

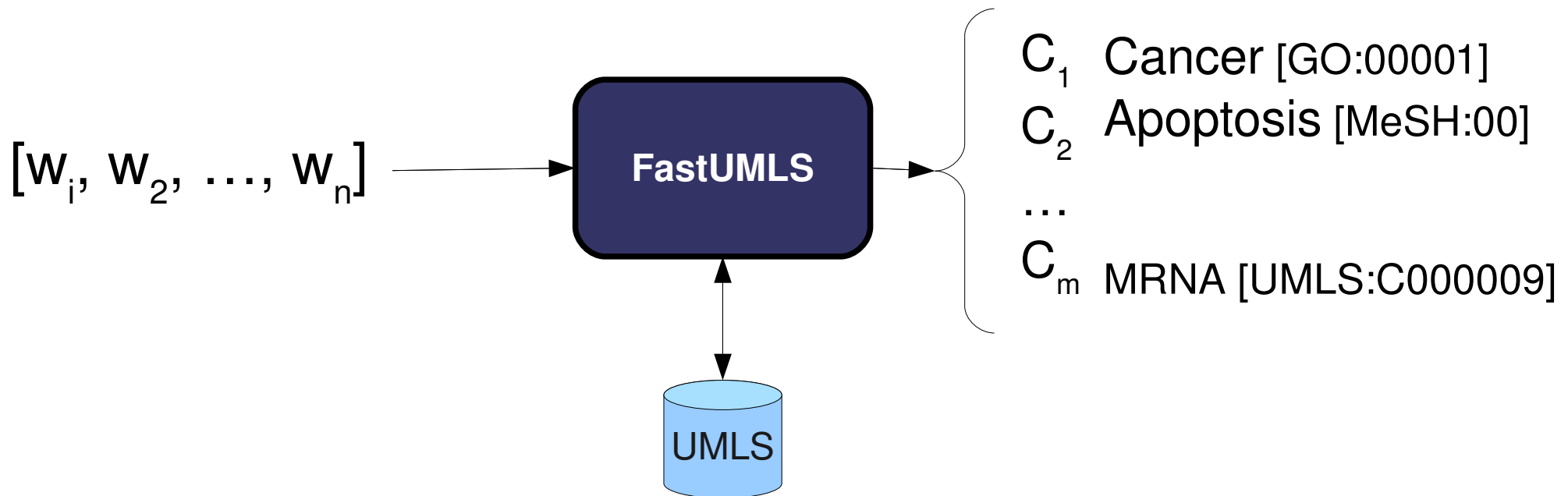
# FastUMLS: Extracción de conceptos en textos biomédicos

