Building Knowledge Graphs For 20 Years:

Should We Keep Doing This?

Prof. Raphael Troncy

INSTITUT CARNOT Télécom & Société numérique

EURECOM
Sophia Antipolis

Introducing the Knowledge Graph: things, not strings

Cross-posted on the Inside Search Blog

Search is a lot about discovery—the basic human need to learn and broaden your horizons. But searching still requires a lot of hard work by you, the user. So today I'm really excited to launch the Knowledge Graph, which will help you discover new information quickly and easily.

Take a query like [taj mahal]. For more than four decades, search has essentially been about matching keywords to queries. To a search engine the words [taj mahal] have been just that—two words.

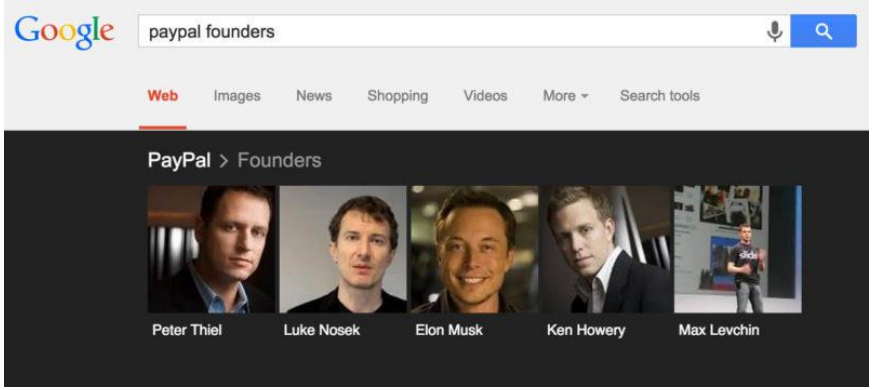
But we all know that [taj mahal] has a much richer meaning. You might think of one of the world's most beautiful monuments, or a Grammy Award-winning musician, or possibly even a casino in Atlantic City, NJ. Or, depending on when you last ate, the nearest Indian restaurant. It's why we've been working on an intelligent model—in geek-speak, a "graph"—that understands real-world entities and their relationships to one another: things, not strings.

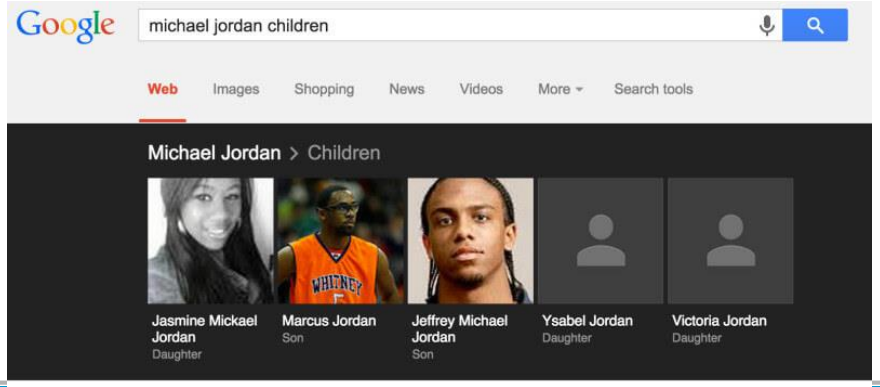
The Knowledge Graph enables you to search for things, people or places that Google knows about—landmarks, celebrities, cities, sports teams, buildings, geographical features, movies, celestial objects, works of art and more—and instantly get information that's relevant to your query. This is a critical first step towards building the next generation of search, which taps into the collective intelligence of the web and understands the world a bit more like people do.

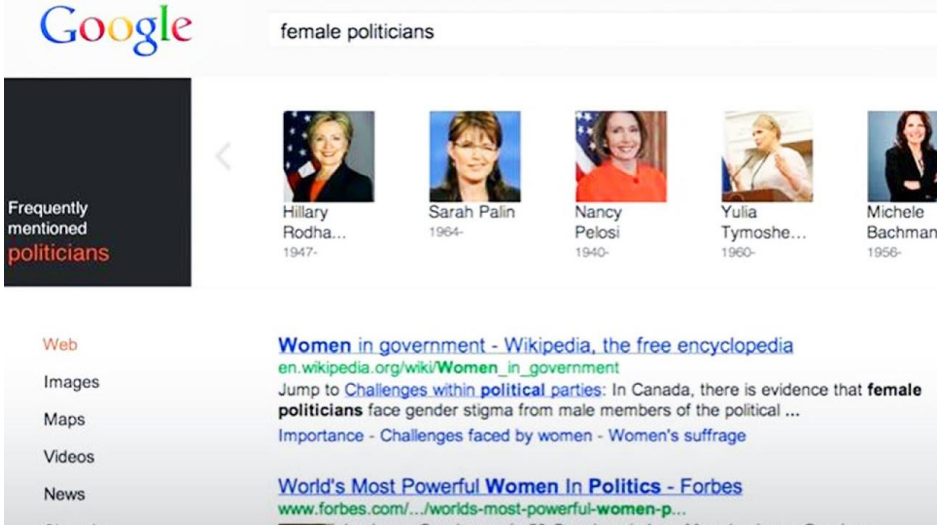
https://blog.google/products/search/introducing-knowledge-graph-things-not/

depth. It currently contains more than 500 million objects, as well as more than 3.5 billion facts about and relationships between these different objects. And it's tuned based on what people search for, and what we find

Google's Knowledge Graph Gains "Carousel"

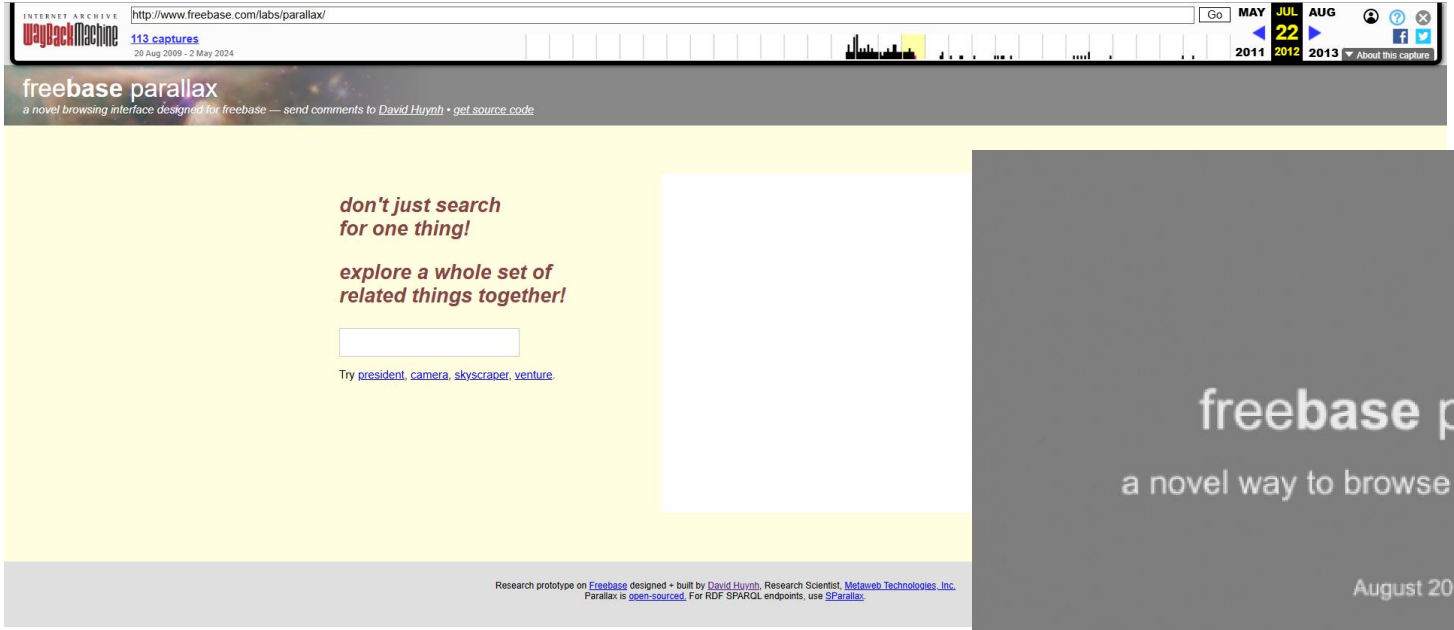






https://searchengineland.com/googles-knowledge-graph-now-worldwide-129948

Don’t just search for one thing!



https://code.google.com/archive/p/freebase-parallax/

https://vimeo.com/1513562

Standing on the shoulders of giants

- **Who are the previous keynote speakers of ESWC?**

Extended Semantic Web Conference

Add languages

Learn to edit
Community portal
Recent changes
Upload file
Special pages

Contents hide
(Top)
List of conferences
References
External links

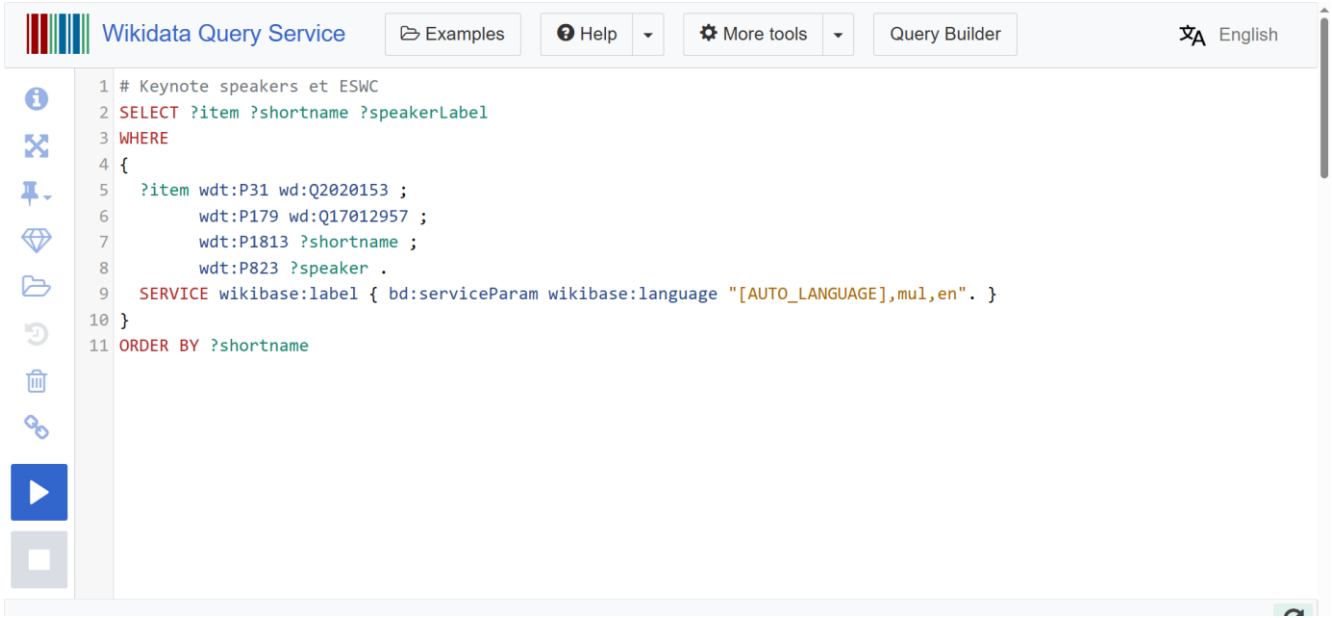
Year	Conference	City	Country	Awards
2025	ESWC2025	Portorož	Slovenia	
2024	ESWC2024	Hersonissos	Greece	
2023	ESWC2023	Hersonissos	Greece	
2022	ESWC2022	Hersonissos	Greece	**Best Paper:** Never Mind the Semantic Gap: Modular, Lazy and Safe Loading of RDF Data by Eduard Kamburjan, Vidar Norstein Klungre, and Martin Giese[4] **Best Student Paper:** The Problem with XSD Binary Floating Point Datatypes in RDF by Jan Martin Keil & Merle Gänßinger[5]
2021	ESWC2021		Online	
2020	ESWC2020	Heraklion	Greece	
2019	ESWC2019	Portorož	Slovenia	**Best Paper:** Hai Huang and Fabien Gandon: *Learning URI Selection Criteria to Improve the Crawling of Linked Open Data*[6] **Best Student Paper:** Matthäus Zloch, Maribel Acosta, Daniel Hienert, Stefan Dietze and Stefan Conrad: *A Software Framework and Datasets for the Analysis of Graph Measures on RDF Graphs*[7]
2018	ESWC2018	Heraklion	Greece	
2017	ESWC2017	Portorož	Slovenia	**Best Paper:** Patrik Schneider, Thomas Eiter and Josiane Xavier Parreira: *Spatial Ontology-Mediated Query Answering over Mobility Streams*[8] **Best Student Paper:** Sarven Capadisli, Amy Guy, Christoph Lange, Sören Auer and Tim Berners-Lee: *Linked Data Notifications*[9]
2016	ESWC2016	Anissaras, Crete	Greece	**Best Paper:** Géraud Fokou, Stéphane Jean, Allel Hadjali, and Michael Baron: *RDF Query Relaxation Strategies Based on Failure Causes*[10]
2013	ESWC2013 Archived 2013-10-12 at the Wayback Machine	Montpellier	France	**Best Paper:** Axel-Cyrille Ngonga Ngomo, Lars Kolb, Norman Heino, Michael Hartung, Sören Auer and Erhard Rahm: *When to Reach for the Cloud: Using Parallel Hardware for Link Discovery*[11] Elena Cabrio, Serena Villata and Fabien Gandon: *A Support Framework for Argumentative Discussions Management in the Web*[12]
2012	ESWC2012	Heraklion	Greece	**Best Paper:** Uta Lösch, Stephan Bloehdorn, Achim Rettinger: *Graph Kernels for RDF Data*[13]

Tools hide
General
What links here
Related changes
Permanent link
Page information
Cite this page
Get shortened URL
Download QR code
Add interlanguage links
Print/export
Download as PDF
Printable version
In other projects
Wikidata item

Appearance hide
Text
Small
Standard
Large
Width
Standard

Standing on the shoulders of giants

- **Who are the previous keynote speakers of ESWC?**



Standing on the shoulders of giants

- **Who are the previous keynote speakers of ESWC?**

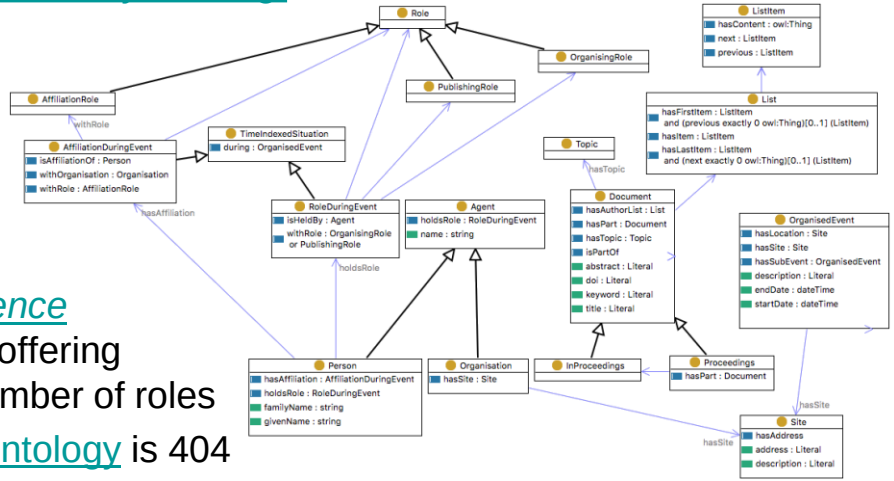
https://w.wiki/EMbz

Table ▾ | 54 results in 57 ms | </> Code | Download ▾ | Link ▾

item	shortname	speakerLabel
wd:Q64020172	ESWC 2006	Frank van Harmelen
wd:Q64020172	ESWC 2006	Eduard Hovy
wd:Q64020250	ESWC 2007	Georg Gottlob
wd:Q64020250	ESWC 2007	Stefano Ceri
wd:Q64020250	ESWC 2007	Ronald J. Brachman
wd:Q64020333	ESWC 2008	Nigel Shadbolt
wd:Q64020333	ESWC 2008	Ricardo Baeza-Yates
wd:Q64020333	ESWC 2008	Claudio Gutiérrez
wd:Q64020447	ESWC 2009	Alan Smeaton
wd:Q64020447	ESWC 2009	Craig Knoblock
wd:Q64021290	ESWC 2010	Wolfgang Wahlster
wd:Q64021290	ESWC 2010	Noshir Contractor
wd:Q64021290	ESWC 2010	Aldo Gangemi
wd:Q64021290	ESWC 2010	Sean Bechhofer
wd:Q42431329	ESWC 2012	Heiko Paulheim
wd:Q42431329	ESWC 2012	Isabelle Augenstein
wd:Q42431329	ESWC 2012	Elena Simperl
wd:Q64020994	ESWC 2013	David Karger
wd:Q64020994	ESWC 2013	Enrico Motta
wd:Q64021052	ESWC 2014	Luciano Floridi
wd:Q64021052	ESWC 2014	Steffen Staab

Standing on the shoulders of giants

- **Who are the previous keynote speakers of ESWC?**
  - https://github.com/scholarly-wikidata/
  - https://zenodo.org/records/10989709
  - http://www.scholarlydata.org/



- The *conference* ontology is offering a limited number of roles
- The *SWC* ontology is 404
- The *SEO* ontology is great … but the EVENTS KG is 404

SCHOLARLY WIKIDATA: POPULATION AND EXPLORATION OF CONFERENCE DATA IN WIKIDATA USING LLMS

Nandana Mihindukulasooriya
IBM Research
New York, USA
nandana@ibm.com

Sanju Tiwari
Sharda University
Greater Noida, India
tiwarisanju18@ieee.org

Daniil Dobriy
Vienna University for Economics and Business
Vienna, Austria
daniil.dobriy@wu.ac.at

Finn Årup Nielsen
DTU Compute, Technical University of Denmark
Lyngby, Denmark
faan@dtu.dk

Tek Raj Chhetri
McGovern Institute for Brain Research, Massachusetts Institute of Technology
| Center for Artificial Intelligence (AI) Research Nepal
Cambridge, USA | Sundarharaincha-09, Nepal
tekraj.chhetri@cair-nepal.org

Axel Polleres
Vienna University for Economics and Business
Vienna, Austria
axel.polleres@wu.ac.at

ABSTRACT

Several initiatives have been undertaken to conceptually model the domain of scholarly data using ontologies and to create respective Knowledge Graphs. Yet, the full potential seems unleashed, as automated means for automatic population of said ontologies are lacking, and respective initiatives from the Semantic Web community are not necessarily connected: we propose to make scholarly data more sustainably accessible by leveraging Wikidata's infrastructure and automating its population in a sustainable manner through LLMs by tapping into unstructured sources like conference Web sites and proceedings texts as well as already existing structured conference datasets. While an initial analysis shows that Semantic Web conferences are only minimally represented in Wikidata, we argue that our methodology can help to populate, evolve and maintain scholarly data as a community within Wikidata.

