

# El filólogo estructurado



Javier G. Sogo



@jgsogo



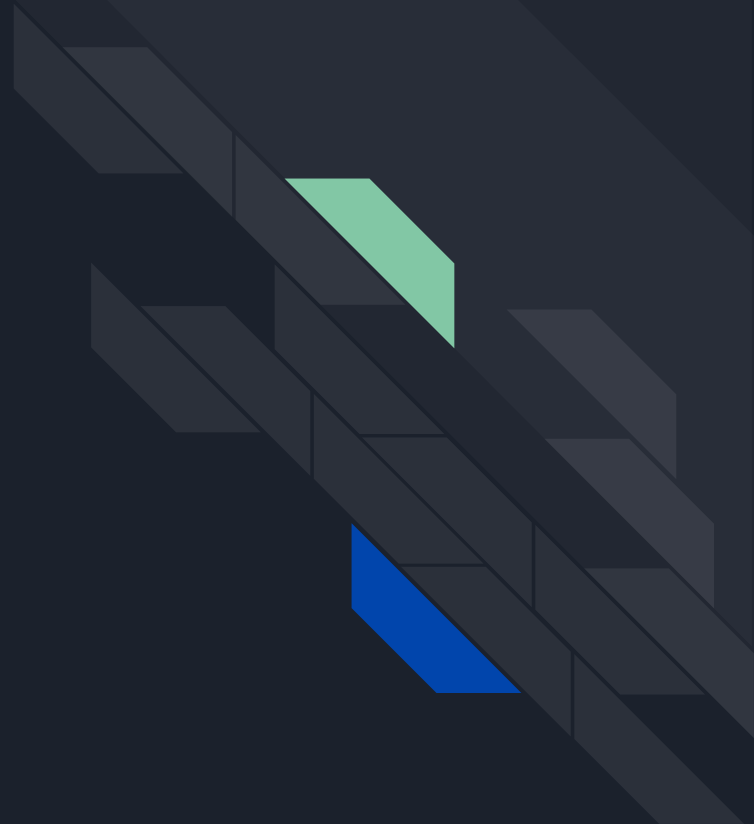
Lingwars



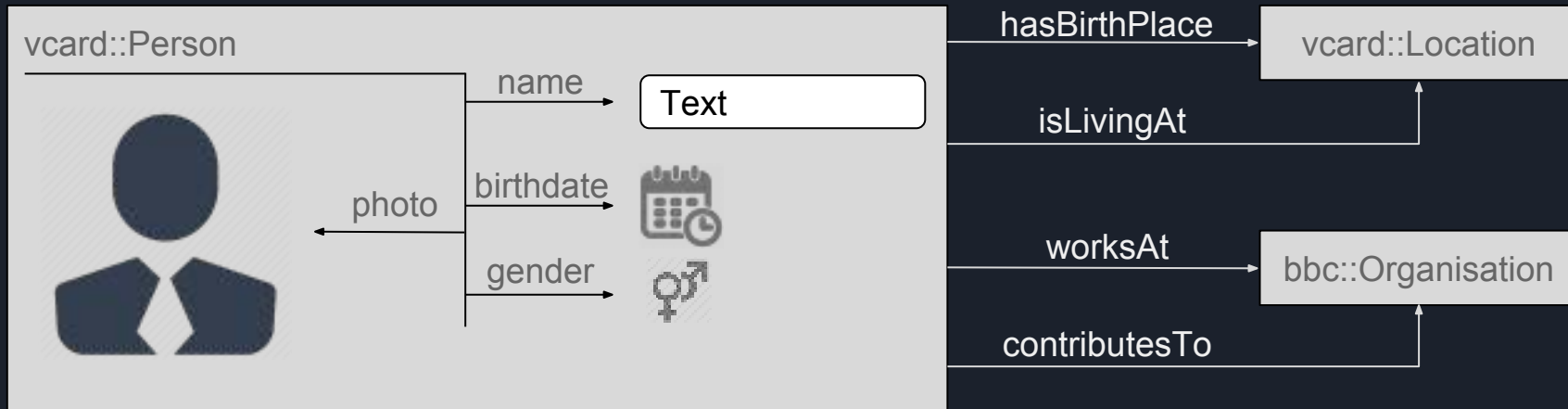
@lingwars



¿Quién es @jgsogo?



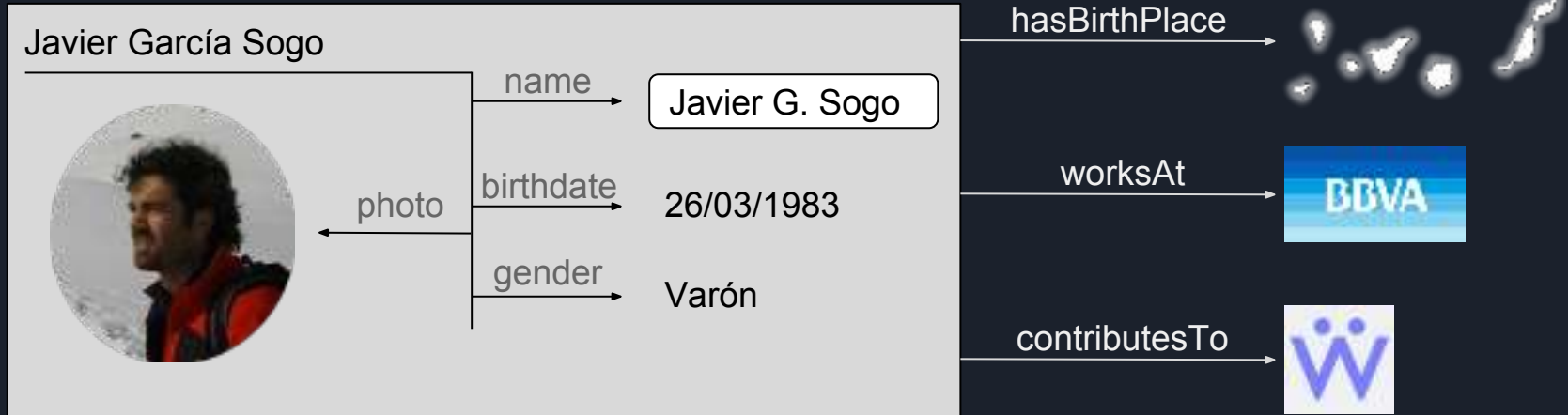
# ¿Quién es @jgsogo?



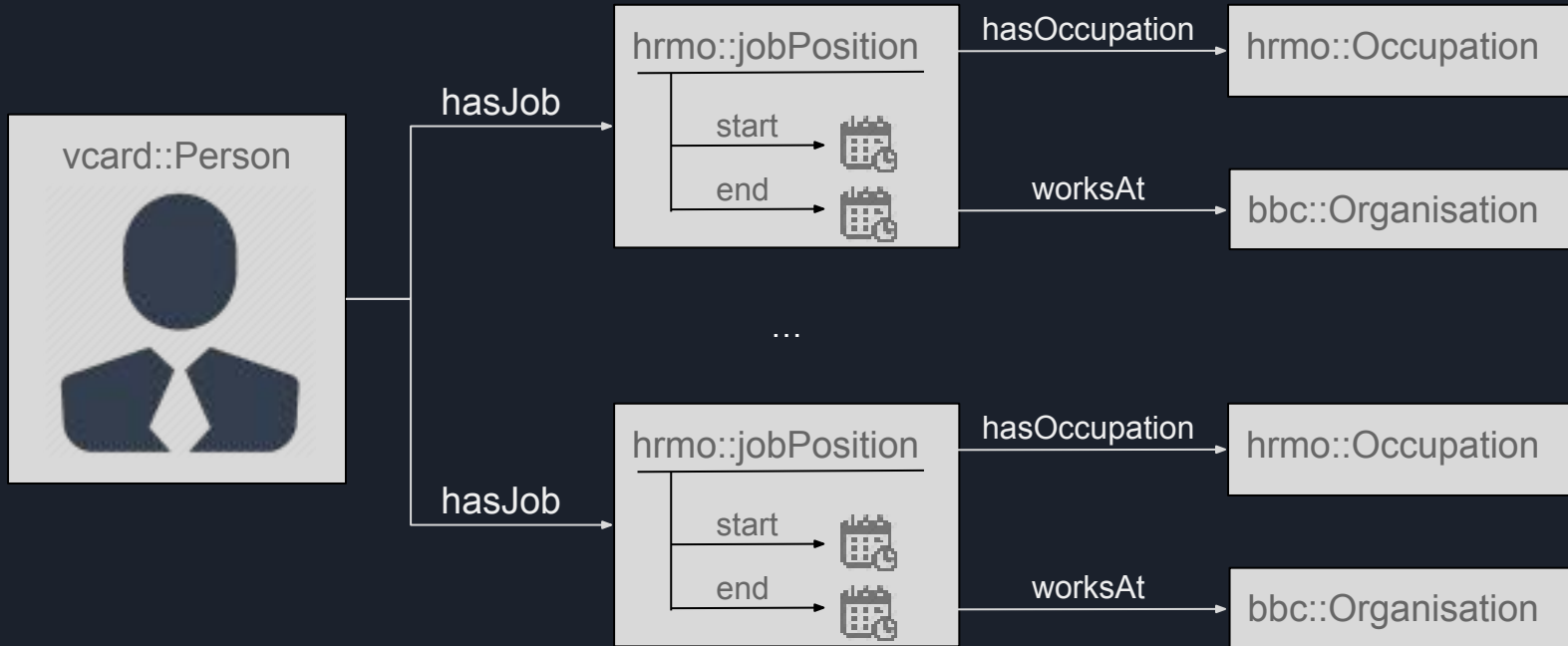
vcard: vCard Ontology (<https://www.w3.org/TR/vcard-rdf/>) mantenida por la W3C

