípticos, estudios o manuales) del Ministerio de Sanidad Pública, Seguridad Alimentaria y Medio Ambiente (Ministerio de Santé Publique, Securité de la Chaine Alimentaire et Environnement), que se ofrecen en formato PDF.[127]
- Conjuntos de datos de la Agencia federal del medicamento y de productos sanitarios (Agence fédérale des médicaments et produits de santé, AFMPS),[128] que incluyen:
    - Resúmenes y características de los medicamentos para uso humano y para uso veterinario, descargables en PDF y en varias lenguas (francés, holandés y alemán).[129]
    - Medicinal Products Database: Banco de datos, con datos en HTML, de medicamentos autorizados para consumo humano y para uso veterinario.[130]

[121] http://www.data.drees.sante.gouv.fr/ReportFolders/reportFolders.aspx?sCS_referer=&sCS_ChosenLang=fr
[122] http://base-donnees-publique.medicaments.gouv.fr/telechargement.php
[123] www.data.gouv.fr/fr/datasets/base-de-donnees-publique-des-medicaments-1/
[124] http://base-donnees-publique.medicaments.gouv.fr/glossaire.php
[125] www.data.gouv.fr/fr/datasets/liste-des-medicaments-et-des-substances-dopantes/#_
[126] https://www.afld.fr/wp-content/uploads/2018/01/Liste-des-interdictions-2018.pdf
[127] https://www.health.belgium.be/fr/publications-et-recherches
[128] https://www.afmps.be/fr
[129] http://bijsluiters.fagg-afmps.be/?localeValue=fr
[130] https://banquededonneesmedicaments.fagg-afmps.be/#/

Estos conjuntos de datos son, a menudo, interesantes porque contienen muestras de los mismos datos en distintos idiomas, de forma que crear RL bilingües o trilingües es mucho más sencillo, o como punto de partida para recursos léxicos dedicados a la traducción.

Por otro lado, en **Canadá** encontramos unos 700 conjuntos de datos en su portal general agrupados bajo la rúbrica de salud, de carácter heterogéneo, y, como en el resto de países, con contenidos estadísticos, indicadores de salud por diferentes tipos de población, o listados de diferentes entidades, lugares, etc. en formatos como CSV o XLM. De ellos, podríamos señalar algunos más interesantes como, por ejemplo:

- Lista de patentes de medicamentos, que incluye una lista alfabética de ingredientes y sus patentes asociadas, disponible en inglés y en francés, en formato CSV.[131]
- Base de datos sobre productos farmacéuticos, principalmente en formato JSON y XML, sobre diferentes aspectos relacionados con los medicamentos.[132]
- Base de datos de productos naturales homologados para uso sanitario, con informaciones en conjuntos de datos separados sobre la lista de ingredientes, nombre de productos, objetivos de productos, entre otros, en diversos formatos y de carácter bilingüe.[133]
- Lista de instrumental médico homologado en vigor, con información específica sobre los productos homologados en vigor o anteriormente homologados, con ficheros en HTML, JSON y XML, y de carácter bilingüe.[134]
- Base de datos Mon Guide Alimentaire, que recoge cantidades y tipos de alimentos de los cuatro grupos alimenticios, en formatos CSV, JSON y en archivos bilingües.[135]

Fuera del portal de carácter oficial podemos encontrar algunos conjuntos de datos relevantes, susceptibles de ser convertidos en RL, como los siguientes:

- Canal de vídeos del Ministerio de la Santé et des Services sociaux de Canadá, con algunos vídeos promocionales sobre salud, que, además, incluyen subtítulos (por tanto, punto fuerte para el desarrollo de algunos tipos de RL).[136]

[131] https://ouvert.canada.ca/data/fr/dataset/1674baf5-6822-4788-afb3-f8659531d7c0
[132] https://ouvert.canada.ca/data/fr/dataset/bf55e42a-63cb-4556-bfd8-44f26e5a36fe
[133] https://ouvert.canada.ca/data/fr/dataset/ef546c83-43a8-4404-943e-ab324164eeb3
[134] https://ouvert.canada.ca/data/fr/dataset/c801a084-210b-4cd2-8513-26a00b66eb6f
[135] https://ouvert.canada.ca/data/fr/dataset/e5f4a98e-0ccf-4e5e-9912-d308b46c5a7f
[136] https://www.youtube.com/channel/UCixMqPbpL6dCQqHxDbgoUfQ

- Algunas guías publicadas en PDF, en varios idiomas, del Ministerio de sanidad.[137] En esta web, bajo diferentes rúbricas, podemos tener acceso a publicaciones en PDF o HTML de diferentes organismos, con los que desarrollar corpus textuales de un ámbito específico.

En el ámbito global europeo, es posible acceder a documentos de ensayos clínicos en el Portal europeo de registro de ensayos clínicos.[138] No obstante, los textos se ofrecen prácticamente en inglés, y el buscador[139] aún no contiene información exhaustiva, ya que esta depende de cada agencia nacional del medicamento.

6.2.3 Justicia

En este dominio, encontramos dentro del **ámbito hispanoamericano** conjuntos de datos destacables, muchos equiparables a los censados en el ámbito peninsular:

- Documentos oficiales equiparables al Boletín Oficial del Estado (en general, es importante subrayar que no es posible la descarga masiva de datos en estos portales):
    - Documentos del Boletín Oficial de la República Argentina[140] e Información Legislativa (HTML y PDF).[141]
    - Documentos del Diario Oficial de la República de Chile[142] en PDF.
    - Documentos de la Imprenta Nacional de Colombia (Diario Oficial en PDF)[143]
    - Documentos del Boletín Oficial de Perú (El Peruano)[144] en PDF.
- Textos (en formato PDF / HTML) de la Biblioteca del Congreso Nacional de Chile,[145] sobre información legal, territorial, historia política y formación cívica, y también audios de algunos contenidos. Licencia Creative Commons Reconocimiento (CC-BY).
- Grabaciones de sesiones plenarias y comités del canal de YouTube (y de TV) de la Cámara de Diputados de Chile.[146] No es posible la descarga masiva de datos, pues los vídeos solo se

[137] *https://www.canada.ca/fr/sante-canada/services/aliments-nutrition/guide-alimentaire-canadien/obtenez-votre-exemplaire.html*
[138] *www.clinicaltrialsregister.eu*
[139] *https://www.clinicaltrialsregister.eu/ctr-search/search*
[140] *www.boletinoficial.gob.ar*
[141] *http://www.infoleg.gob.ar*
[142] *www.diariooficial.interior.gob.cl*
[143] *www.imprenta.gov.co/diariop/diario2.portal*
[144] *https://diariooficial.elperuano.pe/boletinoficial*
[145] *www.bcn.cl*
[146] *http://www.cdtv.cl/* , *https://www.youtube.com/channel/UCYd5k2TyOyOmUJNx0SH17KA/videos*

pueden reproducir en la web, o accediendo mediante el portal de datos abiertos.[147] La licencia es YouTube estándar.

En **Europa**, los textos sobre legislación se pueden obtener en el portal EUR-Lex.[148] A partir de ahí se puede acceder a documentos como el Diario Oficial u otros sobre el Derecho dentro de la UE y la Legislación nacional (por países) en relación a la legislación europea e internacional. Los documentos están en diferentes formatos (generalmente, HTML y PDF). El punto de acceso a las bases de datos con legislación de cada país de la UE se realiza a través del portal N-Lex.[149]

En **Francia,** es posible descargar los textos del boletín del Journal Officiel de la République Française,[150] al que se tiene acceso desde Legifrance.[151] Los documentos datan de 1990 en adelante y se encuentran en varios formatos para su descarga, principalmente, PDF y RDF.
En general, el portal Legifrance[152] es el más interesante en cuanto a datos se refiere, pues en él se centraliza la difusión de todo lo que tiene que ver con el derecho en Francia. Así, entre muchos otros documentos, tenemos acceso a jurisprudencia, en formato RTF,[153] y textos oficiales de diferentes administraciones nacionales e internacionales;[154] o leyes y otros códigos en PDF.[155] En este sitio es interesante también algunos de sus textos y códigos traducidos a distintas lenguas, disponibles en formato PDF (inglés, español, árabe, chino y alemán).[156]

Los datos abiertos del ámbito del derecho se agrupan en el mismo portal,[157] realizado por la DILA (Direction de l'Information Légale et Administrative), aunque nos reenvían en todo momento al portal Legifrance si queremos tener acceso al portal de difusión, o al conjunto de datos abiertos en data.gouv.fr. Este portal es sumamente interesante, pues recoge numerosos recursos del ámbito del derecho, agrupándolos en distintas categorías: datos jurídicos, datos económicos y financieros, datos asociativos (como anuncios relativos a asociaciones o balances anuales de asociaciones, fundaciones, etc.), datos administrativos y datos de debate público (con referencia a información pública, agentes

[147] https://www.camara.cl/camara/opendata.aspx
[148] https://eur-lex.europa.eu/
[149] http://eur-lex.europa.eu/n-lex/index_es
[150] www.journal-officiel.gouv.fr/
[151] www.legifrance.gouv.fr/initRechJO.do
[152] www.legifrance.gouv.fr/
[153] www.legifrance.gouv.fr/initRechJuriConst.do
[154] www.legifrance.gouv.fr/Sites/Juridictions
[155] www.legifrance.gouv.fr/initRechCodeArticle.do
[156] https://www.legifrance.gouv.fr/Traductions/es-Espanol-castellano
[157] http://www.dila.premier-ministre.gouv.fr/repertoire-des-informations-publiques

políticos, síntesis de debates públicos…). La mayoría de los datos que pertenecen a este conjunto son muy interesantes, y serían, sin lugar a dudas, un ejemplo a seguir para la panorámica de datos abiertos de nuestro país. A modo de pincelada, ya que no es el objeto principal de nuestro estudio, destacaremos algunos como:

- La base de datos LEGI con los códigos, leyes y reglamentaciones consolidadas, en formato XML.[158]
- La base de datos del Journal Officiel de la République Française (JORF), por fechas.[159]
- El tesauro de la información pública y de políticas públicas, con un árbol semántico que organiza ambos campos, y que sirve para la indexación de textos, discursos, entrevistas, comunicados y ruedas de prensa, en formato RDF (y formato SKOS).[160]
- El léxico de las palabras clave de la justicia, con la definición de 463 palabras clave del ámbito del derecho, en formato CSV.[161]
- La base de datos CASS con los fallos del Tribunal Supremo, desde 1960, y en formato XML.[162]
- La base de datos KALI con los convenios colectivos a nivel nacional, y también regionales y departamentales, igualmente en formato XML.[163]
- El recurso SARDE, con la indización de búsqueda temática de los textos legislativos y normativos en vigor, que referencia los textos del Boletín oficial (Journal Officiel) y otros boletines dependientes de la DILA. Cuenta con 16.000 descriptores organizados en dos niveles jerárquicos, y se ofrece en formato XML.[164]

En el portal de datos abiertos de **Bélgica** encontramos también toda una serie de datos estadísticos y listas de entidades relativas al ámbito de la justicia. No son demasiados conjuntos de datos, y no parecen de interés como base para la creación de RL. Fuera de él encontramos alguna información más útil para este fin en:

[158] https://www.data.gouv.fr/fr/datasets/legi-codes-lois-et-reglements-consolides/# o https://echanges.dila.gouv.fr/OPENDATA/LEGI/

[159] https://echanges.dila.gouv.fr/OPENDATA/JORF/

[160] www.data.gouv.fr/fr/datasets/thesaurus-information-publique-vie-publique-fr/ , http://jure.juridat.just.fgov.be/JuridatSearchCombined/?lang=fr

[161] https://www.data.gouv.fr/fr/datasets/lexique-mots-cles-de-la-justice-30378293/

[162] https://www.data.gouv.fr/fr/datasets/cass/

[163] www.data.gouv.fr/fr/datasets/kali-conventions-collectives-nationales/

[164] https://www.data.gouv.fr/fr/datasets/sarde-1/

- Los datos referentes al poder judicial se agrupan en la base de datos Juridat,[165] que muestra los autos y decisiones de diferentes tribunales y la jurisprudencia desde 1958 a la actualidad, en holandés, francés y alemán, a través de un buscador, pero solo en formato HTML. Además, posee una licencia específica, no demasiado abierta.
- La legislación consolidada belga y el índice legislativo se pude encontrar en el portal JUSTEL,[166] también a través de un buscador, y solo con datos ofrecidos en HTML.
- Las publicaciones oficiales y públicas difundidas a través del llamado Moniteur Belge,[167] que contiene todos los boletines por fecha y solo con acceso al texto en HTML.