Our main contributions include (a) an analysis of ontologies for representing scholarly data to identify gaps and relevant entities/properties in Wikidata, (b) semi-automated extraction – requiring (minimal) manual validation – of conference metadata (e.g., acceptance rates, organizer roles, programme committee members, best paper awards, keynotes, and sponsors) from websites and proceedings texts using LLMs. Finally, we discuss (c) extensions to visualization tools in the Wikidata context for data exploration of the generated scholarly data. Our study focuses on data from 105 Semantic Web-related conferences and extends/adds more than 6000 entities in Wikidata. It is important to note that the method can be more generally applicable beyond Semantic Web-related conferences for enhancing Wikidata's utility as a comprehensive scholarly resource.

Source Repository: https://github.com/scholarly-wikidata/
DOI: https://doi.org/10.5281/zenodo.10989709
License: Creative Commons CC0 (Data), MIT (Code)

Keywords Scholarly Data · Wikidata · Large Language Model.

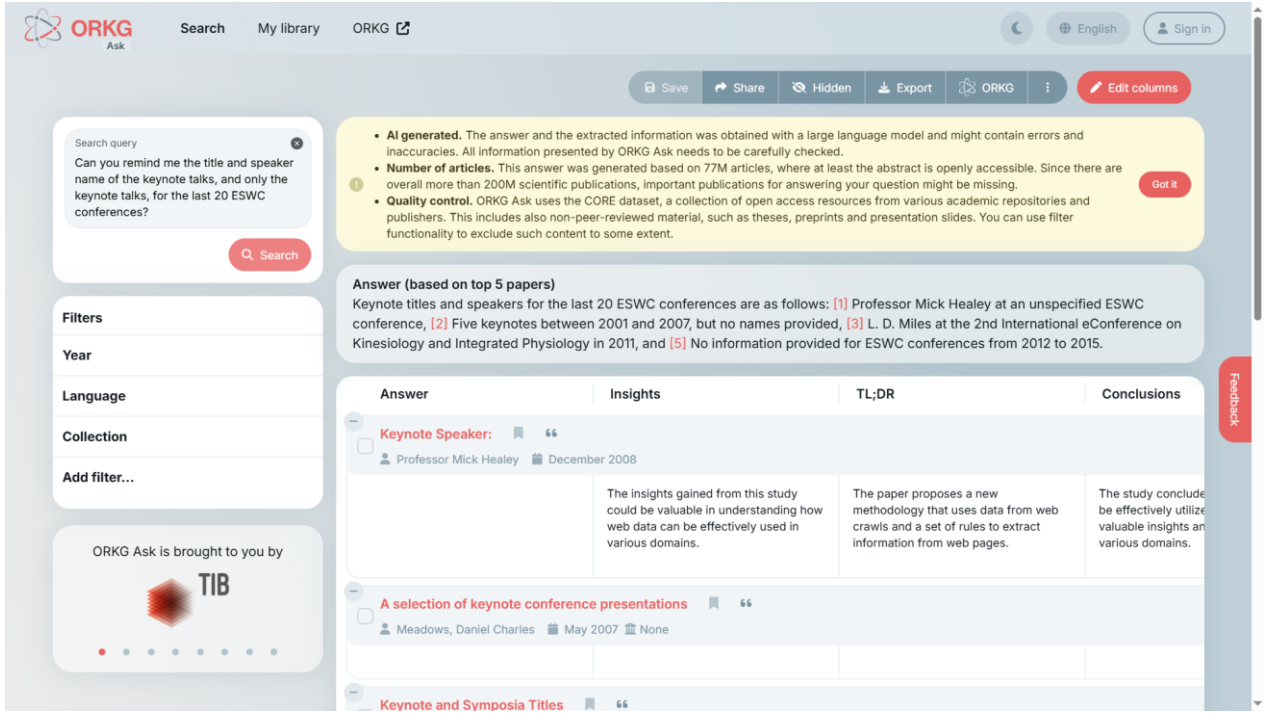
https://arxiv.org/abs/2411.08696

EURECOM
Sophia Antipolis

Standing on the shoulders of giants

- **Who are the previous keynote speakers of ESWC?**

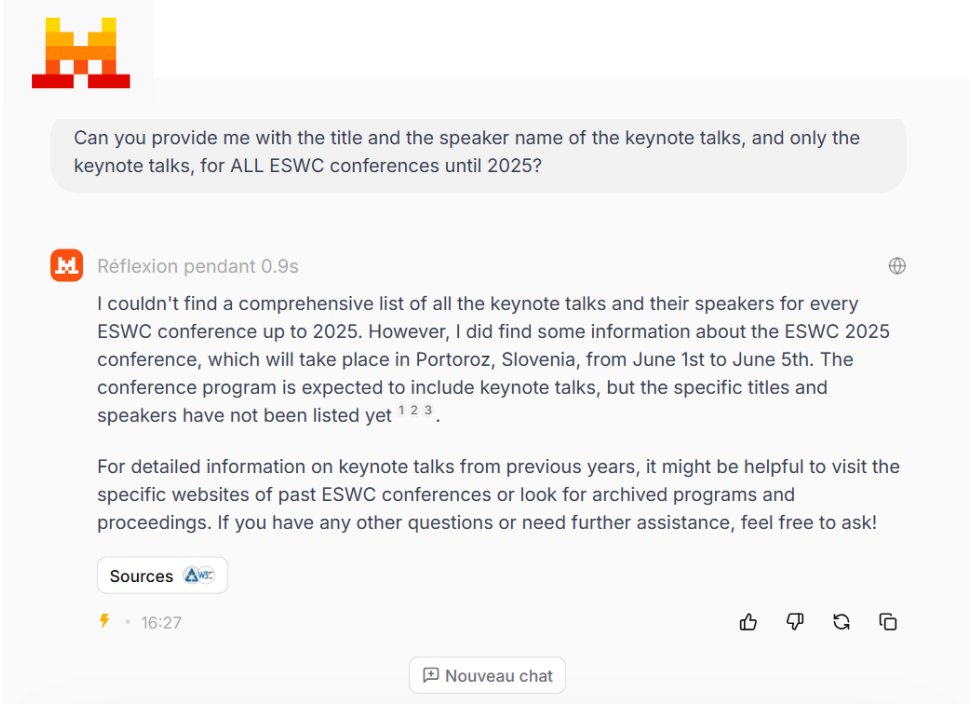
ORKG Ask

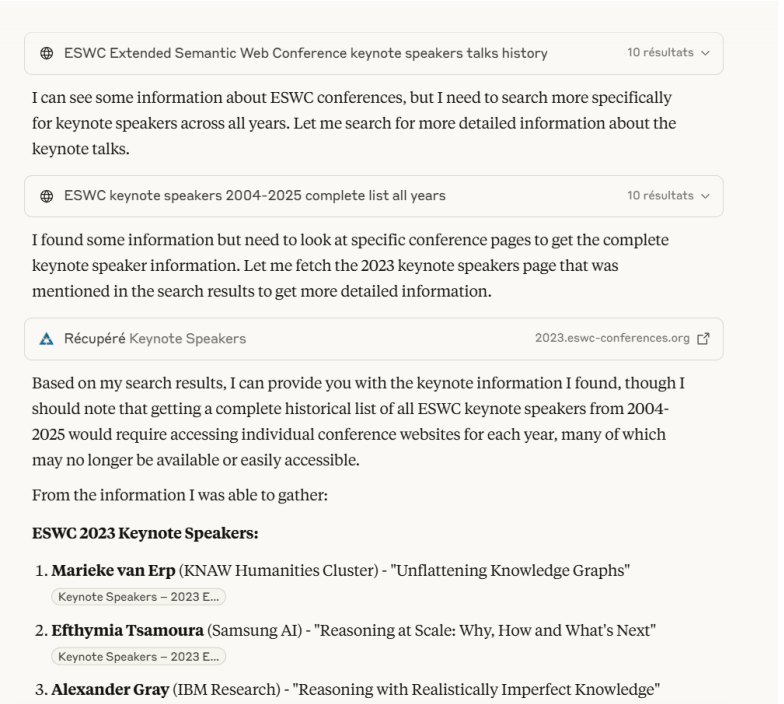


Standing on the shoulders of giants

- **Who are the previous keynote speakers of ESWC?**

Claude





Standing on the shoulders of giants

- **Who are the previous keynote speakers of ESWC?**

ChatGPT

ESWC Keynote Speakers (2004–2025)

2025

- **[Your Name]** – *Building Knowledge Graphs For 20 Years: Should We Keep Doing This?*
- *(Other keynote speakers for 2025 have not been publicly announced yet.)*

2024

- *(Information on keynote speakers for 2024 is not readily available.)*

2023

- **Marieke van Erp** – *Unflattening Knowledge Graphs*
- **Efthymia Tsamoura** – *Reasoning at Scale: Why, How and What's Next*
- **Alexander Gray** – *Reasoning with Realistically Imperfect Knowledge*

2022

- **Matthias Niepert** – *Learning with Discrete Structures*
- **Tova Milo** – *Data Disposal by Design*
- **Axel Ngonga** – *Structured Machine Learning with Multiple Representations*

2021

- **Luc De Raedt** – *From Probabilistic Logics to Neur ...mbolic Artificial Intelligence*

2020

- **Uli Sattler, John F. Sowa, Michael Schmidt** – *(Talk titles not specified)*

2019

- **Peter Haase** – *(Talk title not specified)*

2014

- **Steffen Staab** – *Programming the Semantic Web*
- **Luciano Floridi** – *Coordination, Semantics, and Autonomy*
- **Volker Tresp** – *Machine Learning with Knowledge Graphs*
- **Lise Getoor** – *(Talk canceled)*

2011

- **LarKC Project**
- **LOD2 Project**
- **Ontotext**
- **Pascal2 Project**

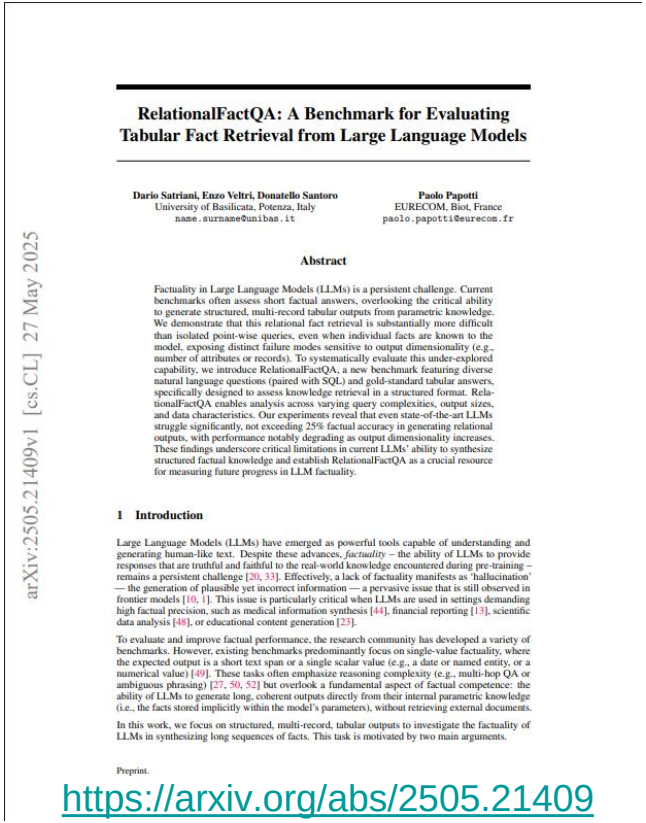
(Specific speaker names and talk titles not provided)

For earlier years (2004–2010) and some intervening years, detailed records of keynote speakers and their talk titles are not readily available in the provided sources.

If you need assistance in locating information for specific years or speakers, feel free to ask!

What did we learn?

- **Multiple sources of truth**
- **Data representation and data quality issues**
- **Maintaining ontologies and knowledge graphs is under-valued**
- **Knowledge preservation is essential**
- **Factuality in Large Language Models is a persistent challenge**



RelationalFactQA: A Benchmark for Evaluating Tabular Fact Retrieval from Large Language Models

Dario Satriani, Enzo Veltri, Donatello Santoro
University of Basilicata, Potenza, Italy
name.surname@unibas.it

Paolo Papotti
EURECOM, Biot, France
paolo.papotti@eurecom.fr

Abstract

Factuality in Large Language Models (LLMs) is a persistent challenge. Current benchmarks often assess short factual answers, overlooking the critical ability to generate structured, multi-record tabular outputs from parametric knowledge. We demonstrate that this relational fact retrieval is substantially more difficult than isolated point-wise queries, even when individual facts are known to the model, exposing distinct failure modes sensitive to output dimensionality (e.g., number of attributes or records). To systematically evaluate this under-explored capability, we introduce RelationalFactQA, a new benchmark featuring diverse natural language questions (paired with SQL) and gold-standard tabular answers, specifically designed to assess knowledge retrieval in a structured format. RelationalFactQA enables analysis across varying query complexities, output sizes, and data characteristics. Our experiments reveal that even state-of-the-art LLMs struggle significantly, not exceeding 25% factual accuracy in generating relational outputs, with performance notably degrading as output dimensionality increases. These findings underscore critical limitations in current LLMs' ability to synthesize structured factual knowledge and establish RelationalFactQA as a crucial resource for measuring future progress in LLM factuality.

1 Introduction

Large Language Models (LLMs) have emerged as powerful tools capable of understanding and generating human-like text. Despite these advances, *factuality* – the ability of LLMs to provide responses that are truthful and faithful to the real-world knowledge encountered during pre-training – remains a persistent challenge [20, 33]. Effectively, a lack of factuality manifests as 'hallucination' — the generation of plausible yet incorrect information — a pervasive issue that is still observed in frontier models [10, 1]. This issue is particularly critical when LLMs are used in settings demanding high factual precision, such as medical information synthesis [44], financial reporting [13], scientific data analysis [48], or educational content generation [23].

To evaluate and improve factual performance, the research community has developed a variety of benchmarks. However, existing benchmarks predominantly focus on single-value factuality, where the expected output is a short text span or a single scalar value (e.g., a date or named entity, or a numerical value) [49]. These tasks often emphasize reasoning complexity (e.g., multi-hop QA or ambiguous phrasing) [27, 50, 52] but overlook a fundamental aspect of factual competence: the ability of LLMs to generate long, coherent outputs directly from their internal parametric knowledge (i.e., the facts stored implicitly within the model's parameters), without retrieving external documents.