bbc: BBC Ontology (<https://www.bbc.co.uk/ontologies>) mantenida por la BBC

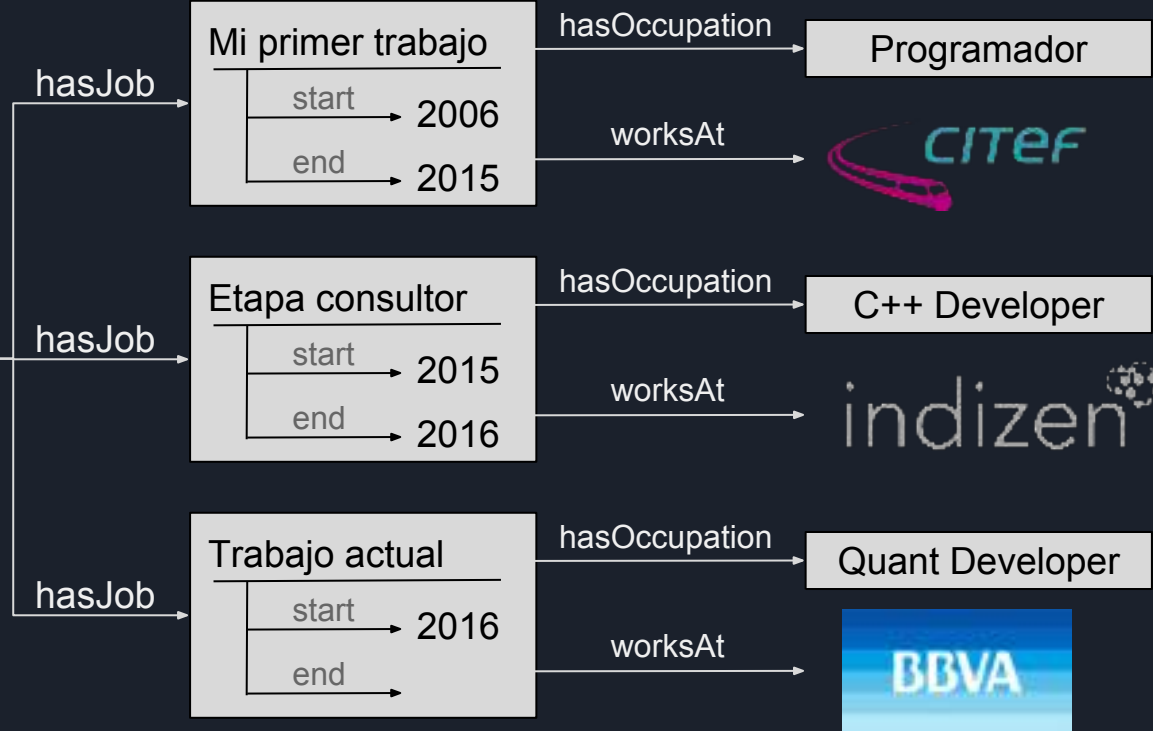
# ¿Quién es @jgsogo?



# ¿Quién es @jgsogo?



# ¿Quién es @jgsogo?



# ¿Y eso de Lingwars?

Grupo de personas (“asociación”)

Objetivo:

- derribar el muro entre ciencias y letras

Somos:


- techies humanistas
- lingüinis filotecnólogos

Organizamos:

