Plan TL
IBERLEGAL

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA

Pasos e iniciativas del Plan TL en el dominio legal

Doaa Samy, Jerónimo Arenas y David Pérez
Oficina Técnica General
Plan TL

Doaa Samy
Oficina Técnica General

Plan TL

11 de diciembre de 2019

Índice

1. Legal-ES: Un conjunto de recursos del dominio legal para el PLN en español
2. Resultados preliminares
3. Casos de uso
4. Futuros pasos

1. Legal-ES

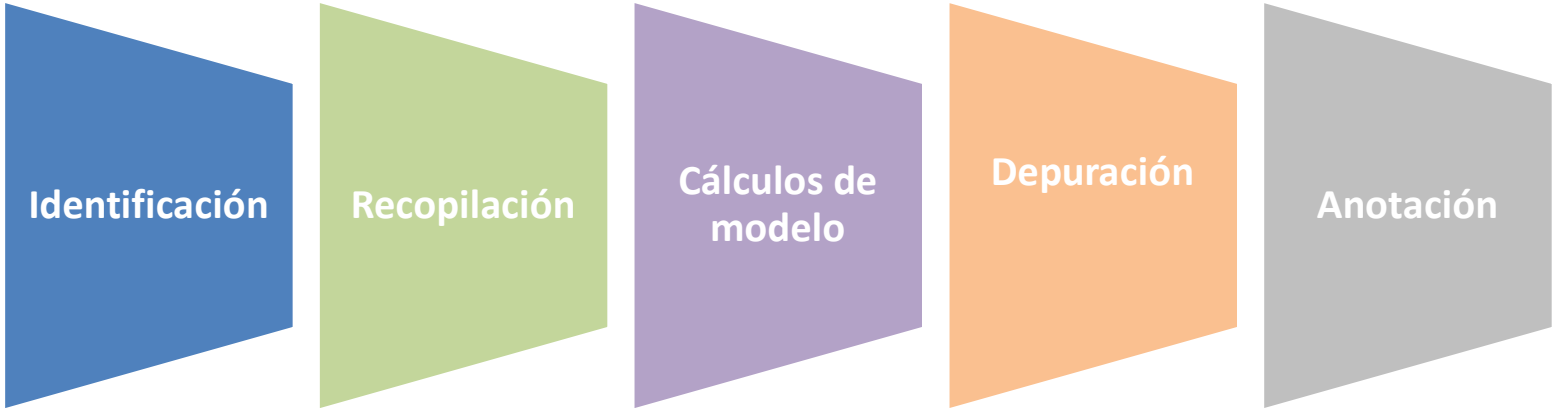
Plan TL
IBERLEGAL

- **Motivación**
  - Un dominio de prioridad para el Plan TL
  - Disponibilidad de grandes cantidades de texto
  - Impacto para la industria, la academia, las Administraciones Públicas y los ciudadanos
  - En línea con estrategias nacionales, europeas e internacionales
- **Estado de la cuestión**
  - En inglés: Disponibilidad de recursos (BlaRC, Cambridge, BYU Legal, *US Supreme Court Opinions*)
    - Eur-Lex
    - Instituto de Euskera: Corpus jurídico (7.2m de palabras)
- **Foros principales: Jurix, Jurisin, ICAIL**

1. Legal-ES

- **¿Qué es Legal-ES?**
  - *Legal-ES es un conjunto de recursos abiertos para el español en el dominio legal.*
- **Legal-ES incluye:**
  - *Un corpus de gran escala de 1500 millones de palabras aproximadamente.*
  - *Modelos de Embeddings calculados del lenguaje en el dominio*
  - *Modelos de tópicos.*
  - *Grafos de documentos*

1. Legal-ES: Fases



Fase 1: Identificación de posibles fuentes (legislativas+adminitrativas)

- Alcance geográfico: internacional, europeo, nacional, autonómico
- Tipología: Legislativo, legal, administrativo, etc.

Fase 2: Recopilación de recursos

- Se ha procedido a la recopilación de un subconjunto de estas colecciones identificadas.

