Encuentro

APORTA

7ª Edición *Madrid, 24 octubre 2017*

El valor de los datos *en el ecosistema global*

Infraestructuras de datos de interés lingüístico

David Pérez Fernández

Infraestructura de Datos Espaciales (IDE): integrado por un conjunto de recursos que permite el acceso y la gestión de conjuntos de datos y servicios geográficos (descritos a través de sus metadatos), disponibles en Internet, que cumple una serie **normas, estándares y especificaciones** que regulan y garantizan la **interoperabilidad** de la información geográfica.

Así mismo es necesario establecer un **marco legal** (INSPIRE Directive, ES IDEs AGE, CCAA) que asegure que los datos producidos por las instituciones serán compartidos por toda la administración y que promueva su uso entre los ciudadanos.

Una IDE es el conjunto de **tecnologías, políticas, estándares y recursos** humanos para adquirir, procesar, almacenar, distribuir y mejorar la utilización de la información geográfica.

Infraestructuras de datos de interés lingüístico: Datos espaciales

- **Conjuntos de datos espaciales**: Nombres Geográficos, Parcela Catastral, Hidrografía, Transporte, Lugares Protegidos (Medio Ambiente), Lugares Protegidos (Patrimonio Histórico-Cultural), Ocupación y uso del Suelo, Edificios, Servicios de Utilidad Pública y Estatales, Regiones Marinas, Clima

- **Servicios de Red**: visualización, localización, descarga (capas + proxy)