- eventos
- talleres
- charlas



Lingwars

 @lingwars

worksAt

vcard::Person



bbc::Organization



subclassOf

bbc::Association

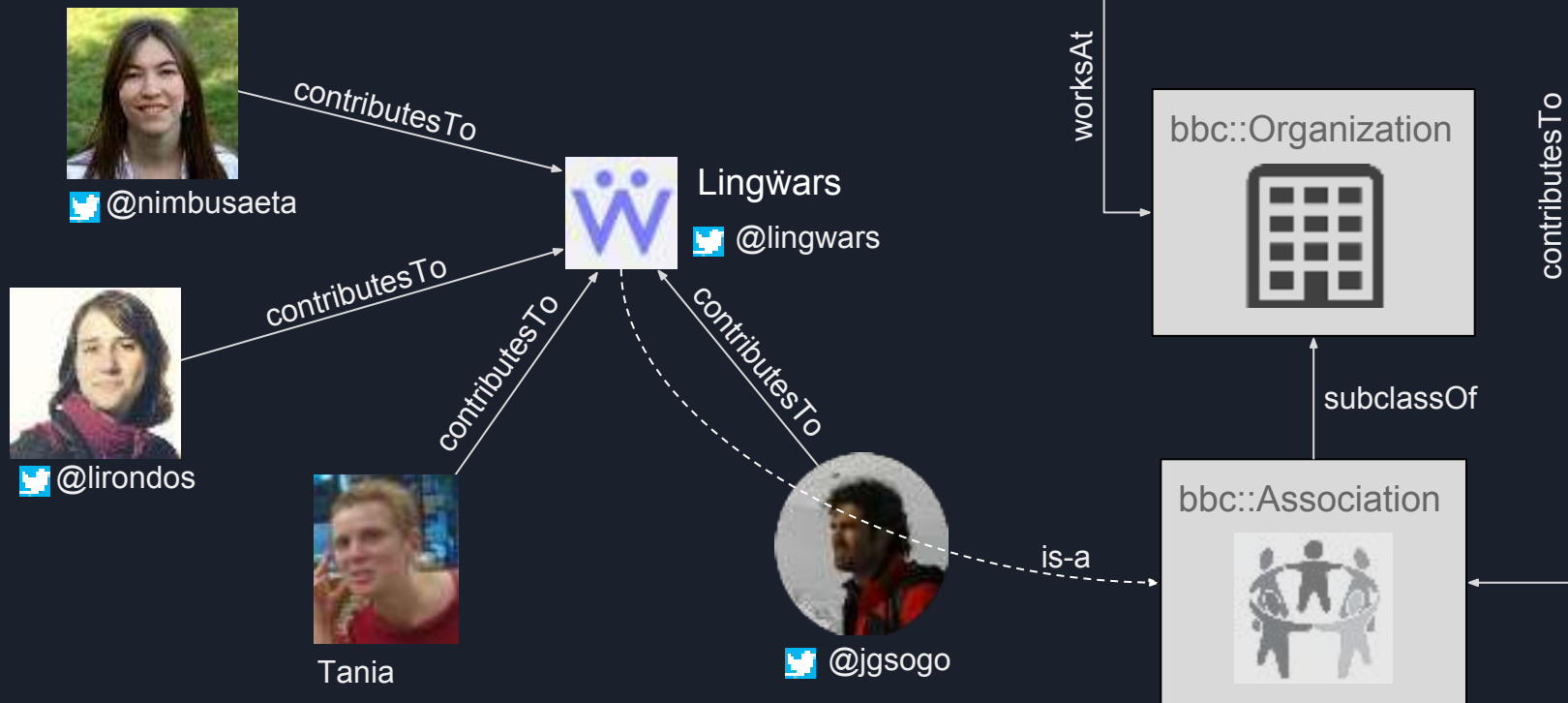


contributesTo

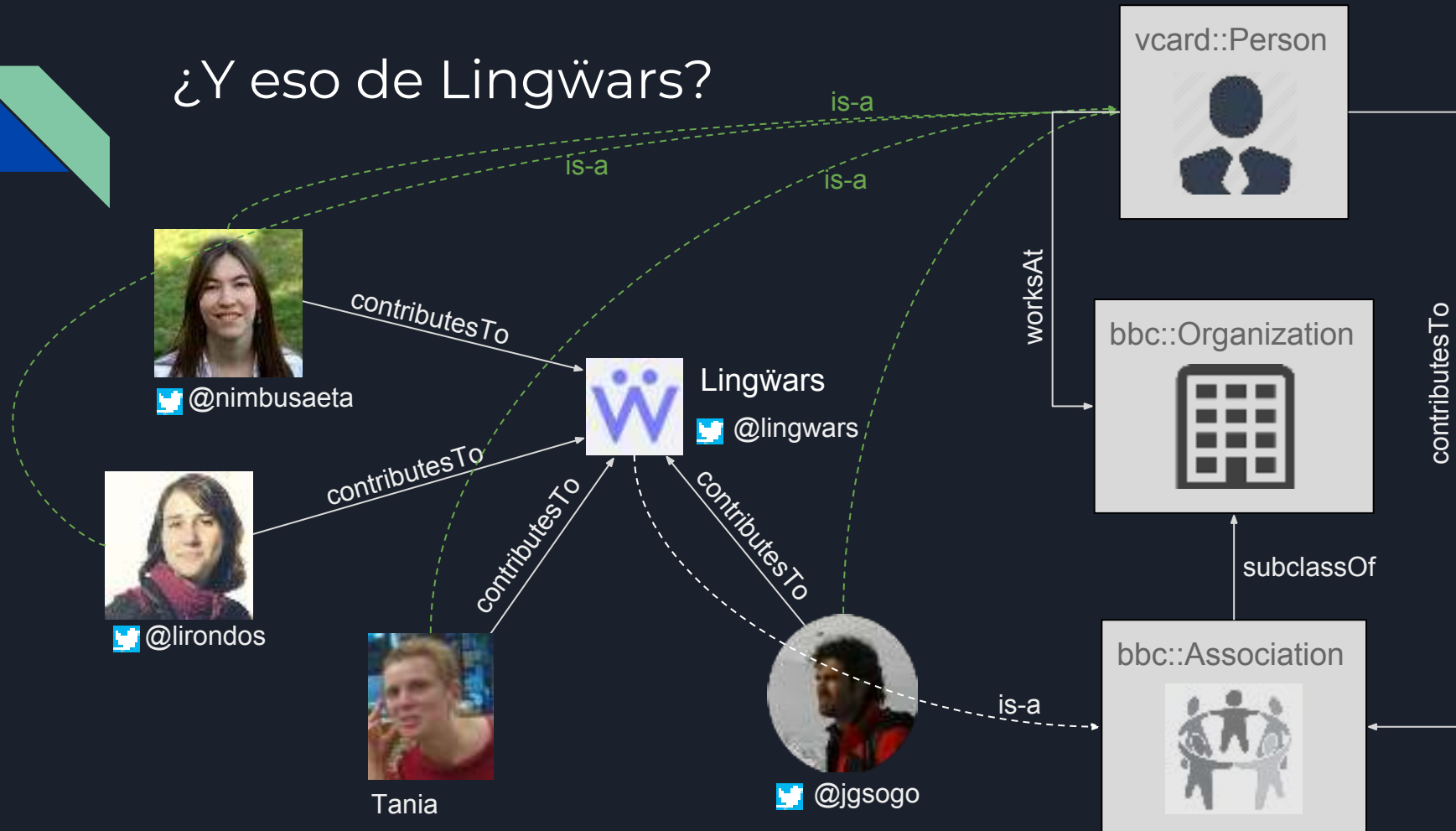
is-a



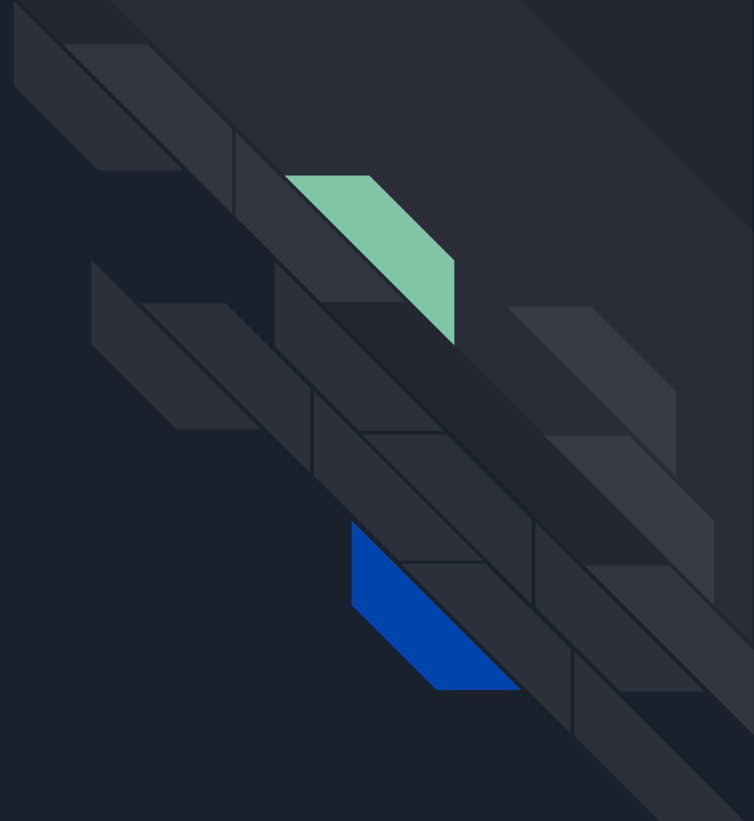
# ¿Y eso de Lingwars?



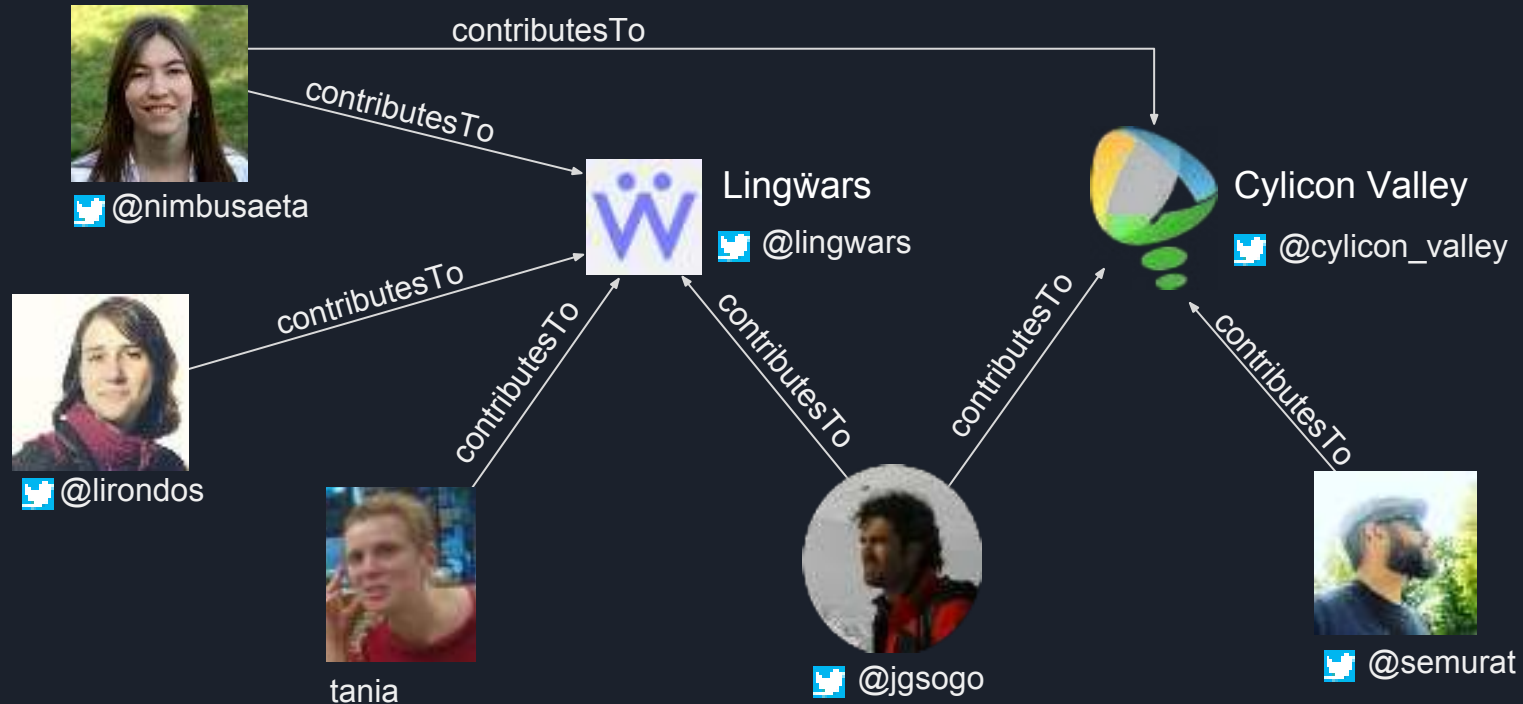
# ¿Y eso de Lingwars?



Más información sobre  
[@jgsogo](#)



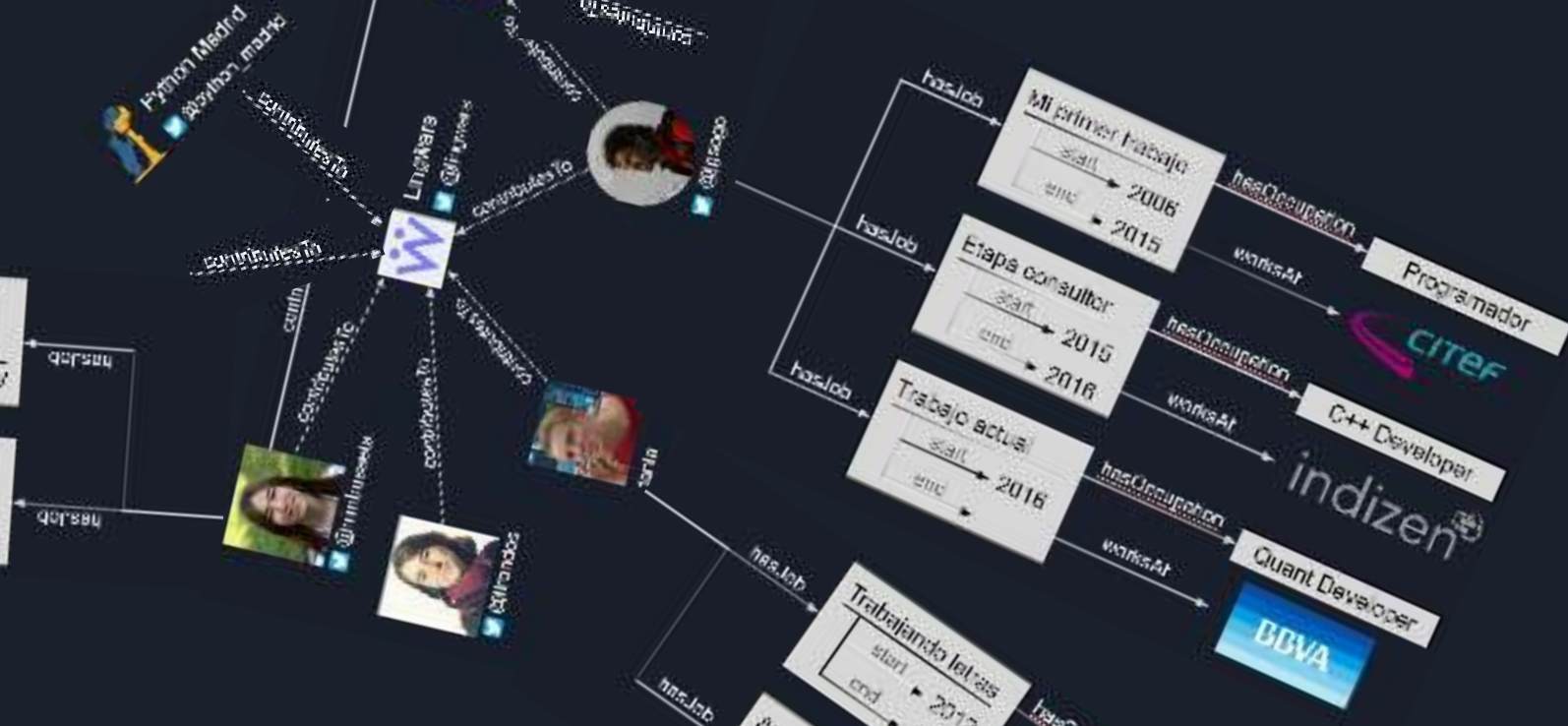
# Más información





- 
- The image shows a resume for a C++ Developer. The resume is tilted and features a dark blue background with white and yellow text. It includes a header with 'CITEF' and 'indizen' logos, a 'Trabajando en' section with a timeline from 2014 to 2016, and an 'Experiencia' section with roles at 'Programador', 'C++ Developer', and 'Quant Developer' at 'BBVA'.

