Estudio sobre datos reutilizables como recursos lingüísticos

(M-009/18-SP)

Antonio Moreno Sandoval,[1] Doroteo Torre Toledano,[1]
Ana Valverde Mateos,[2] Leonardo Campillos Llanos[1]

[1] Universidad Autónoma de Madrid
[2] Unidad de Terminología Médica, Real Academia Nacional de Medicina

25 de septiembre del 2019 – Bilbao – InfoDay

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA
SECRETARÍA DE ESTADO PARA EL AVANCE DIGITAL

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Esquema de la presentación

1. Objetivos del estudio
2. Fases del estudio
3. Ficha de recogida de la información y descripción de conjuntos de datos
4. Metodología para la valoración de la madurez
5. Resultados de la valoración
6. Comparación con otros países
7. Recomendaciones genéricas
8. Entregables

Objetivos del estudio

Objetivos del estudio

- **Identificar y censar conjuntos de datos y documentos susceptibles de reutilización** como **recursos lingüísticos**, publicados en los sitios web de organismos de la Administración General del Estado (AGE), de las Comunidades Autónomas (CCAA), Entidades Locales (EELL) y universidades públicas españolas
- **Elaborar una propuesta de plan de acción** que establezca prioridades y medidas para la conversión de los conjuntos de datos identificados en recursos lingüísticos.

Diferencia entre datos abiertos y recursos lingüísticos

- 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: corpus escritos y orales, lexicones, ontologías, listas de entidades, etc.

- Por **datos abiertos** (DA) se entiende **información de todo tipo generada en el sector público** y que es **susceptible de ser reutilizada**.
  Para que los datos abiertos se conviertan en un recurso lingüístico es necesario recopilarlos y adaptarlos a formatos susceptibles de ser utilizados por las aplicaciones de Tecnología Lingüística.

→ El **objetivo** de este Estudio es **localizar y censar dichos repositorios de datos y valorar el grado de madurez** para su conversión en un RL.

Fases del estudio

Fases del estudio

1. **Establecer el plan de acción y el cronograma** (**EG1**).
2. **Crear la metodología** y la **ficha de recogida** de conjuntos de datos. (**ET1**).
3. **Elaborar el censo de conjuntos de datos (ET2):** Identificar **al menos los 20 conjuntos de datos** más valiosos y de calidad (descritos 24 finalmente).
4. **Analizar y revisar** los conjuntos de datos censados**: Modelo de madurez (ET3).**
5. **Proponer o describir un plan de acción a corto y medio plazo (ET5) para la conversión y aprovechamiento** de los recursos censados de cada tipología.
6. **Elaborar los informes finales (EF1; EF2).**

Procedimiento de búsqueda de los datos reutilizables

- **Recoger conjuntos de datos de áreas de interés: Sanidad, Justicia, Inteligencia competitiva** y **Cultura**
- **Atención a las lenguas cooficiales**

- **Lista inicial de 101 conjuntos de datos** abiertos producidos Administraciones Públicas
- Selección de **24 conjuntos de datos**, algunos compuestos por diferentes tipos de datos y madurez.

Ficha de recogida de la información

Apartados de la ficha

- **Identificación del recurso:** nombre, url, clasificación, lenguas, licencia, etc.
- **Persona de contacto u organización responsable**
- **Creación del recurso:** proveedor, proyecto financiador.
- **Descripción del recurso:** variedad de lengua, niveles de anotación, estándares, tamaño, unidad, formato, dominio, etc.
- **Otros recursos relacionados.**
- **Grado de madurez de datos conforme al modelo desarrollado.**
- **Posibles aplicaciones del futuro recurso lingüístico.**

Metodología para la valoración de la madurez

Metodología

- Ficha con 13 apartados,
- Se completan los relevantes por cada conjunto de datos
- Se hace una estimación entre los puntos totales obtenidos por los puntos totales posibles para cada conjunto
- Tres niveles de madurez

1. Datos de **madurez baja**: -
2. Datos de **madurez media**: *
3. Datos de **madurez alta**: **

Tabla de evaluación del grado de madurez

Aspectos técnicos:
1. Digitalización (conversión a formato procesable).
2. Transcripción (ortográfica, fonológica, …).
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, …).
9. Revisión de contenido (incoherencias, redundancia de datos, …).
10. Anotación conforme a estándares PLN.
11. Presencia de metadatos.
Aspectos legales:
12. Necesidad de anonimización de datos personales.
13. Necesidad de solicitud de permiso de uso.