In this work, we focus on structured, multi-record, tabular outputs to investigate the factuality of LLMs in synthesizing long sequences of facts. This task is motivated by two main arguments.

Preprint.

https://arxiv.org/abs/2505.21409

EURECOM
Sophia Antipolis

Data Integration and Knowledge Graphs

Entertainment (EventMedia)
Tourism (3cixty)
Media Industry (LinkedTV, MeMAD)
Energy (EDF)
ICT (Orange, Huawei)

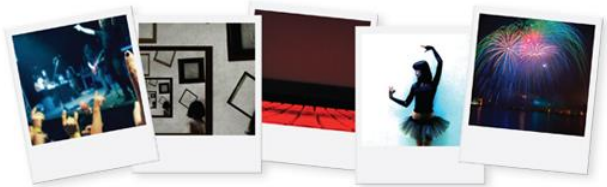
EURECOM
Sophia Antipolis

http://www.flickr.com/photos/crsan/3697785107

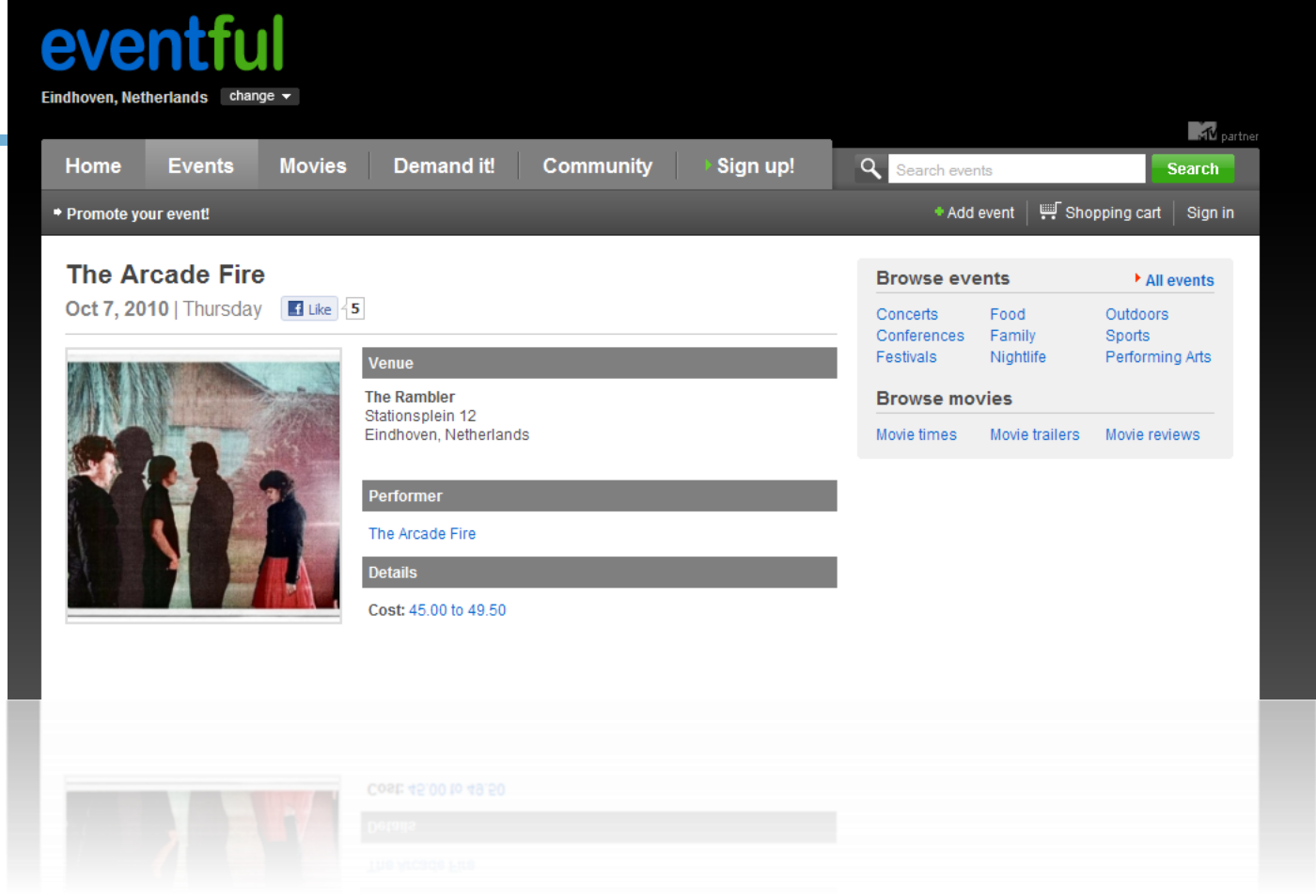
What is an event?

- [Allan J., KAP 2002]:*“Event is something that has a specific time, location, and people associated with it ”*
- [Fialho A., EVENTS 2010]:
  “*Events are observable occurrences grouping*”

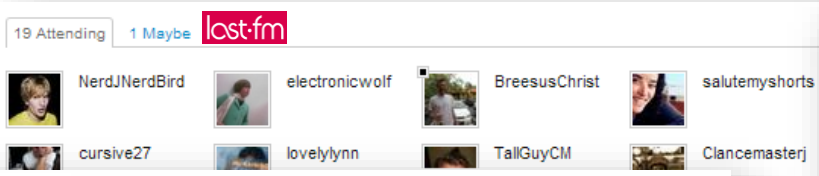
- “*Experiences documented by media*”

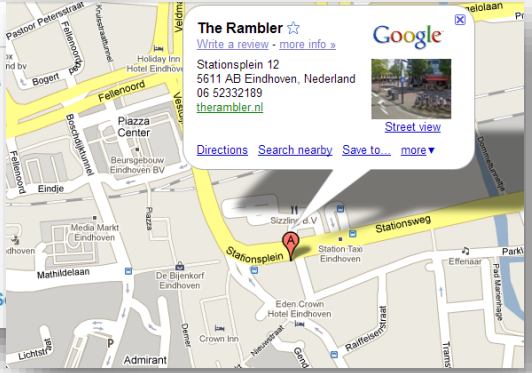


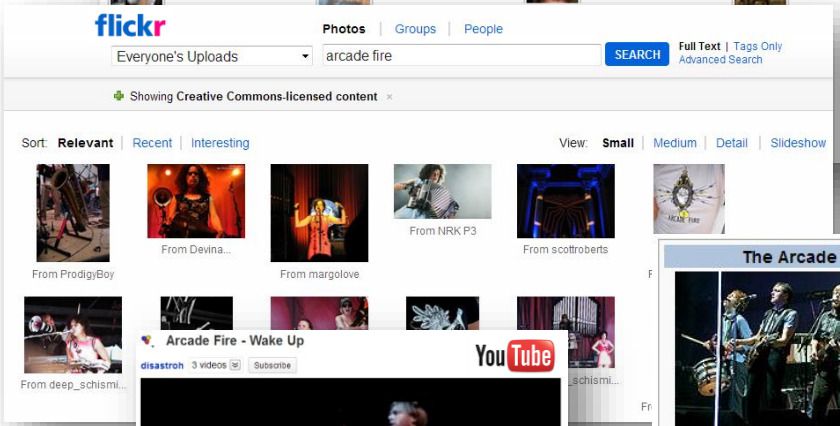
Searching for events



Searching for media illustrating an event







EURECOM
Sophia Antipolis

Exploring and discovering experiences

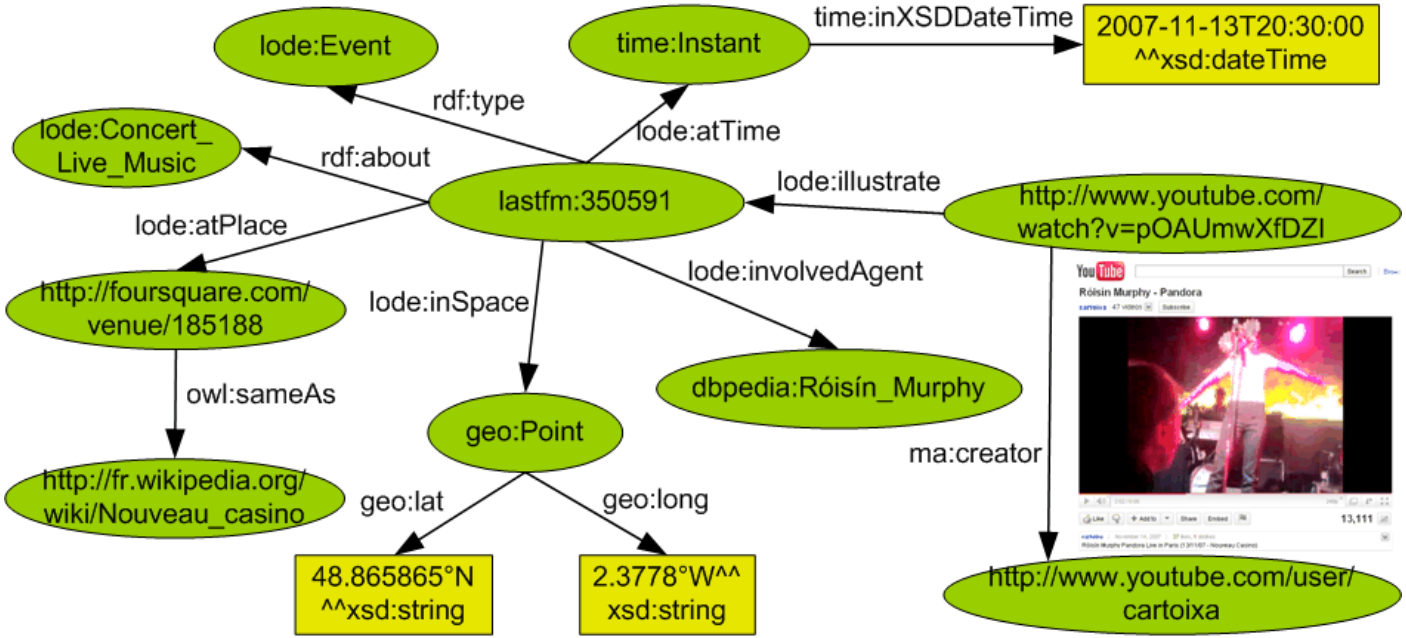
Event Directories

Media Platforms

Social Networks

Representing Events with

Linking Open Descriptions of Events



EURECOM
Sophia Antipolis

Deduplicating noisy data

upcoming
EVENTS & THINGS TO DO

eventful

last·fm

NOV
13

Róisín Murphy PAST EVENT

8 attending | 1 shout | Added by singinanarchist | Export event

Tuesday 13 November 2007 at 8:30pm

Nouveau Casino
109 rue Oberkampf
75011 Paris
France

Show on Map

Tel: 01 43 57 57 40
Web: www.nouveaucasino.net

Flag for review

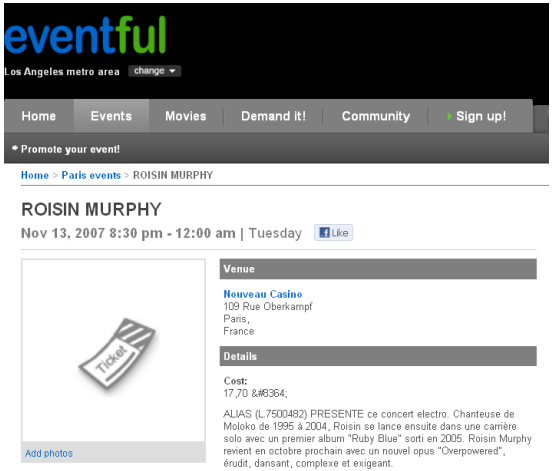
Share this event:

Send | Tweet | Recommend

Play Róisín Murphy Radio

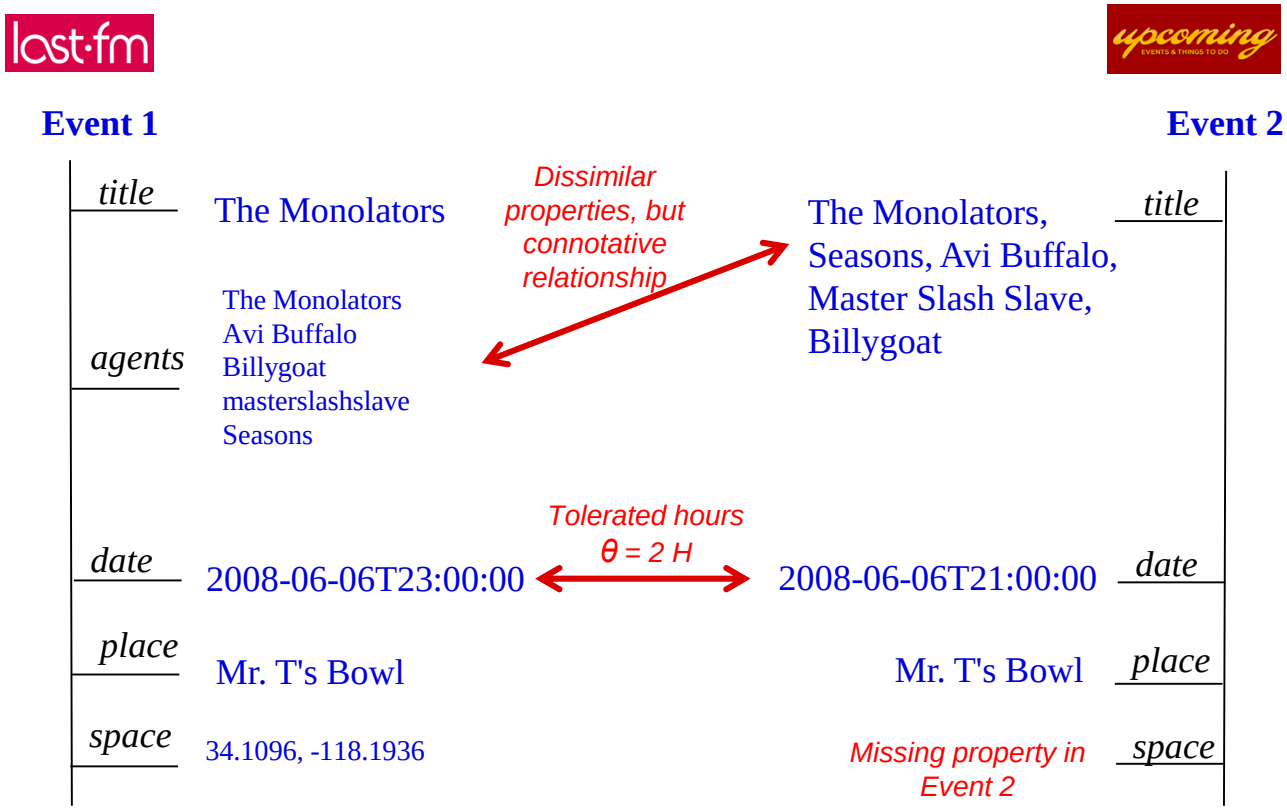
350591

E0-001-005971169-9

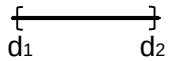
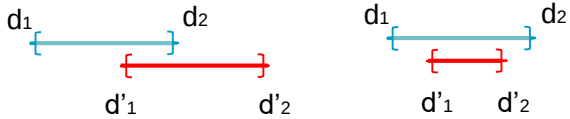


EURECOM
Sophia Antipolis

Deduplication noisy data



Similarity Metrics

Data Type	Existing Metrics	Our Extensions
Date	Temporal distance between two instants 	Temporal Overlap between two intervals 
String	• Token-based (*Jaccard*, *Euclidian*) • Character-based (*Jaro*, *Levenshtein*) • Hybrid Functions	TokenWise = Extension of Jaccard Hybrid • More importance to similar tokens • Penalize unmatched tokens based on the size of the token sets
Numeric	Normalized Difference	-
Geographic	Haversine Formula	-

Matching Events: results

1st step

Blocking keys
AVG (place + time)