Autor: José Luis Marina  
Director: Alberto Pascual

**Facultad de Informática  
Universidad Complutense  
de Madrid**

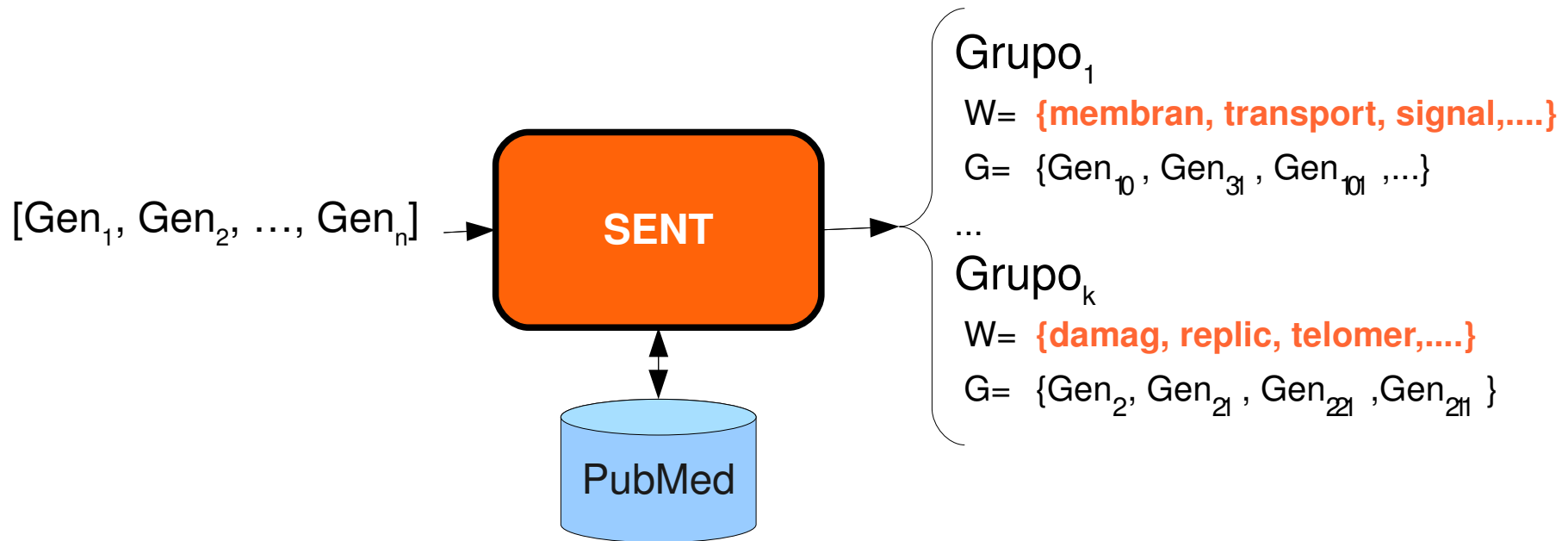


Proporcionar un conjunto de conceptos normalizados a partir de un conjunto de palabras



Dado un conjunto de palabras identificar los que están más relacionados con las palabras de entrada

SENT: Revisa textos relacionados con una lista de GENES y los agrupa de acuerdo a una lista de palabras



¿Podemos resumir esas palabras en uno o dos conceptos?

“Proporcionar un conjunto de conceptos normalizados a partir de un conjunto de palabras”

cancer, tumor, p53, factor, signal,  
target, breast, pathwai, transcript  
factor, growth, wnt, carcinoma,  
line, promot, notch



(NCI) C39202: “Notch and Wnt Signaling Pathway”

(NCI) C39270: “WNT Signaling Pathway”

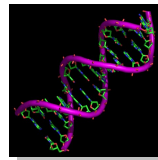
(GO) GO:0016055: “Wnt receptor signaling pathway”

++ De cualquier conjunto de palabras

Activation-induced deaminase (AID) is required for class switch recombination (CSR) and somatic hypermutation (SHM), which are responsible for secondary diversification of antibodies in germinal centers. AID initiates these processes through deamination of cytosines on the immunoglobulin (Ig) locus, a potentially mutagenic activity. AID expression is restricted to germinal-center B cells, but mechanisms that regulate its target specificity are not completely understood. Here, we review the most recent findings on the regulation of AID targeting and discuss how AID activity on non-Ig genes is relevant to the generation of chromosome translocations and to lymphomagenesis

¿Podemos saber cuáles son los principales **conceptos/ideas** sobre los que trata este artículo?

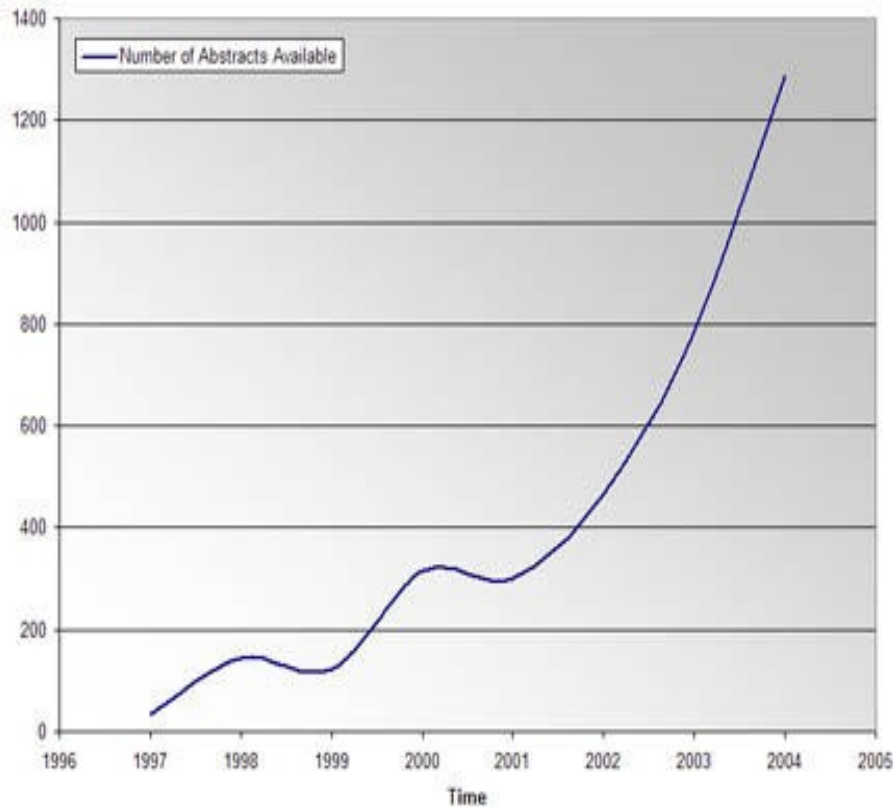
La Bioinformática es el uso de técnicas matemáticas e informáticas para almacenar, gestionar y analizar datos biológicos para responder a problemas de la biología [Kaminski 2000].



