TELECOM Paris

IP PARIS

INAUGURATION

THE NoRDF PROJECT

MODELING AND EXTRACTING COMPLEX INFORMATION FROM NATURAL LANGUAGE TEXT

Mardi 15 décembre 2020

AGENCE NATIONALE DE LA RECHERCHE ANR

MINISTÈRE DES ARMÉES
Liberté
Égalité
Fraternité

DGA

AGENCE INNOVATION DÉFENSE

Converteo
ADLPerformance
DATA FOR BUSINESS ACCELERATION

eDF

GROUPE BPCE

Schlumberger

Défis scientifiques du projet NoRDF

Fabian M. SUCHANEK

Professeur à Télécom Paris,
Institut Polytechnique de Paris

Titulaire de la chaire NoRDF

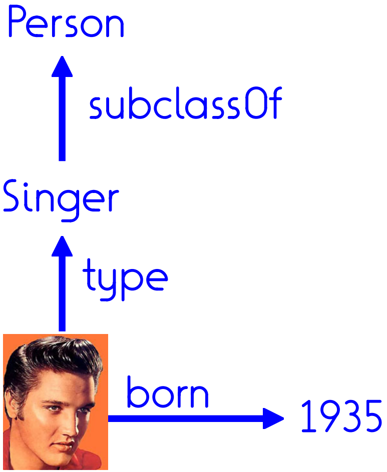
The NoRDF Project

Fabian Suchanek

Amazing! This talk is free of the Corona virus!

(about the speaker, we don't know...)

Knowledge Bases



For us, a knowledge base (KB) is a graph, where the nodes are entities and the edges are relations.

(We do not distinguish T-Box and A-Box.)

Cool knowledge-based applications

When was Elvis born?

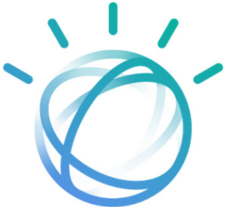
“1935”

Apple Siri

How long was the Thirty Years' War?

Amazon Echo

Discovered 6 kineasis proteins that relate to cancer



IBM Watson

These applications feed from knowledge bases.

There are plenty of knowledge bases

yago
select knowledge

NELL

DBpedia

TextRunner

BabelNet

WIKIDATA

Plus industrial projects at

Google Microsoft ebay facebook IBM

Sponsored Message: YAGO

yago
select knowledge

NELL

TextRunner

We develop YAGO, one of the largest open general purpose KBs.