2nd step

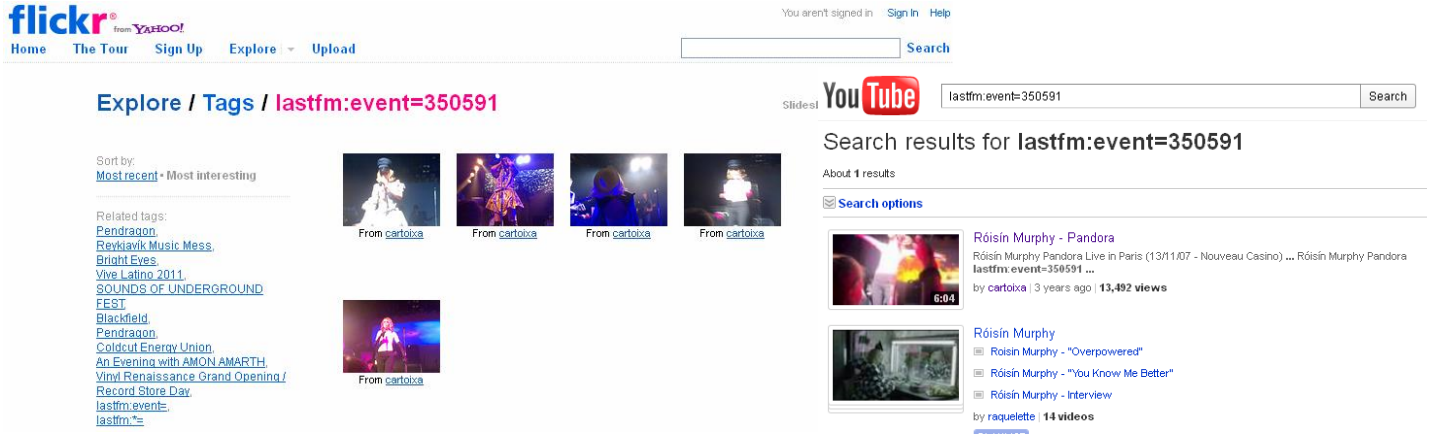
Fair correlation
- LC: Linear Combination
- OR: at least one sim > threshold

P_{source}	P_{target}	*Correlation*	*Coverage*
$time_s$	$time_t$	**1**	**1**
$place_s$	$place_t$	**0.80**	**1**
$title_s$	$title_t$	0.59	**1**
$agent_s$	$title_t$	0.53	**1**
$(lat_s, long_s)$	$(lat_t, long_t)$	(0.43, 0.97)	0.92
$agent_s$	$description_t$	0.24	0.48

	Precision	*Recall*	*F-score*
LC KnoFuss (GA)	0.94	0.74	0.83
LC (PSO)	0.88	0.96	0.92
Two-step LC (PSO)	0.91	0.95	0.93
Two-step OR (PSO)	**0.96**	**0.97**	**0.96**

➡ Missing properties decrease the performance of LC based function

Media explicitly associated with the event



last·fm

upcoming
EVENTS & THINGS TO DO

eventful

APIs

flickr

Machine tags
"lastfm:events"

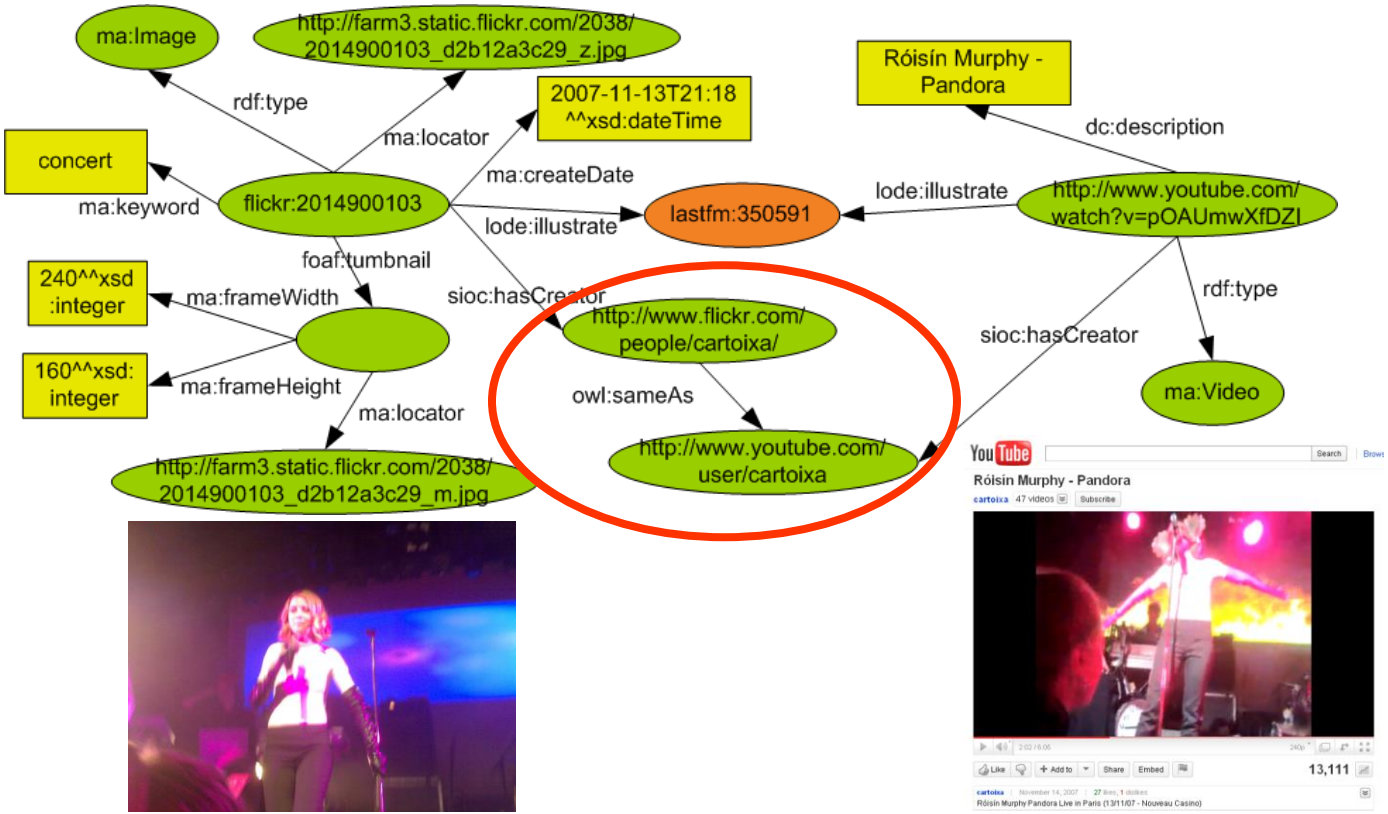
YouTube

4790 photos, 263 videos over 110 events

2+ million images over 110.000+ events

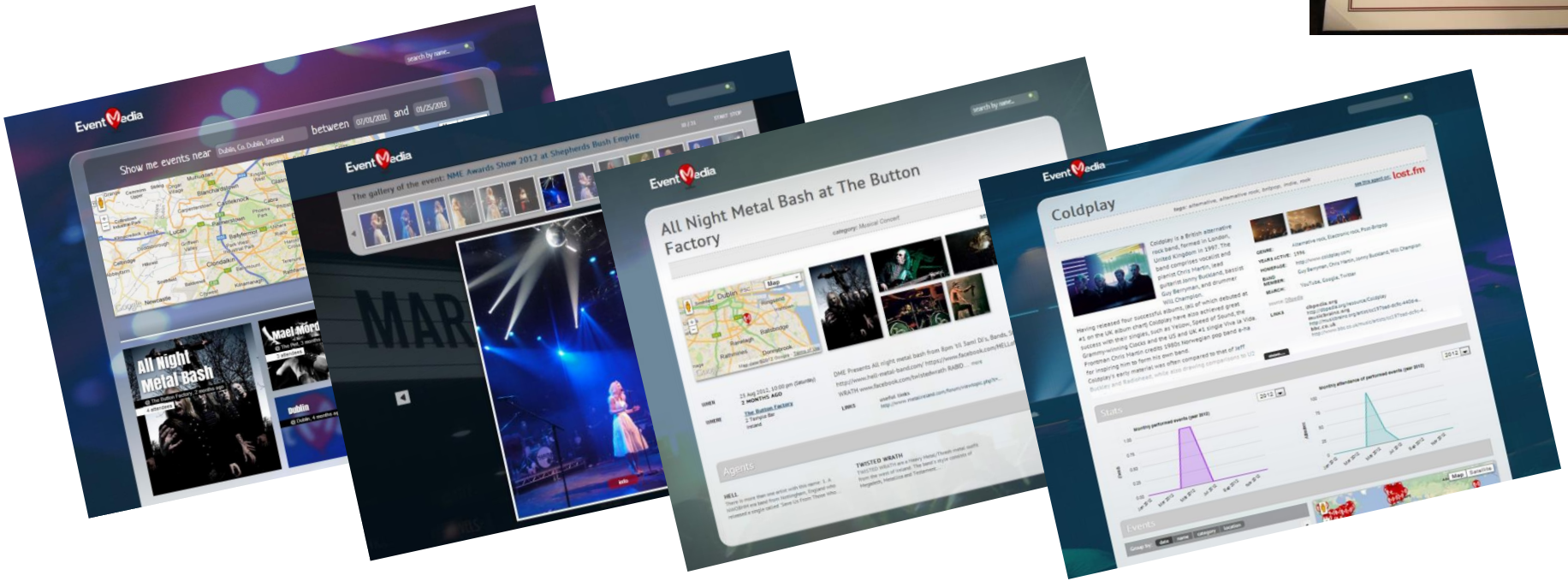
EURECOM
Sophia Antipolis

Representing Media with the W3C Media Ontology



EURECOM
Sophia Antipolis

Live your event



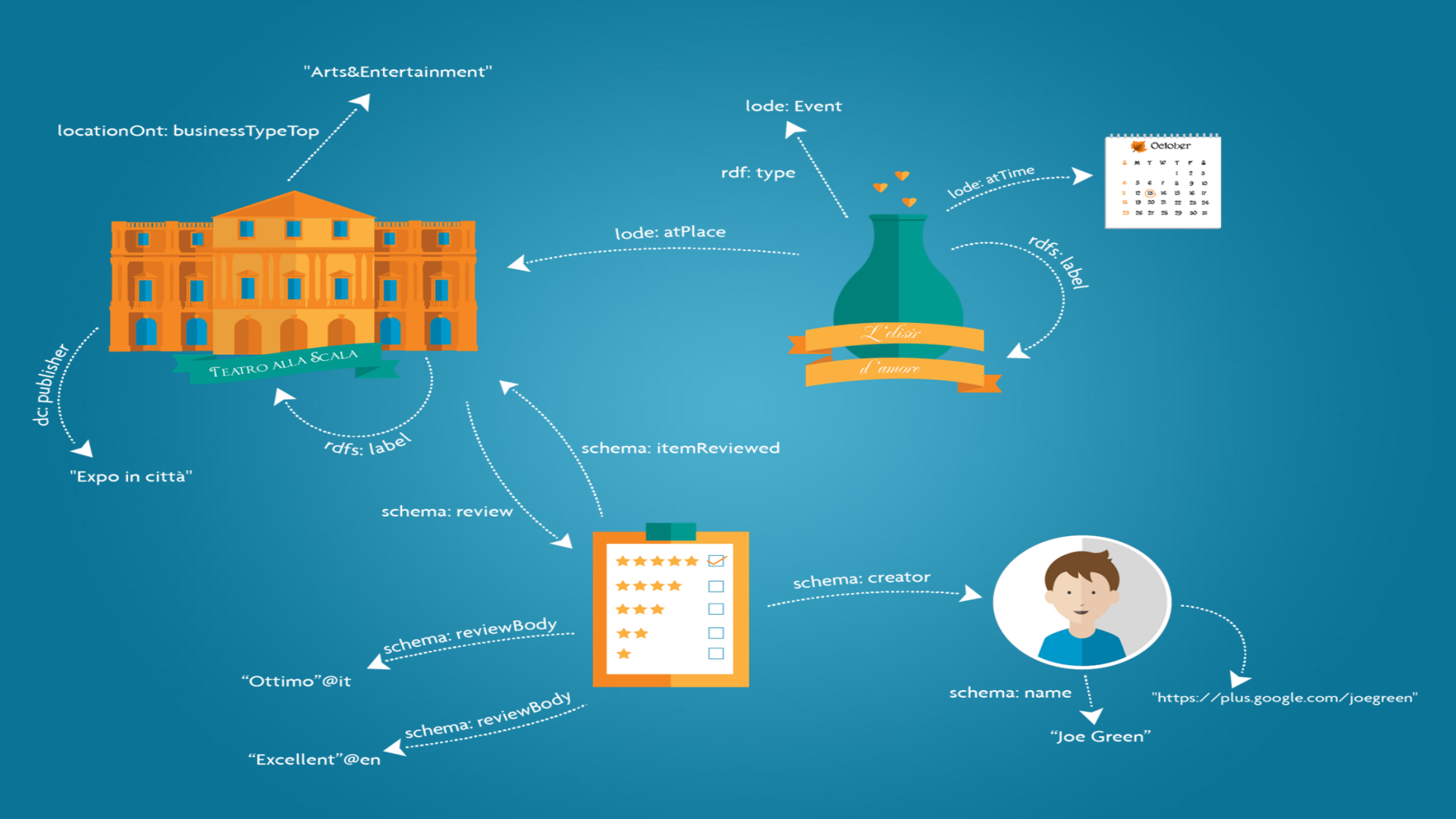
https://www.youtube.com/watch?v=8dVD0SqCB6s

EURECOM
Sophia Antipolis

Building Knowledge Graphs for Tourism

an EIT Digital initiative

EURECOM Sophia Antipolis


"Arts&Entertainment"
locationOnt: businessTypeTop
lode: Event
rdf: type
lode: atTime
October
lode: atPlace
rdfs: label
L'elisir
d'amore
TEATRO ALLA SCALA
dc: publisher
"Expo in città"
rdfs: label
schema: itemReviewed
schema: review
schema: creator
schema: reviewBody
"Ottimo"@it
schema: reviewBody
"Excellent"@en
schema: name
"Joe Green"
"https://plus.google.com/joegreen"

Knowledge Base for Cities

Entschede

30 K places
28 K Events

Time for KB creation divided by 3 between Jan and Nov 2017

Hyper-local sources (OT)

facebook

Google places

yelp

eventful

openagenda

EURECOM
Sophia Antipolis

City Moove Platform R&D

- **Schema.org at the heart of the data model**
- **ETL collectors for places and for events**
  - JSON APIs, XML, RSS, CSV
  - Semantic enrichment:
    - normalization of city names, of address, of geo position
    - prediction of event categories
- **Editorial framework (WordPress based)**
- **Reconciliation engine … based on STEM (paper, code)**

Semantically interconnected data about POIs, Events, Activities, Transportation,

Rizzo G., Troncy R., et al. (2015) 3cixty@Expo Milano 2015: Enabling Visitors to Explore a Smart City. In: 14th International Semantic Web Conference (ISWC'15), Winner of the Semantic Web Challenge

Collecting user trajectory patterns

Data Collection

Twitter Search API

search by "swarmapp.com"

save

Julien	Tate Gallery	Gallery	2017-05-10 16:00
Julien	O'Neill's	Irish Pub	2017-05-10 19:00

Data Enrichment

expand

demographics

contextual

Julien	M	French	40	Design Museum	Gallery	2017-05-10 16:00	Rain
Julien	M	French	40	O'Neill's	Irish Pub	2017-05-10 19:00	Light rain

Path Extraction and user clustering

group per-user PoI categories according to a given time interval

Julien	Gallery, Irish Pub ...	by 8h time interval
Amelie	….	

generate clusters by demographics of collected users

U_1	Julien, Jerome, ….
U_2	Amelie, ...

Path Recommendation

generate paths for each category of venue within a cluster

Per-cluster path learning

[cluster, seed category venue]	category of venue often visited after the seed in the cluster
[U_1, Gallery]	Irish Pub
[U_1, Irish Pub]	Italian Restaurant
[U_1, Art_Museum]	Park
...	