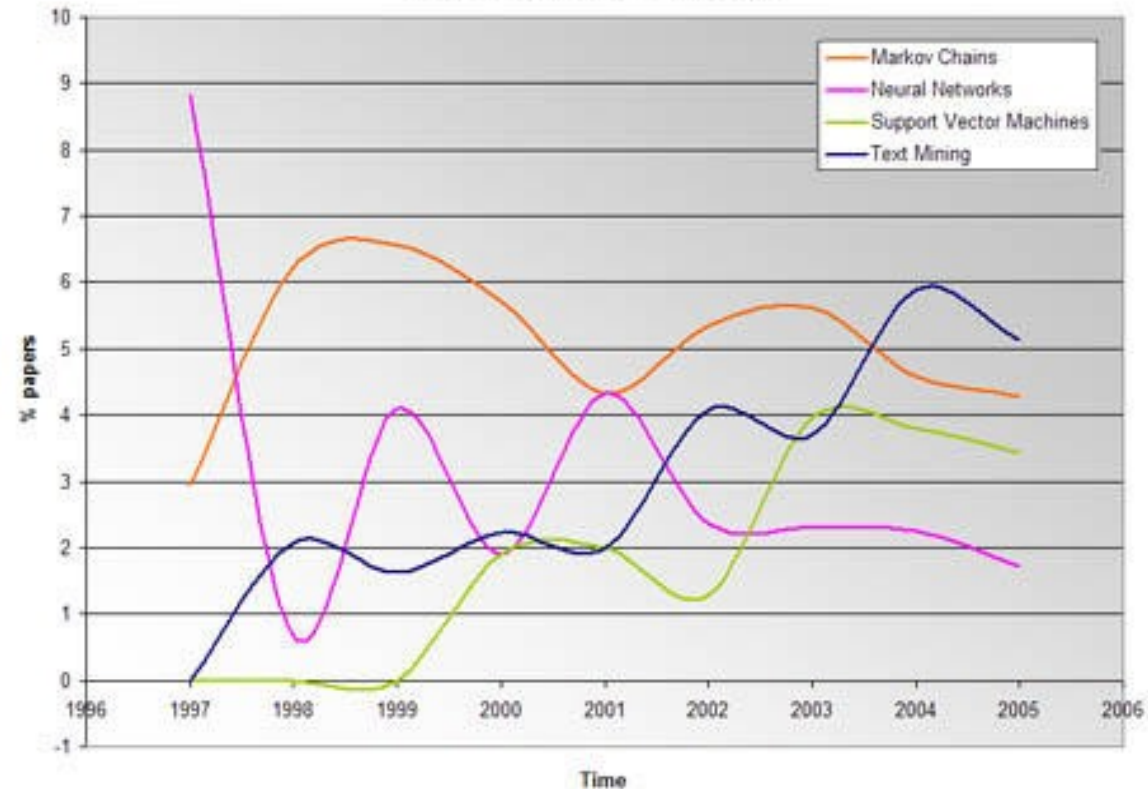
El principal problema es la explosión de datos en el campo de la biomedicina.

Información vs Conocimiento

### Abstracts Available vs Time



### Machine Learning



Una lista que contiene los términos empleados para representar los conceptos, temas o contenidos de los documentos de un dominio, con miras a efectuar una normalización terminológica que permita mejorar el canal de acceso y comunicación entre los usuarios.