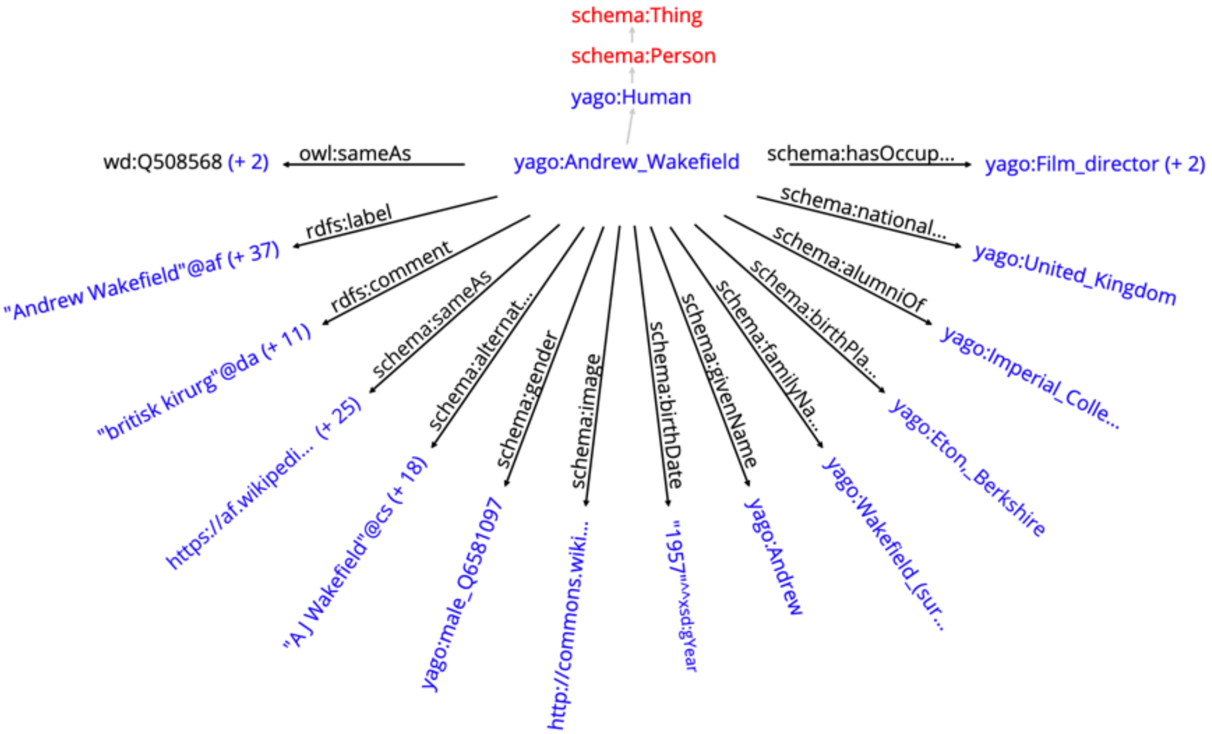
The newest version, YAGO4,

- combines Wikidata and schema.org
- contains 50 million entities and 2 billion facts
- is so clean that it allows for automated reasoning

https://yago-knowledge.org

What's in a knowledge base?

Essentially binary facts ("triples") in the knowledge format "RDF":



From YAGO

What’s in the real world?

In February 1998, Andrew Wakefield published a paper in the medical journal The Lancet, which reported on twelve children with developmental disorders. The parents were said to have linked the start of behavioral symptoms to vaccination. The resulting controversy became the biggest science story of 2002. As a result, vaccination rates dropped sharply. In 2011, the BMJ detailed how Wakefield had faked some of the data behind the 1998 Lancet article.

Beliefs Events Stories

Claims Reasons Falsifications

...none of which is in a knowledge base!

The NoRDF Project: Go Beyond Triples

If we want tomorrow's intelligent applications to be really intelligent, we have to extend their knowledge bases by

Beliefs Events Stories

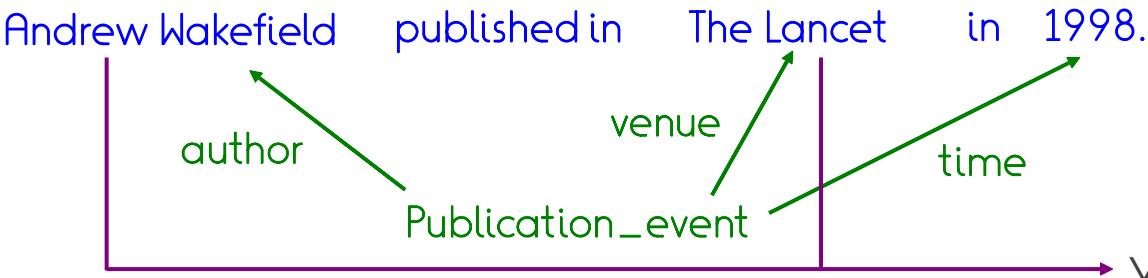
Claims Reasons Falsifications

1) We have to be able to extract complex knowledge from text
   (a process called "Information Extraction", "IE")
2) We have to be able to represent such knowledge and to reason on it

IE: What is possible already

Several cool approaches can extract non-binary information:

- FRED
- K-Parser
- Document spanners
- ClausIE
- StuffIE
- OpenIE
- HighLife
- Advanced Meaning Representation (AMR)



Andrew Wakefield published in The Lancet in 1998.