Fase 3: Cálculos preliminares de modelos

1. Legal-ES

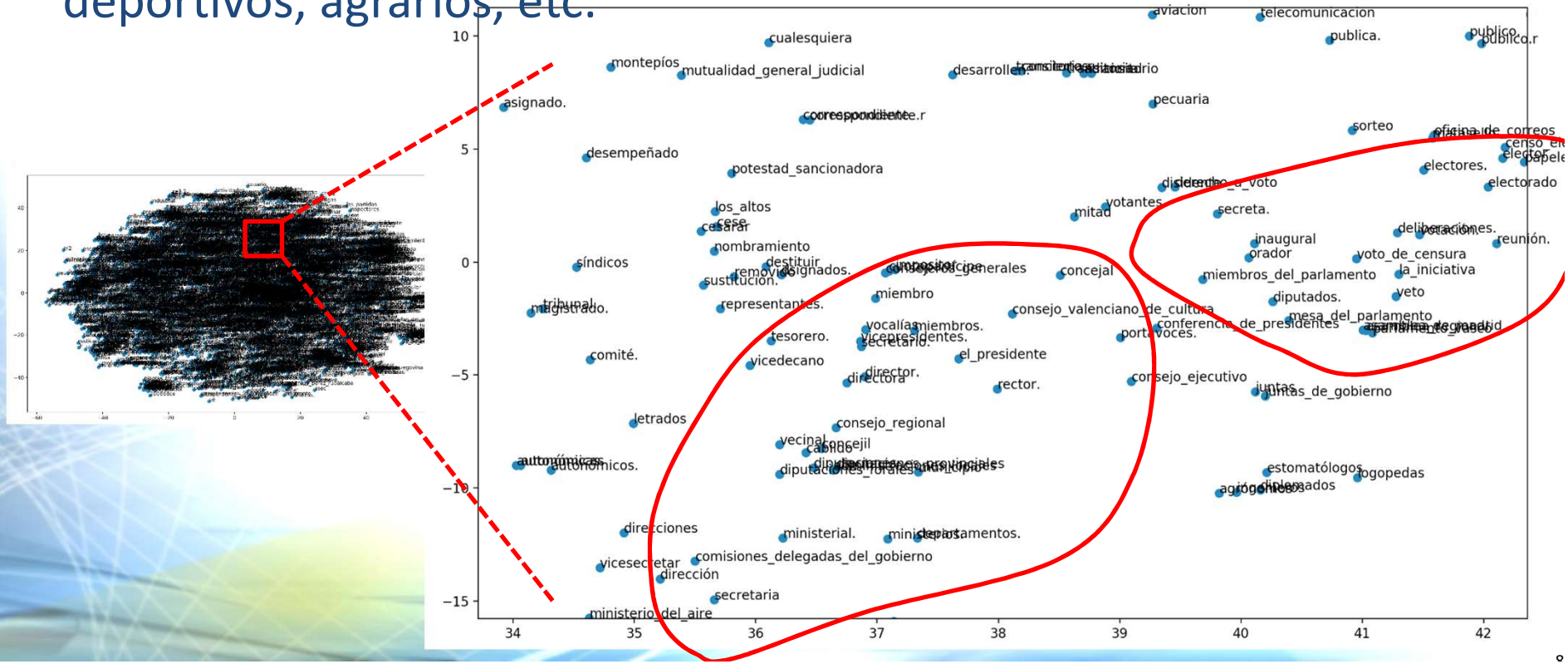
Tabla de recursos identificados:

Fuente	Volumen	Estado
BOE-Legislación	547+ millones de palabras y 216 millones documentos	Descargado
Doctrina Fiscalía	2+ millones de palabras	Descargado
Dictámenes Consejo del Estado	135+ millones de palabras	Descargado
Abogacía del Estado	6+ millones de palabras	Descargado
JRC-Acquis	59+ millones de palabras	Descargado
Eur-Lex	58+ millones de palabras	Descargado
Consultas tributarias	400+ millones de palabras	Descargado
Leyes argentinas	238+ millones de palabras y 142 mil documentos	Descargado
Leyes mexicanas	-----	Identificado
ONU	------	Identificado

Índice

2. Resultados preliminares: BOEEmbeddings

- Embedding de 300 dimensiones colapsado a 2 dimensiones
- Los clusters de cargos unipersonales y términos electorales aparecen cercanos
- Otros clusters se centran en normativas concretas, ámbitos deportivos, agrarios, etc.