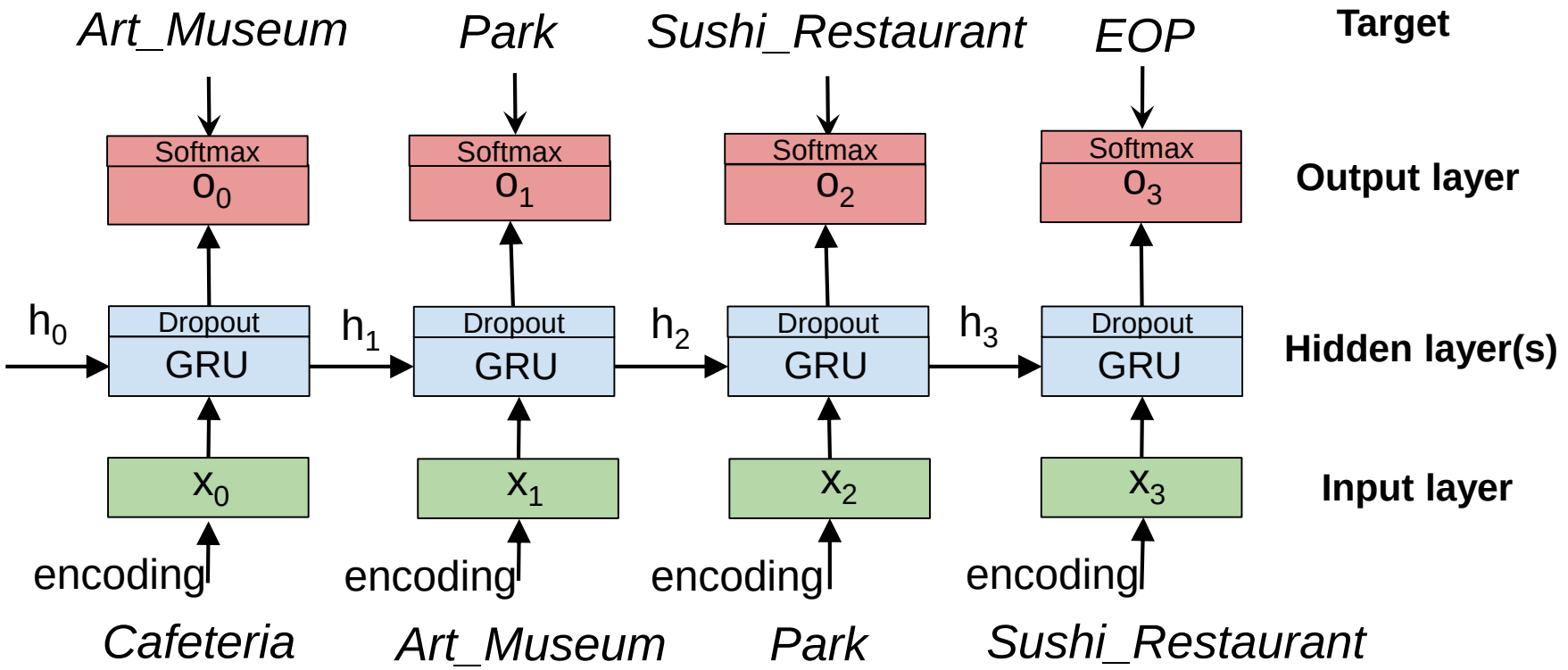
Recommendation

Frédéric, **French man** ... is going to visit **Tate Modern** ...

matches → U_1

recommendations after Tate Modern → Irish Pub, Italian Restaurant

RNN: predicting the next POI category



Learning how to bundle tourism activities

PASTIME

Frédéric, **French man** in his **40**'s, is going to visit **Tate Modern** in **London** this afternoon

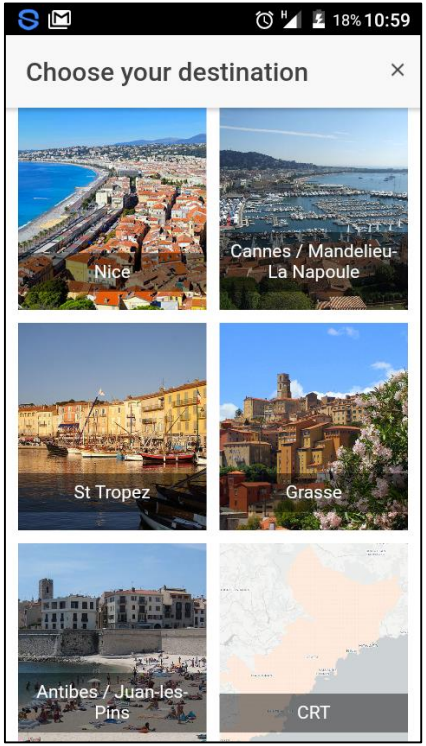
After, he can be interested in

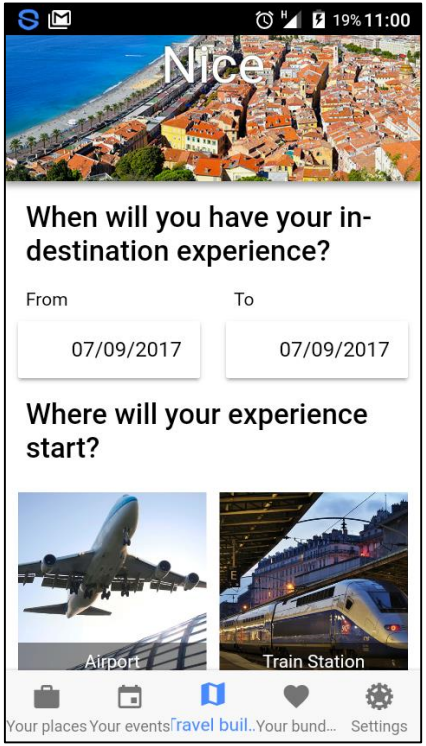
Taking a beer in an **Irish Pub**, first → Then, having a dinner in an **Italian Restaurant** → And, lately, attending an event in a **Jazz Club**

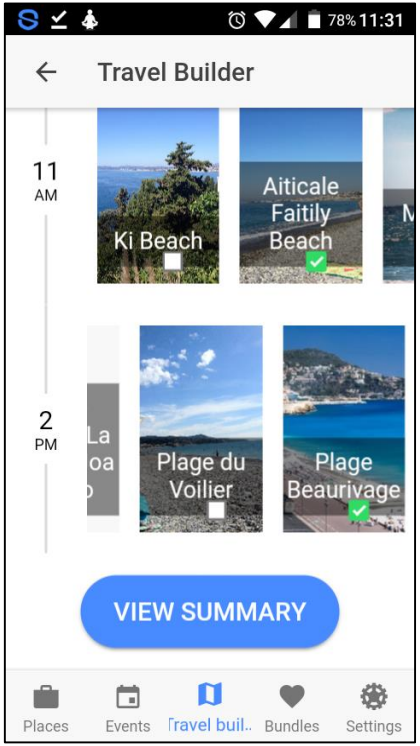
path, ie sequence, of categories of Points of Interest that Frédéric can be interested to go after having seen the Tate Modern gallery

EURECOM Sophia Antipolis

Recommender System







KG on electrical consumption

- **The largest KG about the electrical consumption in France:**
  - Dumps: https://data.edf.eurecom.fr/dumps and https://data.edf.eurecom.fr/sparql
  - 5-star data / spatio-temporal observations
    - http://data.edf.eurecom.fr/graph/housing = housing characteristics from ENEDIS (IRIS, 2023)
    - http://data.edf.eurecom.fr/graph/enedis = daily electrical consumption from ENEDIS (IRIS, 2023)
    - http://data.edf.eurecom.fr/graph/mte-mct = daily electrical consumption from Environment Ministry (2022)
    - http://data.edf.eurecom.fr/graph/rte = electrical consumption (every 15 min) from RTE éCO2mix (PACA region, 2012-2022)
    - http://data.edf.eurecom.fr/graph/cog = geo-shapes of IRIS (2023)
    - http://data.edf.eurecom.fr/graph/meteofrance = weather observations (2023-2024)
    - http://data.edf.eurecom.fr/graph/bpe = public equipment in France (BPE) from Insee (2016)
    - http://data.edf.eurecom.fr/graph/sirene = all corporates in France (SIRENE) from (2022)
  - Ontology, vocabularies and customizable data conversion tools: https://gitlab.pam-retd.fr/colab_eurecom_edf/kg_conso

METEO FRANCE
À VOS CÔTÉS, DANS UN CLIMAT QUI CHANGE

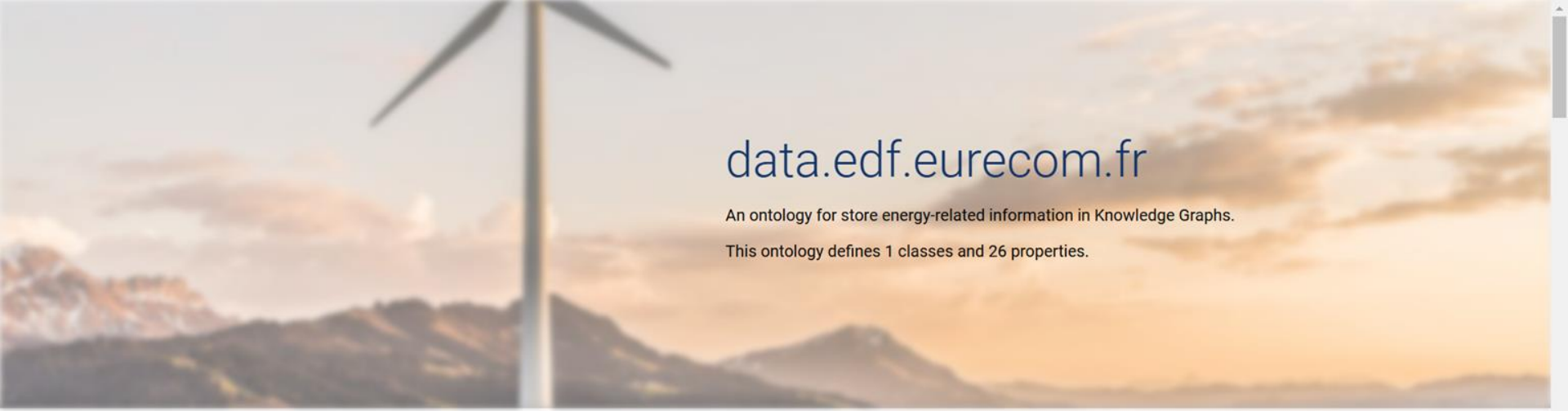
MINISTÈRES TERRITOIRES ÉCOLOGIE LOGEMENT
Liberté
Égalité
Fraternité

Insee

IGN
INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

Rte

ENEDIS



data.edf.eurecom.fr

An ontology for store energy-related information in Knowledge Graphs.

This ontology defines 1 classes and 26 properties.

Summary

Classes

https://data.edf.eurecom.fr/ontology/

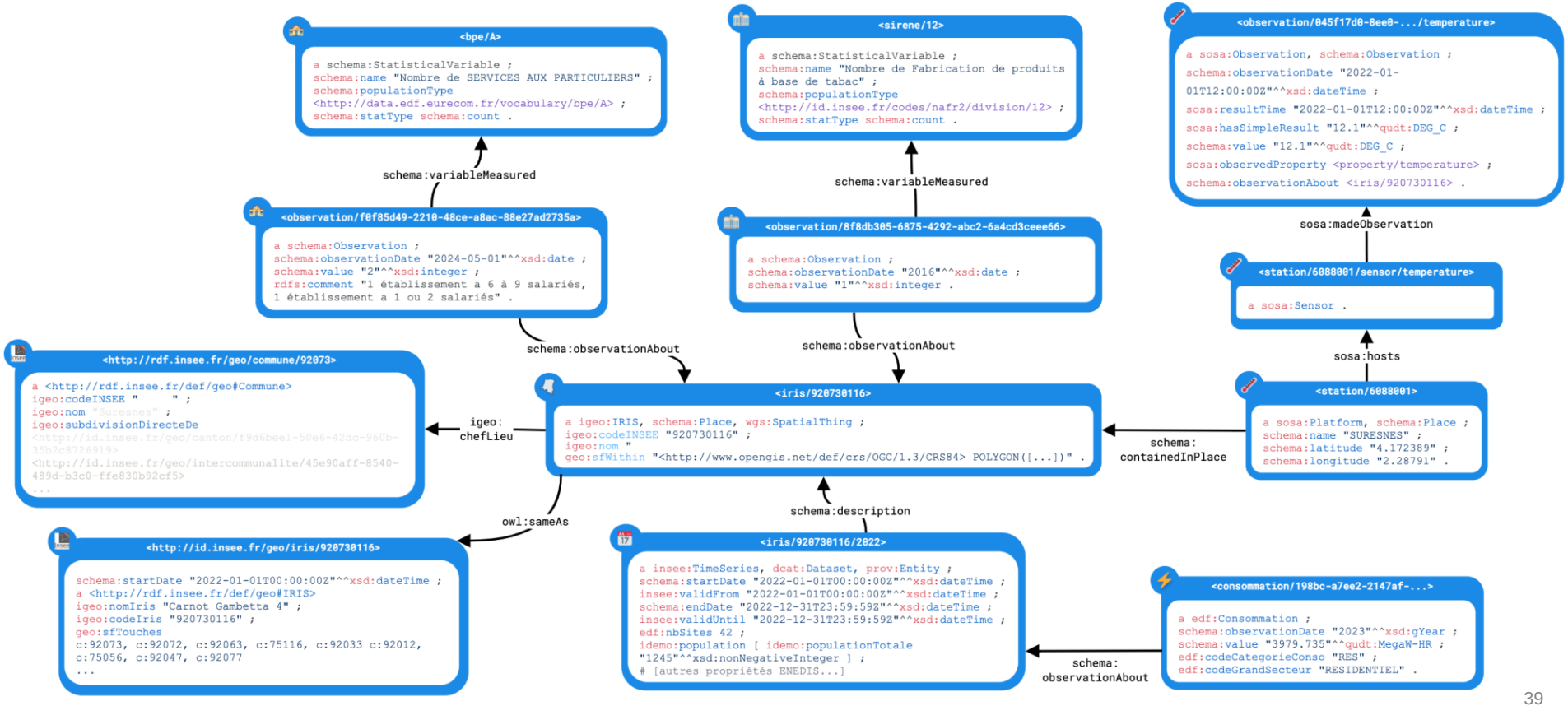
Consommation

Properties

edf:codeCategorieConso | edf:codeGrandSecteur | edf:codeSecteurNaf2 | edf:nbSites | edf:djuATN | edf:djuATR | edf:nbSites | edf:nombreHabitants | edf:partThermosensible | edf:residencesPrincipalesAvant1919 | edf:residencesPrincipalesDe1919a1945 | edf:residencesPrincipalesDe1946a1970 | edf:residencesPrincipalesDe1971a1990 | edf:residencesPrincipalesDe1991a2005 | edf:residencesPrincipalesDe2006a2015 | edf:superficieLogements30a40m2 | edf:superficieLogements40a60m2 | edf:superficieLogements60a80m2 | edf:superficieLogements80a100m2 | edf:superficieLogementsMoins30m2 | edf:superficieLogementsPlus100m2 | edf:tauxChauffageElectrique | edf:tauxLogementsCollectifs | edf:tauxResidencesPrincipales | edf:thermosensibiliteMoyenne | edf:thermosensibiliteTotale

Classes

IRIS as a central geo unit! Description and Temporal Series



Data conversion

- **ENEDIS converter**
  (area = France, year = 2023: 47 GB (301,386,937 triples)
  conversion time ~7h
  - yearly: 0.4 GB (2,628,101 triples)
  - monthly: 1.5 GB (9,454,536 triples)
  - daily: 45 GB (287,575,470 triples)
  - NOTE: considering the period 2011-2023, converting the daily electrical consumption would generate 614 GB (~4 billion triples)
- **SIRENE converter: performance optimization in indexing geospatial data with rtree, conversion time is reduced to ~4h**


Rechercher par code ou nom d'IRIS...
Consommation
Population
Région
Paris
Reims
Saarbrücken
Heilbronn
Karlsruhe
Stuttgart
Strasbourg
Reutlingen
Ulm
Saint-Brieuc
Brest
Rennes
Chartres
Troyes
Nancy
Freiburg im Breisgau
Orleans
Auxerre
Vannes
Angers
Tours
Le Mans
Laval
Basel
Zurich
Bregenz
Nantes
Dijon
Besançon
Bourges
Nevers
Vaduz
VORARLBERG
Bern
SWITZERLAND
FRANCE
Poitiers
Lausanne
Niort
La Rochelle
Geneva
Limoges
Clermont-Ferrand
Lyon
Aosta
Bergamo
Brescia
Angoulême
Saint-Étienne
Novara
Milan
Bay of Biscay
Grenoble
Turin
Valence
Bordeaux
Parma
Genoa
LIGURIA
Avignon
Nîmes
Monaco
Gijón/Xixón
Santander
Toulouse
Montpellier
Oviedo
Bilbao
Marseille
Ligurian Sea
CANTABRIA
Carcassonne
Toulon
PAÍS VASCO
Esri, TomTom, Garmin, FAO, NOAA, USGS | IGN, Esri France
Powered by Esri