author

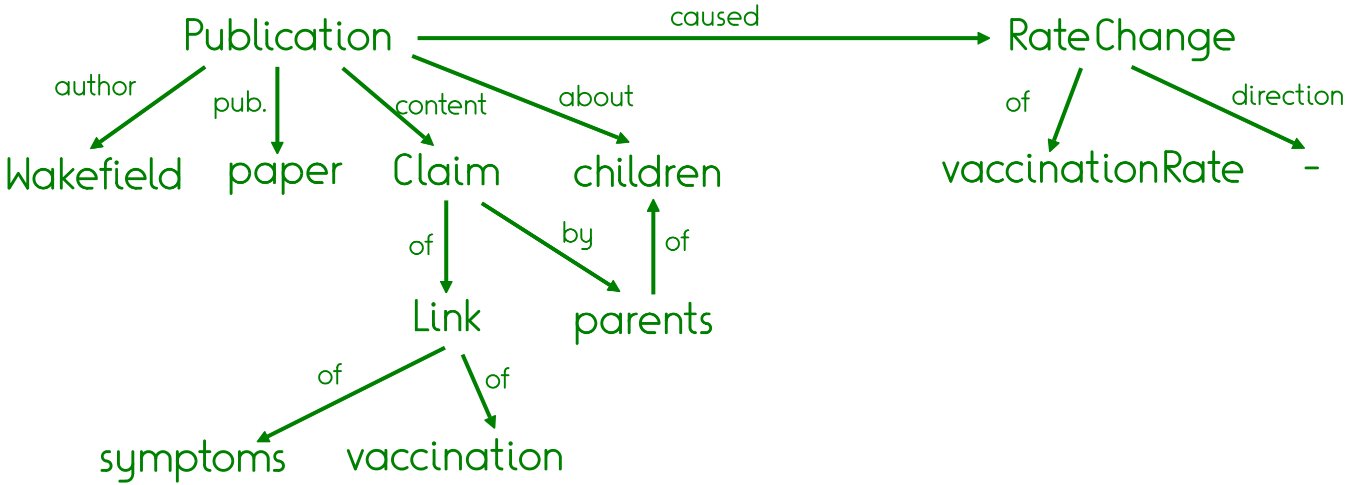
venue

time

Publication_event

IE: What we need

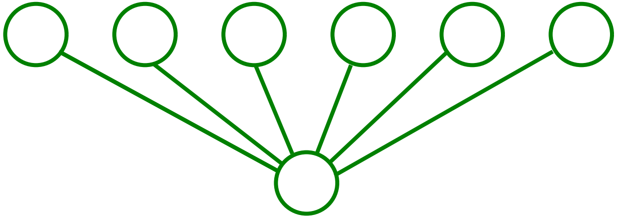
“Wakefield published a paper that reported on children. Their parents were said to have linked the start of behavioral symptoms to vaccination. The resulting controversy caused vaccination rates to fall. …”



Cross-sentence analysis, advanced co-reference resolution, standardized types of frames, relationships between events, negation, hypothetical stances, storylines, …

IE: Why Deep Learning is not enough

"Wakefield published a paper that reported on children. Their parents were said to have linked the start of behavioral symptoms to vaccination. The resulting controversy caused vaccination rates to fall. …"



Did Wakefile publish a paper? √

Who published a paper? √

Were vaccination rates higher before the publication? ?

What caused the controversy? X

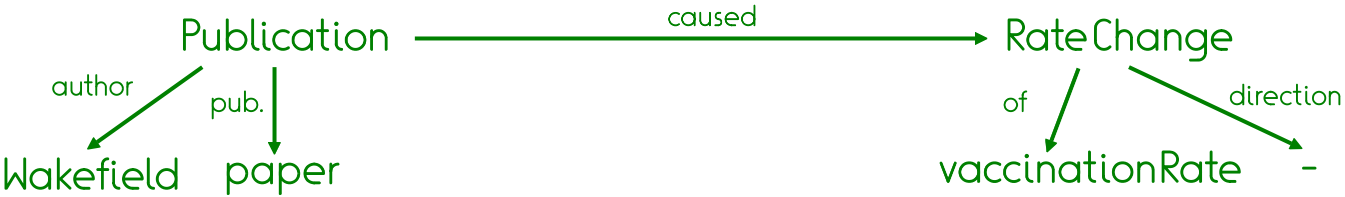
Does vaccination cause autism? X

What nationality is the person who caused the vaccine controversy? X

Reasoning: What we have

As knowledge representation:

- Frames, JSON
- complex objects
- object-relational databases



Reasoning: What we have

As knowledge representation:

- Frames, JSON
- complex objects
- object-relational databases
- Fact identifiers
- RDF*
- Reification

For reasoning:

- RDFS, OWL DL, SHACL
- Description Logic
- Context logics
- Modal logics
- Epistemic logics
- Formal argumentation
- Belief revision
- Provenance and annotated logics

Cannot represent

- "All clients believe that the company delivers a good service"
- "the loss of value on the stock market happened because the public learned of a fraudulent activity by the company"
- "Mary believes everything Paul says, Paul says $X \Rightarrow$ Mary believes X"

... or if they can, they are undecidable

Reasoning: What we need

1) a very simple logic inside a context

First-order logic without ∃ ?

OWL EL?

Datalog?

$\forall\, x\colon scientist(x) \Rightarrow person(x)$ (?)

2) a very simple logic about contexts

Horn Rules?

Datalog?

$\forall\, \phi\colon reads(Mary, \phi)$

$\Rightarrow believes(Mary, \phi)$ (?)

=> a moderately simple logic
in combination

You have a great idea? Let me know!

Vagueness, fuzziness, and probability: orthogonal topics

Applications

- Analysis of fake news / fact checking:
  understand an article about a controversial topic, allow reasoning (who said what when and why, what is the evidence, ...)

- Analysis of the e-reputation of a company:
  extract controversy or beliefs with reasons and supporters, for companies or their products

- Modeling of controversies:
  detect a controversial topic on the Web (in blogs, forums, Twitter), extract opinions, and model different views

Understanding the arguments of the other side
is a prerequisite for refuting them.

›more

Applications

- Flagging of potentially fraudulent activity:
  Detect claims that contradict knowledge, or violate rules.

- Modeling of processes:
  Model sequences of actions, causal relationships, and suggestions.