BOEEmbeddings: búsqueda por similitud

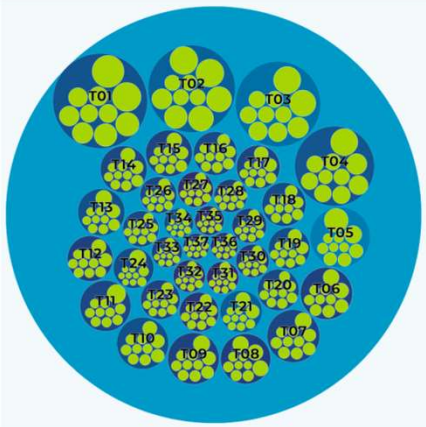
```
[('orden_ministerial', 0.8387089967727661),
 ('reales_decretos', 0.8019422888755798),
 ('decreto', 0.7804737091064453),
 ('real_decreto-ley', 0.7155911326408386),
 ('orden', 0.6945813894271851),
 ('ley', 0.6891162991523743),
 ('real_decreto_ley', 0.6566262245178223),
 ('la_orden', 0.6429240703582764),
 ('rd', 0.5898292064666748),
 ('orden.', 0.5714296698570251)]
```

```
[('irpf.', 0.7073756456375122),
 ('i.r.p.f.', 0.5817404389381409),
 ('impuesto', 0.5020797848701477),
 ('renta', 0.4686171412467956 5),
 ('tributo', 0.4600400030612945 6),
 ('impuestos.',
0.4529829025268554 7),
 ('impuesto_de_sociedades',
0.4449738860130 31),
 ('igic', 0.4337747991085052 5),
 ('impuesto.', 0.4300283193588257),
 ('ite', 0.4253433557891 85)]
```

```
[('chardonnay',
0.9205656051635742),
 ('syrah', 0.8846567869186401),
 ('pinot', 0.8578344583511353),
 ('merlot', 0.8198724985122681),
 ('bianco', 0.8193067312240601),
 ('cabernet_franc',
0.8192765116691589),
 ('sauvignon_blanc',
0.8116549849510193),
 ('cabernet_sauvignon',
0.8089689016342163),
 ('malvasia',
0.7975383996963501),
 ('moscato', 0.7844038605690002)]
```

BOE Topic Models

- Permite encontrar las temáticas más importantes y caracterizarlas por las palabras más frecuentes, su evolución temporal, etc …



T10 - Comptencias y Procedmientos (2.58 %)
autoridad información estados_miembros directiva procedimiento estado_miembro competente nacional ue unión

T11 - Ganadería (2.34 %)
animal veterinario producto carne control especie país sanitario explotación enfermedad

T12 - Agricultura (2.22 %)
producto producción productor vino ayuda semilla variedad explotación agrícola cultivo

T13 - Pesca (2.19 %)
pesca buque zona pesquero captura agua especie población ce cuota

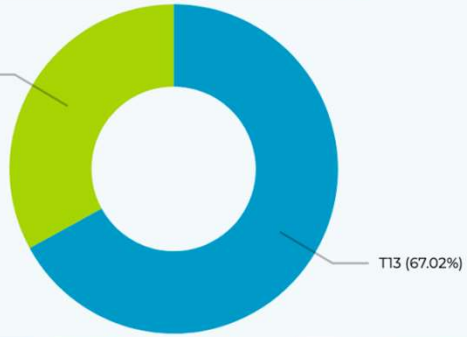
T14 - Eliminación de residuos (2.18 %)
producto sustancia directiva uso residuo ce autorización evaluación riesgo información