- **Metadatos:** conjuntos de datos y de servicios

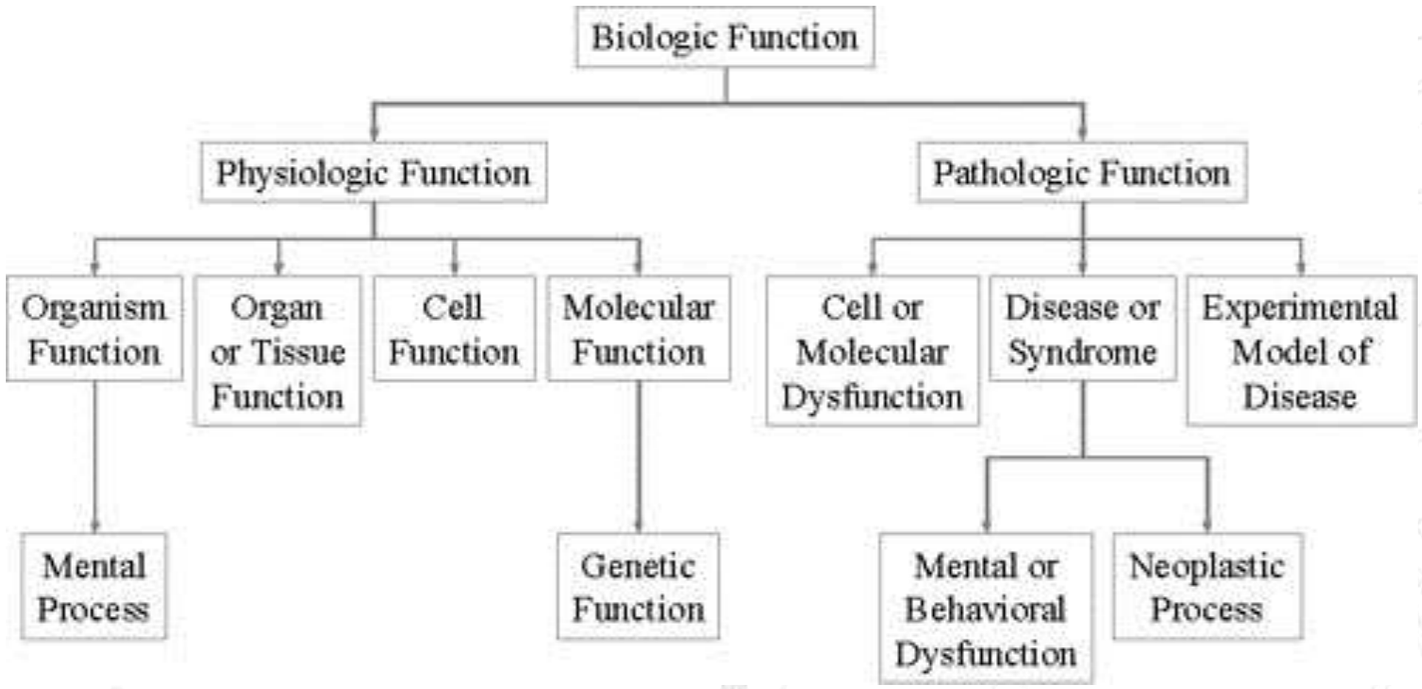
UMLS

- **Metathesaurus**: Gran base de datos de **vocabulario** multilingüe de conceptos biomédicos y de salud. Tesauros, clasificaciones, conjuntos de códigos y listas de términos controlados.

  Provienen de atención al paciente, facturación de servicios de salud, indexación literatura biomédica e investigación básica, investigación clínica y de servicios de salud.

  *(5M términos con ID único, no sólo contiene multitud de vocabularios sino los **mapeos** entre ellos).*

- **Red semántica**: **tipos** semánticos (enfermedades, síndromes, fármacos ...) y **relaciones semánticas** (relaciones útiles que existen entre los tipos semánticos. *Ej. El fármaco X se emplea para tratar la enfermedad Y*).



Unified Medical Language System® (UMLS®)

NIH U.S. National Library of Medicine

Infraestructuras de datos de interés lingüístico: Concepto

- **Corpus textuales**: Patentes (USPTO, EPO, OEPM…), Contratos públicos (TED, Portales nacionales: PlataformaContratacion), publicaciones científicas (SCIELO biomedicina, OpenAir+, Recolecta), Ayudas I+D (Cordis, NSF/NIH, ES?), HCE?, Sentencias judiciales? Crawling, red de enlaces, citas bibliográficas, … ?

- **Memorias de traducción**: DGT-Translation Memory, Six-Language Corpus UN, Acquis Communautaire, Patentes tipo T3, OIL ES?.

- **Transcripciones de audio:** transcripciones de las reuniones del Consejo?, 060 ES

- **Terminología** (IATE EU's multilingual term base, Eurovoc), glosarios, diccionarios

Concepto

- **Taxonomías, clasificaciones, nomenclaturas, vocabularios controlados …**:
  - ✓ **ICD/CIE** International Statistical Classification of Diseases and Related Health Problems
  - ✓ **CPV** classification system for public procurement; poderes adjudicadores, organismos públicos contratantes, áreas geográficas NUTs
  - ✓ **IPC** International patent classification
  - ✓ **Clasificación de Nize** de productos y servicios usada para el registros de marcas
  - ✓ Nomenclatura **UNESCO**, nomenclatura para los campos de la ciencia y la tecnología
  - ✓ **NACE/CNAE** Statistical Classification of Economic Activities in the European

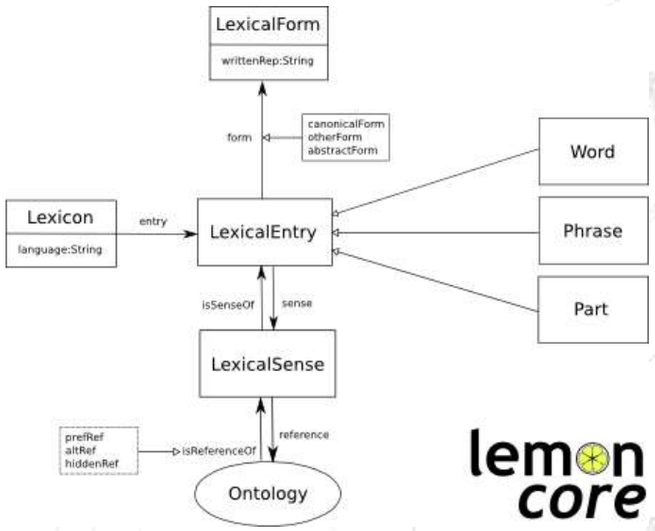
Las clasificaciones **multilingües** son muy interesantes porque actuan de hecho como una estructura semántica multilíngüe. Lamentablente unidimensional, es decir, generalmente sólo establece relaciones de tipo *Is-*

