Words, things and graphs

3IA Flash presentation

Célian Ringwald

Under the supervision of:

Director: Fabien Gandon (Inria, Université Côte d'Azur, CNRS, I3S)

Co-director: Catherine Faron (Université Côte d'Azur, Inria, CNRS, I3S)

Co-supervisor: Franck Michel (Université Côte d'Azur, Inria, CNRS, I3S)

Invited: Hanna Abi Akl (Inria, Data ScienceTech Institute)

3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

UNIVERSITÉ CÔTE D'AZUR

cnrs

Inria

wimmics

2 June 2023

Background

Master SISE
Université Lyon 2
- Erasmus (6months) - Poland
- Intern (6months) - Réuniwatt

2016

2017

Datascientist - DataObserver
ExObserver, SourceBase, MedObserver, DoRTL, Sprayz
& tailor made products

2018

Master 1 Huma-num
Université Lyon 2
TER (6 months) -
ANR LIFRANUM

2019

Master 2 Huma-num
Université Lyon 2
Internship (6months) -
LINCS project - Ottawa DH Lab

2020

Covid time
- Programming Historian
- Mathematic teacher
- Consulting in Datascience

2021

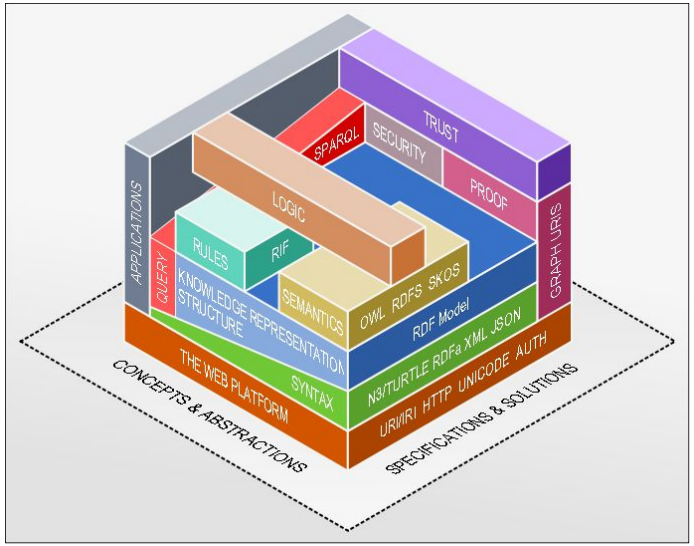
Inria - Wimmics team

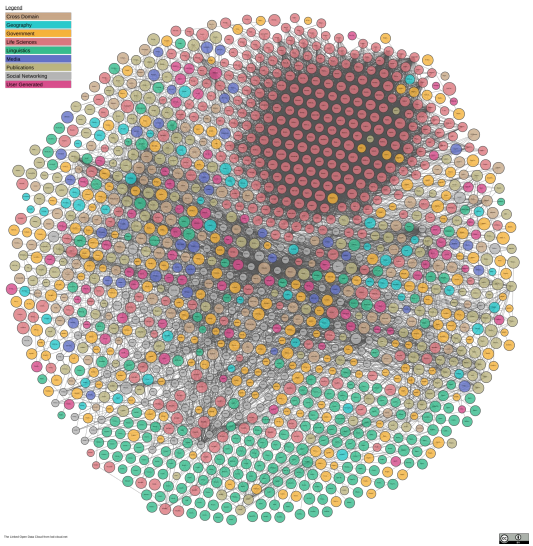
Research engineer (1 year)
- Update of DBpedia FR
- New extractors development
- DBpedia Historique -
Wikidata Integration -
GSOC 2022