Del mismo modo, en **Canadá**, los datos abiertos referidos al ámbito de la justicia se recogen dentro del portal general de datos abiertos. No obstante, la mayoría carecen de interés desde el punto de vista de generación de RL, puesto que se tratan de datos estadísticos o listas concretas de entidades. De ellos, por volumen de datos y utilidad de los mismos, destacaremos:

- La colección de ficheros XML de las leyes y normativas federales del portal de legislación, ofrecidos por el Ministerio de Justicia, también en formato bilingüe, y de actualización quincenal.[168]
- Los textos de leyes y otras normas en PDF, contenidos en la web de legislación, monolingües y bilingües (francés e inglés).[169]
- El tesauro de los temas o asuntos del gobierno de Canadá, un tesauro bilingüe con la terminología del conjunto de ámbitos tratados en los servicios de información del gobierno de Canadá, elaborado por la Biblioteca y el Archivo Nacional de Canadá. Son alrededor de 4800 términos en francés e inglés, con cerca de 2100 descriptores en ambas lenguas, que se pueden obtener en formato RDF, XML y CSV.[170]

6.2.4 Cultura

En el análisis de distintos recursos del ámbito de la cultura, podemos destacar que las Bibliotecas Nacionales de **países hispanoamericanos** tienen recursos valiosos:

[165] http://jure.juridat.just.fgov.be/JuridatSearchCombined/?lang=fr

[166] www.ejustice.just.fgov.be/cgi_loi/loi.pl

[167] http://www.ejustice.just.fgov.be/cgi/welcome.pl

[168] https://ouvert.canada.ca/data/fr/dataset/ff56de85-f8b9-4719-8dff-ecf362adf0af

[169] http://laws-lois.justice.gc.ca/fra/

[170] https://ouvert.canada.ca/data/fr/dataset/d4a0e406-eea9-41a7-bcae-28c31f3b9c65

- Publicaciones Digitales de la Biblioteca Nacional de Argentina (textos en PDF).[171]
- Documentos de la Biblioteca Digital[172] de la Biblioteca Nacional de Chile.[173]
- Documentos digitales de la Biblioteca Nacional de México[174] y Biblioteca Digital Mexicana[175] (se requiere Adobe Flash Player).
- Libros y documentos en versión digital de la Biblioteca Nacional de Perú[176] (se requiere Adobe Flash Player).[177]

Asimismo, **Francia** nos ofrece, para el sector de la cultura, una serie de datos heterogénea dentro del portal general de datos abiertos del país. De todos ellos, aunque no existe ninguno significativo, podríamos destacar algunas opciones interesantes para el tratamiento de las entidades nombradas, como diferentes listas o inventarios de flora[178] y fauna[179]. Otros datos complementarios de este país que podrían servir de base para herramientas de TL o creación de recursos lingüísticos son:

- La colección de los museos de Francia, extracto de la base de datos Joconde, en formato XLM, CSV y JSON, en los que encontramos una lista de datos como su título, autor, fecha, periodo de creación, materiales técnicos, etc.[180]
- El crawler (o la recopilación) de la web electoral, con archivos entre 2002 y 2017, realizado por la Biblioteca Nacional Francesa (BnF), aunque necesitaría de un mayor tratamiento, ya que hablamos de descriptores de sitios web, url, etc., y requeriría una revisión manual experta para tratar de extraer aquellos datos más interesantes.[181]
- Los metadatos de la colección de la Biblioteca Nacional Francesa (BnF)[182], accesibles desde su portal de datos abiertos y en formato RDF, XML o JSON. Se incluirían aquí unos 2 millones de autores y 8 millones de documentos del catálogo general, el archivo y los manuscritos de la BnF. A través de un buscador, tenemos acceso a los metadatos de los autores, obras, así como

[171] *www.bn.gov.ar/colecciones-digitales/publicaciones*
[172] *www.bibliotecanacionaldigital.cl*
[173] *www.bibliotecanacional.cl*
[174] *http://bnm.unam.mx/*
[175] *http://bdmx.mx*
[176] *www.bnp.gob.pe*
[177] *http://memoriaperuana.bnp.gob.pe/*
[178] *https://www.data.gouv.fr/fr/datasets/inventaire-de-la-flore/*
[179] *https://www.data.gouv.fr/fr/datasets/inventaire-de-la-faune-1/*
[180] *https://www.data.gouv.fr/fr/datasets/collections-des-musees-de-france-extrait-de-la-base-joconde-en-format-xml/*
[181] *https://www.data.gouv.fr/fr/datasets/collectes-du-web-electoral-par-la-bnf/*
[182] *http://data.bnf.fr/*

a todos los libros del catálogo, registros sonoros, o documentos digitalizados que tengan relación con nuestra búsqueda.

- Las colecciones de la BnF,[183] digitalizadas en formato imagen o PDF, en el repositorio digital del portal Gallica, En él, se muestran todas las de las obras del catálogo de la BnF. Podemos destacar de su conjunto los casi 4000 libros en formato EPUB.[184] Este EPUB, además, suele presentar un formato accesible (generalmente, para público con problemas de visión), y aquellos que lo son, se indican claramente en los resultados de búsqueda a través de un determinado icono. Todas las obras (libros, revistas, manuscritos, cartas, etc.) se pueden descargar en PDF, tanto por extractos como de forma completa. Por otra parte, algunas de las obras (solo algunos libros determinados de la colección, unos 25.000) se pueden descargar en diferentes formatos, mucho más reutilizables para la creación de RL, como TXT o XML (el TXT, no obstante, no tiene una calidad muy desarrollada para aquellos documentos más antiguos). Del mismo modo, algunas revistas y prensa se pueden descargar también en formato TXT.
- Las publicaciones en formato PDF del Ministerio de Cultura con guías de arte y patrimonio, vocabularios aceptados actualizados (como el vocabulario de las TIC o de biología), y estudios relacionados con este ámbito.[185]

Siguiendo con el repaso a diferentes países de nuestro entorno, en **Bélgica** encontramos pocos datos relacionados con la cultura dentro del portal general de datos abiertos nacional. Se observan, como en otros países del estudio, listas de monumentos o de asociaciones, locales, guías de turismo, listas de préstamos de bibliotecas, etc., pero nada lo suficiente reseñable para ser tenido en cuenta como ejemplo en este estudio.

Una vez más, en **Canadá** encontramos conjuntos de datos de similares características, muchos de ellos a título estadístico o con datos sobre patrimonio y arte. Destacaremos, a modo de breve pincelada, alguno más interesante para la creación de RL, como:

[183] https://gallica.bnf.fr/accueil/?mode=desktop

[184] https://gallica.bnf.fr/services/engine/search/sru?version=1.2&operation=searchRetrieve&query=dc.format%20adj%20%22epub%22%20and%20provenance%20adj%20%22bnf.fr%22%20sortby%20indexationdate/sort.descending

[185] www.culture.gouv.fr/Espace-documentation?types%5B0%5D=Bases+de+donn%C3%A9es+et+sites+multim%C3%A9dia

- La colección del depósito de obras de arte del Consejo de arte de Canadá, con la lista de obras de arte de dicha colección, en formato XML.[186]
- El banco de datos terminológicos y lingüísticos del gobierno de Canadá (TERMIUM Plus©), donde los datos se agrupan en conjunto o separados por secciones o campos de conocimiento (agricultura, arte, construcción, administración, etc. Dichos datos están disponibles en múltiples lenguas (inglés, francés, portugués y español) y se pueden descargar en formato CSV. En XML, por su parte, encontramos el conjunto de la base de datos terminológica bilingüe.[187]
- Canadian Subject Headings (CSH), una lista de 6.000 títulos o índice de materias en inglés, que abarca temas de diferentes dominios, recolectada y actualizada mensualmente por la Biblioteca y Archivos de Canadá. Se ofrece en formato SKOS/RDF,[188] o XML[189] (es posible también encontrar una edición bilingüe en HTML).[190]

A nivel global europeo, es necesario poner de relieve los textos (en formato PDF, RTF o TXT) y las imágenes de obras colecciones digitales del repositorio Europeana.[191] Europeana contiene, en su mayoría, imágenes de las obras, aunque tiene la posibilidad también de buscar por texto, lo que nos muestra también determinados archivos. Dichos textos se pueden descargar, pero siempre en formato PDF.

Lo interesante de este repositorio son sus **metadatos**, ya que posee un esquema propio muy completo, Europeana Metadata Model, basado en RDF Schema y XML, que podría servir de partida para la creación de muchos recursos lingüísticos, aunque, para ello, fuese necesaria su conversión o procesamiento. Estos metadatos se ofrecen bajo licencia Creative Commons CC0 1.0 (dominio público), y pueden descargarse vía API.[192]

Los datos considerados abiertos de Europeana se ofrecen a través de lo que se conoce por Linked Open Data,[193] y como hemos indicado anteriormente, se puede obtener vía API, después de un contacto con

[186] https://ouvert.canada.ca/data/fr/dataset/bb905c07-36ba-4df7-ad46-2fd6e2c15b23
[187] https://ouvert.canada.ca/data/fr/dataset/94fc74d6-9b9a-4c2e-9c6c-45a5092453aa
[188] http://www.collectionscanada.gc.ca/obj/900/f11/040004/csh.rdf
[189] https://www.bac-lac.gc.ca/eng/services/canadian-subject-headings/Pages/canadian-subject-headings.aspx#csh
[190] https://rvmweb.bibl.ulaval.ca/nouveautes/-/nouveaute/repertoire/rvm/index/nvng
[191] https://www.europeana.eu/portal/es
[192] https://pro.europeana.eu/page/sparql
[193] https://pro.europeana.eu/page/linked-open-data

la entidad. Sería interesante, no obstante, que se pudiese obtener bajo otros formatos y con datos abiertos (no enlazados), de manera que fuese mucho más fácil su reutilización.

También es accesible el modelo de ontología del esquema de metadatos de Europeana en formato OWL, aunque se necesita de una comunicación con Europeana para el uso de su contenido.[194]

7 PLAN DE ACCIÓN PARA LA CREACIÓN DE RECURSOS LINGÜÍSTICOS

7.1 INTRODUCCIÓN

Como en la sección precedente, es necesario realizar un análisis de iniciativas similares al Plan Estratégico en otros países, así como de distintos planes de acción que tengan por objetivo la reutilización de datos públicos. Este análisis se realizó junto al de datos abiertos en otros ámbitos geográficos. Así, es conveniente recordar aquí que se recogieron las iniciativas promovidas en países hispanoamericanos y en Europa, con algunas incursiones en EEUU y Canadá. Otros países importantes en el desarrollo de TL como China, Japón o India quedaron, sin embargo, fuera de este estudio.

Así, es importante destacar que lo más parecido a un plan de acción para la reutilización de datos públicos o de contenidos de páginas webs en administraciones públicas para la creación y mantenimiento de RL es el portal de la European Language Resource Coordination, [4] que proporciona una lista de los recursos existentes en distintos países de Europa, y que, además, tienen carácter abierto.

Del análisis de estas iniciativas, y a modo de resumen previo, se puede concluir que España se encuentra bien situada, y mejor en datos abiertos que en RL abiertos. Es decir, encontramos numerosos datos abiertos, de diferente tipología y envergadura, y bajo distintos formatos reutilizables, pero no tantos recursos lingüísticos listos para su uso en TL como en otros países.

[194] https://github.com/europeana/corelib/blob/master/corelib-edm-definitions/src/main/resources/eu/rdf/edm.owl

7.2 SITUACIÓN DE ESPAÑA EN EL CONTEXTO INTERNACIONAL EN CUANTO A DOTACIÓN DE RECURSOS LINGÜÍSTICOS

España se encuentra entre los países de cabeza en cuanto a la existencia de RL en formato libre y disponible para los investigadores y desarrolladores de PLN y TL. A nivel europeo, Francia y Reino Unido son los referentes en cuanto a RL, aunque como ya se ha comentado en este informe, la Comisión Europea ha advertido del retraso con respecto al inglés y a las compañías americanas, que han *capitalizado* el uso de los Big Data lingüísticos. Esto está permitiendo que las grandes multinacionales tecnológicas como Google, Apple, Microsoft, Amazon o IBM ofrezcan servicios lingüísticos en muy diferentes dominios y lenguas.