Infraestructuras de datos de interés lingüístico: Concepto

- **Recursos semánticos**: WordNet, DB Pedia, Yago, Babelnet, SNOMED-CT, Datos enlazados en la BNE. Estructura RDF y reutilización de ontologías.

- **Colecciones de entidades nombradas**: Localizaciones geográficas (Geonames), nombres de personas, personajes públicos (BNE, US Library of Congress), organizaciones (TED), nombres y autores de obras artísticas (Eurpeana, BNE, Library of Congress US), nombres de empresas, marcas, etc.

META-SHARE



lemon core

Infraestructuras de datos de interés lingüístico: Usos

- **Corpus textuales:** anotación y entrenamiento de procesadores lingüísticos.
- **Memorias de traducción:** entrenamiento de motores de traducción automática.
- **Transcripciones de audio:** entrenamiento de sistemas de transcripción voz a texto aplicaciones móviles, sistemas conversacionales, …
- **Terminología y vocabularios controlados**: representación simplificada de documentos, agrupación y distancia entre documentos.
- **Colleciones de entidades nombradas**: NERC, representación simplificada de documentos, recuperación de información.
- **Taxonomías**: clasificación automática de documentos, construcción semiautomática de KB, distancia entre documentos.
- **Recursos semánticos**: análisis semántico de documentos, distancia entre documentos, SRL, detección de eventos, WSD.

Infraestructuras de datos de interés lingüístico: ideas

- Abrir recursos en las **Administraciones Públicas**; alta calidad, recursos lingüísticos realmente completos y sostenibles.
- **Repositorios** communes y **censo** de recursos.
- **Servicios sindicados** (INSPIRE Idea, datos provenientes de multiples fuentes que cubren diferentes capas y áreas geográficas). Múltiple dominios y tipos de información lingüística relacionada.
- Como cualquier otra área de datos abierta: **normalización e interoperabilidad** (actualmente 90 normas de anotaciones lingüísticas).
- Protección de **datos personales** (HCE y otros AAPP, difícil de anonimizar).
- Algunos recursos son caros, tienen un desarrollo y mantenimiento muy largo. **Esfuerzos internacionales** conjuntos, organizaciones internacionales, administraciones públicas y asociaciones.
- Los recursos lingüísticos se **superponen**; terminología, semántica, diccionarios... Definir las **autoridades** por dominio e idiomas

Infraestructuras de datos de interés lingüístico: Política pública

Como los datos **geospaciales** (INSPIRE EU, National Spatial Data Infrastructure EEUU (NSDI), IDEE ES) o la infraestructura de información **médica** (UMLS EEUU, Unified Medical Language System) …

Datos [abiertos] de interés lingüístico son otra infraestructura de datos

Debe ser:

- **Coherente**; Coordinación entre las diferentes lenguas, dominios de información y diferentes tipos de recursos lingüísticos, con autoridades bien definidas.
- **Sostenible**; importante trabajo de la AAPP; también es necesaria la asociación público privada en algunos sectores (ej. Industrial, Automóvil).
- **Interoperable**; unificación de la estandarización. Es necesaria la cooperación del ETSI de la UE.
- **Distributed**; distintos servicios deben coordinarse para distribuir de forma coherente recursos en distintos idiomas, dominios y de distita tipología.

CORPUS

- Compartir **conjuntos de datos completos** normalizados, no sólo servir registros individuales.
- Colocar los **conjuntos de datos cerca de los recursos computacionales** (idea AWS: conjuntos de datos públicos; ej. 500Tb crawling).
- Almacenar y compartir los **procesamientos intermedios**; las anotaciones lingüísticas son costosas para grandes corpus. **Reutilizar** el trabajo de **investigación** (Idea CLARIN ERIC).