- ¿Qué personas actualmente están trabajando en BBVA y que contribuyan en asociaciones de Valladolid?







Internet.  
Buscando información







Búsqueda por “palabras clave”

10.000.0000.00000000 resultados

**Recuperación de información\*  $\Rightarrow$  Documentos**

\* Information retrieval



**Extracción de información\*  $\Rightarrow$  Datos**

\*Information extraction

Imagen: Papou "Shouting in despair" en DevianArt



# El objetivo final en PLN

**Convertir el lenguaje en datos estructurados**, de tal forma que una máquina sea capaz de trabajar con él:

- Buscar datos concretos
- Generar lenguaje
- Resumir contenidos

Will a computer program ever be able to convert a piece of English text into a programmer friendly data structure that describes the meaning of the natural language text? Unfortunately, no consensus has emerged about the form or the existence of such a data structure. Until such fundamental

Collobert *et al.* 2011. Natural Language Processing (Almost) from Scratch



# El objetivo final en PLN

Utilizar Internet como una **gran base de datos global**:

- Convertir todo en datos estructurados
- Realizar inferencias sobre los datos

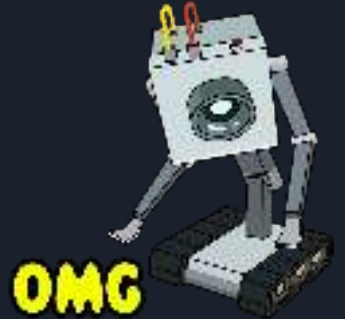
The Semantic Web is an **extension** of the current Web in which information is given well-defined meaning, better enabling computers and people to work in **cooperation**. It is based on the idea of having data on the Web defined and **linked** such that it can be used for more effective discovery, automation, integration, and reuse across various applications. For the

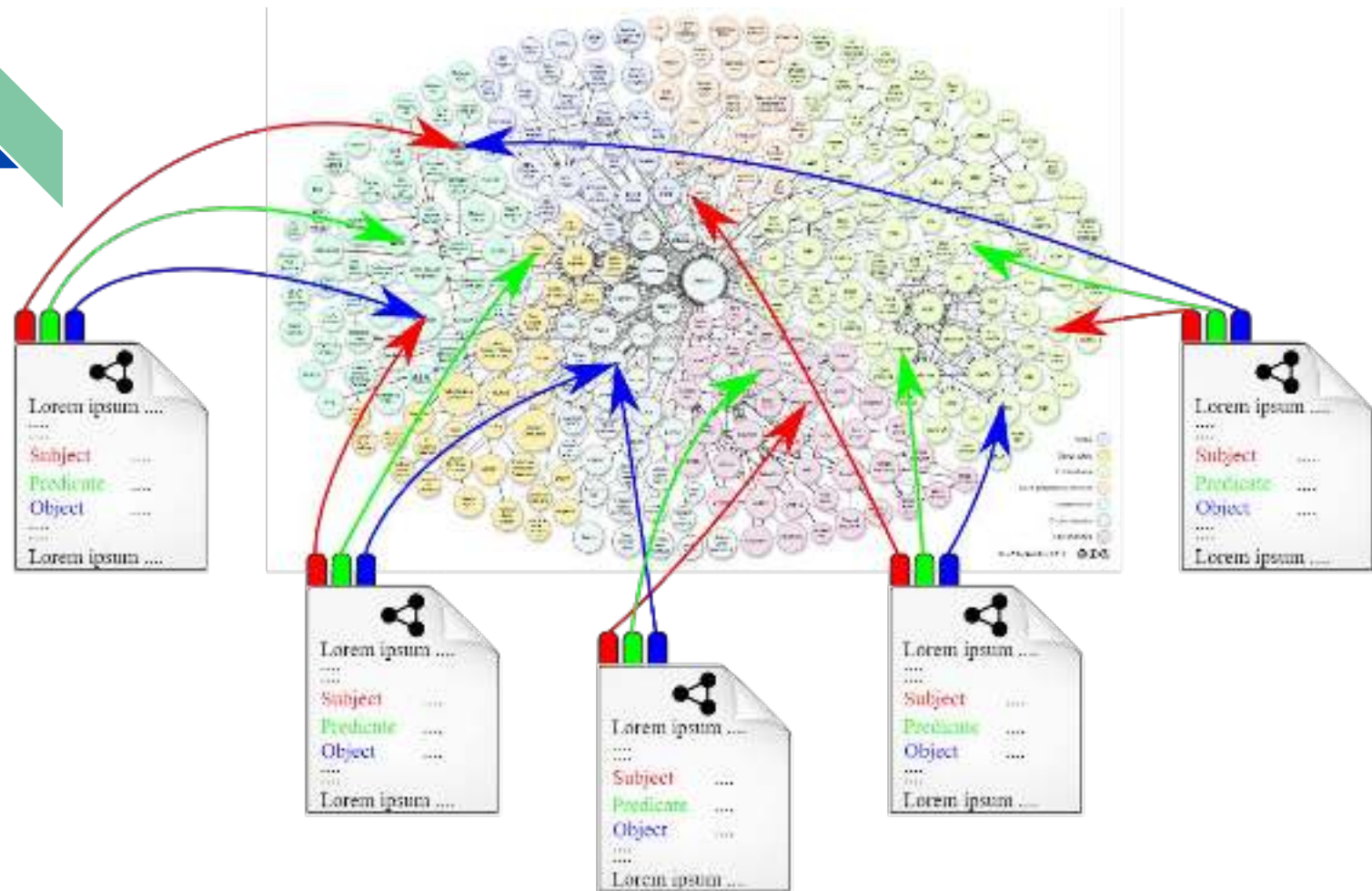
Hendler, J., Berners-Lee, T., Miller, E. Integration Applications on the Semantic Web, 2002

# La Web Semántica (Web 3.0)

Los datos están representados de tal forma que las máquinas pueden entenderlos:

- Encontrar datos concretos en un documento
- Acceder a información relacionada con los datos
- Realizar inferencias
- Asistentes personales







Linked Open Data. 2014

Linked Datasets as of April 2014

Linked Open Data. 2014



Lo que piensa....

...el ingeniero







Lo que piensa....

...el lingüista

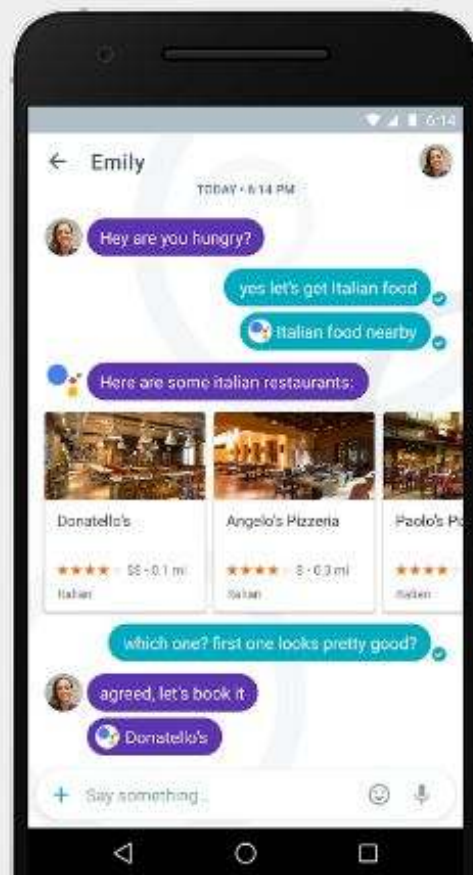


La realidad es...

... un chatbot para  
pedir comida

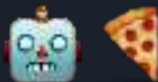


Get answers in your chat from  
**the Google assistant**

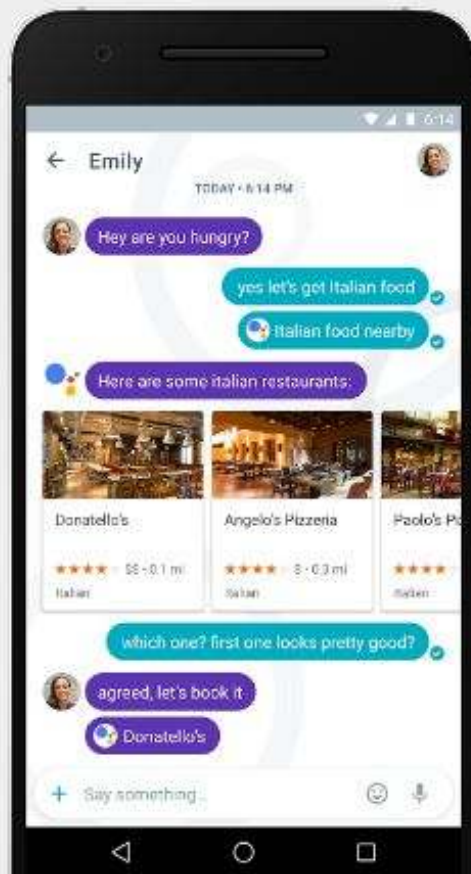


La realidad es...

... un chatbot para  
pedir ~~comida~~ pizza



Get answers in your chat from  
the Google assistant

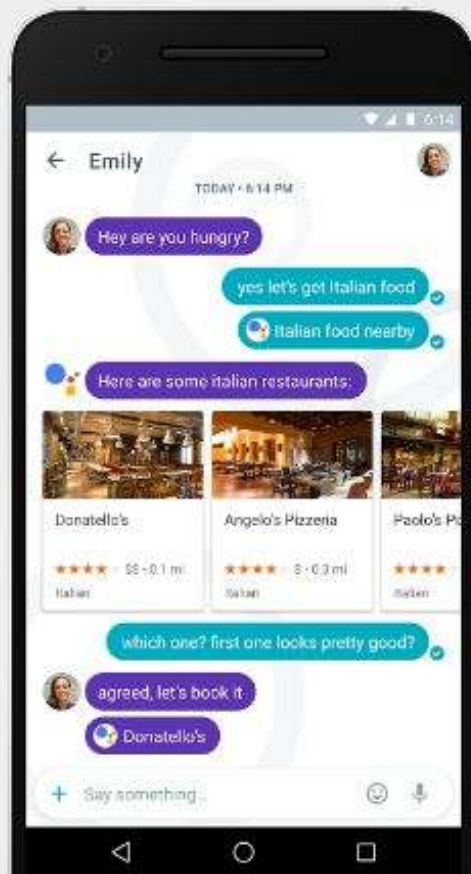


La realidad es...