**Ontología = Conceptos + Relaciones → Representación Computacional**

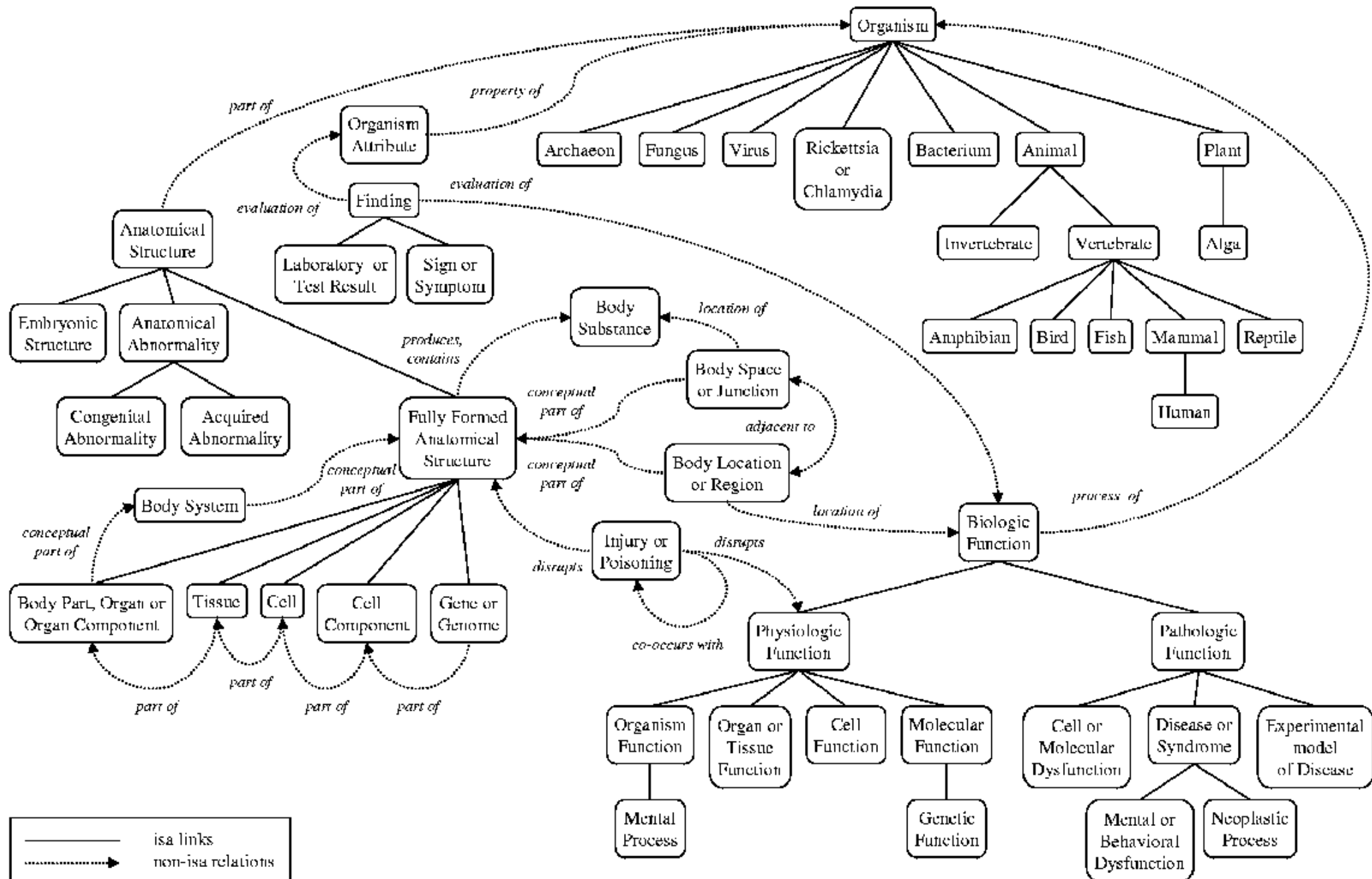
Funciones

Integrar en un Vocabulario Común

Almacenar Conocimiento

Permite Preguntar por Conocimiento

# Conceptos: Ontologías



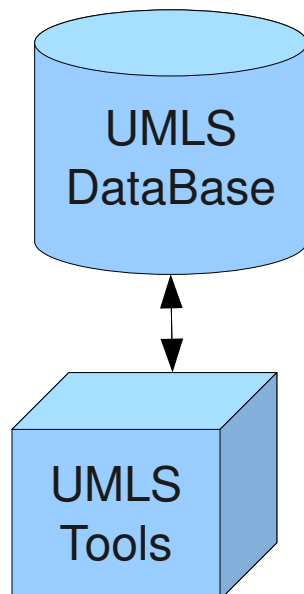
## ¿Cómo se crean y mantienen las Ontologías?

“Curators”: Expertos en el dominio del conocimiento



- Identifican los conceptos.
- Comprueban su corrección.
- Identifican las relaciones entre ellos.
- Detectan duplicados
  
- Revisión de artículos y literatura relacionada.
- Ayudados por herramientas automáticas.

El propósito de UMLS - Unified Medical Language System® es facilitar el desarrollo de sistemas informáticos que se comporten como si entendieran el significado del lenguaje utilizado en Biología y en Medicina.



- Varios orígenes de datos “curados”.
- Conceptos, nombres, cadenas, orígenes.
- Relaciones originales y nuevas.
- Se puede cargar en un RDBMS como MySQL.

- Herramientas de PLN
- Java UMLS Database Navigators

## Conceptos y CUI

Un concepto es un significado, que puede tener distintos **nombres**, pero no está (no debe) estar repetido. “Dolor de Cabeza”, “Cefalea”

## Términos e Identificadores Léxicos - LUI

Agrupar un conjunto de variaciones léxicas de un mismo significado.  
“adenoidectomy”, “ADENOIDECTOMY” y “Adenoidectomies” → L0001425

## Nombres de Conceptos y Cadenas - SUI

Cada nombre de concepto o cadena en cada lenguaje en el Metathesaurus tiene un identificador único y permanente o SUI.

## Átomos e Identificadores de Átomos - AUI

Cada vez que una cadena aparece en un vocabulario se le asigna un identificador único o AIU (“Atom Unique Identifier”). Si la misma cadena aparece en muchos vocabularios diferentes se le asigna a cada ocurrencia un AUI distinto, y todos esos AUIs estarán relacionados con el mismo SUI.