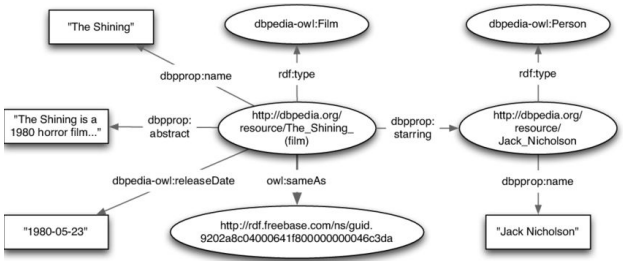
The Semantic Web,

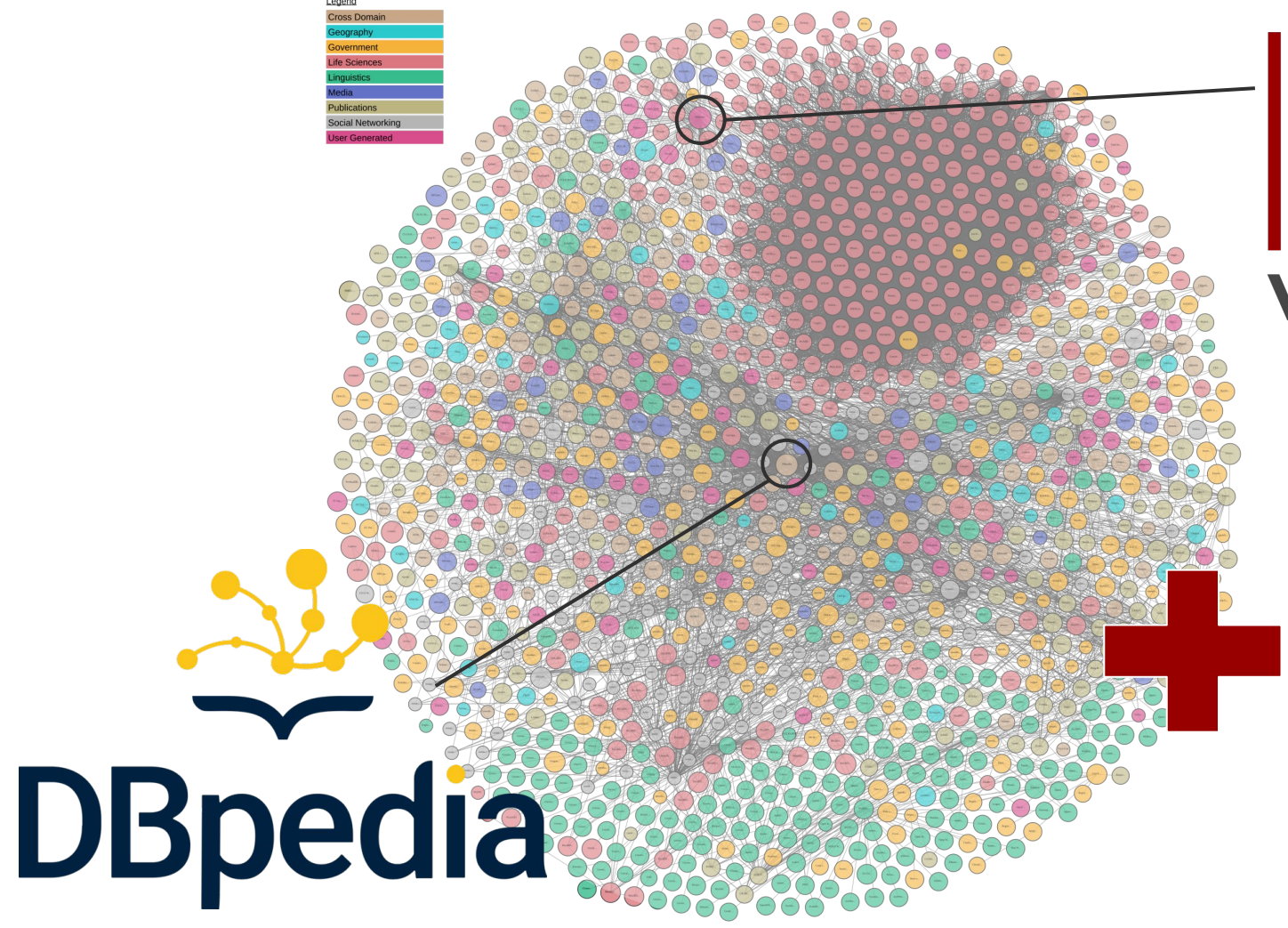
the Linked Open Data

and the Knowledge Graph…



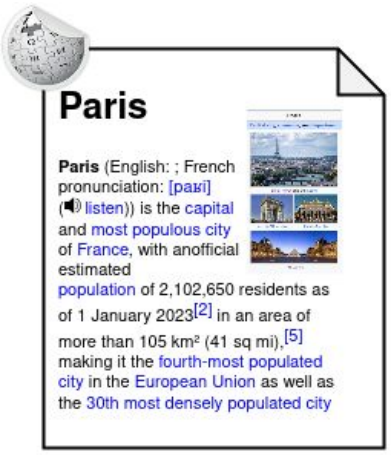





Legend
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated
WIKIDATA
DBpedia
WIKIPEDIA
The Free Encyclopedia
The Linked Open Data Cloud from lod-cloud.net

Learning RDF pattern extractors from **natural language** and **knowledge graphs** :

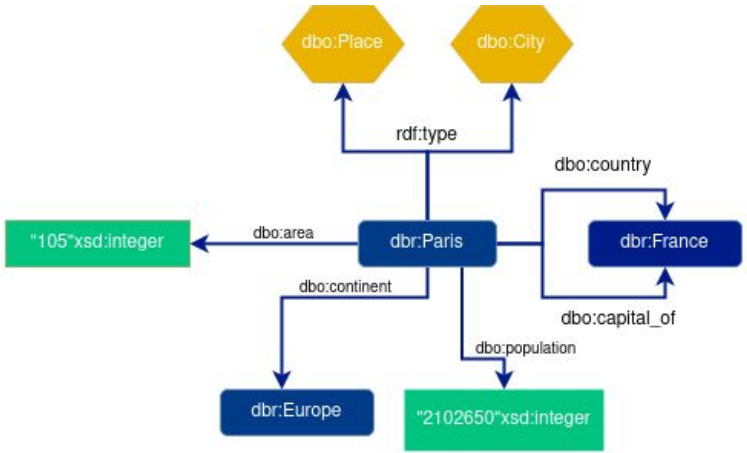
application to **Wikipedia** and the **LOD**



Paris

Paris (English: ; French pronunciation: [paʁi] (listen)) is the capital and most populous city of France, with an official estimated population of 2,102,650 residents as of 1 January 2023[2] in an area of more than 105 km² (41 sq mi),[5] making it the fourth-most populated city in the European Union as well as the 30th most densely populated city

(1) Wikipedia article



(2) Extracted RDF graph

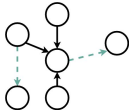
```
@prefix dbr: <http://dbpedia.org/resource/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