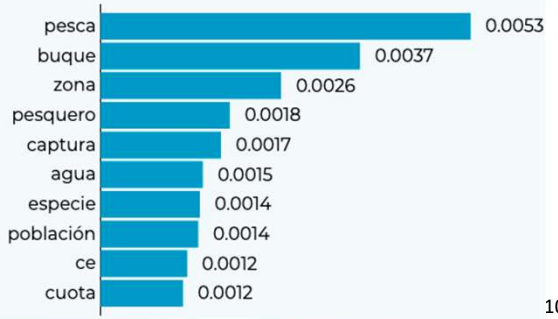
- Permite caracterizar documentos, e implementar búsqueda “semántica”

Conflicto positivo de competencia número 1.109/1986, planteado por el Gobierno Vasco, en relación con la Orden de 13 de junio de 1986, del Ministerio de Agricultura, Pesca y Alimentación.

T21 - Legislacion CCAA (1.75 %)

T13 - Pesca (2.19 %)





2. Resultados preliminares: Grafos de Documentos

Plan TL
IBERLEGAL

- Cálculo de grafos semánticos

Índice

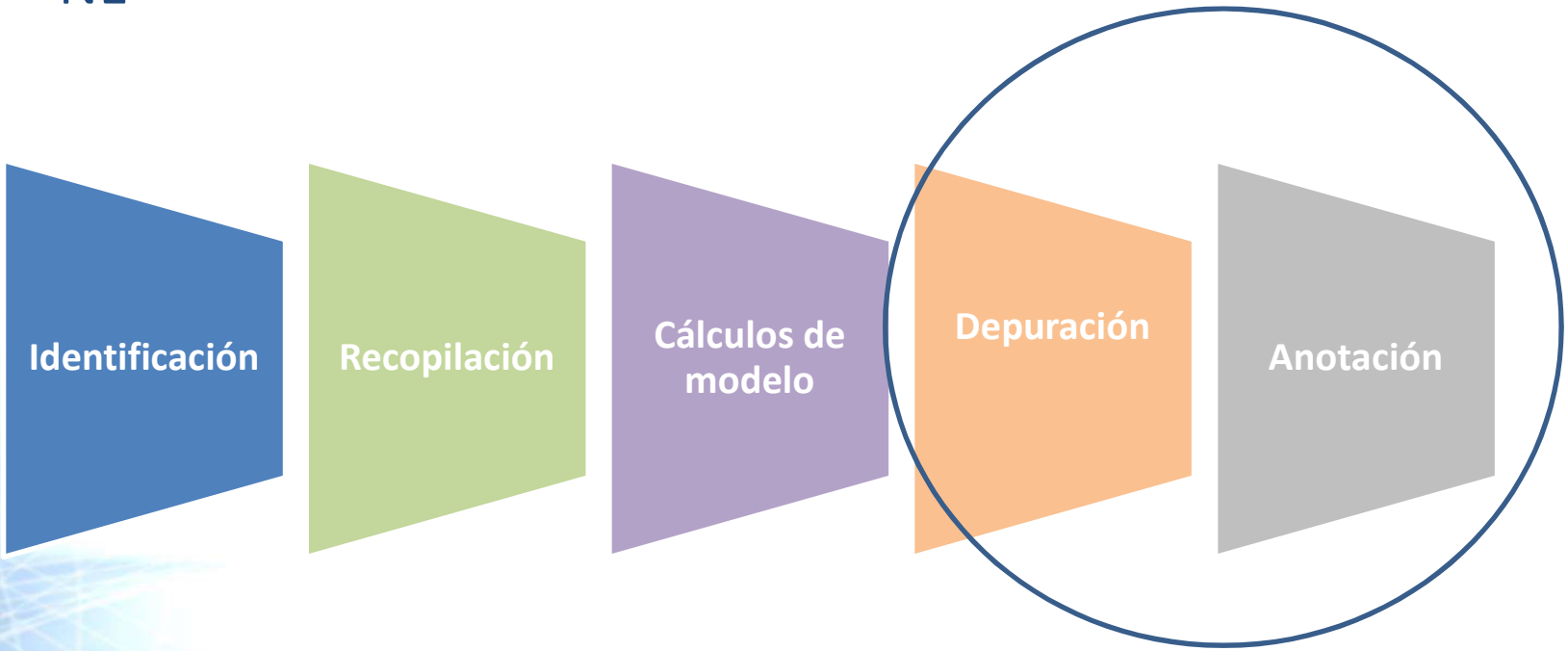
3. Casos de uso

- Extracción de terminología
- Enriquecimiento de glosarios (sinónimos, etc.)