Así las cosas, en este contexto, es clara la oportunidad y la necesidad de continuar en la senda de dos de los objetivos centrales del Plan español de TL que se han nombrado ya en este informe:

- **Eje 1: Desarrollo de infraestructuras lingüísticas:** Dirigido a aumentar el número, calidad y disponibilidad de las infraestructuras lingüísticas (recursos, procesadores y campañas de evaluación) de propósito general en español y lenguas cooficiales.

- **Eje 3: La Administración Pública como impulsor de la industria del lenguaje**: Dedicado a la mejora de la calidad y capacidad del servicio público incorporando las tecnologías de procesamiento de lenguaje natural (incluyendo la traducción automática y los sistemas conversacionales), y actuando, además, como tractor de la demanda. Este aspecto apoya, también, la generación, estandarización y difusión de recursos lingüísticos creados en el contexto de la actividad de gestión pública propia de la Administración en el marco de la política de Reutilización de la Información del Sector Público (RISP).

Pese a la apertura y disponibilidad aparente de recursos, es importante destacar que **la comunidad investigadora en PLN en España necesita una RL de calidad y en cantidad suficiente para poder desarrollar aplicaciones competitivas en el mercado internacional**, con la amenaza potencial de que otras empresas de fuera de nuestras fronteras puedan ocupar el espacio del PLN en las lenguas del Estado.

En este punto, y en relación a la competitividad en diferentes mercados y a la creación de aplicaciones que engloban RL, queremos destacar también que con la actual Ley de Datos abiertos financiados por dinero público, los grupos de investigación y empresas que desarrollen RL con proyectos públicos están obligados a dar acceso a dichos RL. Este aspecto, a priori interesante, dado el carácter abierto de los

recursos, puede ser también un factor no incentivador para que los investigadores desarrollen procesos de conversión de datos en RL. Las razones son claras: el coste de creación de RL es muy alto para luego dejar el RL a disposición de la competencia. En ese sentido, es obvio que muchos equipos o grupos de investigación que optan por la financiación pública prefieren desarrollar aplicaciones de PLN usando RL ya existentes a generar nuevos recursos que luego estarán rápida y fácilmente disponibles. Además, la concepción de RL no suele producir, de manera inmediata, la misma cantidad de publicaciones científicas de impacto que el desarrollo de algoritmos o prototipos. El beneficio de los RL se ve, por consiguiente, a medio plazo, cuando se empiezan a utilizar por otros equipos o en competiciones abiertas de evaluación (*shared tasks*). Esto implica que la urgencia por conseguir publicaciones de impacto en el mundo académico no incentive el desarrollo de RL, que, para ser de calidad, requieren tiempo y personal cualificado. Una solución desde la perspectiva académica, como señalan Berden y Langendoen (2010)[195], es un reconocimiento explícito en la evaluación de méritos de investigación del valor de la creación de RL, así como del enriquecimiento y la anotación de los conjuntos de datos. Por todo ello, es necesario que las Administraciones financien adecuadamente la creación de RL de uso libre, insistiendo especialmente en las áreas de interés ya señaladas (inteligencia competitiva, sanidad, justicia, cultura).

Para entender el proceso de creación de RL, veamos un esquema de las fases de su desarrollo:

1. **Diseño del RL**: estructura, muestreo, análisis de cuestiones de propiedad intelectual, niveles de anotación, herramientas de anotación, formatos, metadatos, acceso a contenido y datos.
2. **Recogida de los datos**: recolección, limpieza del ruido y normalización.
3. **Anotación de los datos:** uso, adaptación y desarrollo de anotadores por niveles lingüísticos; adaptación al dominio o la lengua.
4. **Revisión de los datos anotados:** control de calidad de la anotación (acuerdo entre anotadores), creación de los estándares de referencia (*gold standard*).
5. **Gestión de los datos:** creación del portal de acceso libre (p. ej., en página web propia o repositorio tipo github), mantenimiento del RL.

Además, hay que tener en cuenta que **la utilidad de un RL** se mide por varios factores, todos ellos relevantes:

[195] *Bender, E. & Langendoen, T. (2010): "Computational Linguistics in Support of Linguistic Theory". Linguistic Issues in Language Technology. 3 (2), págs 1-31. Accesible en https://journals.linguisticsociety.org/elanguage/lilt/article/view/661.html.*

- **Calidad de los datos**: relevantes para el dominio y la aplicación, limpios de ruido, variados y representativos.
- **Cantidad de los datos**: mejor cuantos más datos y más variados, criterios esenciales para el aprendizaje automático.
- **Disponibilidad de los datos**: acceso y documentación completa, para que los investigadores interesados puedan usar los RL abiertos de la manera más directa y fácil posible.
- **Estatuto legalidad de uso o propriedad intelectual**: muchos RL pueden haberse implementado con datos abiertos, pero con derechos de propiedad intelectual e industrial vigentes. Por ello, el uso de datos públicos sin estos requisitos es una fuente especialmente interesante para el desarrollo de RL.

Finalmente, es importante incidir en la necesidad del uso de **estándares técnicos en la creación de RL**. Lógicamente, para reutilizar los datos en diferentes aplicaciones y herramientas de PLN, la generación de RL tiene que estar ajustada a formatos y anotaciones usadas por la comunidad PLN. Este objetivo no es sencillo de conseguir, porque los equipos de investigadores y las empresas desarrolladoras tienen fuertes motivaciones para defender sus métodos de anotación frente a los de la competencia. Por ello, las evaluaciones competitivas, donde se mide la actuación y la calidad los diferentes sistemas en una tarea concreta, son un importante factor de normalización. Las competiciones DARPA (diseñadas por el Gobierno americano) son un claro ejemplo de cómo se pueden impulsar estándares para RL. En Europa, por el momento, se cuenta con CLEF, o con las tareas organizadas por la SEPLN y las diferentes Redes temáticas de Tecnologías del Habla[196] y de Tratamiento de Información Multilingüe y Multimodal.[197]

Como ejemplo de características que deben contener los RL para garantizar la interoperabilidad entre datos y procesadores podemos mencionar algunos de los requisitos definidos por CLARIN:

- Codificación de caracteres en UTF-8.
- Segmentación del texto en unidades (palabras, sintagmas, frases o párrafos).
- Lenguajes de marcado para codificar la anotación lingüística: XML, RD, tabular vertical, JSON.
- Código estándar de lengua.
- Metadatos con la descripción de los recursos.
- Formatos estándar para codificar memorias de traducción: TMX y XLIFF.

[196] www.rthabla.es

[197] http://sinai.ujaen.es/timm/

- Formatos estándar para codificar diccionarios y recursos léxicos: LMF y TBX.

7.3 RECOMENDACIONES PARA EL DESARROLLO DE UN PLAN DE ACCIÓN

En la sección 5 se presentaron las conclusiones del análisis de los datos considerados. En esta sección se pretende describir un plan de acción a corto y medio plazo para la conversión y aprovechamiento de los recursos censados y analizados con el objeto de establecer prioridades, así como sugerir nuevos conjuntos de datos y comentar lecciones aprendidas para el desarrollo futuro de mejoras y ampliaciones de RL. Se articula, así, en torno a dos tipos de recomendaciones. Las primeras son de carácter genérico, y las segundas están enfocadas a estrategias concretas sobre diferentes recursos.

7.3.1 Recomendaciones Genéricas

1. Garantizar la disponibilidad y el acceso universal a los datos abiertos para RL en todas las lenguas del Estado a través de un portal común y único. La disparidad de iniciativas en las distintas administraciones públicas ha generado un conjunto de recursos para RL muy disperso. Es necesario que se agrupen en un portal específico de datos abiertos para generar RL. Otra opción es que la tarea se dividiera por dominios temáticos (sanidad, justicia, etc.). Con esta iniciativa se conseguiría que los usuarios de RL tuvieran un punto de acceso fácil y directo, favoreciendo la transparencia y la igualdad de oportunidades a todos los actores. Sin embargo, esta recomendación tiene aspectos complejos de gestión. Por ejemplo, el uso de los metadatos específicos de cada portal, que funcionan con normas y convenciones particulares, hace que la unificación sea difícil y costosa. Además, nos encontramos también con el problema del mantenimiento, una vez que un RL forma parte de este punto de acceso. ¿Quién se hace cargo de su actualización y gestión? ¿El creador original o el portal que lo gestiona? Tal vez, la solución intermedia sería que un portal general recogiera la información sobre los RL disponibles y remitiera a la página del RL concreto, gestionado por sus creadores; o que se incluyera algún tipo de etiqueta dentro de un portal existente que hiciera referencia a su uso como RL o a su posibilidad de conversión a RL. En este sentido, se podría utilizar el Catálogo Nacional de datos abiertos albergado en datos.gob.es como sitio de agregación de este tipo de recursos.

Como referencias internacionales tenemos diferentes instituciones y portales. En Europa, como ya hemos señalado al principio de este estudio, ELDA/ELRA (European Language Distribution/Resource Association) [4,5] son los referentes en distribución de RL, desde hace 20 años. En octubre de 2018, el catálogo de ELDA supera los 1080 RL, divididos entre 588 corpus (escritos, orales y multimedia) y 492 recursos léxico-conceptuales. El español (con 358 RL) es la segunda lengua después del inglés (479 RL) en el catálogo de ELDA, muy por delante del francés (223), el alemán (198) y el chino (90). Esto nos

indica que, por una parte, el español es una lengua muy activa en el desarrollo de RL y TL y, por otra, que ELDA es el gran repositorio de datos lingüísticos para el español en la actualidad. No se puede decir lo mismo de otras lenguas cooficiales del Estado, como el catalán (16) y el euskera (7).

El equivalente americano de ELDA/ELRA es el Linguistic Data Consortium, [9] gestionado por la Universidad de Pennsylvania. Agrupa a un consorcio de universidades, bibliotecas, empresas e instituciones de investigación creado para proporcionar acceso a RL. El LDC es la asociación pionera en el mundo en estos temas (existe desde 1992), y su evolución muestra el avance y la importancia de los RL en la investigación y desarrollo de TL. De hecho, se creó inicialmente como un repositorio de recursos y ha evolucionado hacia una organización que genera y distribuye una gran variedad de RL, muy estrechamente relacionada con la organización de evaluaciones competitivas de TL. Los RL en español no llegan a 70 y, mayoritariamente, son de variantes americanas.

Otra asociación internacional dedicada a los RL es la Open Language Archives Community (OLAC)[198], creada en 2000 por un consorcio internacional, y que en 2016 se unió al Linguistic Linked Open Data Cloud (LLOD)[199]. Ambas asociaciones están orientadas hacia la promoción de estándares y buenas prácticas para la compartición de recursos para la Web Semántica. Dentro de esta comunidad, se ha desarrollado un conjunto de criterios para la publicación de RL que destaca, por ejemplo, que: los datos tienen que estar abiertos con licencias CC (Creative Commons); tienen que estar identificados por una URI y, por lo tanto, tienen que seguir los estándares de la web (HTML, RDF y JSON); y finalmente, que deben contener enlaces de otros recursos, de forma que permitan descubrir nuevos RL. A pesar de su enorme interés e impacto, esta estrategia solo se ha desarrollado para aplicaciones de la web semántica, dejando fuera RL que interesan a las otras aplicaciones de la TL.

En el campo de la Traducción Automática (TA) existe también el repositorio ELRC-SHARE[200], gestionado por la ya mencionada European Language Resource Coordination, que se encarga de mantener y coordinar los RL en todas las lenguas oficiales de la UE para que funcionen en la plataforma CEF eTranslation.[201] Por otra parte, ofrecen el servicio de TA en línea de la Comisión Europea, denominado MT@EC. Este genera traducciones para cualquier lengua oficial de la UE y para todas las administraciones públicas de cualquier país de la Unión. Sin embargo, hay que tener en cuenta que ambos servicios tienen su ámbito de aplicación restringido: no se trata de TA general (como la que

[198] http://www.language-archives.org/index.html
[199] http://linguistic-lod.org/
[200] https://elrc-share.eu/
[201] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

proporcionan Google Translate o DeepL) sino que está adaptada a la terminología específica de las administraciones públicas (por ejemplo, textos legales, administrativos o médicos). Si a ello añadimos que solo están disponibles para administraciones públicas, que no contienen ningún recurso sobre las lenguas cooficiales del Estado[202], y que el grado de estandarización de los RL traductológicos es muy poco uniforme, nos encontramos ante un repositorio limitado para los intereses de la comunidad investigadora y desarrolladora de TL en español, catalán, vasco o gallego.