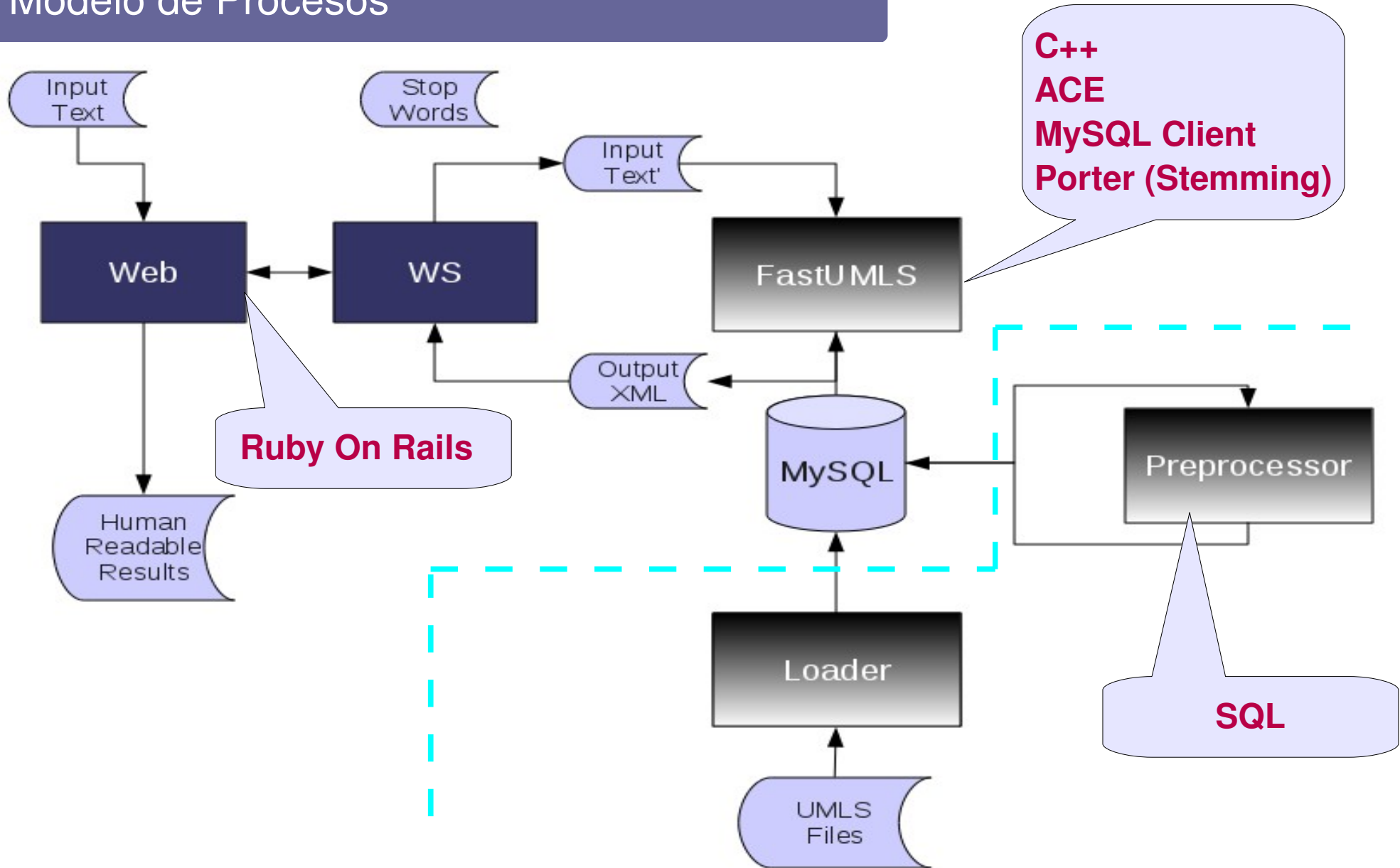
## Ejemplos

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
<b>C0004238</b> Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	<b>L0004238</b> Atrial Fibrillation (preferred) Atrial Fibrillations	<b>S0016668</b> Atrial Fibrillation (preferred)	<b>A0027665</b> Atrial Fibrillation (from MSH)
		<b>S0016667</b> Atrial Fibrillation (from PSY)	
		<b>S0016669</b> Atrial Fibrillations (from MSH)	
	<b>L0004327</b> (synonym) Auricular Fibrillation Auricular Fibrillations	<b>S0016899</b> Auricular Fibrillation (preferred)	<b>A0027930</b> Auricular Fibrillation (from PSY)
		<b>S0016900</b> (plural variant) Auricular Fibrillations	<b>A0027932</b> Auricular Fibrillations (from MSH)