dbr:Paris rdfs:type dbo:Place, dbo:City ;
    dbo:country dbr:France ;
    dbo:capital_of dbr:France ;
    dbo:continent dbr:Europe ;
    dbo:area "105"^^xsd:integer ;
    dbo:population "2102650"^^xsd:integer .
```

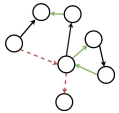
(3) Extrated triples written with RDF turtle synthax

WHY ?

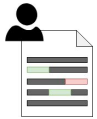
For Information retrieval



For Knowledge base completion



For Knowledge base correction



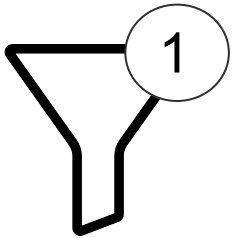
For Fact checking

To avoid Large Language Models hallucinations

A first dataset

1 833 493 dbo:Person

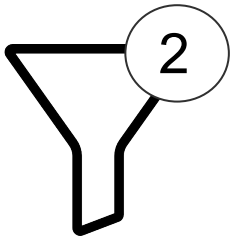
> Focus on datatype properties : dbo:birthdate / dbo:deathdate and rdfs:label



Constraints related to the KB :

- must have a dbo:birthdate
- could have a dbo:deathdate
- must have one rdfs:label

> 967 500 dbo:Person



Constraints related to the Wikipedia article :

- The value of the relation must be findable in the text

> 553 899 dbo:Person

A first dataset :