- Extracción de Información
- Visualización
- Anonimización
- Enlazar documentos

Índice

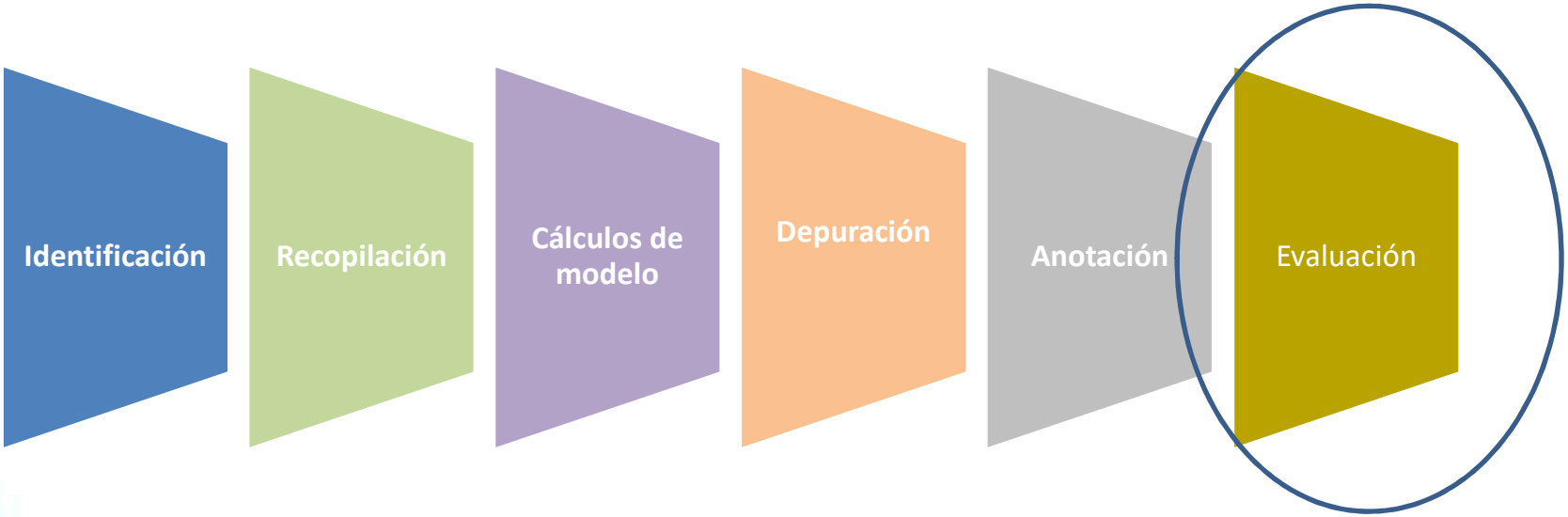
4. Futuros pasos: Depuración y anotación

- Depuración y anotación de un subconjunto
  - POS
  - NE



4. Futuros pasos: Tareas de evaluación

- NERC en textos legales (Iberlef?)



Plan TL
IBERLEGAL

Estudios

- TL en Justicia
- Terminología: SNOMED → SNOLEX

Talleres e infodays

- LT4Gov: LREC 2020 → Mayo 2020
- Infoday → Marzo 2020

Colaboraciones

- ¿RAE?
- ¿MPR-CPAGE?
- ¿BOE?

Gracias

www.PlanTL.es

PlanTecnologiasLenguaje@mineco.es