RECURSOS SEMÁNTICOS

- Los **recursos semánticos** son muy útiles pero **caros** y tienen un largo tiempo de desarrollo y mantenimiento (ej SNOMED-CT). Esfuerzos **internacionales sostenidos**.
- Construir verdaderos **recursos semánticos**: multidimensionales y reutilización de recursos existentes.

TERMINOLOGÍA

- Mantenimiento **costoso** para grandes corporaciones. Detección **automática** de nuevos términos. Alineación automática de la terminología multilingüe.
- Uso de **técnicas modernas**: WordEmbeddings, modelado de tópicos,...
- Obtener terminología y diccionarios de dominio de las **Administraciones Públicas**

CORPUS PARALELOS, MEMORIAS DE TRADUCCIÓN

- Compartir **memorias de traducción**; mejora SMT y NMT (zeroshot AT)
- **Anonymize** texto; permite traducirlo a través de servicios cloud o interadministrativos (CEF. AT). Necesario en el caso de algunos corpus; sentencias, HCE.
- Herramientas automáticas de **alineación de documentos**. También para cuerpos rastreados por web.
- **Sindicar glosarios, diccionarios;** organizaciones distintas pueden cubrir muchos idiomas y dominios.

CLASIFICATIONES Y TAXONOMÍAS

- **Mapping taxonomies;** realmente difícil se requieren técnicas semiautomática modernas; modelado de tópicos, WordEmbeddings, FCA.

NOMBRES Y TERMINOLOGÍA

- **Administración Pública** dispone de mucha información: lugares, personas y personalidades, organizaciones, marcas, nombres de fármacos, etc. Los países multilingües también han traducido estos recursos en sus lenguas regionales.

NIVEL REGIONAL

- **Coordinación** con iniciativas europeas y nacionales. Cooperación necesaria para la sostenibilidad a largo plazo; muchos de los grandes servicios son competencia autonómica.

NIVEL NACIONAL

- Incluir normalización y reutilización por medio de una etapa de control en los **programas de investigación y desarrollo** también en **contratación pública**.
- Mantener un **censo** de recursos lingüísticos.
- Proyectos PLN: Desarrollar proyectos orientados a **campañas de evaluación** (gold standars abiertos; métricas uniformes)

NIVEL EUROPEO

- Considerar las tecnologías del lenguaje como una línea en el programa marco de **H2020**, **CEF** y en general en la estrategia digiltal, **Digital Single Market**
- ETSI **standardization** roadmap; incluir los estándares de tecnologías del lenguaje en su plan de trabajo.

NIVEL EUROPEO

- Incluir procesadores, herramientas y repositorios de **PLN en infraestructuras comunes**: CEF y otras infraestructuras de investigación, coordinación entre CEF AT y otros CEF DSI con iniciativas nacionales.
- Incluir los datos lingüísticos abiertos entre las tareas específicas de las próximas acciones dirigidas por la Comisión para promover la **reutilización de la información del sector público**.
- Promover **plataformas** para el almacenamiento y procesamiento de recursos; especialmente orientadas a investigadores y PYMES innovadoras.

NIVEL INTERNATIONAL

- **Cooperación internacional** para el desarrollo de infraestructuras lingüísticas multilingües internacionales (ej. IHTSDO en el desarrollo SNOMED-CT)
- Acuerdo necesario entre EU (proyectos EU de estandarización EAGLES, PAROLE, SIMPLE, LIRICS) e ISO en materia de **inciativas internacionales de estandarización** (ej. WG37 "Terminology and other language and content resources").

Encuentro

APORTA

7ª Edición *Madrid, 24 octubre 2017*

El valor de los datos *en el ecosistema global*

¡Muchas Gracias!

David Pérez Fernández
dperezf@minetad.es