... un chatbot para  
pedir ~~comida~~ pizza  
en inglés



Get answers in your chat from  
**the Google assistant**



Convertir el lenguaje en  
datos estructurados



# Convertir el lenguaje en datos estructurados

¿Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta?

1. Tokenizar
2. Pos-tagging
3. Resolver correferencias
4. Reconocimiento de entidades
5. Semántica
6. Codificación de dependencias



# Paso 1. Tokenizar

Identificar oraciones y palabras

¿ Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta ?

- En **español** es sencillo separar palabras. Hay muy pocas excepciones.
- ¿Y separar las oraciones? Atención a abreviaturas, acrónimos,...
- **Ambigüedad**: el contexto nos indica qué tiene sentido considerar una palabra

en dos mil quince

en dos mil quince

en dos mil quince

## Paso 2. Etiquetar pos-tagging

Rol gramatical (Part Of Speech) de cada palabra

¿	Qué	persona	trabajó	en	Indizen	en	2015	y	contribuye	en	la	misma	asociación	que	@nimbusaeta	?
DET	NOUN	VERB		NOUN		NUM		VERB					NOUN		NOUN	

- Análisis morfológico de toda la vida.
- ¿Qué conjunto de etiquetas utilizar?
  - EAGLES ⇒ castellano
  - Universal dependencies ⇒ conjunto válido para todas las lenguas
- **Ambigüedades.**



# Paso 3. Resolver correferencias

Identificar términos que hacen referencia al mismo ente

¿ Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta ?

DET

NOUN

VERB

NOUN

NUM

VERB

NOUN

NOUN



One morning I shot an elephant in my pajamas.  
How he got into my pajamas I'll never know.

(Groucho Marx)

izquotes.com

Tokenización

Pos-tagging

Correferencias

## Paso 4. Identificar entidades nombradas

Named Entity Recognition (NER): personas, lugares, organizaciones,...

¿ Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta ?

DET

NOUN

VERB

NOUN

NUM

VERB

NOUN

NOUN

indizen<sup>®</sup>



- Personas: nombres propios, nicks,...
- Mayúsculas (español),...
- Búsqueda en bases de datos (DBPedia)

Tokenización

Pos-tagging

Correferencias

Entidades

# Paso 5. Desambiguar semánticamente

Utilizar el contexto para elegir el significado

¿ Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta ?

DET

NOUN

VERB

NOUN

NUM

VERB

NOUN

NOUN

indizen<sup>8</sup>



- “Encontrar en el diccionario” la entrada correspondiente a cada palabra (¡lematizar!).
- Persona:
  - 1. f. Individuo de la especie humana
  - 8. f. Gram. Categoría gramatical inherente en algunos pronombres [...]

Tokenización

Pos-tagging

Correferencias

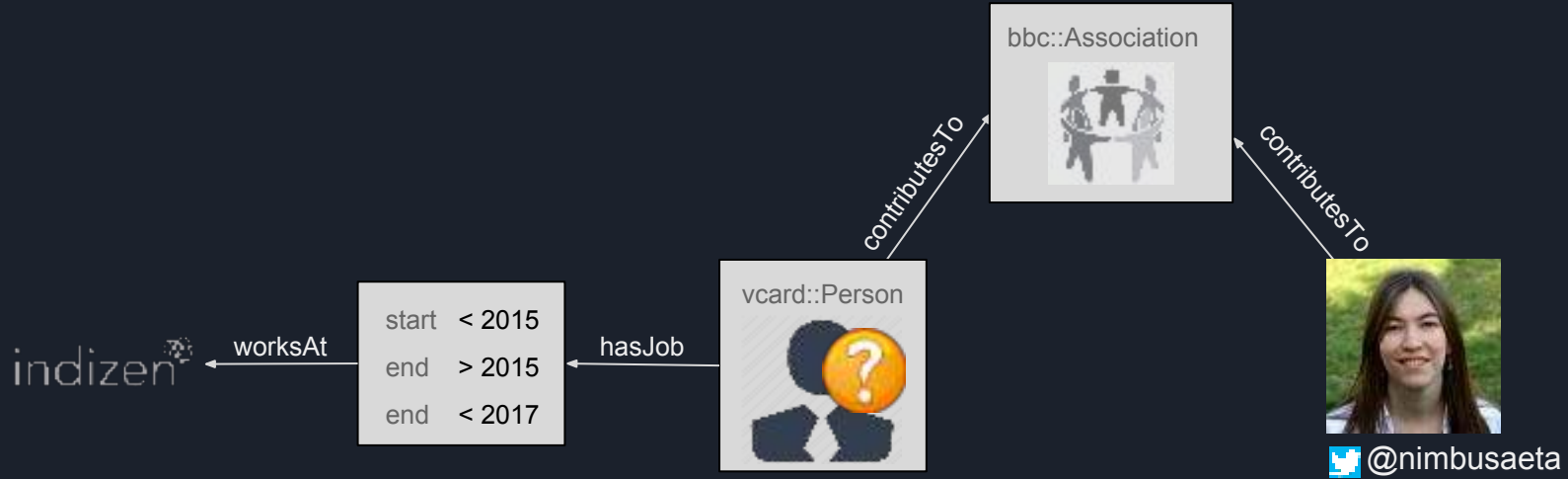
Entidades

Semántica

## Paso 6. Codificación en una estructura

Codificar el lenguaje en una estructura inteligible para la máquina

¿Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta?



Tokenización

Pos-tagging

Correferencias

Entidades

Semántica

Dependencias

# Paso 6. Codificación en una estructura

Codificar el lenguaje en una estructura inteligible para la máquina

¿Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta?



# Ontologías





# Inteligencia artificial

Fracaso tras fracaso

1315. Raimundo Lulio. Razonamiento mecánico

1950. Alan Turing. Juego de imitación

Redes neuronales

1960 - 1974. Sistemas basados en reglas -- Problemas de búsqueda

Representación del conocimiento

Procesamiento de lenguaje natural (Guerra Fría)

1980 - 1987. Sistemas expertos

Redes neuronales (de verdad)

1993 - 2001. Algoritmia

2001. Internet: Big Data, Deep learning, buzz words....





# Ontología. Definición

Gruber (1993). Explicit specification of a conceptualization.

Borst (1997). Formal specification of a shared conceptualization.

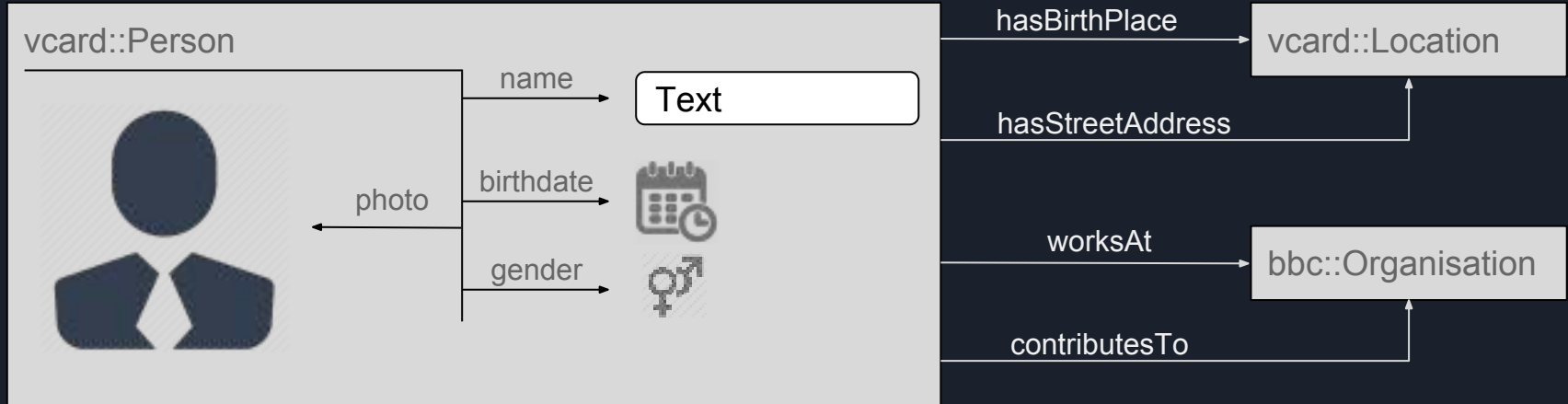
Studer (1998). An ontology is a formal, explicit specification of a shared conceptualization.