Saint-Nicolas

Saint-Nicolas

Aperçu | Logements | Équipements

Consommation | Similaires

Logements par superficie en 2023

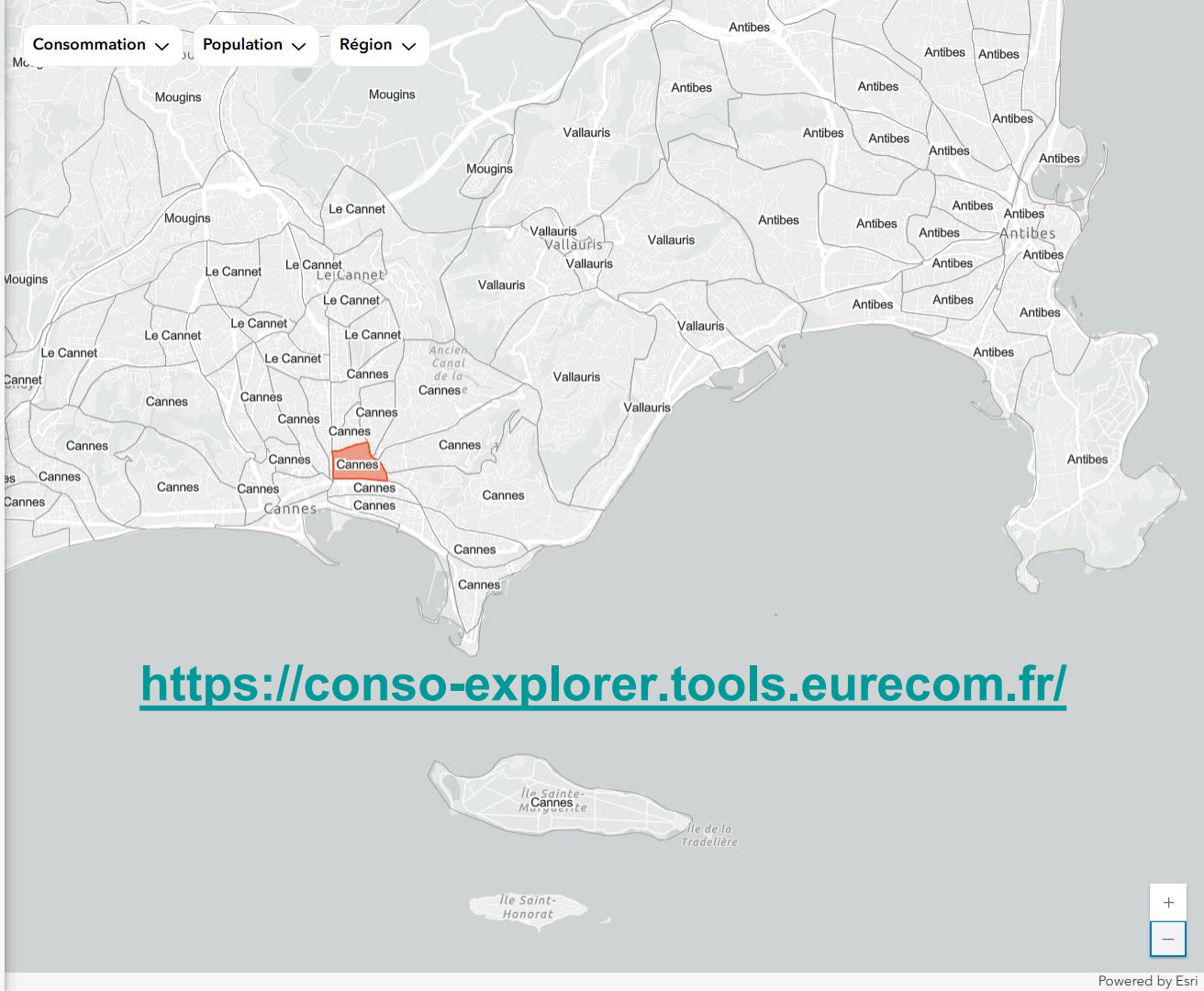
< 30m²	30-40m²	40-60m²	60-80m²	80-100m²	> 100m²
13.49%	12.28%	28.36%	27.12%	11.81%	6.94%

Structure du parc de logements en 2023

Logements collectifs	Résidences principales	Chauffage éléctrique
97.97%	50.78%	52.10%

Taux de résidence principale par période de construction en 2023

< 1919	1919-1945	1946-1970	1971-1990	1991-2005	2006-2015	> 2016
18.41%	22.23%	22.84%	12.55%	15.84%		



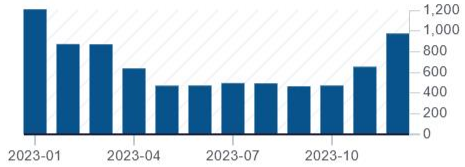
https://conso-explorer.tools.eurecom.fr/

Saint-Nicolas

Histogramme de la consommation (Résidentiel)

8103.10 kWh consommés en 2023

Cliquez sur un mois pour afficher les détails

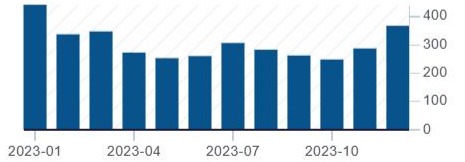


Estimé selon les coefficients des profils de consommation ENEDIS

Histogramme de la consommation (Professionnels)

3668.50 kWh consommés en 2023

Cliquez sur un mois pour afficher les détails



Estimé selon les coefficients des profils de consommation ENEDIS

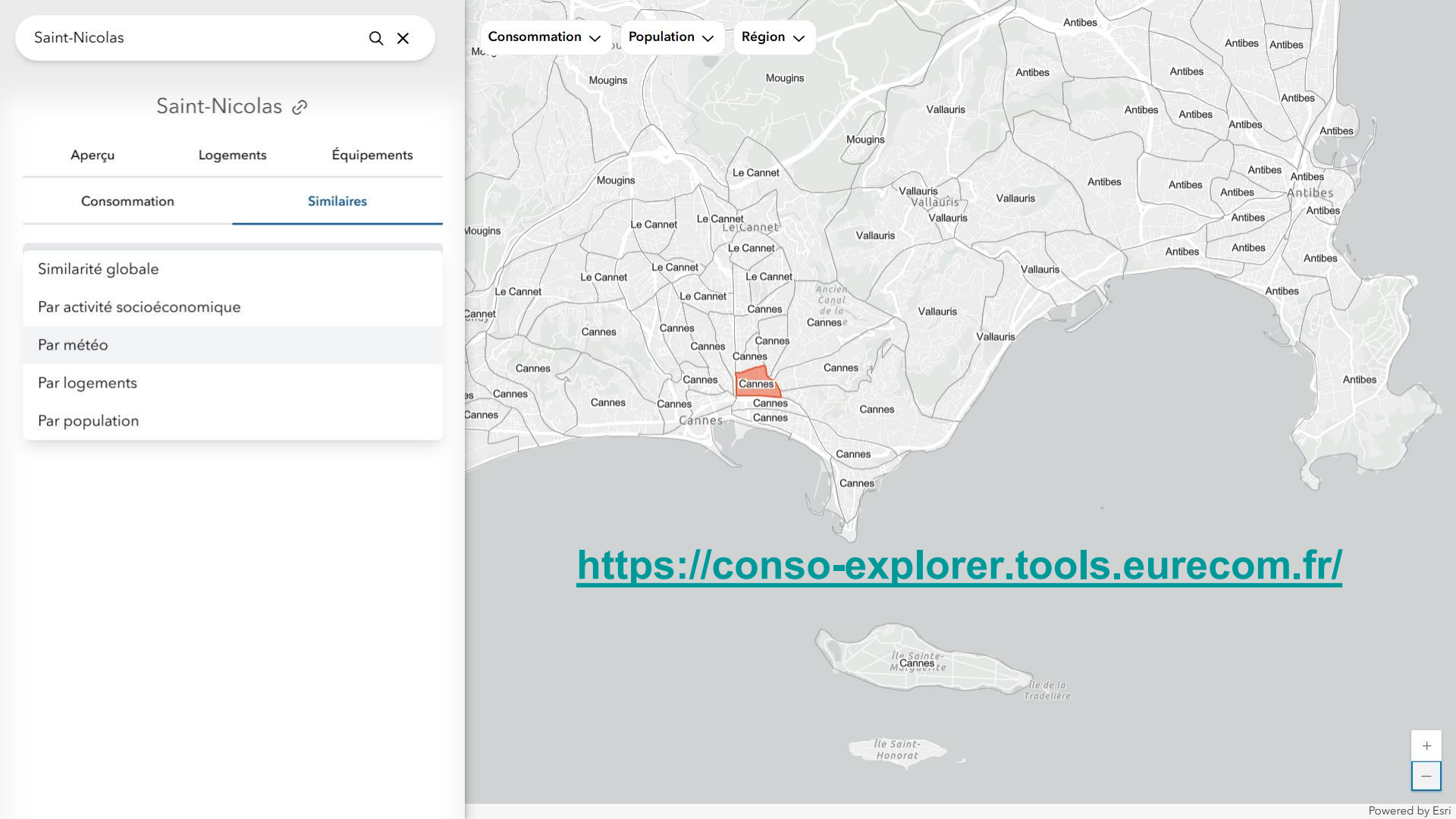
Histogramme de la consommation (Entreprises)

6355.65 kWh consommés en 2023

Consommation
Population
Région

https://conso-explorer.tools.eurecom.fr/

Powered by Esri


Saint-Nicolas
Saint-Nicolas
Aperçu
Logements
Équipements
Consommation
Similaires
Similarité globale
Par activité socioéconomique
Par météo
Par logements
Par population
Consommation
Population
Région
https://conso-explorer.tools.eurecom.fr/
Powered by Esri

About: http://data.edf.eurecom.fr/iris/060290109

Goto Sponge NotDistinct Permalink

An Entity of Type : schema:Place, within Data Space : data.edf.eurecom.fr associated with source document(s)

Type: schema:Place New Facet based on Instances of this Class

Attributes	Values
rdf:type	https://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing Iris schema:Place
sameAs	http://id.insee.fr/geo/iris/060290109
type d'Iris	Iris d'habitat
ogcgs:sfTouches	http://data.edf.eurecom.fr/iris/060290103 http://data.edf.eurecom.fr/iris/060290104 http://data.edf.eurecom.fr/iris/060290105 http://data.edf.eurecom.fr/iris/060290106 http://data.edf.eurecom.fr/iris/060290108 http://data.edf.eurecom.fr/iris/060290110 http://data.edf.eurecom.fr/iris/060290112
chef-lieu	Cannes
code Insee	060290109
nom	Saint-Nicolas
ogcgs:sfWithin	POLYGON((7.0250528601963 43.55451455142,7.0250817563957 43.554326090878,7.0250985245159 43.554035377161,7.0249036307976 43.554047113167,7.0214267328022 43.554261649267,7.0184087025035 43.554236553262,7.0170881664498 43.554217895227,7.0170330088826 43.554226244089,7.017013059423 43.554967523403,7.0169670445137 43.556008867641,7.0169469832388 43.556465466872,7.0169148471996 43.557245036419,7.0169144294475 43.55725676362,7.0169718094507 43.557244729006,7.0170461326558 43.557244679603,7.017407383517 43.557255626511,7.0178579333989 43.557285787489,7.0182641980247 43.557373442982,7.0183170362006 43.557384999371,7.0186726037547 43.557491648418,7.0190800273602 43.557613492392,7.0197658676544 43.557801595672,7.0200131417836 43.55786540616,7.0205443577701 43.557968243041,7.0209792778737 43.558040412254,7.0221445002966 43.558227848831,7.0222700204432 43.557844814416,7.0222809661086 43.557806570527,7.0225598805539 43.556961088993,7.0226314110023 43.556815192595,7.0226787686162 43.556748571307,7.0227830144168 43.556627587925,7.0229835454607 43.556484114764,7.0236422778426 43.556232655649,7.02383607183 43.556134475999,7.024010411563 43.55604152216,7.0241232887867 43.555955352552,7.0242111397772 43.555865606674,7.0244237060291 43.555628890925,7.0245826255885 43.555422093123,7.0247396531141 43.555188338099,7.0248618829169 43.554988306548,7.0249303288323 43.554851532031,7.0249498577344 43.554812068492,7.0250528601963 43.55451455142))

is schema:observationAbout of http://data.edf.eurecom.fr/observation/23ee704b-b493-5e0f-89c4-412785a38015
http://data.edf.eurecom.fr/observation/32eedf79-df6e-5547-bef8-93be95a13ef7
http://data.edf.eurecom.fr/observation/7cd29628-e135-500a-8360-781ce02bcb09
http://data.edf.eurecom.fr/observation/7d4c4ab0-4bec-5aa5-9835-18f90bff47e9
http://data.edf.eurecom.fr/observation/83270b6d-c9d2-5901-a3c4-fe57181cf250
http://data.edf.eurecom.fr/observation/92da5489-2b19-5a22-90f6-9c7d10412f1b
http://data.edf.eurecom.fr/observation/b048e023-7e53-50b5-844d-14bcce26e4a1
http://data.edf.eurecom.fr/observation/30831e60-4f4f-5d99-a212-49f06d69143a
http://data.edf.eurecom.fr/observation/3b3bd38e-72c6-53e7-93ff-5f94c66309c2
http://data.edf.eurecom.fr/observation/41aab329-8d8c-5b29-a080-f36c0b49354a
»more»

What did we learn?

- Ontologies are mostly flat schemas (schema.org)
- Controlled vocabularies carry most of the semantics
- KG construction from API and structured data
- Reconcile entities / de-duplicate records
- The value of the KG depends on the freshness of the data (daily updates require engineering effort)
- Triple stores do scale
- The KG is hidden and it powers specialized user interfaces and recommender systems

Building Knowledge Graphs for Cultural Heritage

Music: DOREMUS
Textile: SILKNOW
Smell experiences: ODEUROPA

EURECOM
Sophia Antipolis

Building the **DOREMUS graph**

pic: https://pxhere.com/it/photo/1523259

1801
Composition of *Sonata quasi una fantasia* by Beethoven

1802
First publication of music score

1802
Premiere

2014
Studio **Performance**

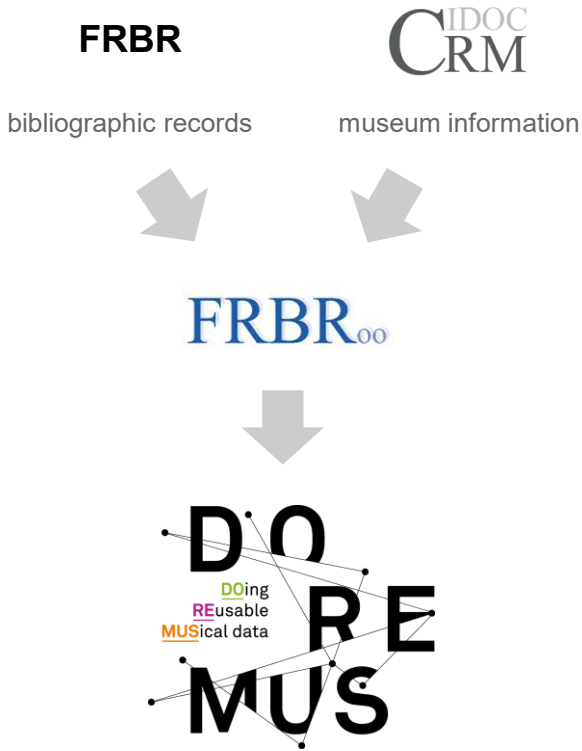
2014
Studio Performance **Recording**

2014
Editing tracks of Recording

2015
CD publication including
« Sonata quasi una Fantasia »

The DOREMUS Model

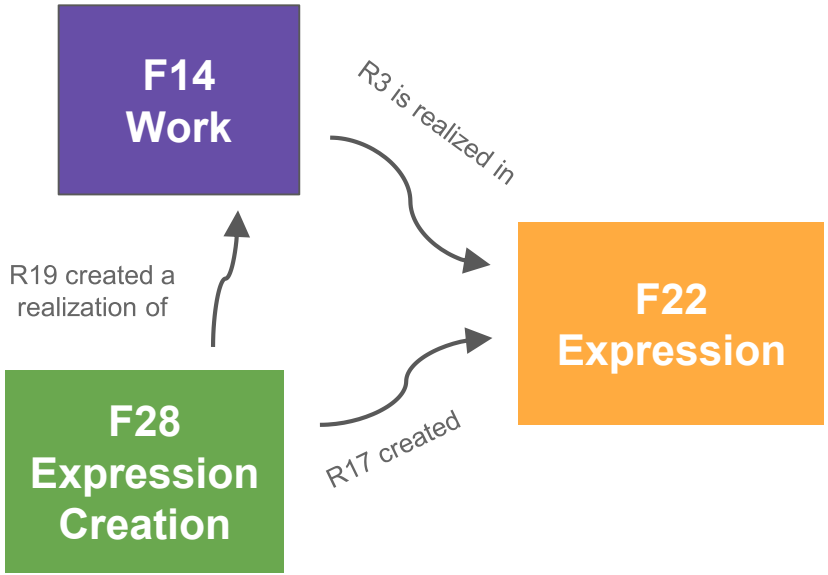
- Music specific **extension of FRBRoo**
- **Dynamic**: it is made up of autonomous combined modules
- Relies on **Linked Data** principles (everything is an URI, RDF model)