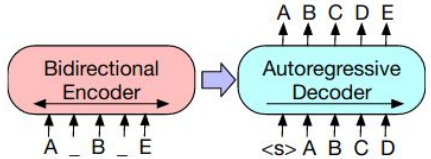
dbr:Thomas_Bragg : Thomas Bragg (November 9, 1810 – January 21, 1872) was an American politician and lawyer who served as the 34th Governor of the U.S. state of North Carolina from 1855 through 1859. During the Civil War, he served in the Confederate States Cabinet. He was the older brother of General Braxton Bragg. They were direct descendants of Thomas Bragg (1579–1665) who was born in England and settled in the Virginia Colony. Born in Warrenton, [...]

```
dbr:Thomas_Bragg rdfs:label "Thomas Bragg" ;
    dbo:birthDate "1810-11-09" ;
    dbo:deathDate "1872-01-21".
```

Transformer architectures and pre-trained models

(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



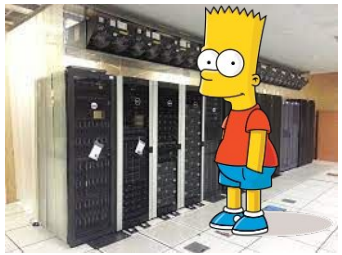
(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure 1: A schematic comparison of BART with BERT (Devlin et al., 2019) and GPT (Radford et al., 2018).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 7871–7880. DOI:https://doi.org/10.18653/v1/2020.acl-main.703

First little experiment

with BART-base
on the Inria NEF cluster



Name	sampled	State	Run time	F1 micro	loss	prec micro	recall micro
model3	false	crashed	**10 hours**	90.15	0.003	99.63	82.32
model2	**false**	**crashed**	**10 hours**	**90.15**	**0.003**	**99.63**	**82.32**
model1	**5% of the dataset**	**finished**	**26 minutes**	**89.37**	**0.015**	**98.37**	**81.88**

Qualitative analysis : type of errors

1. The fact isn’t in KB (FP)
2. The value in the text and in the KB are differents (FP)
3. The model hallucinates because of incomplete data in text (FP)
4. The model hallucinates because text doesn’t support the fact (FN)

Qualitative analysis : type of errors

1. The fact isn't in KB

René Descartes (/deɪˈkɑːrt/ or UK: /ˈdeɪkɑːrt/; French: [ʁəne dekaʁt] (listen); Latinized: **Renatus Cartesius**;[note 3][17] 31 March 1596 – 11 February 1650)[18][19][20]:58 was a French philosopher, scientist, and mathematician, widely considered a seminal figure in the emergence of modern philosophy and science. Descartes was Roman Catholic.[22][23]

DBpedia

```
dbr:René_Descartes rdfs:label "René Descartes" ;
                   dbo:birthDate "1596-03-31" .
```

```
dbr:René_Descartes rdfs:label "René Descartes" ;
                   dbo:birthDate "1596-03-31" ;
                   dbo:deathDate "1650-02-11".
```

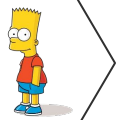
Qualitative analysis : type of errors

2. The value in the text and in the KB are differents

John Wesley Gilbert (July 6, 1863[1] – November 18, 1923) was an American archaeologist, educator, and Methodist missionary to the Congo. Gilbert was the first graduate of Paine College, its first African-American professor, and the first African-American to receive an advanced degree from Brown University.[2][3]

DBpedia

dbr:John_Wesley_Gilbert rdfs:label "John Wesley Gilbert" ;
dbo:birthDate “1863-07-06" ;
dbo:deathDate “1923-11-19” .



dbr:John_Wesley_Gilbert rdfs:label "John Wesley Gilbert" ;
dbo:birthDate "1863-07-06" ;
dbo:deathDate "1923-11-18".

Qualitative analysis : type of errors

3. The model hallucinates because of incomplete data in text

Mary Grace Quackenbos Humiston (née Winterton) (1869–1948) was the first female Special Assistant United States Attorney.[1] She was a graduate of the New York University School of Law and was a leader in exposing peonage in the American South.

dbr:Mary_Grace_Quackenbos rdfs:label "Mary Quackenbos" ;
dbo:birthDate "1869-09-17" .

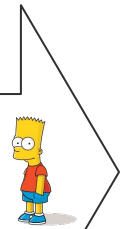
dbr:Mary_Grace_Quackenbos rdfs:label "Mary Quackenbos" ;
dbo:birthDate "1869-06-18".