<b>Domain</b>	Ceñida a un dominio de conocimiento
<b>Explícita</b>	Sin ambigüedades
<b>Formal</b>	La máquina tiene que poder entenderlo
<b>Compartida</b>	Tiene que haber consenso



# Ontología. Componentes

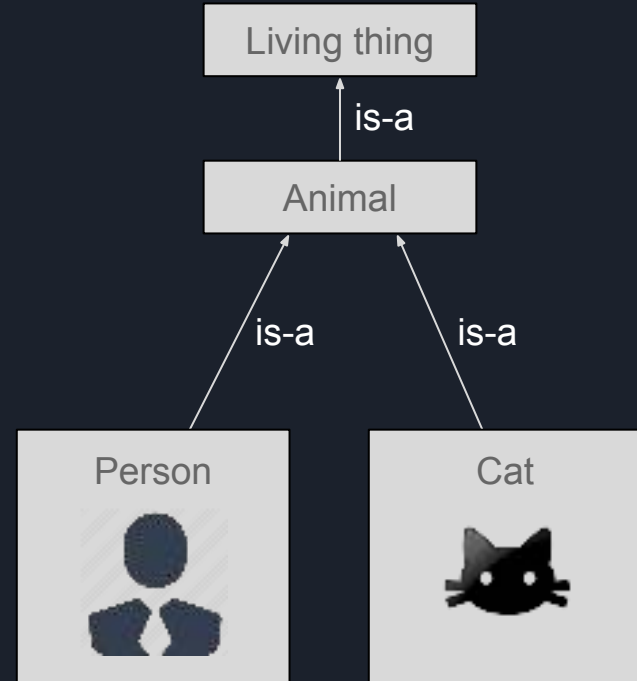
1. Clases - Conceptos
2. Atributos - Propiedades (de un concepto)
3. Relaciones (entre conceptos)



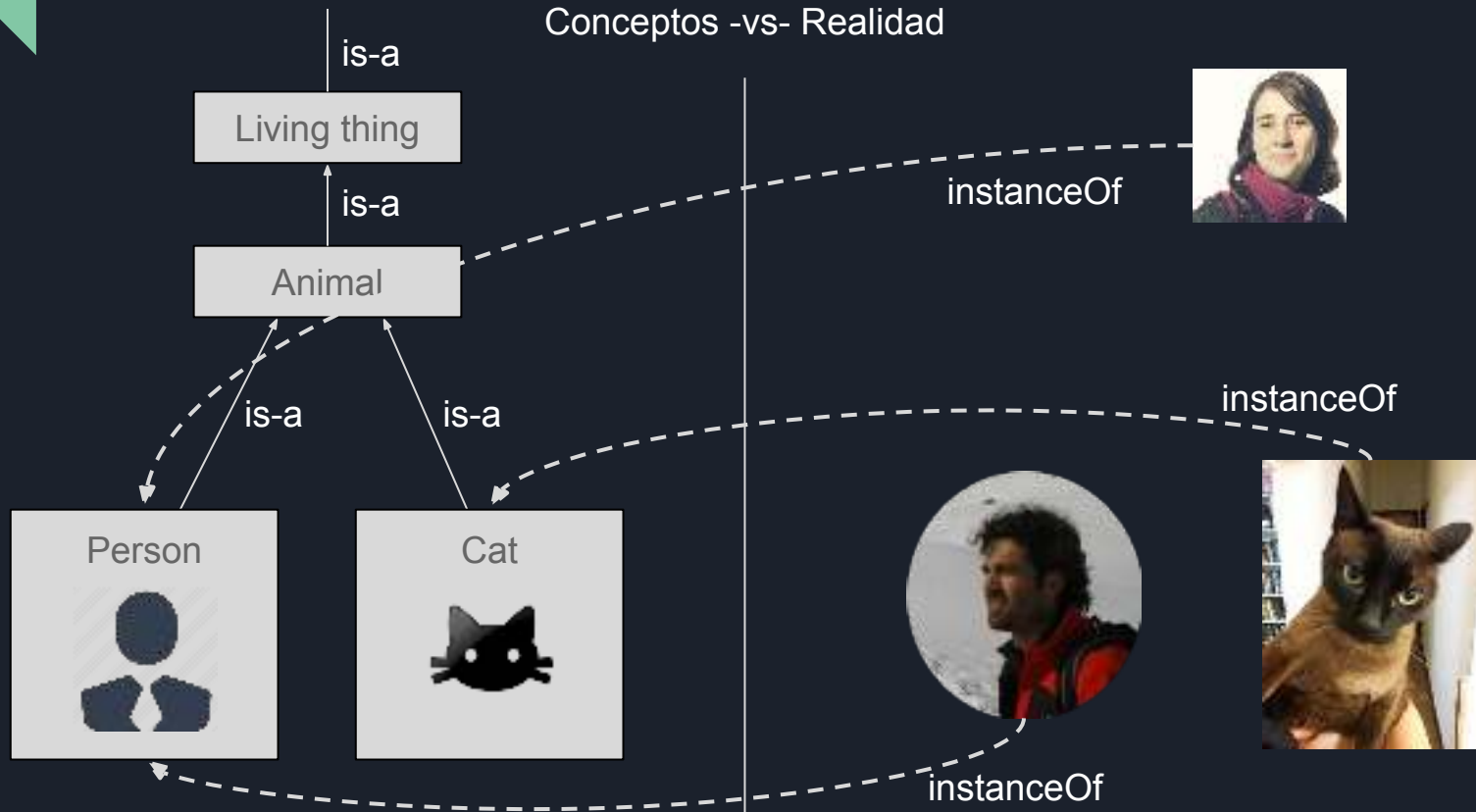
# Ontología. Conceptos

Son los nodos de la red:

- Se definen por sí mismos (o con atributos)
- Se organizan en taxonomías
  - relación is-a: hiperonimia/hiponimia
- Propiedades:
  - completitud
  - conocimiento disjunto

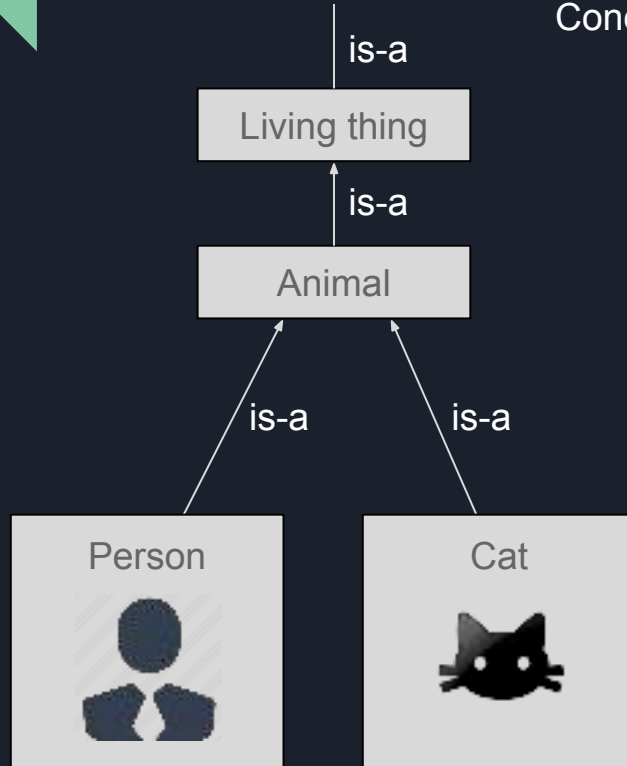


# Ontología. Clases - Conceptos



# Ontología. Clases - Conceptos

## Conceptos -vs- Realidad



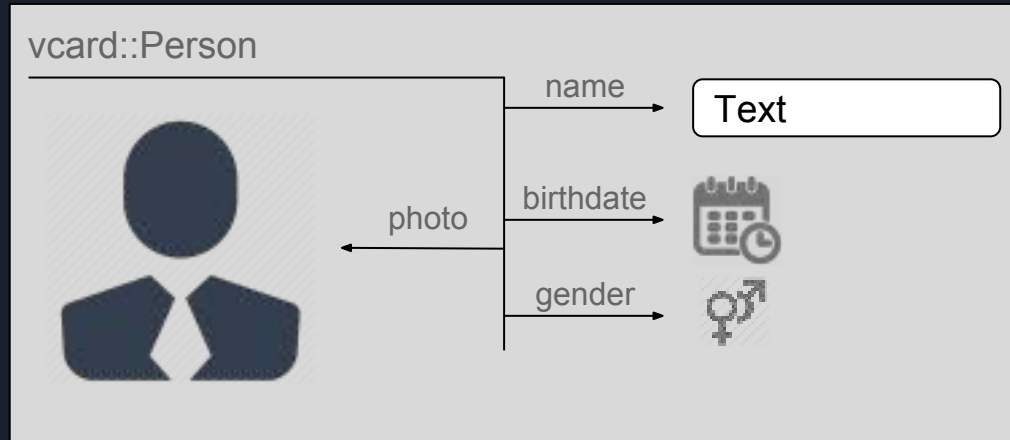
Completo: todo el dominio de conocimiento

Disjunto: nada puede ser dos cosas



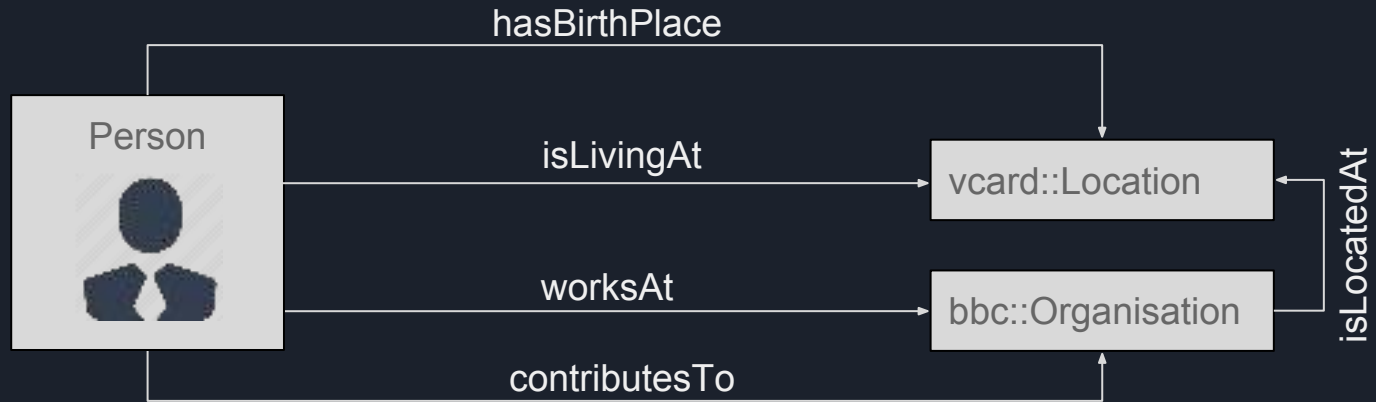
# Ontología. Propiedades - Atributos

- Definen características de un único concepto
- Información concreta



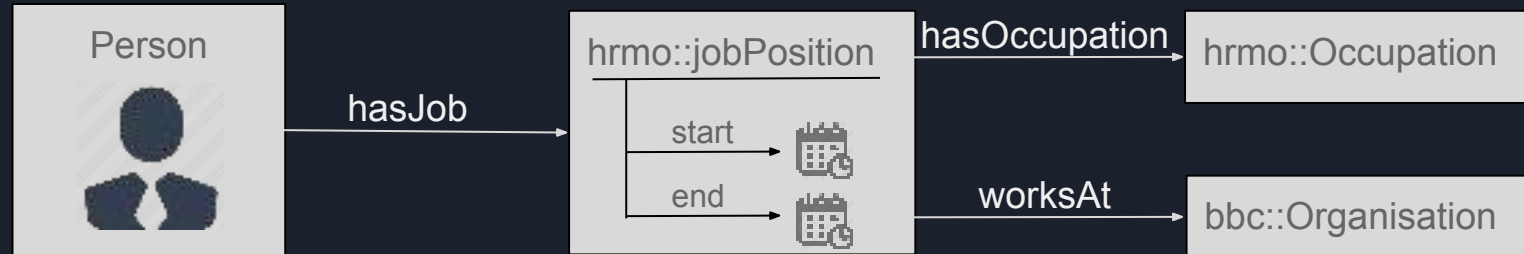
# Ontología. Relaciones

Pueden existir varias relaciones entre dos mismos conceptos, cada una de ellas con significado diferente