Resultados de la valoración de la madurez

Resultados

- **Madurez alta**:
  - Diccionarios terminológicos de TERMCAT
  - Índices de clasificación de los catálogos de la BNE (Datos.BNE.es)
  - Orphadata (INSERM)
  - Memorias de traducción del Instituto Vasco de Administración Pública (IVAP)
  - Memorias de traducción de la Diputación Foral de Gipuzkoa
  - Nomenclátor de Prescripción (Centro de Información de Medicamentos, CIMA)

Resultados

- **Madurez media:**
  - Oficina Española de Patentes y Marcas (OEMP)
  - Padrón del Instituto Nacional de Estadística (INE)
  - Topónimos del Instituto Geográfico Nacional (IGN)
  - Archivo Audiovisual del Congreso de los Diputados de España
  - Publicaciones del Instituto de Salud Carlos III (monografías y revistas) y vídeos

Resultados

- **Madurez media** *(cont.)*:
  - Publicaciones en repositorio SciELO (Scientific Electronic Library Online)
  - Guías de Práctica Clínica (GPC) del portal Guía Salud
  - Vídeos del portal web de TV del Gobierno Vasco relacionados con la Salud
  - Fichas de medicamentos y prospectos, y publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS)
  - Textos de jurisprudencia del CENDOJ
  - Boletín Oficial del Estado: Diario, Códigos electrónicos, y Legislación

Resultados

- **Madurez baja:**
  - Biblioteca Digital Hispánica
  - Publicaciones periódicas digitalizadas de la Hemeroteca Digital
  - Archivo de RTVE y RTVE A la carta
  - Patentes multilingües digitalizadas en PATSTAT de European Patent Office (EPO) y patentes en ámbito Latinoamericano
  - Grabaciones de vistas judiciales del Consejo General del Poder Judicial

Comparación con otros países

Metodología

- Enfoque en países Latinoamericanos y Europa. Canadá y EEUU
- Análisis de iniciativas similares al Plan Estratégico
- Estado de los datos abiertos por países

Resultados

- **España** se encuentra **bien situada**, **mejor** en **datos abiertos** que en RL
- **Francia** y **Reino Unido** son los países **más avanzados**
- Buen nivel pero **lejos del desarrollo del PLN en EE.UU.**, que beneficia a las grandes compañías.

Recomendaciones genéricas

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**

2. **Conversión** de los millones de **páginas digitalizadas en PDF o imagen en texto plano**.

Recomendaciones genéricas

3. **Descarga masiva** de grandes ficheros en **formatos apropiados (XML, CSV, JSON).**

4. Mejorar la **visibilidad** de los conjuntos de datos en cuanto a su **disponibilidad y madurez**

5. **Transcripción de ficheros multimedia**

Recomendaciones genéricas

6. **Estimular** la adopción de **licencias de libre uso y acceso** a los datos

7. **Estimular la reutilización de datos** organizando **competiciones tecnológicas**

8. **Facilitar el acceso a capacidad de cómputo y almacenamiento**

Entregables

Entregables

- **Informe final**, entregado en enero 2019, revisado en septiembre de 2019. 188 páginas
- Listado detallado de 24 recursos: formato Excel
- Listado de 101 recursos: formato Excel
- Disponibles en la página del Plan TL: https://www.plantl.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/documentos-reutilizables-como-recurso-linguitico.aspx

Gracias por su atención

¿Preguntas?

Condiciones iniciales de los 20 conjuntos de datos censados

≥ 20 de documentos
≥ 50 % total: Sanidad, Justicia e Inteligencia Competitiva
1 de todos ellos: Sanidad, Justicia e Inteligencia Competitiva
≥ 4 de las 7 categorías de documentos.
≥ 3 documentos en 2 lenguas cooficiales.
Provenientes de webs de AGE, CC.AA., EE.LL., universidades públicas.
Hincapié en portales de datos abiertos o páginas web de consejerías de las áreas de salud, justicia e innovación.
Encontrar ≥ 3 documentos que no estén referenciados en datos abiertos.