Siguiendo con este repaso de iniciativas, encontramos de nuevo a CLARIN[203], que como ya indicamos previamente en este estudio, es la infraestructura de investigación europea para recursos y tecnologías lingüísticas. Su público son académicos, investigadores y estudiantes de Humanidades y Ciencias Sociales. La red ofrece acceso en línea a datos y herramientas en formato digital que están distribuidas en distintos centros, de manera interoperable, es decir, que se pueden combinar tanto los datos como las herramientas de diferentes fuentes de manera encadenada para obtener consultas complejas. A pesar de su alcance limitado en temática, CLARIN ofrece buenos ejemplos de estandarización de metadatos, anotaciones y herramientas de consulta y procesamiento. Parecida a CLARIN, tenemos que destacar a META-NET, una red de excelencia similar a la anterior, pero enfocada al desarrollo industrial. Además, es otra red con un portal de recursos abiertos y distribuido (META-SHARE) de interés.[204]

Fuera del ámbito más cercano, en septiembre de 2018, se presentó públicamente el Google Dataset Search, herramienta para "descubrir" conjuntos de datos estructurados en la web. Los requisitos para que Google indexe estos datos pasan por: uno, que estén estructurados; y dos, que se presenten en alguno de estos formatos: JSON-LD, microdatos o RDFa[205]. Además, el vocabulario empleado en la descripción de los metadatos debe basarse en schema.org[206], que reúne un conjunto de términos compartidos por las principales compañías de desarrollo de buscadores (Google, Microsoft, Yahoo y Yandex). Esta propuesta se aleja del alcance de los objetivos de este estudio, pero no podemos dejar de mencionarla, pues muestra las tendencias y acuerdos de las diferentes comunidades de desarrolladores. Más adelante, en el punto 4, ampliaremos la información sobre formatos y estándares.

[202] *En la consulta realizada el 5 de octubre de 2018 solo había 20 recursos disponibles en español. Ninguno en catalán, vasco o gallego. Por contraste, en inglés había 168, 49 en francés, y 45 alemán.*

[203] *https://www.clarin.eu/*

[204] *www.meta-net.eu/meta-share*

[205] *https://developers.google.com/search/docs/guides/intro-structured-data#structured-data-format*

[206] *https://schema.org/*

En resumen, no se pueden proponer estándares para todos los recursos lingüísticos debido a su variada tipología y las características específicas requeridas para cada tipo de utilización. Lo que sí resulta útil es proponer un estándar para información y catalogación de recursos lingüísticos, por ejemplo basado en el Universal Resource Catalog de ELDA.[207]

2. Impulsar la conversión de los millones de páginas digitalizadas en PDF o imagen en texto plano. La mayor parte de los datos lingüísticos en acceso abierto en Internet están en formato no directamente procesable. Se recomienda, pues, una inversión en su tratamiento para convertirlo en texto plano. Esta tarea no es trivial, porque hay que adaptar los actuales modelos de lenguaje de los OCR a nuevos dominios, estructuras documentales y, sobre todo, a variantes gráficas de otras épocas. Pero el impacto en términos de cantidad y variedad de datos lingüísticos previsto es enorme, lo que produciría una emergencia de nuevas aplicaciones e impulsaría los trabajos en Humanidades Digitales. Además, aquellos OCR ya existentes podrían servir para el entrenamiento y mejora de nuevos modelos de reconocimiento, pudiendo usarse así conjuntos de datos de las distintas administraciones. Por otra parte, es imprescindible mencionar que este tipo de conversiones tendrían como objetivo nuevas formas de texto que mejorarían la accesibilidad para usuarios con discapacidad. Ejemplos concretos serían la transformación a formatos interesantes para invidentes o para su uso en sistemas de conversión texto–voz, así como entrenamiento para la subtitulación automática de audios y vídeos.

Las líneas concretas de acción pasarían, por un lado, por convertir progresivamente los documentos en PDF, EPUB o las imágenes digitalizadas a formato TXT, CSV o XML. Es importante recordar que algunos recursos reseñados en este informe, como la BDH, ya proporcionan una primera conversión a texto, pero que todos ellos necesitan una revisión. Ciertas herramientas disponibles proporcionan resultados de calidad creciente; por ejemplo, GROBID[208] permite convertir artículos en PDF al formato XML TEI, y se podría aplicar a artículos con licencia Creative Commons del Instituto de Salud Carlos III o de revistas del repositorio SciELO. Por otro lado, en los organismos productores de recursos de interés y con potencial de uso para PLN, se puede comenzar a distribuir dichos recursos en los formatos idóneos aquí sugeridos, y a la vez, adoptar licencias flexibles de uso. En este sentido, podría ser interesante la creación de un Servicio legal para resolver las dudas de los organismos de cara a la apertura de sus datos

3. Mejorar la visibilidad de los conjuntos de datos en cuanto a su disponibilidad y madurez. Nuestro informe ha revelado la enorme variedad y el diferente estado de madurez de los conjuntos de datos.

[207] http://www.elra.info/en/catalogues/universal-catalogue/
[208] https://github.com/kermitt2/grobid

Una mejor visibilidad permitiría un uso de los datos más certero. Así, siguiendo el modelo del portal de datos abiertos de Francia, se podría hacer una clasificación por código de colores sobre la situación del recurso, gracias al que, sin necesidad de pinchar en el enlace de descarga del recurso o conjunto de datos, se pudiera observar si está disponible (en color verde), temporalmente no disponible (amarillo) o no lo está (en color rojo).

Figura 3: Indicación de disponibilidad de datos abiertos en el portal www.data.gouv.fr

Además, de este mismo portal, podríamos retomar otras formas de visualización de los conjuntos de datos, que podrían ser de ayuda a la hora de su selección como base de futuras aplicaciones. En el portal francés, cada recurso posee un cuadro con información concreta de los mismos, como su licencia, frecuencia de actualización, fecha de última actualización, palabras clave, etc., que, de un simple vistazo, nos permite rápidamente saber mucho sobre ellos. Lo observamos en un ejemplo en la siguiente ilustración:

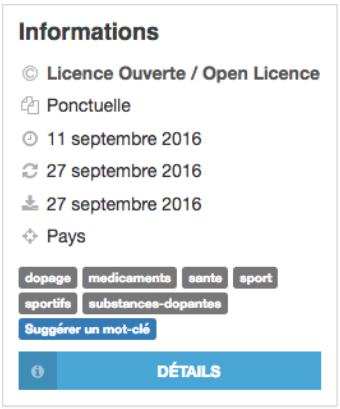


Figura 4: Información detallada sobre los datos abiertos en el portal www.data.gouv.fr

Un esquema similar, que facilita de nuevo la reutilización de los datos, lo encontramos en el portal de datos abiertos del Gobierno de Estados Unidos (http://data.gov), en el que se incluyen palabras clave asociadas al recurso que facilitan mucho su búsqueda, o las búsquedas de conjuntos de datos asociados a una determinada palabra clave, como se puede ver en el siguiente ejemplo de palabras clave asociadas al recurso "U.S. Chronic Disease Indicators (CDI)":

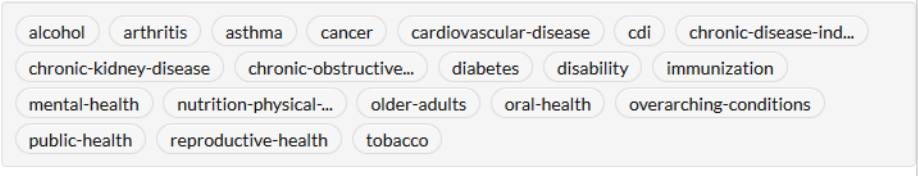


Figura 5: Palabras clave asociadas al recurso "U.S. Chronic Disease Indicators (CDI)" en el portal de datos abiertos de Estados Unidos http://data.gov

Además, en el caso del portal de datos abiertos del Gobierno de Estados Unidos, en la ficha de los datos se incluyen directamente un buen número de metadatos adicionales, como aparece en la Figura 6 para el mismo recurso:

Resource Type	Dataset
Metadata Created Date	April 5, 2018
Metadata Updated Date	August 20, 2018
Publisher	Centers for Disease Control and Prevention
Unique Identifier	https://data.cdc.gov/api/views/g4ie-h725
Maintainer	DPH Public Inquiries
Maintainer Email	PublicInquiriesDPH@cdc.gov
Public Access Level	public
Bureau Code	009:20
Metadata Context	https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld
Metadata Catalog ID	https://healthdata.gov/data.json
Schema Version	https://project-open-data.cio.gov/v1.1/schema
Catalog Describedby	https://project-open-data.cio.gov/v1.1/schema/catalog.json
Data Dictionary	https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725
Harvest Object Id	3fc0fa5f-195d-4e66-8ff1-b94216fc523a
Harvest Source Id	15c538b5-31a0-474e-8ba5-c85ee421cb4d
Harvest Source Title	Healthdata.gov
License	http://opendefinition.org/licenses/odc-odbl/
Data Last Modified	2018-07-09
Program Code	009:020
Related Documents	https://www.cdc.gov/mmwr/pdf/rr/rr6401.pdf
Source Datajson Identifier	True
Source Hash	61b4f1830dd19a7c515c1ae0efc7ee450777bc64
Source Schema Version	1.1

Figura 6: Metadatos asociados al recurso "U.S. Chronic Disease Indicators (CDI)" en el portal de datos abiertos de Estados Unidos http://data.gov

Ambos ejemplos de portales de datos abiertos incluyen informaciones muy relevantes como el conjunto de palabras clave, la situación del recurso y ciertos metadatos asociados al conjunto de datos, que sería muy importante considerar para futuros desarrollos de sitios genéricos, y que el portal web de datos abiertos del Gobierno de España (http://datos.gob.es) todavía no incluye.

4. **Facilitar la descarga masiva de grandes ficheros en formatos apropiados (texto plano, XML, CSV, JSON, RDF).** Como se ha puesto de manifiesto en este estudio, algunos repositorios contienen grandes colecciones de datos que se pueden descargar, pero solo de manera manual y descargando muchos ficheros pequeños que únicamente son accesibles mediante la exploración de páginas web, muchas veces, de navegación compleja. Esto consume mucho tiempo y, además, no es exhaustivo. Se propone, por tanto, que los datos que estén en abierto se puedan descargar de manera directa y, siempre que sea posible, agrupados en grandes ficheros que contengan toda una colección o partes sustanciales (ej. datos de un mes o un año) de la colección. El mejor ejemplo de esta característica son las memorias de traducción de la DG de Traducción de la Comisión Europea, uno de los RL más descargados y usados en PLN. En España, el catálogo de autores, materias, títulos de la BNE o el Nomenclátor de prescripción del CIMA son excelentes recursos que cuenta ya con esta aproximación. De hecho, no podemos dejar de destacar que, en enero de 2019, la Biblioteca Nacional de España ha hecho un esfuerzo muy considerable para poner a disposición pública y gratuita de los recursos de la Biblioteca Digital Hispánica, en diferentes formatos y con una conversión de la imagen a texto mediante OCR.

En relación con lo anterior, la actualización periódica de los RL descargables también es un aspecto esencial, así como la posibilidad de contar con un formato procesable directamente por las aplicaciones de PLN (en general, CSV, XML, JSON, RDF).

Como ejemplo de características que deben contener los RL para garantizar la interoperabilidad entre datos y procesadores podemos mencionar algunos de los requisitos definidos por CLARIN:

- Codificación de caracteres en UTF-8.
- Segmentación del texto en unidades (palabras, sintagmas, frases o párrafos).
- Lenguajes de marcado para codificar la anotación lingüística: XML, RD, tabular vertical, JSON.
- Código estándar de lengua.
- Metadatos con la descripción de los recursos.
- Formatos estándar para codificar memorias de traducción: TMX y XLIFF.
- Formatos estándar para codificar diccionarios y recursos léxicos: LMF y TBX.