# Ontología. Conceptos complejos

Codifican relaciones complejas entre más de dos conceptos:



# LIKE A BOSS

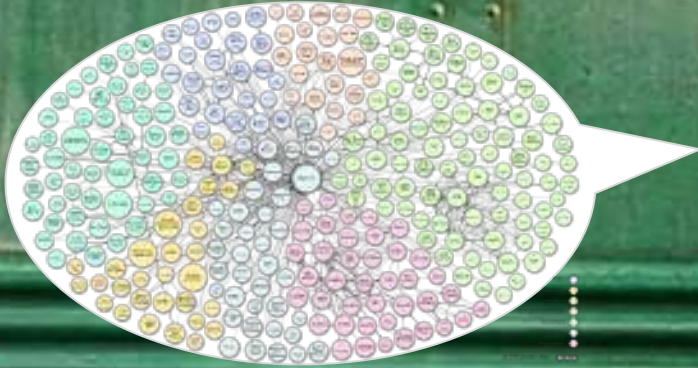
Es tu ontología y defines:

- lo que quieres
- como quieres





ANKS CITY TRANSIT SYSTEM



**“Happiness is only real when shared”**



# Ontología. Ejemplos

## Linked Open Data

- DC: The Dublin Core
- FOAF: Friend of a friend
- Geonames
- RDFCalendar
- ....
- WordNet

- BabelNeet
- Gazetteer

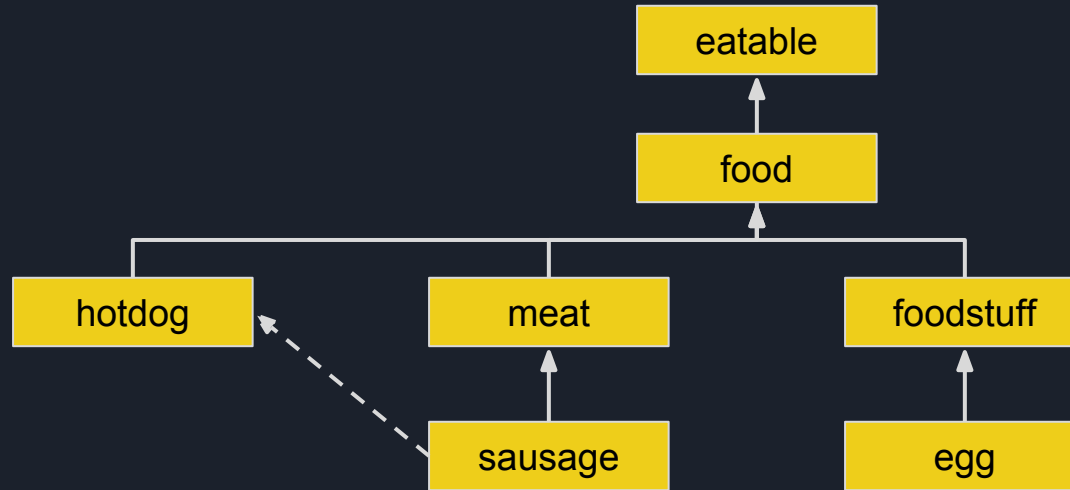
## Privadas:

- Chatbots
- Buscadores
- ¿?

# WordNet.

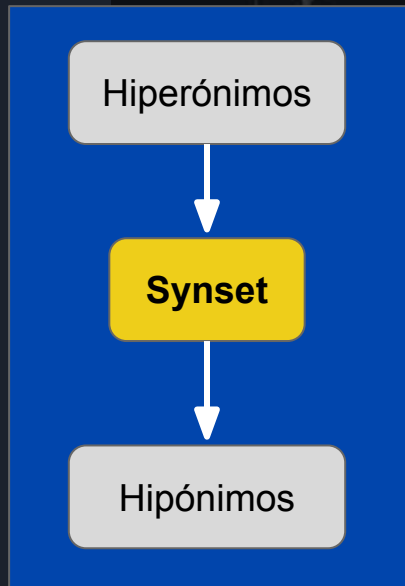
## La ontología del lenguaje

Red de **conceptos** (synsets) con relaciones entre ellos: sinonimia, hiperonimia, hiponimia,...



# Hiperonimia - Hiponimia

Concreción en significado



Animal

**Ser orgánico que...**

Animals

Mamífero

**Dicho de un animal: Del grupo de...**

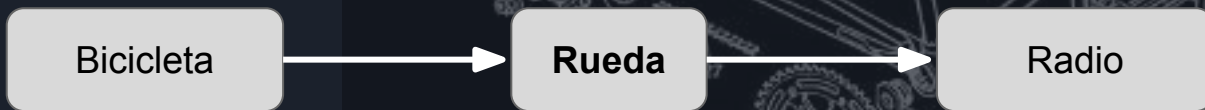
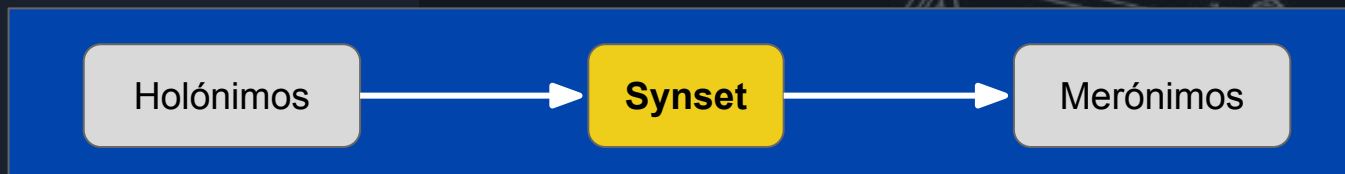
Cetáceo

**Dicho de un mamífero: Del grupo de...**



# Holonomia - Meronimia

El todo y la parte



# El filólogo estructurado





# El filólogo estructurado

Recoger todo el conocimiento de un dominio

Resumirlo en conceptos y relaciones

Codificarlo para una máquina (base de datos)

Explotar su utilización

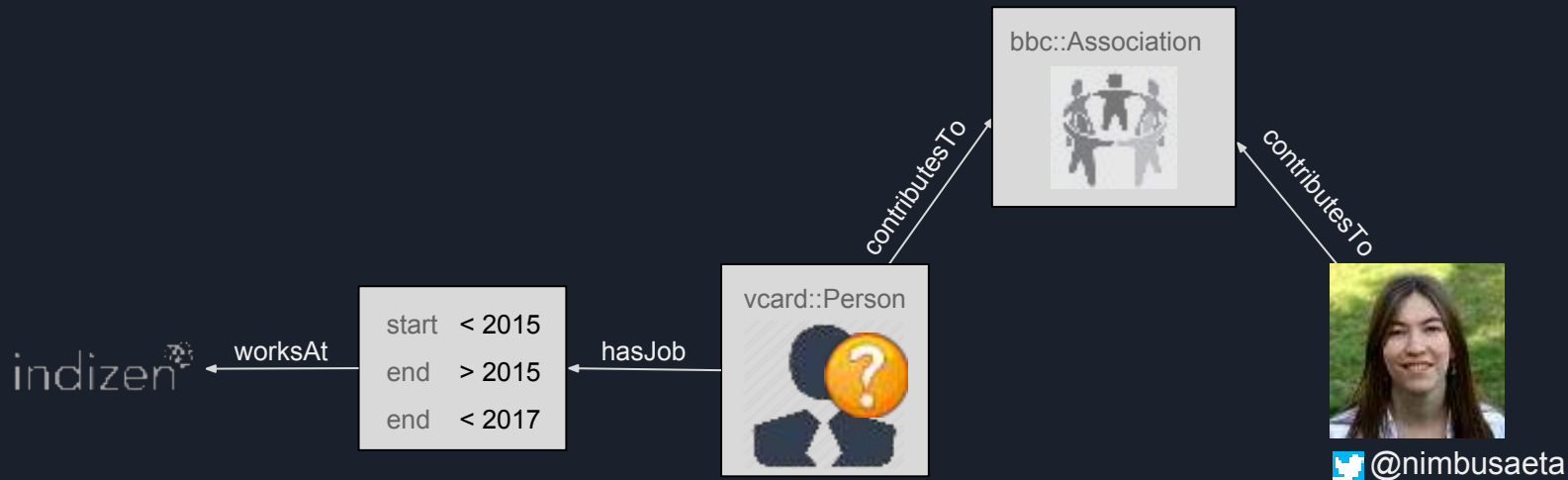
Imagen: Picasso. Retrato de Wilhelm Uhde. 1910



# ¿Por qué un filólogo?

Nuestro objetivo es parsear una oración y convertirla en datos estructurados. **DESAMBIGUAR**

¿Qué persona trabajó en Indizen en 2015 y contribuye en la misma asociación que @nimbusaeta?