- Smarter chatbots:
  Allow dialogues that go beyond single-shot questions.

- Legal text understanding:
  Analyze a law, a regulation, or a contract, and derive
  what is permitted and what is obligatory for which party.

Our project "NoRDF"

Our project "NoRDF" aims to extract and model complex information from natural language text. The project runs for 4 years, supported by:

ANR AGENCE NATIONALE DE LA RECHERCHE

TELECOM Paris

IP PARIS

eDF

Schlumberger

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DES ARMÉES

GROUPE BPCE

Converteo ADLPerformance
DATA FOR BUSINESS ACCELERATION

Your company name here?

Our project “NoRDF”: Who’s there?

Fabian Suchanek

Professor at Télécom Paris, DIG team

Knowledge Bases, Reasoning, NLP

Chloé Clavel

Professor at Télécom Paris, S^{2}A team

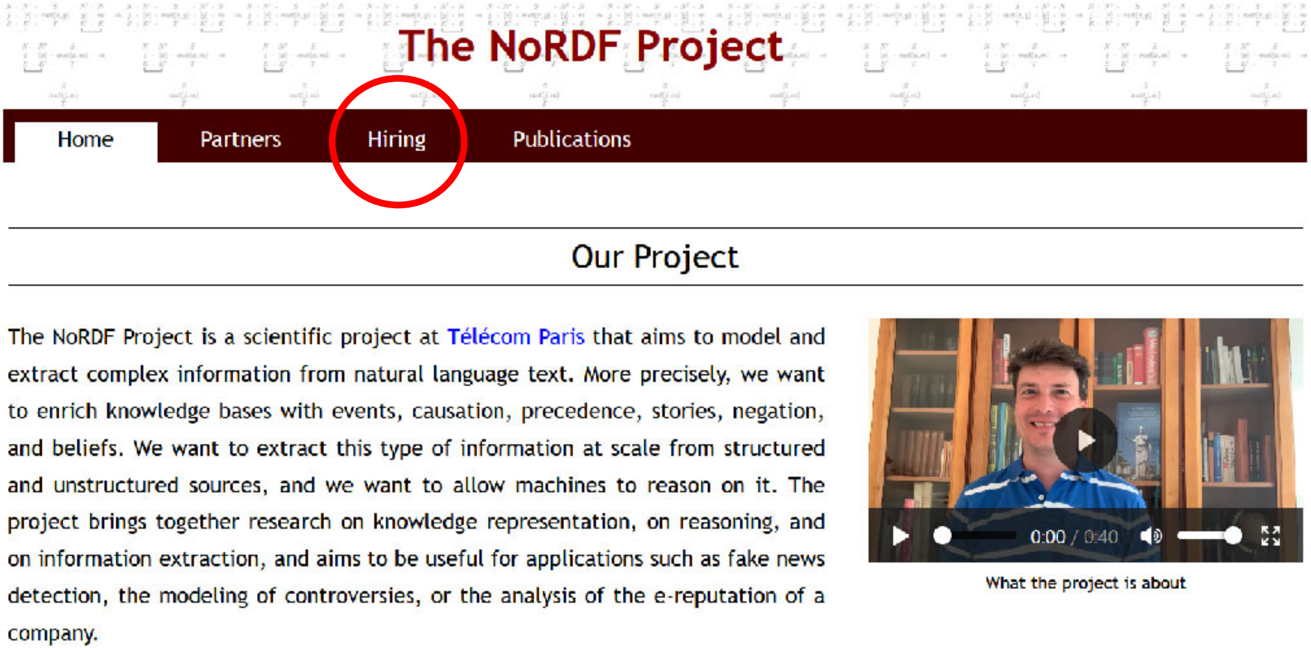
Affective Computing, Sentiment Analysis

We hired

- Pierre-Henri Paris (CNAM) as a postdoc
- Chadi Helwe (American Univ. of Beirut) as PhD student
- Cyril Chhun (Polytechnique) as a PhD student
- Saif Ghribi (Télécom Paris) as engineer

And we are hiring!

We are hiring PhD students, postdocs, and engineers, for the project or anything that has to do with NLP, knowledge bases, and reasoning!



The NoRDF Project

Our Project

The NoRDF Project is a scientific project at Télécom Paris that aims to model and extract complex information from natural language text. More precisely, we want to enrich knowledge bases with events, causation, precedence, stories, negation, and beliefs. We want to extract this type of information at scale from structured and unstructured sources, and we want to allow machines to reason on it. The project brings together research on knowledge representation, on reasoning, and on information extraction, and aims to be useful for applications such as fake news detection, the modeling of controversies, or the analysis of the e-reputation of a company.

What the project is about

Join our team! https://suchanek.name -> NoRDF