Qualitative analysis : type of errors

4. The model hallucinates because text doesn't support the fact

Ankushrao Tope was an Indian politician, elected to the Lok Sabha, the lower house of the Parliament of India as a member of the Indian National Congress.[1][2][3]

dbr:Ankushrao_Tope rdfs:label "Ankushrao Tope" ;
dbo:birthDate "1942-09-18" ;
dbo:deathDate "2016-04-03".



dbr:Ankushrao_Tope rdfs:label "Ankushrao Tope" ;
dbo:birthDate "1946-11-02".

Challenges

Datasets :

RQ1.1. How to support fact extraction relying on different document granularity?

RQ1.2. What is the best strategy to extract rare relations?

RQ1.3. How to represent a fact that spans several sentences and vice-versa?

Challenges

Datasets :

RQ1.1. How to support fact extraction relying on different document granularity?
RQ1.2. What is the best strategy to extract rare relations?
RQ1.3. How to represent a fact that spans several sentences and vice-versa?

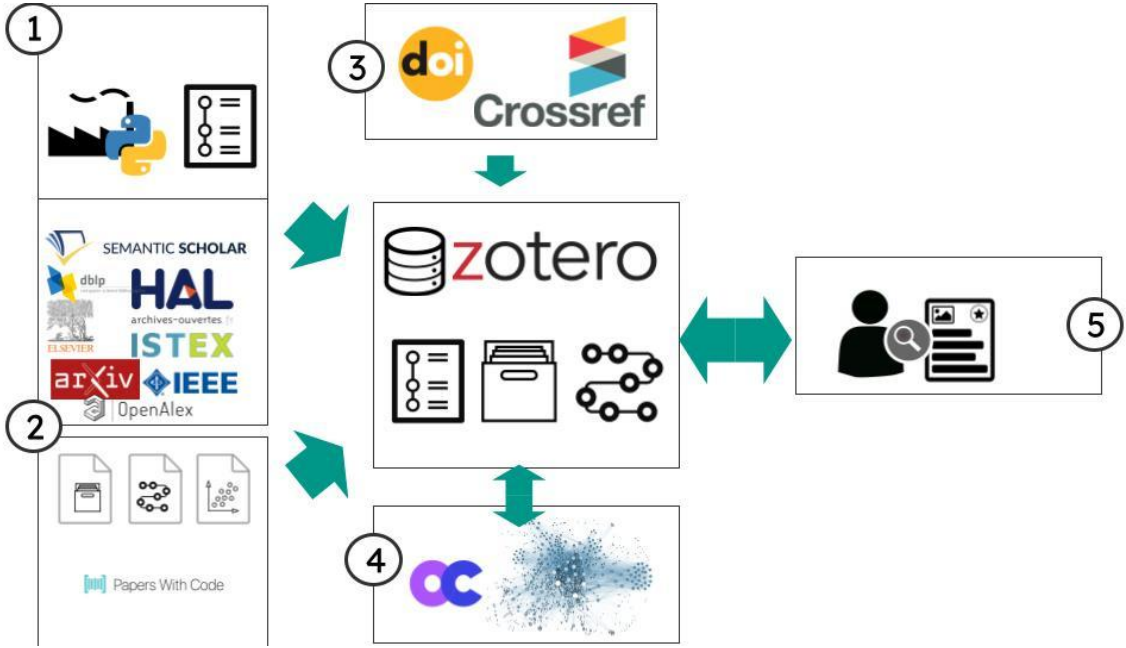
Modelling :