En este sentido, se puede sugerir que, de los formatos interesantes para el PLN, el más básico es el texto plano en codificación UTF-8. Es el punto de acceso inicial al procesamiento lingüístico. Por tanto, los documentos en formatos como PDF o Word tienen que ser convertidos a TXT para que puedan ser procesados después por un sistema PLN. Un escalón más allá en la anotación es CSV, pues permite el acceso a los datos clasificados, pero sin una anotación específica. De esta manera, el usuario de los datos que trabaja con CSV puede elegir la codificación y la anotación que mejor se ajuste a sus

herramientas de procesamiento. De la misma manera, cabe destacar la importancia de emplear estándares de codificación de caracteres uniformes y de alta cobertura; en concreto, y como ya hemos señalado, UTF-8.

Continuando con los formatos, XML es el más estructurado de ellos, pero requiere una clara documentación (DTD)[209]. Muchas veces, el procesamiento de XML no es fácil para las aplicaciones de PLN, dado que se requieren librerías específicas para leer y extraer el contenido de los elementos (*parsing*). Por tanto, solo se recomienda la preferencia por XML en las aplicaciones donde esté ya muy estandarizado, como memorias de traducción (TMX) o bases de datos terminológicas (TBX), o donde se pueda obtener, junto con los datos, la DTD del archivo.

En cuanto a JSON, se queda en un punto intermedio entre ambos formatos. Por un lado, es una alternativa a XML por su mayor facilidad para acceder al contenido de elementos específicos en la jerarquía de datos, especialmente cuando se procesan altos volúmenes de información. Por otro, es un formato recomendable para aplicaciones donde se emplee habitualmente javascript, lo que es idóneo en servicios web; y se integra fácilmente con python. No obstante, JSON es el formato preferido por Google Dataset Search.

Otro de los formatos más estandarizado es RDF (Resource Description Framework), que es un modelo para intercambiar datos en la web semántica. Originariamente se diseñó para ser un estándar de codificación de metadatos dentro de XML, pero posteriormente evolucionó a una manera de codificar información sobre relaciones entre entidades del mundo real. Su utilidad principal es la posibilidad de conectar y trabajar con información distribuida en diferentes fuentes. Por todo ello, es muy popular entre la comunidad de datos abiertos entrelazados (LLOD). Como característica esencial, podemos destacar que emplea URIs (identificador de recursos uniforme) para identificar de forma unívoca recursos en una red.

Para el caso particular de los datos multimedia (audio / vídeo) existen herramientas de conversión muy potentes (por ejemplo, ffmpeg), por lo que el formato del audio/vídeo rara vez supone un problema añadido. En cualquier caso, sería aconsejable usar un formato extendido y eficiente en la compresión (como MP3 para audio y MP4 para vídeo) a fin de facilitar la descarga eficiente. En cuanto a los datos

[209] *En Francia, para muchos conjuntos de datos en XML se puede descargar también la DTD. Por ejemplo: Référentiel de DTD: DTD LEGIFRANC, que es una DTD común a todas las bases de datos jurídicas (https://www.data.gouv.fr/fr/datasets/legi-codes-lois-et-reglements-consolides/).*

asociados al audio/vídeo (anotaciones), es válido lo comentado anteriormente para los datos textuales, siendo deseables formatos como TXT, CSV o XML.

Tipo de recurso	Formato recomendado	Formatos adecuados
Corpus textuales	Anotación en XML o TXT en codificación UTF-8	JSON, CSV; no es conveniente PDF
Corpus de audios:	WAV 16 bits, 16 KHz. (voz) o 44.1 KHz (música, audio)	FLAC; MP3 (de alta calidad); otros formatos convertibles (con posible pérdida de calidad)
Corpus de vídeos	MPEG-4 (MP4) de alta calidad	H.264; cualquier otro formato de alta calidad convertible
Corpus memorias de traducción	TMX	CSV
Entidades nombradas y recursos léxicos	Anotación en XML o TXT en codificación UTF-8	JSON, CSV, RDF

Tabla 29. Formatos recomendados para cada tipo de recurso

5. Creación de recursos anotados:

En el apartado anterior solo se han tratado los formatos más usados, sin entrar en modelos de etiquetado de información lingüística. La evolución de las TL ha demostrado a lo largo del tiempo que las normas, guías y esquemas de anotación se suceden unas a otras. Esto es debido a que las teorías subyacentes, los conceptos y los usos cambian con los años. Además, la comunidad investigadora no acepta de manera mayoritaria ningún modelo de anotación, ni siquiera un conjunto de etiquetas. Pensemos en los múltiples etiquetarios de PoS (categorías sintácticas) o de maneras de tratar las MWE (*multiword expressions*, expresiones multipalabra o pluriverbales). Los autores de este informe son partidarios de dejar a la libertad del investigador la elección de las etiquetas y de los criterios de asignación de estas. La uniformidad forzada daría como resultado la imposibilidad de comprobar la validez de representaciones alternativas al modelo de anotación estandarizado. Sin embargo, un requisito obligado de cualquier RL anotado es que aporte la documentación exhaustiva del conjunto de etiquetas y de los criterios de asignación. De esta manera, cualquier usuario del RL podrá conocer el alcance descriptivo del recurso y realizar una nueva anotación si no le satisface.

Lo mismo ocurre con los recursos derivados de RL, es decir, aquellos que se han generado a partir de datos o recursos lingüísticos previos. Nos referimos a listados de palabras con frecuencias (como las

palabras del CREA)[210] o modelos de n-gramas (como el Ngram Viewer[211] basado en Google Books). También, más recientemente, los denominados *word embeddings,* modelos de representación de palabras en espacios vectoriales, entrenados con algoritmos neuronales y/o métodos distribucionales a partir de enormes colecciones de datos; por ejemplo, Google[212] distribuye embeddings entrenados con Word2Vec, y existen otros entrenados mediante fastText[213] o GloVE.[214] Todos estos recursos derivados son muy útiles para el PLN estadístico o de aprendizaje automático, y deberían estar claramente identificados en el portal del RL propuesto aquí.

6. Facilitar herramientas:

Enlazando con el apartado anterior, nos referiremos a la necesidad de contar con herramientas específicas para preparar los RL, entre otros:

- Herramientas de segmentación de oraciones y secciones (ej. cabecera del documento).
- Programas de segmentación de palabras (tokenización).
- Herramientas de normalización, desambiguación y expansión de abreviaturas y acrónimos (especialmente necesarios en el dominio técnico y biomédico).
- Herramientas de anonimización de datos personales.
- Anotadores morfológicos (*Pos taggers*) y sintácticos (*parsers*), y de sintagmas (*chunkers*).
- Reconocedores de entidades, especialmente, para demarcar sus límites, y con convenciones como el uso de offsets o etiquetas BIO (*Beginning, Inside, Outside*).

Es importante insistir en que, aunque hay numerosas aplicaciones disponibles, cada tipo de texto puede requerir un formateo previo especial. Por ejemplo, un corpus de informes médicos tendrá unas abreviaturas y convenciones de nombres de proteínas o medicamentos muy distinto del equivalente en un corpus de leyes, de patentes, o de textos históricos. Por ello, se recomienda que el portal de RL incluya también estos programas de uso gratuito para su procesamiento previo.

7. Proporcionar la transcripción de ficheros multimedia. En las cadenas públicas de radio y televisión, así como en las grabaciones de vistas orales de los juicios, existe una enorme cantidad de datos para ser utilizados en procesamiento de habla de todos los dominios estudiados (salud, inteligencia

[210] http://corpus.rae.es/lfrecuencias.html
[211] https://books.google.com/ngrams
[212] https://code.google.com/archive/p/word2vec/
[213] https://fasttext.cc/
[214] https://nlp.stanford.edu/projects/glove/

competitiva o justicia). Para ello, sería muy conveniente proporcionar una transcripción automática o semiautomática de los contenidos, y que fuera descargable fácilmente, para facilitar su procesamiento posterior. Como es sabido, la transcripción manual de todos los contenidos sería inviable por tiempo y coste. Sin embargo, se puede partir de una transcripción manual (y, por tanto, de calidad) de una pequeña parte de los contenidos y emplear esa transcripción manual para adaptar un sistema de transcripción a estos datos, de modo que se pueda obtener una primera transcripción automática de todos los contenidos similares. Posteriormente, esta transcripción automática se puede revisar manualmente, para lograr una mejora progresiva del sistema de transcripción. Es decir, que se vaya entrenando y mejorando con las sucesivas transcripciones realizadas.

8. Estimular la adopción de licencias de libre uso y acceso a los datos. El empleo de licencias de tipo Creative Commons resulta muy práctico y, en términos generales, agiliza la obtención de los RL, así como su mención explícita y fácilmente visible en los repositorios de datos. Actualmente, el uso de algunos recursos de alto potencial o de valor competitivo requiere contactar con las entidades correspondientes o citar expresamente la fuente de datos, o ni siquiera lo anterior, pues es posible que detenten derechos protegidos.

En este sentido, es preciso hacer hincapié en todos los niveles de la administración en la necesidad de tener en cuenta los aspectos legales sobre los conjuntos de datos para su reutilización, concienciando de que la mera puesta a disposición del público no implica de por sí que los datos puedan ser empleados y reutilizados sin problemas. Un ejemplo paradigmático, con el que hemos tropezado en numerosas ocasiones en la realización de este estudio, lo encontramos en la publicación de contenidos de vídeo en YouTube. Muchas administraciones de las Comunidades Autónomas y locales publican vídeos en YouTube, que siendo producidos por las propias administraciones podrían publicarse (también en YouTube) como datos abiertos sin restricciones bajo una licencia Creative Commons. Sin embargo, en la mayoría de los casos, se publican bajo la "Licencia Estándar de YouTube", que, básicamente, permite la reproducción del vídeo en YouTube, pero no la descarga, ni la reutilización del contenido, salvo que se obtenga el permiso explícito del creador del vídeo. Dado que el contenido ha sido generado por una administración pública, que debería velar por la posibilidad de reutilizar los datos, entendemos que, probablemente, se deba a que la Licencia Estándar de YouTube es la licencia por defecto a la hora de publicar un vídeo en YouTube, y que, por tanto, no se ha prestado atención a la posibilidad de reutilización de estos datos a la hora de publicarlos.

9. Estimular la reutilización de los datos mediante la organización de competiciones tecnológicas basadas en los mismos. El DARPA (Defense Advanced Resarch Projects Agency), en colaboración con

el NIST (National Institute of Standards and Technology), ambos de Estados Unidos, llevan más de tres décadas organizando evaluaciones tecnológicas competitivas en áreas de interés para el Gobierno americano, consiguiendo focalizar la atención de los investigadores de medio mundo en los problemas de investigación que quieren resolver. Para conseguirlo, han generado los conjuntos de datos necesarios para una evaluación, establecido los estándares de evaluación comparables para todos los participantes, y, finalmente, evaluado los resultados de los participantes estableciendo una clasificación de los mismos. Posteriormente, se suele organizar un workshop/seminario en el que se comparten y comparan los avances tecnológicos.

La parte más costosa para la preparación de una evaluación consiste en obtener los datos y generar los metadatos necesarios para la misma. Pero en el contexto de este informe, los conjuntos de datos ya existen, por lo que la realización de una evaluación requiere, fundamentalmente, generar los metadatos necesarios para llevar a cabo la misma.

En Europa se están empezando también a crear iniciativas similares. Por ejemplo, AIRBUS ha lanzado recientemente una evaluación competitiva en el ámbito del reconocimiento de voz "AIRBUS Air Traffic Control (ATC) Automatic Speech Recognition (ASR) Challenge"[215]. En España no nos quedamos atrás y se están gestando iniciativas significativas para este ámbito. Precisamente, durante la elaboración de este informe, se está llevando a cabo una campaña de evaluación de varias tecnologías (reconocimiento de voz, búsqueda de palabras clave, segmentación de locutores) sobre uno de los conjuntos de datos encontrados como prioritarios en este estudio: las Evaluaciones ALBAYZIN 2018[216] que, en la edición de este año, incluyen datos de RTVE. Igualmente, las recientes campañas de evaluación IberEval[217] han abordado tareas de PLN enfocadas al análisis de redes sociales o textos biomédicos.