### Concepto C0004057 → “Aspirin”

Aspirin  
 aspirin  
 Acetylsalicylic Acid  
 Acetylsalicylic acid  
 acetylsalicylic acid  
 Acid, Acetylsalicylic  
 Acetysal  
 Acylpyrin  
 Colfarit  
 Easprin  
 Ecotrin  
 Endosprin  
 Magnecyl  
 Micristin  
 Polopiryna  
 Zorprin  
 2-(Acetyloxy)benzoic Acid  
 Benzoic acid, 2-(acetyloxy)-  
 Aspergum  
 Empirin  
 Entericin  
 St. Joseph  
 Measurin  
 ACIDE ACETYLSALICYLIQUE  
 ACETYLSALICYLIQUE, ACIDE  
 ASPIRINE  
 ASPIRIN  
 ASPIRINA  
 ACIDO ACETILSALICILICO

## Modelo de Procesos



## Preprocesado: Modelo de Datos del que partimos

### MRCONSO

STR	LUI	SUI	AUI	SAB
Primary Adrenal Insufficiency	L0494851	S5907334	A12807945	NCI
Primary Adrenal Insufficiency	L0494851	S5907334	A6975965	MSH



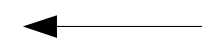
### MRXNS\_ENG (Cadenas Normalizadas)

LAT	NSTR	CUI	LUI	SUI
ENG	adrenal insufficiency primary	C0001403	L0494851	S5907334

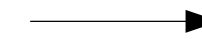


### MRXNW\_ENG (Palabras Normalizadas)

LAT	NWD	CUI	LUI	SUI
ENG	adrenal	C0001403	L0494851	S5907334
ENG	insufficiency	C0001403	L0494851	S5907334
ENG	primary	C0001403	L0494851	S5907334



"adrenal"



LUIs o Términos

## Preprocesado: Modelo de Datos que queremos

### FU\_MRXSNEWCONSO\_ENG

Field	Type	Null	Key	Default	Extra
CUI	varchar(8)	NO		NULL	
LUI	varchar(8)	NO	MUL	NULL	
SNWD	varchar(100)	NO	MUL	NULL	
SNWD_CARD	int(11)	NO		NULL	
LUI_CARD	int(11)	NO		NULL	
LUI_WEIGHT	double	NO		NULL	
SNWD_WEIGHT	double	NO		NULL	

"adrenal"



"adren"



CUI	LUI	SNWD	SNWD_CARD	LUI_CARD	LUI_WEIGHT	SNWD_WEIGHT
C1403891	L0368421	adren	1770	3	7.79257425770879	3.90412658114857
C0001403	L0278071	adren	1770	4	9.15101108465032	3.90412658114857
C0001403	L0278357	adren	1770	3	7.35481996167742	3.90412658114857
C0300942	L0278357	adren	1770	3	7.35481996167742	3.90412658114857
C0001403	L0494851	adren	1770	3	7.28466969560075	3.90412658114857
C0001613	L0001613	adren	1770	2	5.76357355441712	3.90412658114857
C0001613	L0683397	adren	1770	3	6.60864930712491	3.90412658114857
C0001614	L0001614	adren	1770	3	6.64237861631983	3.90412658114857
C0001614	L1191251	adren	1770	3	7.03641625689895	3.90412658114857
C0001614	L6329909	adren	1770	3	6.70017421906099	3.90412658114857

## Proceso FastUMLS



	$e_1$	$e_2$	...	$e_i$		$e_n$
$L_1$	1	0		0		0
$L_2$	1	1		0		0
...		0		1		0
$L_i$	1	1		0		0
...	0	0		0		1
$L_m$	0	1		0		1

### Términos Ordenados por Peso:

- Promiscuidad de las palabras
- Palabras de la entrada en el Término
- Palabras totales del Término

Un Término que tiene asociadas palabras muy genéricas – palabras que están presentes en muchos otros conceptos – tiene mayor probabilidad de salir que otro relacionado con palabras menos generales.

## Proceso FastUMLS

$$E = \{e_1, e_2, \dots, e_m\}$$

$$U = \{p_1, p_2, \dots, p_u\}$$

$$T = \{l_1, l_2, \dots, l_T\}$$

$$l_j = \{p_{j1}, p_{j2}, \dots, p_{jx}\}$$

$$M, m_{ij} \in [0, 1]$$

	$e_1$	$e_2$	...	$e_i$		$e_n$
$L_1$	0	1		0		1
$L_2$	1	1		0		1
...		0		0		1
$L_i$	0	1		0		0
...	0	0		0		0
$L_m$	0	1		0		0

$$w_p = |\log(P(p))| = \left| \log\left(\frac{\text{card}(p)}{T}\right) \right|$$

$$w_{li} = \sqrt{w_{pi1}^2 + w_{pi2}^2 + \dots + w_{pix}^2}$$

$$W_{lj}(E) = \frac{m_{j1} \cdot w_{e1}^2 + m_{j2} \cdot w_{e2}^2 + \dots + m_{jm} \cdot w_{em}^2}{\sqrt{w_{pj1}^2 + w_{pj2}^2 + \dots + w_{pjx}^2} \cdot \sqrt{w_{e1}^2 + w_{e2}^2 + \dots + w_{em}^2}}$$

$$W_{lj}(E) = \frac{m_{j1} \cdot w_{e1}^2 + m_{j2} \cdot w_{e2}^2 + \dots + m_{jm} \cdot w_{em}^2}{w_{lj} \cdot w_E}$$

Ponderar la probabilidad de que hubieran salido seleccionados si el conjunto de entrada se hubiera escogido al azar.