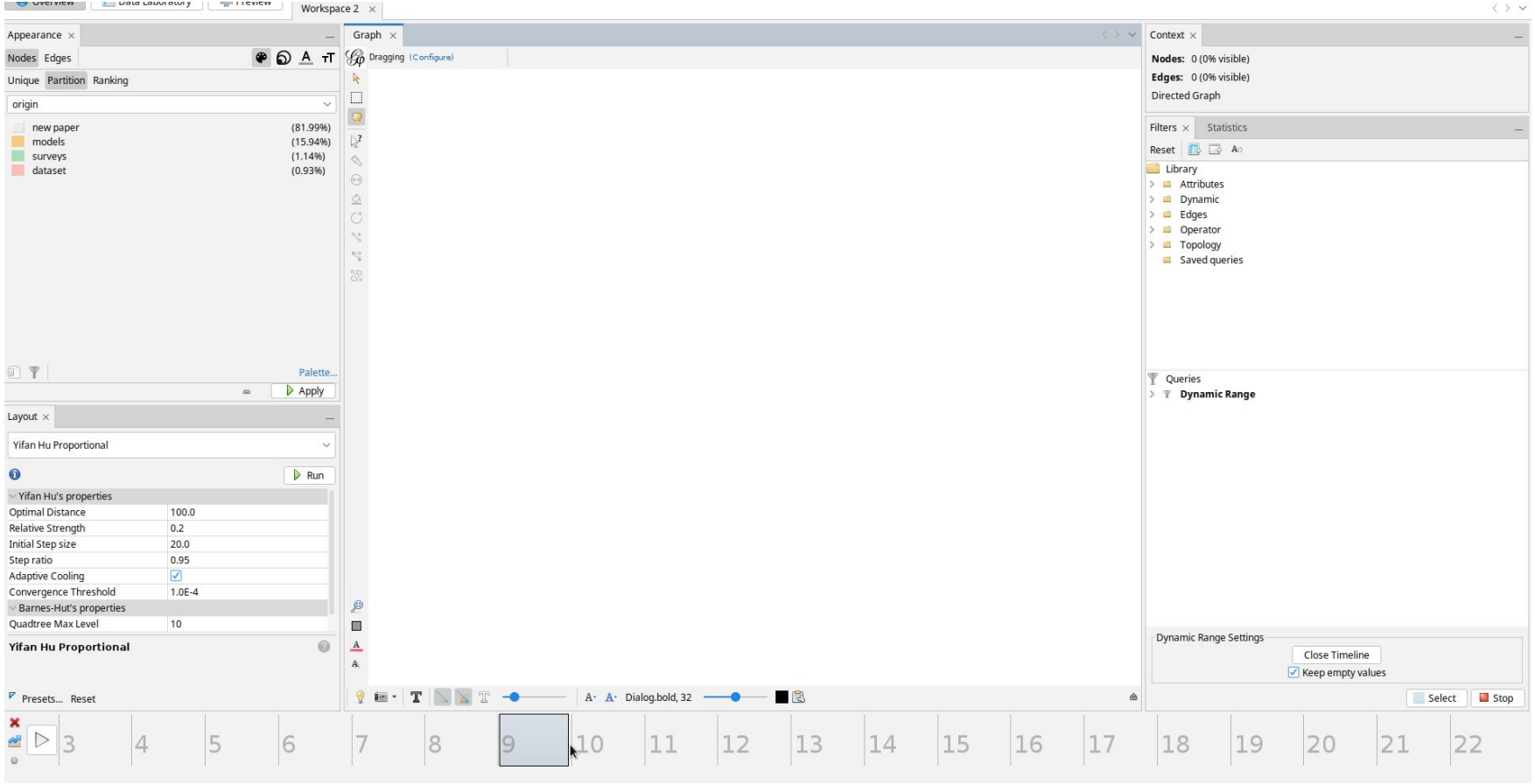
RQ2.1. How to restrict language model to only facts that are supported by our text and our knowledge base?
RQ2.2. In our context, what is the best adaptation design for pre-trained models?
RQ2.3. Can we design an iterative procedure to produce consistent RDF triples ?

How to approach a 35 years old research domain ?

SciLEx

A scriptbox
for Science
Literature Exploration
in Machine learning





Watching the science literature construction on the last 10 years

https://datalogism.github.io/

https://github.com/datalogism

celian.ringwald@inria.fr

Or maybe see you at
the Generative Modeling Summer School
Copenhagen, June 26th to 30th, 2023
https://gemss.ai/

Discussions

____ ?
____!

_____?
____, _____ .
_____, __,__ & __. __, ____.

_____, _____. ____ ! ___ _____. ____,
____ ?

_, _____ ?
_____. ___, _____. ___, ____.

Thank you

3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

UNIVERSITÉ CÔTE D'AZUR

cnrs

Inria

wimmics