Las iniciativas de este tipo logran hacer avanzar una tecnología en concreto y, sobre todo, consiguen que la tecnología se adapte al conjunto de datos que se desee procesar. Por ejemplo, este mismo esquema se podría emplear para hacer avanzar la tecnología de reconocimiento de caracteres (OCR) para digitalizar textos históricos de la Biblioteca Nacional de España. De este modo, se conseguiría avanzar hacia una solución de OCR que quizás no sea la mejor en general, pero que, desde luego, se

[215] https://aigym.airbus.com
[216] http://iberspeech2018.talp.cat/index.php/albayzin-evaluation-challenges/
[217] https://sites.google.com/view/ibereval-2018

podría adaptar mucho a la problemática concreta que se quiere afrontar (la digitalización de textos históricos).

Por todo ello, creemos que las competiciones tecnológicas pueden tener un papel protagonista en la reutilización de datos, y también como estrategia para afrontar problemas complejos en los que la adaptación a los datos concretos sea esencial, como lo son la mayor parte de los grandes retos tecnológicos en los ámbitos del procesamiento del lenguaje humano, en cualquiera de sus modalidades.

10. Facilitar el acceso a capacidad de cómputo y almacenamiento. La evolución de la tecnología en los ámbitos del procesamiento del lenguaje humano en la actualidad se apoya, cada vez más, en la disponibilidad de grandes cantidades de datos y de acceso a grandes capacidades de cálculo. En muchos casos, tareas como la participación en una evaluación tecnológica como las que se comentaban en el punto anterior requieren el manejo de grandes cantidades de datos y de ingentes capacidades de cálculo, a los que solo pueden acceder equipos de investigación consolidados. En otros casos, los conjuntos de datos son tan grandes de por sí, que, simplemente, descargar dichos datos resulta una tarea ardua. Por ejemplo, el archivo de RTVE contiene más de 10.000 horas de vídeo solo en Telediarios, un conjunto que, obviamente, no es fácil de tratar con medios computacionales sencillos. De igual modo, el entrenamiento y desarrollo de los recientes modelos de aprendizaje basados en redes neuronales requiere unidades de cálculo con capacidad por encima de las unidades centrales de procesamiento (CPU) tradicionales. Para procesar eficientemente datos que superan las centenas de millones de palabras, se recurre a unidades de procesamiento gráfico (GPU). El coste de inversión en GPUs no está aún al alcance de los grupos de investigación o PyMEs, lo que les aleja del paradigma del *Big Data* lingüístico.

En España, no obstante, contamos con la Red Española de Supercomputación (RES) coordinada por el Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS), que dispone de grandes cantidades de recursos de cálculo y almacenamiento, y que parece el lugar adecuado para, por un lado, almacenar los grandes conjuntos de datos en un repositorio de datos de gran tamaño, y, por otro lado, procesarlos de forma eficiente. De este modo, se reducirían las operaciones de descarga y replicado de los conjuntos de datos que se considerasen enormes, y, por otra parte, estos conjuntos de datos serían accesibles y se conservarían en el que es, muy posiblemente, el mejor entorno en el que procesarlos.

7.3.2 Estrategias concretas para recursos seleccionados

En este apartado vamos a destacar una selección de recursos concretos, que pueden servir de líneas prioritarias en la aplicación del Plan de Acción.

7.3.2.1 Los recursos digitales de la BNE

Este estudio ha analizado tanto la Hemeroteca Digital como la Biblioteca Digital Hispánica. En ambos casos, nos encontramos ante conjuntos de datos de enorme interés lingüístico y cultural, de gran variedad temática, temporal y geográfica. Además, tienen a su favor que están ya digitalizados en PDF, son de acceso libre, no tienen impedimentos legales de copyright, y su uso no requiere la solicitud expresa de un permiso escrito. Su principal inconveniente es el bajo grado de madurez para ser convertidos en RL.

Sin embargo, hay que hacer una importante distinción: la BDH contiene una conversión a TXT mediante OCR de la mayoría de los textos de su colección y, además, recientemente se han incluido todos los metadatos (autor, título, editorial, lengua, etc.) en varios formatos (JSON, CSV, XML), que permiten la descarga masiva. A partir de esos metadatos se puede acceder a la versión digital y a la versión texto desde OCR. Naturalmente, la calidad del texto convertido varía mucho de un documento a otro. Por tanto, es necesaria la revisión de la conversión. En el caso de la HD, esta conversión de imagen a texto no está disponible, ni tampoco la descarga masiva de los metadatos.

La apuesta por la conversión de los recursos digitales de la BNE tiene, por tanto, los siguientes puntos fuertes:

- Es la continuación natural del esfuerzo de digitalización del Patrimonio Bibliográfico español conservado en la BNE, como institución de referencia. También se uniría al esfuerzo nacional y europeo de recopilar y dar acceso a la cultura digital a través de los portales Hispana y Europeana. La conversión de los documentos en PDF en formato texto es el siguiente paso para poner a disposición de los investigadores colecciones de datos lingüísticos muy valiosas. Este paso ya se ha dado con los textos de la BDH, pero no con la Hemeroteca. Además, está por revisar la conversión realizada en los libros digitalizados.
- Los primeros beneficiados de la conversión a RL en formato TXT serían los investigadores en Humanidades Digitales. En los últimos 10 años se ha visto una explosión de nuevas metodologías en las Humanidades, las Ciencias Sociales y las Artes gracias a la aplicación de métodos computacionales. Ejemplos de excelencia son el Lancaster University Digital

Humanities Hub[218] (donde colaboran conjuntamente especialistas en humanidades, lingüística de corpus y PLN) o el Digital Humanities at King's College London[219], departamento que ocupa el primer puesto en el ranquin de este campo en el Reino Unido. En Alemania, el Gobierno Federal lleva una década financiando TextGrid[220], que combina un repositorio de datos y herramientas de código libre, y que ha evolucionado a DARIAH-DE[221], una infraestructura que apoya la investigación y docencia en Humanidades y Cultura mediante el uso de recursos y métodos computacionales. Así, el proyecto DKPro-Wrapper es un ejemplo excelente de cómo emplear diferentes anotadores de PLN para analizar textos literarios.

- La disponibilidad de estos RL textuales supondría una aportación significativa para la comunidad PLN en España. Efectivamente, la variedad de dominios, épocas y lenguas (nacionales y europeas) que contiene la Biblioteca Digital Hispánica permitiría desarrollar y probar nuevos anotadores (morfológicos, sintácticos, semánticos), construir lexicones, generar listados de frecuencias, n-gramas, modelos vectoriales, etc.

En resumen, en la era de los datos, la conversión a texto digital procesable por herramientas PLN supondría un claro empuje, tanto a las TL, como a las disciplinas que estudian y elaboran contenidos culturales.

7.3.2.2 Archivo de RTVE y RTVE a la carta

Este estudio ha analizado tanto la página Web de RTVE a la carta[222] como el archivo de RTVE.[223] En ambos casos nos encontramos con datos (documentos en formato vídeo y audio) de gran interés lingüístico, cultural y tecnológico, con gran variedad temática. En el caso del Archivo de RTVE, se incluyen programas antiguos (desde la década de 1960) que se van digitalizando, y que, en muchos casos, constituyen verdaderas joyas históricas y culturales. En lo referente a RTVE a la carta, se van incluyendo los contenidos más modernos, y destaca por su extensión, conteniendo cerca de 100.000 horas de televisión de alta calidad, entre los que destacamos por su variedad y su interés histórico, los informativos con más de 10.000 programas del Telediario.

Uno de los principales puntos débiles de este recurso es que los programas no se distribuyen con una licencia que permita su reutilización. La licencia solo posibilita reproducir los contenidos en el

[218] http://wp.lancs.ac.uk/dighum/
[219] https://www.kcl.ac.uk/artshums/depts/ddh/index.aspx
[220] https://textgrid.de/en/
[221] https://de.dariah.eu/en/startseite
[222] http://www.rtve.es/alacarta/
[223] http://www.rtve.es/television/archivo/

reproductor de la página web, pero no es posible ni tan siquiera descargarlos. Poner a disposición de la comunidad investigadora y del entorno productivo un recurso tan extenso como este conseguiría grandes avances en tecnologías de recuperación de información y de procesado de información de audio y vídeo, como son los sistemas de reconocimiento de voz, de segmentación de locutores, de resumen de vídeos, etc. Para ello, sería necesario que RTVE permitiese la descarga (preferiblemente masiva, en lugar de vídeo a vídeo) y la reutilización de los contenidos de RTVE a la carta, o, al menos, de una selección lo suficientemente extensa de ellos (sobre los que RTVE tuviese los derechos necesarios). Para desarrollar recursos de TL también sería válido disponer de segmentos desordenados de los mismos.

Una vez conseguido el primer paso de poner los contenidos a disposición de la comunidad, sería necesario un segundo paso, consistente en generar metadatos para poder evaluar las tecnologías que se desease potenciar. Por ejemplo, si RTVE u otra Administración estuviese interesada en potenciar la tecnología de subtitulado automático, sería primordial proporcionar, al menos, un conjunto relativamente pequeño de datos (en este caso, podrían ser suficientes algunas decenas o cientos de horas) a los que se añadirían los metadatos adecuados para entrenar y evaluar dichos sistemas. Estos datos podrían ser la base de una competición tecnológica para hacer avanzar la tecnología, y adaptarla a la tarea específica.

En el caso concreto de RTVE, parece que esa es la línea de ejecución que se persigue, porque, en el momento de realización de este informe, se está llevando a cabo una primera evaluación tecnológica competitiva en los ámbitos del reconocimiento de voz, segmentación de locutores y detección de palabras clave, todas ellas sobre datos facilitados por RTVE (con una licencia negociada *exprofeso* y válida solo temporalmente). Estas evaluaciones, ya citadas anteriormente, se están llevando a cabo dentro de una iniciativa de evaluaciones competitivas mantenida por la Red Temática en Tecnologías del Habla[224] y conocidas como Evaluaciones ALBAYZIN,[225] de carácter bianual desde 2010. En la edición de 2018 todas las evaluaciones ALBAYZIN incluyeron datos de RTVE.[226]

En el caso de RTVE, esta será la primera evaluación competitiva sobre sus datos, pero es necesario hacer hincapié en que no debería ser la única, ya que la tecnología se podrá ir adaptando a la tarea solo con el trabajo de varias ediciones y con la disponibilidad de los datos y metadatos acumulados durante varias ediciones. Es más, la tecnología va evolucionando año a año, por lo que tiene sentido

[224] www.rthabla.es

[225] http://lorien.die.upm.es/~lapiz/rtth/evaluacion.php

[226] http://iberspeech2018.talp.cat/index.php/albayzin-evaluation-challenges/

evaluar periódicamente la tecnología sobre un tipo de datos. Por poner un ejemplo (quizás un poco extremo), el DARPA y el NIST de Estados Unidos han venido organizando evaluaciones de reconocimiento del locutor desde el año 1996 anualmente hasta el año 2006, y a partir de ese momento, las sigue organizando bianualmente, y los resultados de estas evaluaciones han venido indicando de forma muy detallada la evaluación del estado del arte en estas tecnologías.

7.3.2.3 Patentes de ámbito iberoamericano o registradas en la OEPM y PATSTAT

El potencial de los recursos (patentes, modelos de utilidad e informes técnicos) de los repositorios que hemos censado no se corresponde con el grado de madurez óptimo para tareas de PLN. Baste imaginar el grado de mejora de acceso a la información al ingente volumen de patentes que existen: el desarrollo de herramientas de PLN permitiría acelerar el tiempo necesario para su consulta, facilitar el acceso a creadores autónomos sin formación profunda en propiedad intelectual industrial, aumentar el volumen de consulta de datos, y de manera ideal, proporcionar dichas facilidades en varias lenguas. El desarrollo de estas herramientas requiere primero compilar recursos lingüísticos.