# ¿Por qué un filólogo?

## Trabajar:

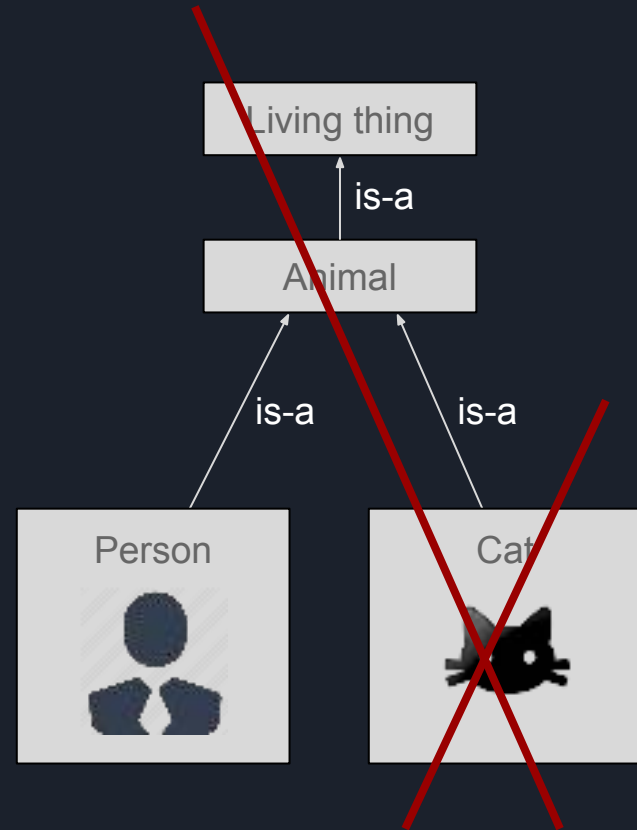
2. tener una ocupación remunerada en una empresa, una institución, etc.. (DLE)

# ¿Por qué un filólogo?

## Trabajar:

2. tener una ocupación remunerada en una empresa, una institución, etc.. (DLE)

- Agente: Animal >> Person

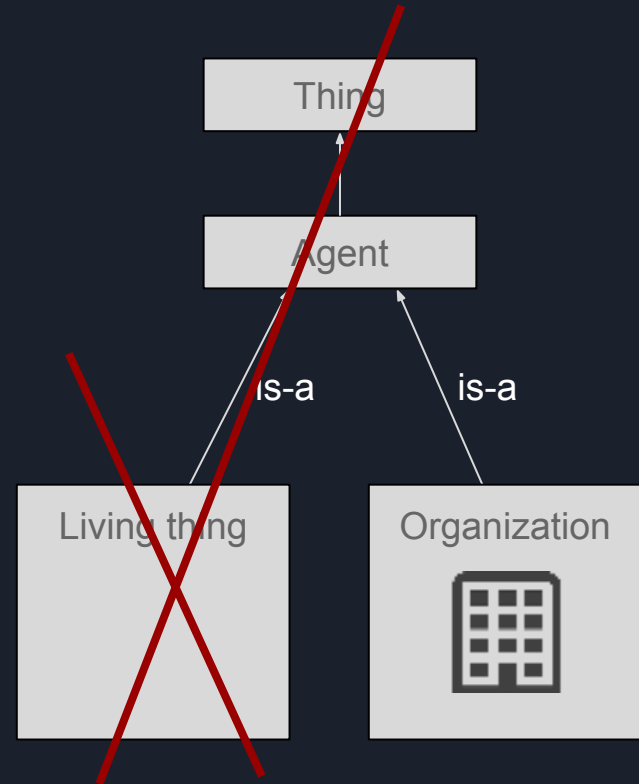


# ¿Por qué un filólogo?

## Trabajar:

2. tener una ocupación remunerada en una empresa, una institución, etc.. (DLE)

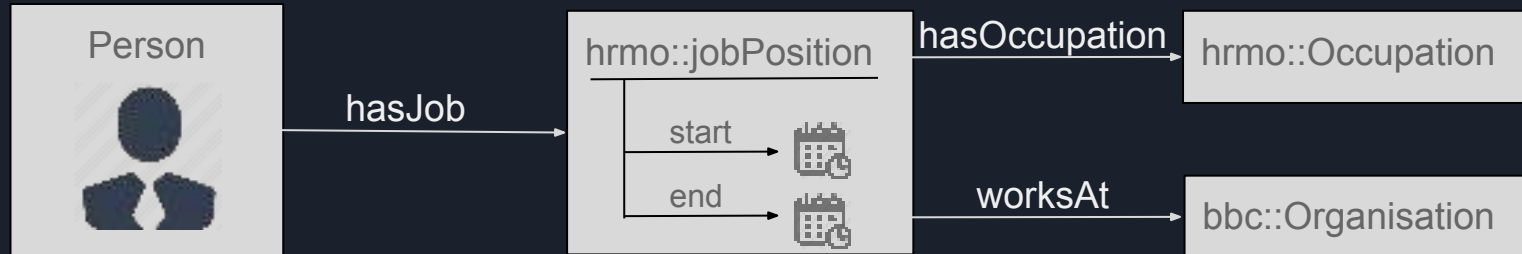
- Agente: Animal >> Person
- C1: Thing >> Agent >> Organization



# ¿Por qué un filólogo?

## Trabajar

2. tener una ocupación remunerada en una empresa, una institución, etc.. (DLE)



# ¿Por qué un filólogo?

## Viajar

1. Trasladarse de un lugar a otro, generalmente distante, por cualquier medio de locomoción. (DLE)





**Una base de datos con la mente del filólogo**



**PLN consiste en resolver un puzzle**



# Resolver ambigüedades

	Cristina Vela	trabaja	en	la	Universidad de Valladolid	
--	---------------	---------	----	----	---------------------------	--

en

la

---- *stop words*

trabajar

Tengo 18 significados posibles

U. Valladolid

DBPedia me dice que es una [Thing >> Agent >> Organization]

Cristina Vela

⇒ tiene que ser una [Agent >> Living thing >> Animal >> Person]





**¡No hace falta resolver todo el puzzle!**

# PLN. ¿Cuál es el problema fundamental?

Tokenizar

Pos-tagging

Resolver correferencias

Reconocimiento de entidades

Codificación en una estructura



Ontologías

Knowledge databases



# PLN. ¿Cuál es el problema fundamental?

Tokenizar

Pos-tagging

Resolver correferencias

Reconocimiento de entidades

Codificación en estructura

Ontologías

Knowledge databases

**DESAMBIGUAR**



# PLN. ¿Qué aporta más valor?

## Algoritmia:

- Librerías NLP gratuitas: CoreNLP, NLTK, spaCy, Freeling,...
- TensorFlow: SyntaxNet -- código abierto noviembre de 2015
- word2vec (2013): word embedding

## Bases de datos y corpus:

- WordNet: gratuito, última versión de 2005
- ¿Corpus etiquetados?
- Google compra Freebase (700 M\$) en 2010 y lo cierra

El filólogo ha muerto.  
¡Larga vida al filólogo!





**Programamos para no trabajar mañana**



# ¡Larga vida al filólogo!

- Número de lenguas
- El lenguaje evoluciona
- Hay que definir cómo codificar el lenguaje
- Hay mucho por codificar

**¡Hacen falta especialistas!**



# Muchas gracias

## ¿Preguntas?



Javier G. Sogo



@jgsogo



Lingwars



@lingwars



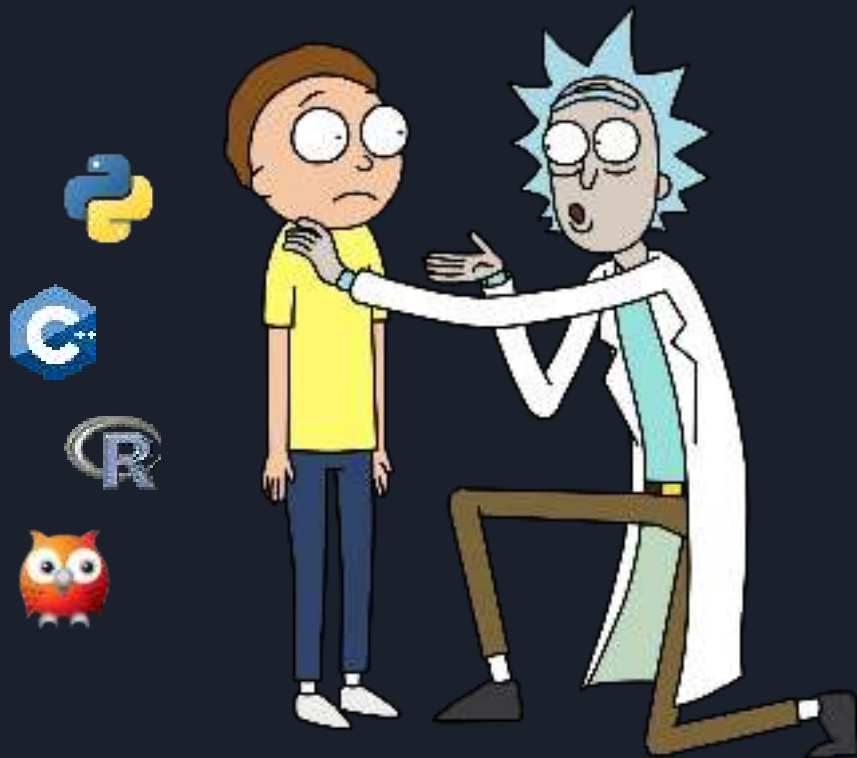




# ¿Por dónde empezar? (1/2)

- UVa - Grado en español (plan 441)

- 41746. Latín
- 41748. Español
- 41749. Inglés
- 41750. Francés
- 41751. Alemán
- 41752. Italiano
- 41756. Latín vulgar
- 41757. Griego
- 41758. Árabe



## ¿Por dónde empezar? (2/2)

- ¡Ponte manos a la obra!
- Internet es tu amigo:
  - Codecademy
  - Tutoriales
  - Blogs
- Busca sinergias en la comunidad (cerca y lejos)
  - Lingwars >> GAPLEN
  - Cylicon Valley