Penalización de términos con alta probabilidad - debido a lo genérico de sus palabras.

## Estado del Arte: ¿Hay Procesos Similares?

### Técnicas de Procesamiento del Lenguaje Natural PLN:

Procesan textos completos buscando párrafos, sentencias y proposiciones.  
Tratan una a una las proposiciones y muestran los conceptos relacionados.

- + Identifican significados, por ejemplo las negaciones. (“No agresivo”)
- No relacionan palabras en párrafos o frases diferentes.
- Suelen ser lentos porque demandan mucha capacidad de proceso.

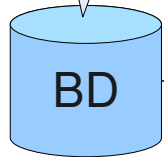
### Técnicas de Permutación de Palabras (Como FastUMLS)

IndexFinder: Procesa todas las palabras de un texto buscando para cada palabra los conceptos relacionados, y muestra conceptos con **todas** las palabras.

- + Utiliza palabras en párrafos o frases diferentes
- + Rápido
- No Identifica significados. (“No agresivo”)
- No muestra conceptos sin acierto en Todas las palabras

## Rendimiento

14 millones de registros  
3 millones de conceptos



FastUMLS

Palabras en el Texto	Elementos Matriz	Segundos
200	50.000	< 20
12	5.000	< 2

	$e_1$	$e_2$	...	$e_i$		$e_n$
$L_1$	0	1		0		1
$L_2$	1	1		0		1
...		0		0		1
$L_i$	0	1		0		0
...	0	0		0		0
$L_m$	0	1		0		0

## Pruebas Comparativas (PLN - Interactive MetaMap IM)

### Identificación de Conceptos en Genes Anotados por Expertos

**Objetivo: Comparar FU contra una herramienta y resultados anotados por expertos.**

Expertos han anotado Genes con conceptos de Gene Ontology (GO)

Se han procesado los textos de descripción de varios genes buscando esos conceptos.

IM:

No ha encontrado en ningún caso ninguno de los conceptos.

No utiliza palabras muy alejadas en el texto.

Conceptos muy genéricos y por frase.

FastUMLS:

30 a 40% de los conceptos (Posición 200 del array)

Muestra conceptos de más vocabularios (no sólo GO)

Parece coherente revisar los resultados con expertos (\*)

## Pruebas Comparativas (PLN - Interactive MetaMap IM)

### **Comparación de Resultados Directa.**

Objetivo: Comparar resultados con una herramienta utilizada y referenciada.  
Se procesan la misma frases (200) según las identifica IM dentro de un texto científico.  
Comparamos si los conceptos son los mismos y si están en las 20 posiciones iniciales.

### FastUMLS:

En el 62% de los casos se recuperan los mismos conceptos.  
En el 81% de los casos se recuperan más de la mitad de los conceptos.  
Tiende a penalizar conceptos genéricos (“enfermedad” frente “fiebre”)

### IM:

Identifica negaciones y evita algunos resultados (“not cancerigenous gen”)

## Pruebas Supervisadas – Investigadores en Biomedicina

### Revisión de los resultados por Expertos – Test (5 investigadores, 23 artículos)

**Objetivo: Comprobar la idoneidad o no de los conceptos ofrecidos.**

Varios investigadores procesan el resumen de artículos escritos por ellos mismos.

De acuerdo a los conceptos ofrecidos resultados contestan a un test por cada artículo.

**Para cada uno de los artículos y atendiendo sólo a los 20 primeros conceptos valore cada una de las siguientes preguntas:**

#### 1.- Los conceptos ofrecidos son correctos...

- a) En un 100%.
- b) Entre 80% y 100%
- c) Entre 50% y 80%
- d) En menos de un 50%

#### 3.- Respecto a la especificidad de los conceptos ofrecidos... (un concepto demasiado general sería "enfermedad" en un artículo sobre "la enfermedad Alzheimer")

- a) Son bastante específicos en su mayoría.
- b) En general hay conceptos específicos y genéricos a partes iguales.
- c) En general son demasiado genéricos.