El primer obstáculo que sería adecuado solventar es, si fuera necesario, negociar con cada institución la disponibilidad de acceso o distribución de los datos. En el caso de las patentes de ámbito hispanoamericano, nos referimos, en concreto, a documentos proporcionados por el Instituto Nacional de la Propiedad Intelectual de Chile, el Instituto Mexicano de la Propiedad Intelectual de México, la Superintendencia de Industria y Comercio de Colombia, o la corporación PROSUR de los 13 países participantes. En el caso de las patentes españolas, la Oficina Española de Patentes y Marcas permite el acceso y reutilización de los datos, y las correspondientes versiones de documentos en inglés (incluidos en la PATSTAT de la European Patent Office) están sometidas a leyes de propiedad intelectual, aunque se permite su uso citando la fuente. El interés estratégico de dichos recursos nos lleva a sugerir, no obstante, el contacto con cada institución previo a su uso.

El segundo paso sería facilitar la descarga de los documentos, proporcionando conjuntos de datos en ficheros comprimidos (ej. ZIP o RAR), que pueden agruparse, por ejemplo, por año y mes de solicitud de patente, o por área de aplicación según la Clasificación Internacional de Patentes, CIP (en el caso de los documentos indexados según dicho tesauro). Actualmente, los archivos solo se pueden recuperar mediante consulta por palabras clave (ej. número de solicitud, título o país de los autores de la patente) en los buscadores correspondientes (EspaceNet, de ámbito mundial global, o Latipat-EspaceNet, reducido al ámbito iberoamericano). La descarga de documentos, a fecha actual, requiere experiencia en el dominio y un tiempo considerable. En el caso de los documentos que no incluyen

metadatos o no están clasificados conforme al tesauro de la CIP, recomendamos adicionalmente su clasificación para facilitar y homogeneizar el acceso a los mismos.

El tercer paso concierne la conversión de formatos PDF o imágenes escaneadas a formatos óptimos como XML, CSV, JSON o TXT, aplicando técnicas avanzadas de OCR y empleando codificaciones de caracteres homogéneas y estándar (ej. UTF8). Dado que dichas tareas requieren un tiempo importante para llevarse a término, además de personal para revisar la calidad de la conversión, se puede comenzar por un subconjunto que sea prioritario. Se puede atender a su contenido (ámbito de ingeniería industrial, informática, biomedicina, agricultura y ganadería, alimentación, producción textil, etc.) o fecha de solicitud (p. ej. las patentes más recientes de los últimos 10 años).

Alcanzar un nivel más alto de madurez requeriría tareas propias de PLN como segmentación en secciones del documento (cabeceras o pies de página con texto administrativo, campos de datos numéricos, secciones como descripción de modelos, etc.), segmentación en frases y tokenización, reconocimiento de entidades (ej. organismo o empresa depositaria, país, fecha, personas autoras, etc.), y en el caso de emplear textos comparables en español e inglés, alineamiento de segmentos bilingües.

Por último, el acceso a los recursos de calidad óptima debería disponerse en un repositorio de fácil visibilidad (en cada institución o en una web que agrupe todos los archivos), mencionando explícitamente el tipo de licencia (idealmente, Creative Commons).

7.3.2.4 Recursos de dominio médico

En este campo, encontramos recursos de un grado alto de madurez (los bancos de datos de OrphaData y el Nomenclátor de prescripción del CIMA), con un modelo de licencia, distribución y riqueza de anotaciones que contrastan con el resto de recursos. OrphaData, asimismo, es un recurso plurilingüe.

Los recursos con menor madurez a los que nos referimos son las publicaciones científicas (informes, folletos o revistas) del Instituto de Salud Carlos III (ISCIII) o de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), junto a las alertas médicas y farmacológicas que este organismo publica, y a las guías de práctica clínica del portal Guía Salud. Para dichos documentos, se requiere, en primer lugar, la conversión del formato PDF o HTML a formatos más adecuados para tareas de PLN (p. ej., TXT, CSV, XML o JSON), asegurando la adopción de codificaciones de caracteres estándar (p. ej., UTF8). Las guías de práctica clínica de Guía Salud en varias lenguas requerirían, además, alinear los segmentos traducidos en cada lengua para obtener un corpus paralelo o comparable.

Una vez realizadas dichas tareas, se podría aspirar a alcanzar un grado mayor de madurez enriqueciendo los textos con términos clave de terminologías de referencia o códigos de nomenclaturas internacionales (como códigos del CIE-10 o CUIs del UMLS). Por ejemplo, las publicaciones de la AEMPS o el ISCIII pueden enriquecerse con descriptores bibliográficos ya disponibles (MeSH, traducido en español por el IBECS). De igual modo, las alertas farmacéuticas de la AEMPS pueden anotarse con términos relativos a los efectos secundarios, siguiendo la terminología MedDRA, y las sustancias farmacológicas, clases terapéuticas o productos genéricos pueden incluir códigos de la ATC (*Anatomical Therapeutic Classification*). En el caso concreto de las alertas medicamentosas, la anotación de dicha información sería un proceso necesario para automatizar el procesamiento y mejorar el acceso a los datos de farmacovigilancia recogidos mediante el sistema NotificaRAM[227]; extraer información de las alertas publicadas en Internet y redes sociales; y analizar o comparar datos de dichas fuentes.

Sugerimos que la manera idónea de realizar dichas tareas de anotación puede partir de un pequeño conjunto de datos anotado a mano por personas cualificadas (a nivel de entidades o relaciones como "causa"), que puede usarse para entrenar sistemas de anotación basados en modelos de aprendizaje automático, y que, finalmente, se puede emplear y distribuir en competiciones (*shared tasks*) de ámbito nacional.

Por lo que respecta a los vídeos del ISCIII y del portal de TV de gobierno vasco del dominio de salud, nuestras recomendaciones van en la misma línea que lo expuesto para el conjunto documental de RTVE. El potencial de dichos datos para desarrollar reconocedores de habla, sistemas de subtitulado o de conversión de texto a habla del dominio médico se ve frenado por el estado actual de los datos y el acceso a los mismos. Reiteramos, por tanto, la necesidad de convertir los contenidos a formatos idóneos (MP3, WAV, MP4) y permitir la descarga de las grabaciones de vídeo y audio, a ser posible de manera masiva. Asimismo, es conveniente adoptar licencias que faciliten el uso y distribución de datos; por ejemplo, los vídeos del ISCIII tienen derechos de propiedad intelectual y es preciso contactar con la institución. Por último, insistimos en que dichas tareas son solo un primer paso que ha de completarse con tareas de subtitulación (semi)automática, revisión por personal cualificado y enriquecimiento con metadatos.

[227] www.notificaram.es

8 CONCLUSIONES

El análisis de portales web, conjuntos de datos y recursos lingüísticos, tanto nacionales como internacionales, muestra la importancia de la conversión de datos públicos en recursos reutilizables a disposición de investigadores que trabajan en TL en cualquiera de sus modalidades, tanto la comunidad PLN como la de tecnologías del habla. El liderazgo del actual Plan Estratégico es esencial para que los datos públicos sean accesibles universalmente, de una calidad contrastada, en un número suficiente para que puedan aplicarse técnicas de minería de datos y aprendizaje automático, y en formatos aceptados y usados de manera generalizada en la academia y en la industria.

España se encuentra bien situada en el contexto europeo, pero como destaca el mencionado informe CONNECT, las lenguas necesitan una infraestructura para el PLN. Todo desarrollo del PLN depende de esta infraestructura, que es costosa de crear y mantener, pero que es necesaria para cualquier lengua y para diferentes entornos de uso. Nuestro estudio ha demostrado que el país cuenta, al menos, con dos fuentes riquísimas: en el campo textual, los millones de documentos digitalizados de la Biblioteca Nacional de España; en el campo multimedia, los archivos de RTVE. En las áreas de Sanidad, Inteligencia competitiva y Justicia, también existe un enorme volumen de datos en PDF (legislación, patentes o documentos científicos) con alto potencial para convertirse a recurso lingüístico, pero requiere la generación de textos a formatos fácilmente procesables por la comunidad PLN y personal cualificado para su procesamiento.

El *Informe sobre el estado de las tecnologías del lenguaje en España* de 2015 [1] recogía una interesante reflexión sobre la distancia entre el tejido industrial de las PYMES y la investigación académica:

> *Un obstáculo importante para el pleno desarrollo de esta industria es la poca visibilidad que las tecnologías de PLN y TA tienen en el tejido industrial español de las TIC. (...) La falta de demanda de productos mantiene a las empresas españolas del sector en unas dimensiones reducidas (de entre 1 y 10 trabajadores, con alguna excepción), sin conciencia sectorial ni una estructura de representación que agrupe sus intereses.*
> *(...) En la academia, con más de 30 grupos de investigación, la falta de crecimiento del sector causa que la mayor parte de doctores formados en España se vean obligados a buscar oportunidades en multinacionales extranjeras, con la consiguiente pérdida de inversión en formación de especialistas altamente cualificados.*

Los patrimonios documental, jurídico, cultural y lingüístico del país están muy exhaustivamente representados en diferentes conjuntos de datos abiertos y la comunidad I+D está esperando con

fundadas expectativas que esos recursos se conviertan en RL. La disponibilidad de estos recursos debería permitir a la industria española de TL poder competir con las empresas internacionales en un Mercado Único Digital basado en el tratamiento multilingüe. De esta manera, España se podría convertir en un referente en TL a nivel europeo junto a Francia e Reino Unido, si se potencia el desarrollo y las infraestructuras de RL.

9 ANEXO 1: TIPOLOGÍA DE RECURSOS LINGÜÍSTICOS

Sitio Web	Corpus textuales	Corpus multimodales	Memorias de traducción	Entidades nombradas	Recursos léxicos
A					
B					
C					

Tabla 29: Tipología de recursos lingüísticos

10 ANEXO 2: FICHA TÉCNICA UTILIZADA PARA EL CENSO DE RECURSOS

Modelo de ficha técnica utilizada para el censo (los ejemplos entre paréntesis son simplemente ilustrativos, aunque se refieren, en este caso, a RL reales):

1. ***Identificación del recurso.***

- *Nombre:* (ej. C-ORAL-ROM).
- *Clasificación por tipo de recurso: (ej. corpus textual, memoria de traducción...).*
- *Clasificación por número de lenguas:* (ej. monolingüe/bilingüe).
- *Lenguas:* (ej. español, catalán, euskera... así como las posibles variantes del español)
- *Descripción del recurso:* (ej. Forma parte de un corpus multilingüe de lengua espontánea en las cuatro lenguas romance principales: francés, italiano, portugués y español. El proyecto fue financiado por la UE bajo el V Framework Programme (IST-2000-26228).).
- *Fecha de comienzo de creación (del recurso)*: (ej. desde 1998).
- *Fecha de finalización de creación (del recurso):* (ej. 2016).
- *Frecuencia de actualización*: (ej. anual).
- *Fecha de última actualización*: (ej. mayo de 2018).
- *Versión:* (ej. final (2004)).
- *Identificador del recurso (ISLRN, ISSN, ISBN, DOI u otro) y/o URL:* (ej. doi.org/10.1075/scl.15).
- *Tipo de licencia:* (ej. comercial, ELDA; libre con restricciones, Creative Commons, tarifa, si aplica, DPI o IPR).
- *Descarga masiva disponible:* SI/NO.

2. ***Persona de contacto u organización responsable***

- *Nombre y correo electrónico:*
- *Nombre organización (abreviatura, dpto., URL):*

3. ***Creación del recurso***

- *Proveedor y/o creador:* (ej. LLI-UAM).
- *Proyecto(s) financiador(es):* (ej. C-ORAL-ROM).

4. ***Descripción del recurso***

- *Variedad de la lengua (estándar, dialecto, argot, otro).*
- *Niveles de anotación lingüística:* (ej. POS, lematización, prosodia).

- *Conforme a los estándares (EAGLES, PAROLE, CONLL, TMX, etc.).*
- *Tamaño:* (ej. 300.000 palabras, 210 textos).
- *Unidad (términos, entradas, textos, oraciones, otro):* (ej. palabras, enunciados).
- *Formato (CSV, HTM, etc.):* (ej. texto, UTF-8).
- *Dominio (economía, legislación, etc.):* (ej. abierto).
- *Género (crónica, publicidad, oficial, etc.):* (ej. formal, informal, medios de comunicación).
- *Tipo de texto: (académico, blog, etc.):* (ej. habla espontánea).
- *Tipo de documento: (artículo, manual, etc.):* (ej. monólogos y conversaciones).
- *Información adicional* (URL con información relacionada, etc.).

5. *Otros recursos relacionados*

- *Identificación y URL de recursos relacionados.*

6. *Grado de madurez de datos conforme al modelo de la metodología*

- *Necesidades de procesamiento (manual o automático):* bajas/medias/altas (ej. conversión de formatos, alineamientos, anotación, transcripción, verificación, etc.).

7. *Plantilla de evaluación de la madurez de los datos para conversión en recurso lingüístico*

Puntos que considerar para evaluar el grado de madurez	**Valores: -,*, ** - (Desconocido o N/A)**	**Observaciones**
Aspectos técnicos (necesidad de procesamiento manual o automático)		
1. Digitalización (conversión de PDF a formato procesable, OCR a TXT, o procesamiento de formato HTML/DOC a XML…).		
2. Transcripción (ortográfica, fonológica, suprasegmental…).		
3. Alineación vídeo/sonido y texto		
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1).		
5. Anotación morfológica y/o sintáctica.		
6. Anotación de entidades nombradas.		

7. Otros tipos de anotación (semántica, pragmática, palabras clave…).		
8. Revisión de aspectos formales (ortografía, formato de etiquetado, homogeneización de formato, estilo…).		
9. Revisión de contenido (incoherencias, redundancia de datos, mapeo a ontología o base de datos, selección de subconjunto de datos, o tareas que requieran un revisor experto).		
10. Anotación conforme a estándares de la comunidad PLN.		
11. Presencia de metadatos.		
Aspectos legales		
12. Necesidad de anonimización de datos personales.		
13. Necesidad de solicitud de permiso de uso (contactar con los distribuidores, solicitar permiso de grabación o transcripción…).		

Tabla 30: Plantilla para la evaluación de la madurez como RL de un recurso

8. *Posibles aplicaciones del futuro recurso lingüístico*

- *Ejemplo de aplicaciones posibles:* (ej. entrenamiento y evaluación de sistemas de reconocimiento de habla; modelo para el desarrollo de sistemas conversacionales).

  Recomendaciones: (ej. Conversión de pdf a texto mediante OCR y revisión manual…)

11 ANEXO 3: FORMATOS RECOMENDADOS

Tipo de recurso	Formato recomendado	Formatos adecuados
Corpus textuales	Anotación en XML o TXT en codificación UTF-8	JSON, CSV; no es conveniente PDF
Corpus de audios:	WAV 16 bits, 16 KHz. (voz) o 44.1 KHz (música, audio)	FLAC; MP3 (de alta calidad); otros formatos convertibles (con posible pérdida de calidad)
Corpus de vídeos	MPEG-4 (MP4) de alta calidad	H.264; cualquier otro formato de alta calidad convertible
Corpus memorias de traducción	TMX	CSV
Entidades nombradas y recursos léxicos	Anotación en XML o TXT en codificación UTF-8	JSON, CSV, RDF

12 ANEXO 4: RECOMENDACIONES DE ACTUACIÓN

1. **Garantizar la disponibilidad y el acceso universal** a los datos abiertos para RL en todas las lenguas del Estado **a través de un portal común y único**.
2. Impulsar **la conversión de** los millones de **páginas digitalizadas en PDF o imagen en texto plano.**
3. Mejorar **la visibilidad de los conjuntos de datos** en cuanto a su **disponibilidad y madurez**.
4. Facilitar la **descarga masiva de grandes ficheros en formatos apropiados** (texto plano, XML, CSV, JSON, RDF).
5. **Creación de recursos anotados** de utilidad general y de disponibilidad abierta**.**
6. Facilitar **herramientas de conversión de datos abiertos a RL.**
7. Proporcionar la **transcripción de ficheros multimedia.**
8. Estimular la **adopción de licencias de libre uso y acceso a los datos.**
9. Estimular la reutilización de los datos mediante **la organización de competiciones tecnológicas** basadas en los mismos**.**
10. Facilitar el **acceso a capacidad de cómputo y almacenamiento** para grandes volúmenes de datos.
11. Impulsar la **publicación de conjuntos de datos anonimizados**, esencial en los documentos médicos o legales.
12. En los **recursos de traducción**, identificar las **lenguas fuente y meta**, así como el **alineamiento** de las “unidades de traducción”.

13 REFERENCIAS

[1] N. Bel, G. Rigau *et al*., "Informe sobre el estado de las tecnologías del lenguaje en España", España 2015. [https://www.plantl.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf]

Fundación Orange y Arvo Consultores (2014): Datos abiertos en las Comunidades Autónomas y sus mayores ayuntamientos 2014 Resumen ejecutivo [www.proyectosfundacionorange.es/docs/eE2014/Datos_Abiertos_2014_resumen_ejecutivo.pdf]

[2] A. Soroa *et al.,* "Plataformas y Sistemas de procesamiento lingüístico de alto rendimiento", Estudio preparado por la UPV-EHU para el Plan TL, España, 2017. [https://www.plantl.gob.es/tecnologias-lenguaje/actividades/Estudios%20tcnicos%20y%20de%20gobernanza/Estudio%20de%20plataformas%20y%20sistemas%20de%20procesamiento%20ling%C3%BC%C3%ADstico/informe_nlpar.pdf]

[3] Norma ISO/IEC 25012 [http://iso25000.com/index.php/normas-iso-25000/iso-25012]

[4] European Language Resource Coordination. [http://www.lr-coordination.eu]

[5] European Language Resources Association (ELRA). [http://www.elra.info]

[6] European Language Distribution Association (ELDA). [http://www.elda.fr]

[7] CLARIN. [https://www.clarin.eu]

[8] Dublin Core. [https://en.wikipedia.org/wiki/Dublin_Core]

[9] Linguistic Data Consortium. [https://www.ldc.upenn.edu]

[10] META-NET [http://www.meta-net.eu]

[11] G. Aguado de Cea et al., "Inventario de recursos lingüisticos de la Administración Pública para Traducción Automática", Estudio preparado por la UPM para el Plan TL, España, 2016. [https://www.plantl.gob.es/tecnologias-lenguaje/actividades/Estudios%20tcnicos%20y%20de%20gobernanza/Inventario%20de%20recursos%20para%20traducci%C3%B3n%20autom%C3%A1tica/inventario-recursos-traduccion-Retele.pdf]

[12] Zabala Innovation Consulting, "Análisis de la situación del sector de las tecnologías del lenguaje en Europa", Análisis preparado por Zabala Innovation Consulting S.A. para el Plan TL, España, 2018. [https://www.plantl.gob.es/tecnologias-

lenguaje/actividades/Estudios%20del%20sector/Sector%20de%20tecnolog%C3%ADas%20del%20lenguaje%20en%20Europa/informe_final_UE.pdf]

14 GLOSARIO DE SIGLAS Y ACRÓNIMOS

SSCC	Sistemas Conversacionales
AEMPS	Agencia Española de Medicamentos y Productos Sanitarios
AETS	Agencia de Evaluación de Tecnologías Sanitarias
AMPS	Agencia Federal de Medicamentos y Productos Sanitarios de Francia
API	Application Programming Interfaces
ASR	Automatic Speech Recognition
ATC	Anatomical Therapeutic Classification
BDH	Biblioteca Digital Hispánica
BNE	Biblioteca Nacional de España
BNF	Biblioteca Nacional de Francia
BOE	Boletín Oficial del Estado
BOPI	Boletín Oficial de la Propiedad Intelectual
BORM	Boletín Oficial del Registro Mercantil
CC	Creative Commons
CCAA	Comunidades Autónomas
CDI	Chronic Disease Indicators
CEF	Connecting Europe Facility
CENDOJ	Centro de Documentación Judicial
CGPJ	Consejo General del Poder Judicial
CIMA	Centro de Información de Medicamentos
CIP	Clasificación Internacional de Patentes
CLARIN	Common Language Resources and technology INfrastructure
CNTI	Centro Nacional de Tecnologías de la Información de Venezuela
CONACYT	Consejo Nacional de Ciencia Y Tecnología de México
CONLL	Conference on Natural Language Learning
CONNECT	Communications Networks, Content and Technology
CPC	Clasificación Cooperativa de Patentes

CPU	Unidad Central de Procesamiento
CREA	Corpus de Referencia del Español Actual
CSV	Comma-Separated Values (formato de archivo)
CUI	Concept Unique Identifier
DA	Datos Abiertos
DARPA	Defense Advanced Research Projects Agency
DC	Dublin Core
DCI	Denominación Común Internacional
DG	Dirección General
DGLFLF	Dirección General de la Lengua Francesa y de las Lenguas de Francia
DILA	Dirección de Información Legal y Administrativa - Direction de l'Information Légale et Administrative
DOI	Digital Object Identifier
DPI	Derechos de propiedad intelectual
DTD	Definición de Tipo de Documento (Document Type Definition)
ECLA	European Company Lawyers Association
ELDA	European Language Distribution Association
ELI	European Language Infrastructure
ELRA	European Language Resources Association
ENMT	Escuela Nacional de Medicina del Trabajo
EPATRAS	European Patent Translation System
EPO	European Patent Office
EPRSC	Engineering and Physical Sciences Research Council
EPUB	Electronic PUBlication (formato de archivo)
ET	Entregable Técnico
EUIPO	European Union Intellectual Property Office
GPC	Guías de Práctica Clínica
GPU	Unidad de Procesamiento Gráfico
GROBID	GeneRation Of Bibliographic Data
HLT	Human Language Technology
HTML	HyperText Markup Language
IACS	Instituto Aragonés de Ciencias de Salud
IBECS	Índice Bibliográfico Español en Ciencias de la Salud

IBM	International Business Machines Corporation
ILDA	Iniciativa Latinoamericana por los Datos Abiertos
INALCO	Institut National des Langues et Civilisations Orientales
INE	Instituto Nacional de Estadística
IPR	Intellectual Property Rights
ISCIII	Instituto de Salud Carlos III
ISLRN	International Standard Language Resource Number
IVAP	Instituto Vasco de Administración Pública - Herri Arduralaritzaren Euskal Erakundea
IZO	Servicio Oficial de Traductores del Gobierno Vasco
JORF	Journal Officiel de la République Française
JSON	JavaScript Object Notation
LDC	Linguistic Data Consortium
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LOPD	Ley Orgánica de Protección de Datos
LREC	Language Resources and Evaluation Conference
MEDDRA	MEDical Dictionary for Regulatory Activities
MINECO	Ministerio de Economía y Empresa
MPR	Ministerio de Presidencia, Relaciones con las Cortes e Igualdad
MWE	MultiWord Expressions
NIST	National Institute of Standards and Technology
OAI	Open Archives Initiative
OCR	Optical Character Recognition
OEPM	Oficina Española de Patentes y Marcas
OLAC	Open Language Archives Community
OPIC	Oficina de la Propiedad Intelectual de Canadá
OWL	Web Ontology Language
PDF	Portable Document Format
PLN	Procesamiento del Lenguaje Natural
PoS	Part Of Speech
PYME	Pequeña Y Mediana Empresa
RAR	Roshal Archive (archivo con formato de compresión sin pérdida)
RD	Real Decreto

RDF	Resource Description Framework
RES	Red Española de Supercomputación
REST	Representational State Transfer (API-REST)
ReTeLe	Red de Tecnologías del Lenguaje
RGPD	Reglamento General de Protección de Datos
RISP	Reutilización de la Información del Sector Público
RL	Recurso(s) lingüístico(s)
RTVE	Radio Televisión Española
SciELO	Scientific Electronic Library Online
SEAD	Secretaría de Estado para el Avance Digital
TA	Traducción Automática
TBX	TermBase Exchange
TED	Tenders Electronic Daily
TEI	Text Encoding Initiative
TERMCAT	Centro de Terminología de la Lengua Catalana
TL	Tecnologías del Lenguaje
TMX	Translation Memory eXchange
TTS	Text-to-Speech (conversión de texto a voz)
TXT	Archivo de Texto simple o de texto plano
UE	Unión Europea
UMLS	Unified Medical Language System
URI	Identificador de Recursos Uniforme (Uniform Resource Identifier)
URL	Localizador de Recursos Uniforme (Uniform Resource Locator)
UTF-8	8-bit Unicode Transformation Format
WAV	Waveform Audio Format
XML	eXtended Markup Language
ZIP	Formato de compresión sin pérdida ZIP