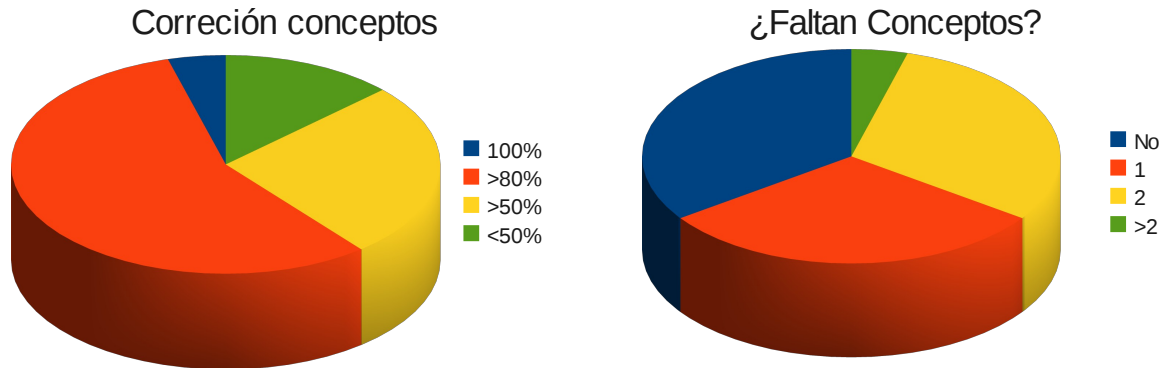
#### 2.- Respecto a los conceptos ofrecidos, ¿se echa de menos algún concepto relacionado con el artículo?

- a) No, ninguno.
- b) Sí, uno
- c) Sí, dos.
- d) Sí, más de dos.

#### 4.- La valoración general de los resultados es:

- a) Muy buena.
- b) Buena, pero necesita mejorar.
- c) Media, necesita mejorar bastante.
- d) Mala.

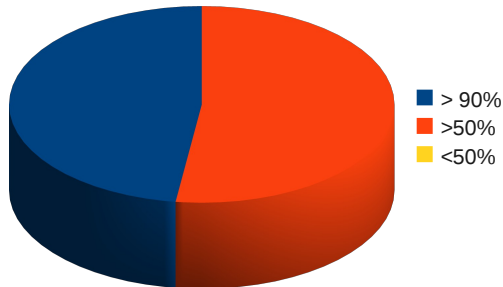
## Pruebas Supervisadas – Investigadores en Biomedicina



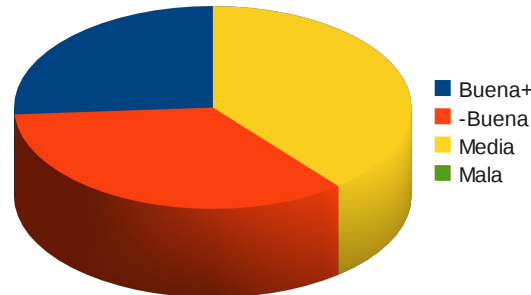
- **60%** - Más de un 80% de los conceptos son adecuados.
- **85%** - Más de la mitad de los conceptos ofrecidos son correctos.
- **35%** - No se ha echado en falta ningún concepto.
- **95%** - El número de conceptos a incorporar no es mayor que dos.

## Pruebas Supervisadas – Investigadores en Biomedicina

¿Son Específicos?



Valoración General



- **48%** - Los conceptos son específicos en un 90%
- **100%** - La mitad de los conceptos o más son específicos.
- **60%** - la herramienta ofrece resultados buenos o más que buenos.
- **75%** - los resultados son buenos pero necesitan mejorar.

## Conclusiones

FastUMLS es eficaz y preciso a la hora de ofrecer conceptos de UMLS relacionados con un texto corto o una lista de palabras, permutando las mismas y asignado pesos.

FastUMLS es eficaz y útil para la anotación automática de textos largos que traten sobre un mismo tema.

Asignar pesos a los términos en base a la probabilidades se muestra como mejor estrategia que mostrar sólo los términos con aciertos totales.

Preprocesar los datos y los cálculos de pesos incrementa notablemente el rendimiento.

FastUMLS debe mejorar en aspectos como: La redundancia de conceptos, en el preprocesado de cadenas cortas (“10”) y en la identificación de negaciones,

## Enriquecer las búsquedas

Incorporar al flujo de proceso las relaciones entre conceptos y los grupos semánticos que ofrece UMLS, y así incrementar el número de conceptos a tratar y a ponderar

## Creación de Bases de Datos de Textos anotados

Procesar los numerosos artículos científicos de PubMed para asociar conceptos a cada uno. Utilizar esta bases de datos para mejorar las búsquedas de documentos relacionados con uno dado a través de sus conceptos.

## Identificación de Grupos de Conceptos y Palabras o Textos

Utilizar técnicas de bi-clustering para identificación de grupos en la matriz de palabras por conceptos que ofrece FastUMLS, y poder inferir relaciones entre los conceptos de cada grupo.

---

# ¿Preguntas?

---



Extracción de Conceptos en Textos Biomédicos

**Facultad de Informática  
Universidad Complutense  
de Madrid**