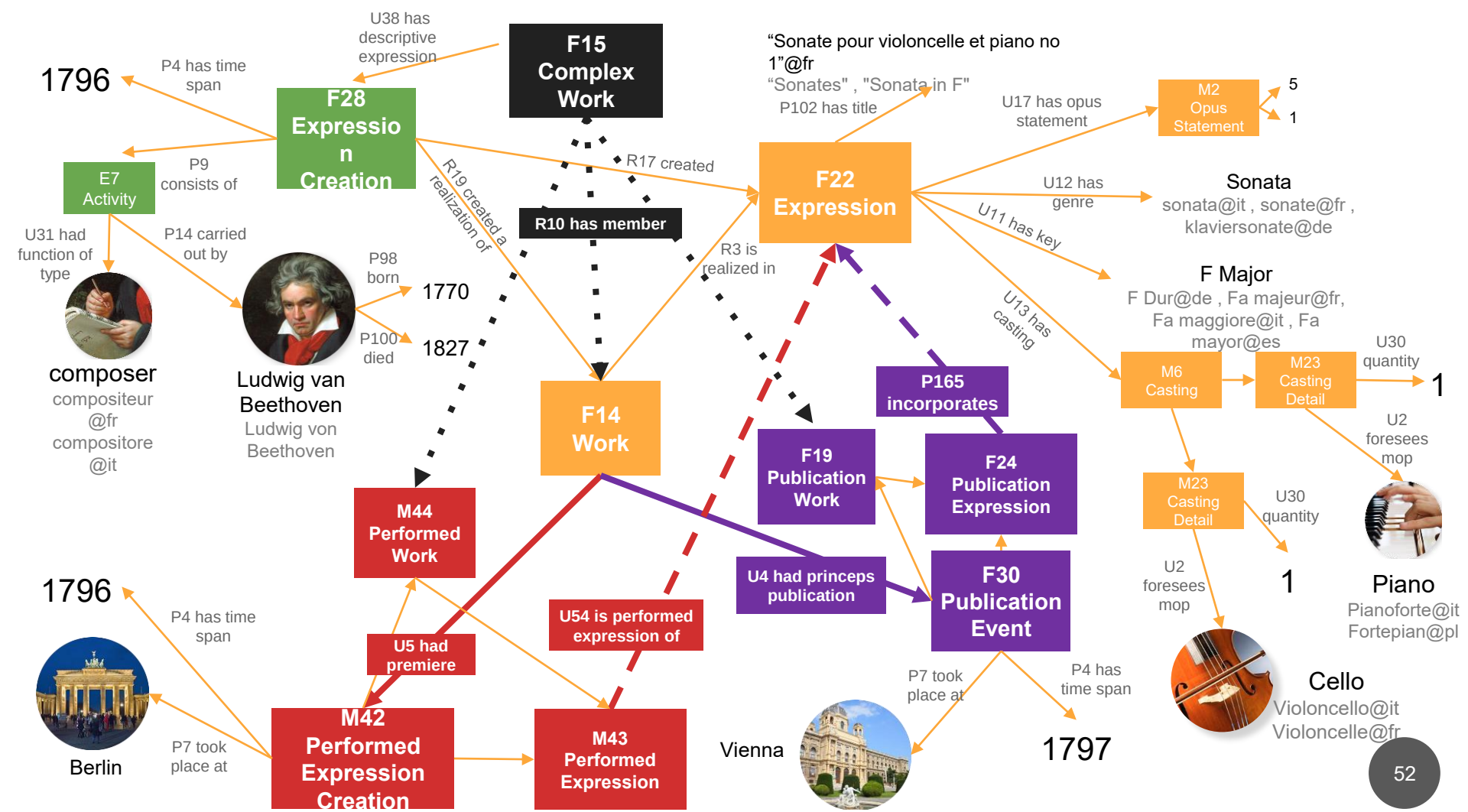


P. Choffé and F. Leresche (2016). ***DOREMUS: connecting sources, enriching catalogues and user experience.*** In 24th IFLA World Library and Information Congress.

The building blocks

Work-Expression-Event




1796
P4 has time span
U38 has descriptive expression
F15 Complex Work
“Sonate pour violoncelle et piano no 1”@fr
“Sonates" , "Sonata in F"
P102 has title
F28 Expression Creation
E7 Activity
P9 consists of
R19 created a realization of
R17 created
F22 Expression
U17 has opus statement
M2 Opus Statement
5
1
U12 has genre
Sonata
sonata@it , sonate@fr , klaviersonate@de
U11 has key
F Major
F Dur@de , Fa majeur@fr, Fa maggiore@it , Fa mayor@es
U31 had function of type
P14 carried out by
R10 has member
R3 is realized in
U13 has casting
composer
compositeur @fr
compositore @it
P98 born
1770
P100 died
1827
Ludwig van Beethoven
Ludwig von Beethoven
F14 Work
P165 incorporates
M6 Casting
M23 Casting Detail
U30 quantity
1
U2 foresees mop
F19 Publication Work
F24 Publication Expression
M44 Performed Work
M23 Casting Detail
U30 quantity
1
U2 foresees mop
Piano
Pianoforte@it
Fortepian@pl
U4 had princeps publication
F30 Publication Event
1796
P4 has time span
U5 had premiere
U54 is performed expression of
P7 took place at
P4 has time span
Cello
Violoncello@it
Violoncelle@fr
Berlin
P7 took place at
M42 Performed Expression Creation
M43 Performed Expression
Vienna
1797

Controlled Vocabularies What?

Alternate labels

“Sax”@en

“Saxophone”@en

“Saxofone”@pt

“Sassofono”@it

“Saxophone”@fr

Alternate languages

Notes

“English term is preferred globally”

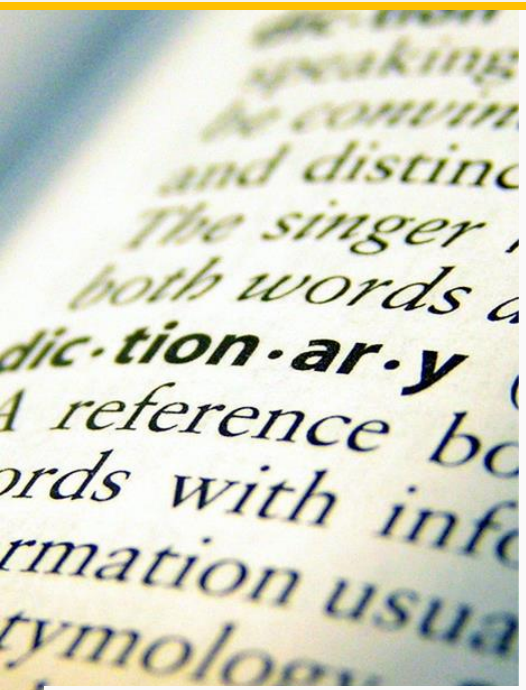
“Woodwinds”@en
“Legni”@it

“Baritone Saxophone”@en

Hierarchy

Example: http://data.doremus.org/vocabulary/iaml/mop/wsa

Controlled Vocabularies Which ones?



GENRES
Diabolo
IAML
Itema3
Redomi
RAMEAU

INTERLINKED

Medium of performance
MIMO
Itema3
IAML
Diabolo
RAMEAU
Redomi

INTERLINKED

Modes

Catalogues

Derivation types

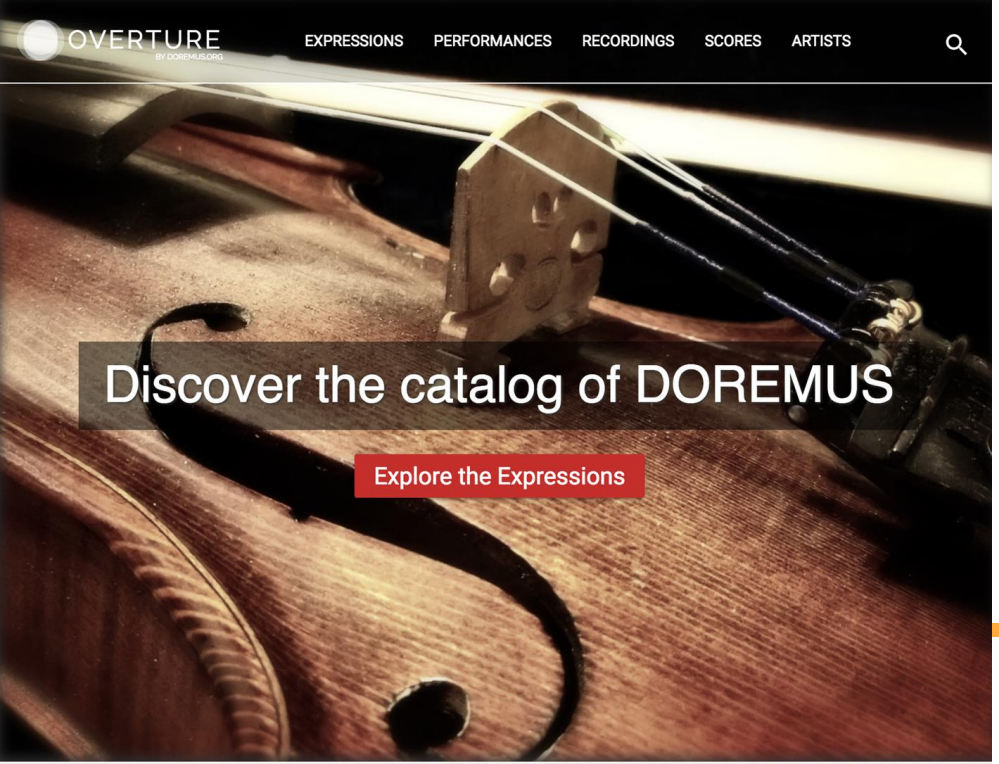
Musical keys

Functions

more available at
http://data.doremus.org/vocabularies

23 families of vocabularies · 11,000+ concepts · 610 links between terms

P. Lisena et al. (2018). ***Controlled Vocabularies for Music Metadata.***
In *19th International Conference on Music Information Retrieval (ISMIR)*. Paris, France.



OVERTURE allows you to explore data about classical and jazz works coming from the French National Library (BnF), the Philharmonie of Paris and Radio France. This data is also available in our triplestore at data.doremus.org/sparql.

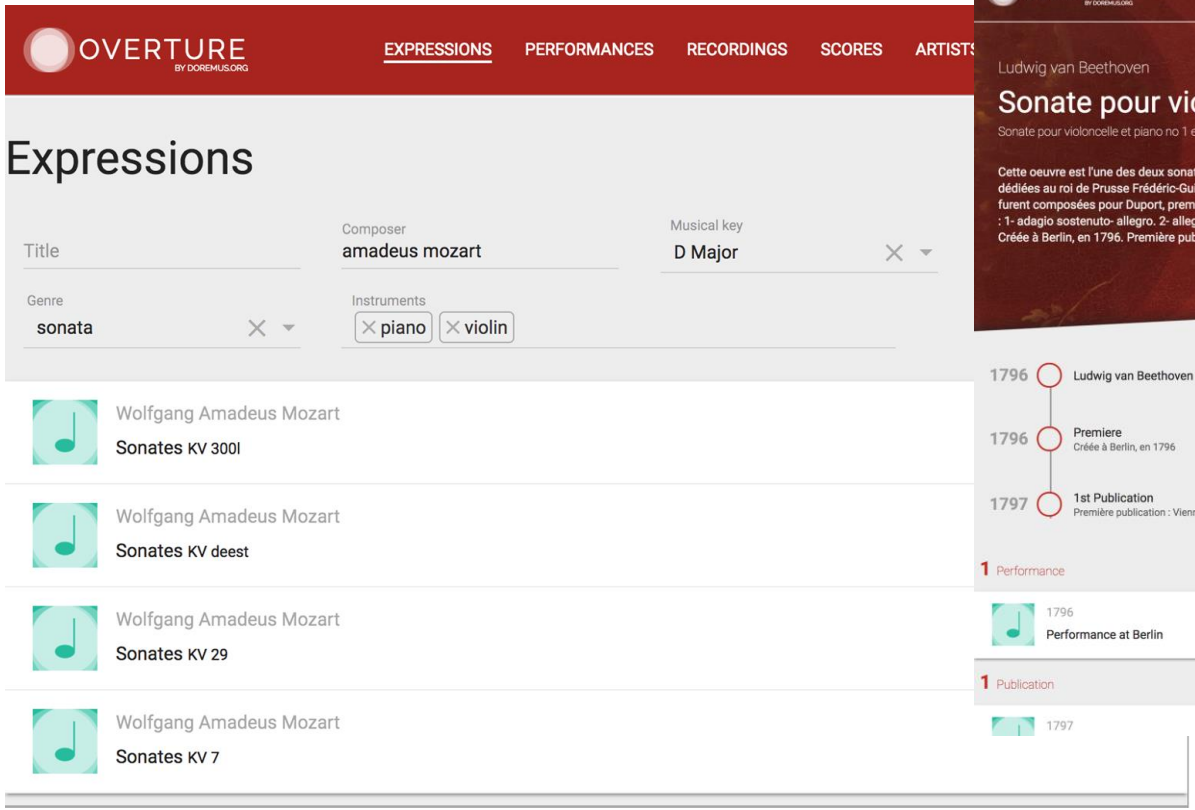
OVERTURE
BY DOREMUS.ORG

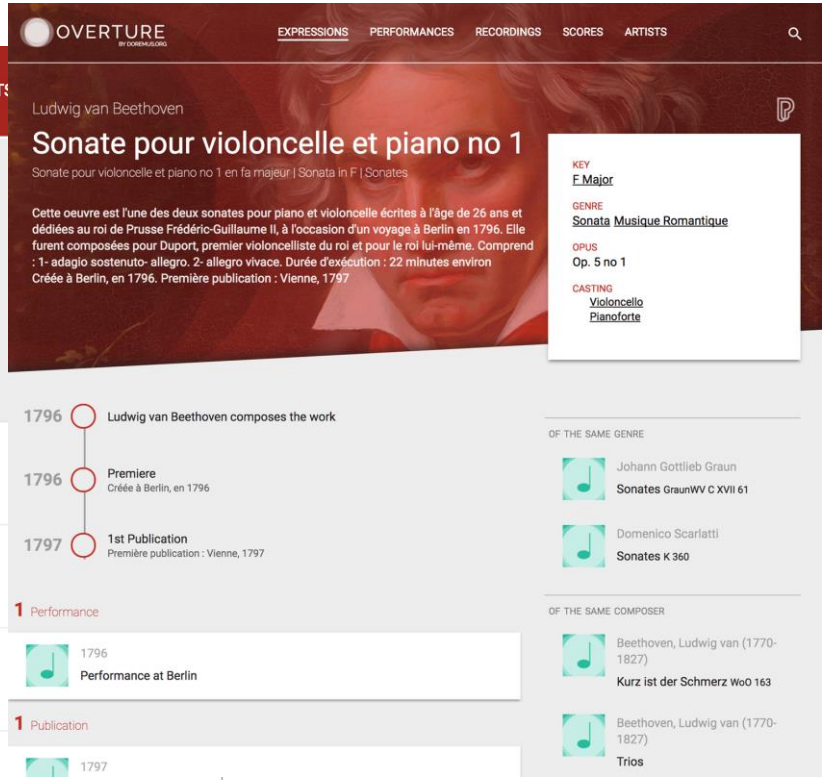
Demo
http://overture.doremus.org/

Repository
https://github.com/DOREMUS-ANR/overture

Exploratory Search Engine

http://overture.doremus.org





WORKS PERFORMANCES RECORDINGS SCORES ARTISTS

Johann Sebastian Bach

Bach Jean Sebastien | Jean Sébastien Bach | Jean Sebastien Bach

Johann Sebastian Bach (31 March [O.S. 21 March] 1685 – 28 July 1750) was a German composer and musician of the Baroque period. He enriched established German styles through his skill in counterpoint, harmonic and motivic organisation, and the adaptation of rhythms, forms, and textures from abroad, particularly from Italy and France. Bach's compositions include the Brandenburg Concertos, the Goldberg Variations, the Mass in B minor, two Passions, and over three hundred cantatas of which around two hundred survive. His music is revered for its technical command, artistic beauty, and intellectual depth. EN

BIRTH
Eisenach, 1685-03-21

DEATH
Leipzig, 1750-07-28

To Artist Info

Similar artists

Played MoP | Period | Casting | Function | Genre | Key

set weights:

Francesco Durante
1684 - 1755

Antonio Vivaldi
1678 - 1741

Domenico Scarlatti
1685 - 1757

Pietro Paolo Bencini
1670 - 1755

Tomaso Albinoni
1671 - 1751

Giuseppe Sammartini
1695 - 1750

Louis-Nicolas Clérambault
1676 - 1749

The data here visualized come from the musical libraries of:

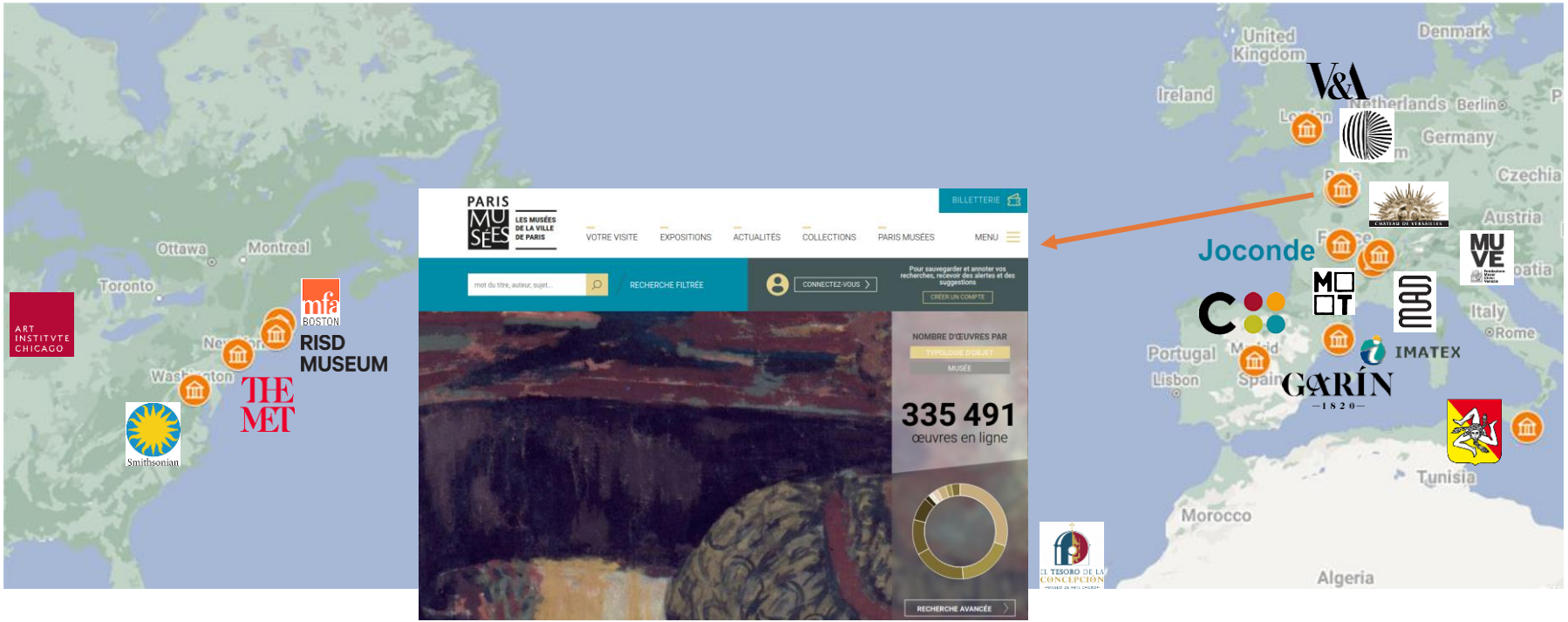
OVERTURE lives inside the DOREMUS project, for describing, publishing, connecting and contextualizing music catalogues on the web of data.

ABOUT CREDITS

Language
English

The European silk heritage is in danger - How can we help to preserve it?

Method: Historians and museum logbook



- Manually searching for candidate museums based on their silk collections
- Only select museums with relevant silk items and image illustrations

Competency questions

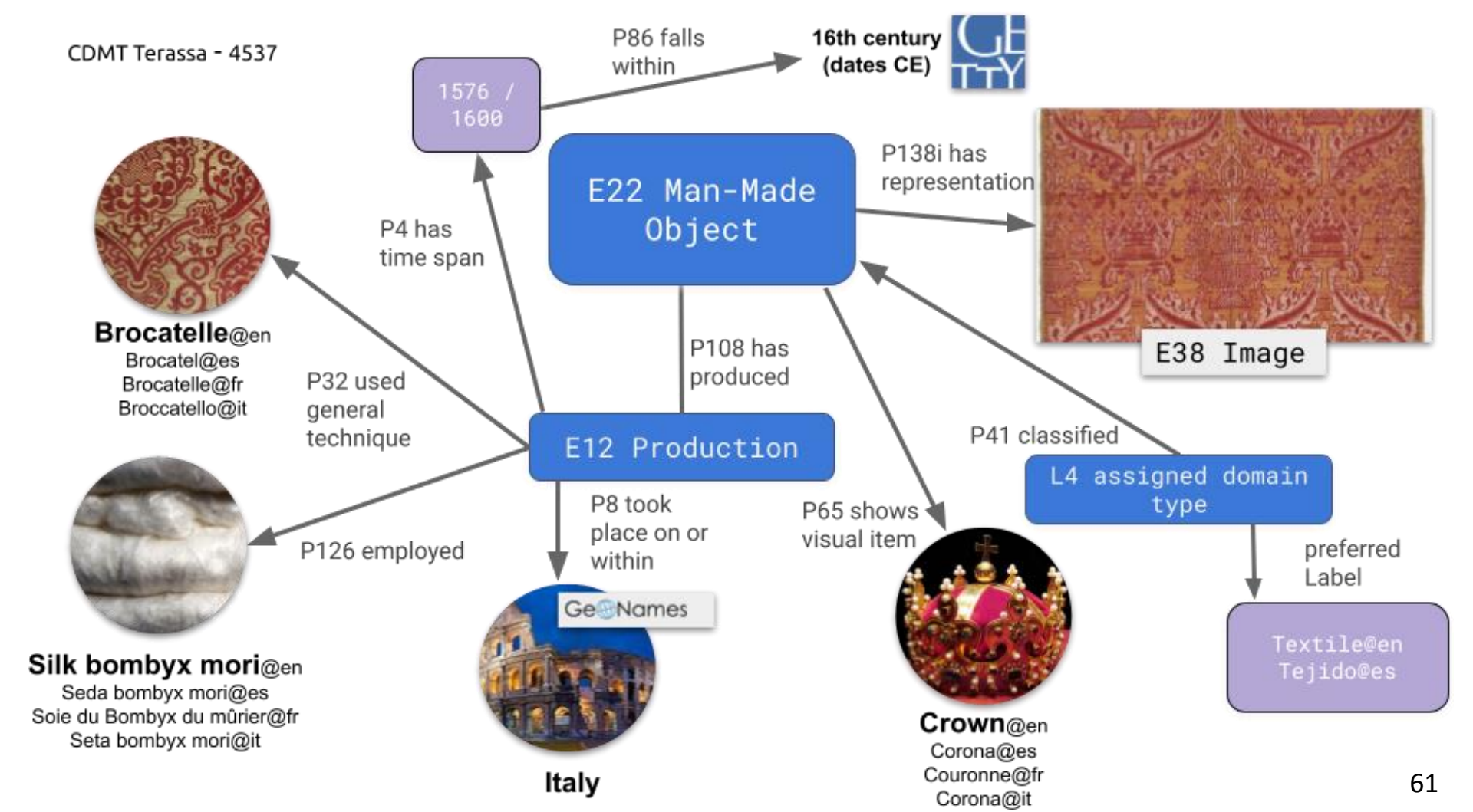
[*Bezerra et al., 2003*; *Wisniewski et al., 2019*]

- Experts formulated 63 questions in English and 15 in Spanish (mostly translations)
- Aim to scope the domain and the future data model

Examples from SILKNOW:

1. **Which items were produced in [Spain], during the [16th century]?**
2. **Which items were produced with [silk] and [silver]?**
3. **Give me all the information you have on [Philippe de la Salle]**
4. **Give me all the items depicting [flowers]**
5. **Who is the [Revel style] name after?**

Bezerra, C., Freitas, F. and Santana, F.: Evaluating ontologies with competency questions. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* , p. 284-285, 2003.
Wiśniewski, D., Potoniec, J., Ławrynowicz, A. and Keet, C.M.: Analysis of Ontology Competency Questions and their formalizations in SPARQL-OWL. *Journal of Web Semantics*, 2019.


CDMT Terassa - 4537
1576 / 1600
P86 falls within
16th century (dates CE)
E22 Man-Made Object
P138i has representation
E38 Image
P4 has time span
Brocatelle@en
Brocatel@es
Brocatelle@fr
Broccatello@it
P32 used general technique
P108 has produced
E12 Production
P41 classified
L4 assigned domain type
preferred Label
Textile@en
Tejido@es
P126 employed
P8 took place on or within
P65 shows visual item
Silk bombyx mori@en
Seda bombyx mori@es
Soie du Bombyx du mûrier@fr
Seta bombyx mori@it
GeoNames
Italy
Crown@en
Corona@es
Couronne@fr
Corona@it

MET - 37.80.1

sixteenth century
(dates CE)

P4 has
time span

E22 Man-Made
Object

P138i has
representation

E38 Image

P108 has
produced

O8 observed

S4 Observation

"This tapestry is one of seven hangings at The Cloisters that depict the hunt of the unicorn....."

P3 has
note

E12 Production

P32 used
general
technique

?

Wool@en
Lana@es
Laine@fr
Lana@it

P126 employed

P8 took place on
or within

France

KG fuels all kinds of machine learning experiments

SILKNOW
Knowledge Graph

Only property values that are linked with the Thesaurus

Easy Data Extraction for Training Data

Text

Integration of predictions

Description
Dress fabric of wool and silk, Great Britain, 1845-1860

Dress fabric of figured wool, possibly a border to a skirt or shawl. White silk warp and woollen weft, loosely woven in tabby and with a woollen pattern weft in dark brown-red, dark green, light green, lilac, purple and pink. From a floral meander of the hem, graceful stems extend 18 inches upwards. The pattern wefts are cropped on the back. The pattern is woven to shape at the right edge and cut at the left. The hem is neatly made with a strip of the plain silk and wool material at the back.

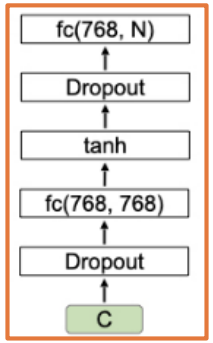
Timespan	Technique	Material
XIX	[NA]	[NA]
[NA]	[NA]	Animal fibre
[NA]	[NA]	Animal fibre
[NA]	Embroidery	Metal thread

Multimodal

Image

Integration of predictions

Supervised Text Classification - Methodology

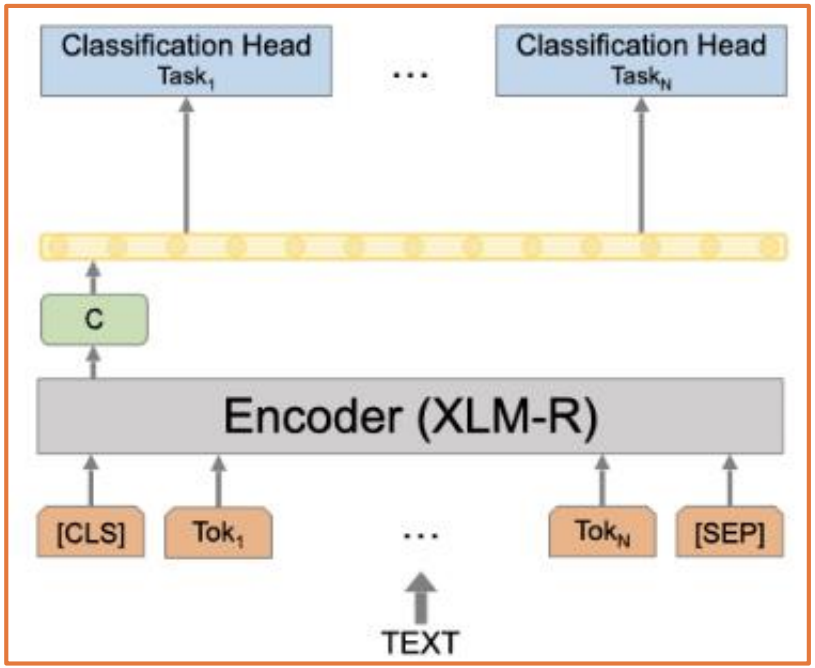


Training:

Examples of objects' text descriptions with the property it needs to learn to infer

Evaluation:

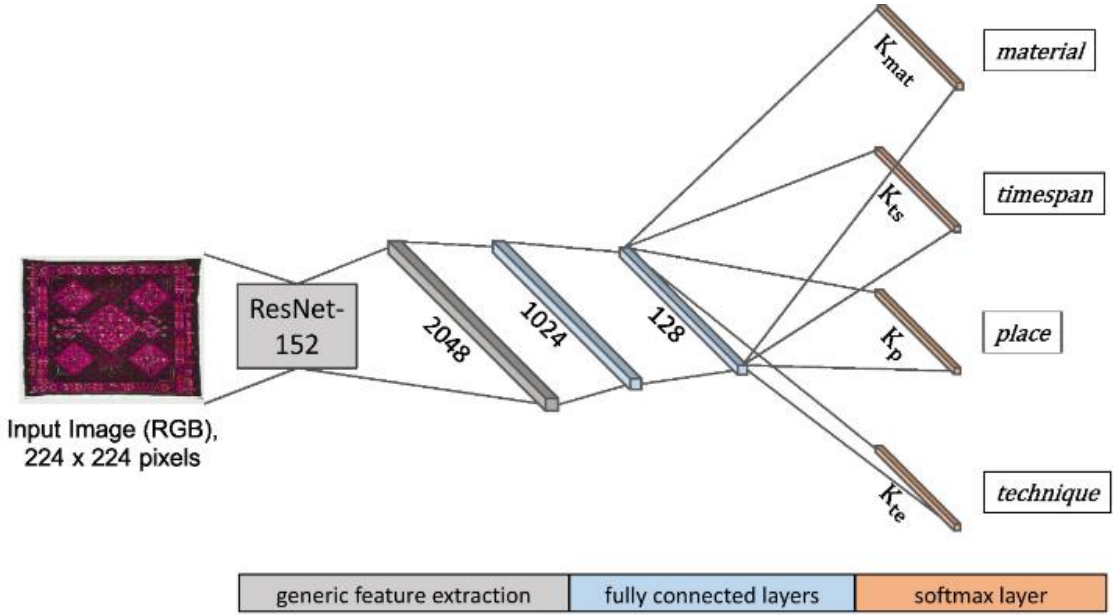
We feed it text but hold out the property value and see if the model guesses it correctly



Based on a shared fine-tuned **XLM-R encoder**, due to preliminary architecture comparisons and because it provides cross-linguality

Supervised Image Classification - Methodology

- Multi-task CNN (ResNet 152[*He et al., 2016*]), due to its proven success with image classification, pre-trained on ImageNet
- 4 output branches, each for one semantic property
- Training on examples with both an image and the semantic property it needs to learn

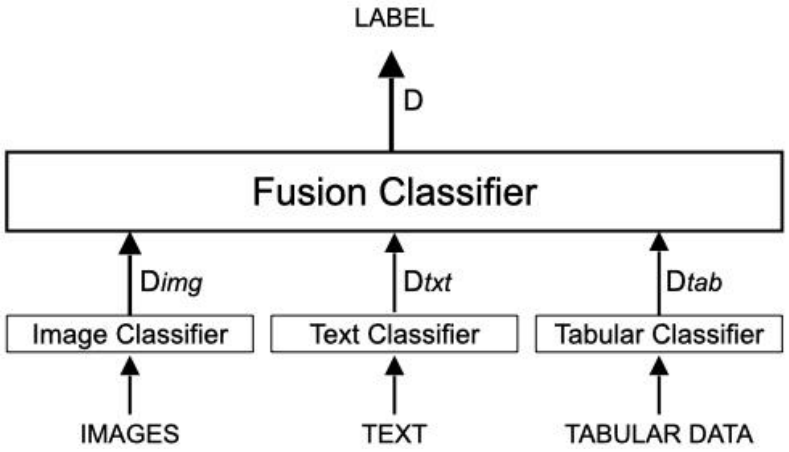


K. He, X. Zhang, S. Ren, and J. Sun. 2016. Identity mappings in deep residualnetworks. InEuropean Conference on Computer Vision (ECCV). 630–645.

Multimodal and tabular classification

- Gradient Boosted Decision Trees in both cases
- XGBoost implementation

Target Variable	Target Value	Feature				
		museum	place	timpespan	technique	material
place	FR	risd	-	[NA]	[NA]	animal fibre
timespan	XVIII	met	[NA]	-	embroidery	animal fibre
technique	other	garin	ES	XX	-	vegetal fibre
material	vegetable fibre	vam	GB	XIX	embroidery	-



Tabular classification input with one example per row task

Results

Variable	Nr. of Classes	Train Set 60%	Valid. Set 20%	Test Set 20%	Image	Text	Tabular	Multimodal
place	9	10,435	3,456	3,470	**38.0**	65.0	46.2	**77.6**
timespan	5	8,819	2,975	2,949	**49.2**	55.6	58.6	**74.2**
technique	4	4,813	1,663	1,675	73.5	**41.0**	68.3	**83.6**
material	3	12,865	4,263	4,351	46.5	**37.4**	49.4	**61.3**

Harmonized F1 scores in % for evaluation on the test set

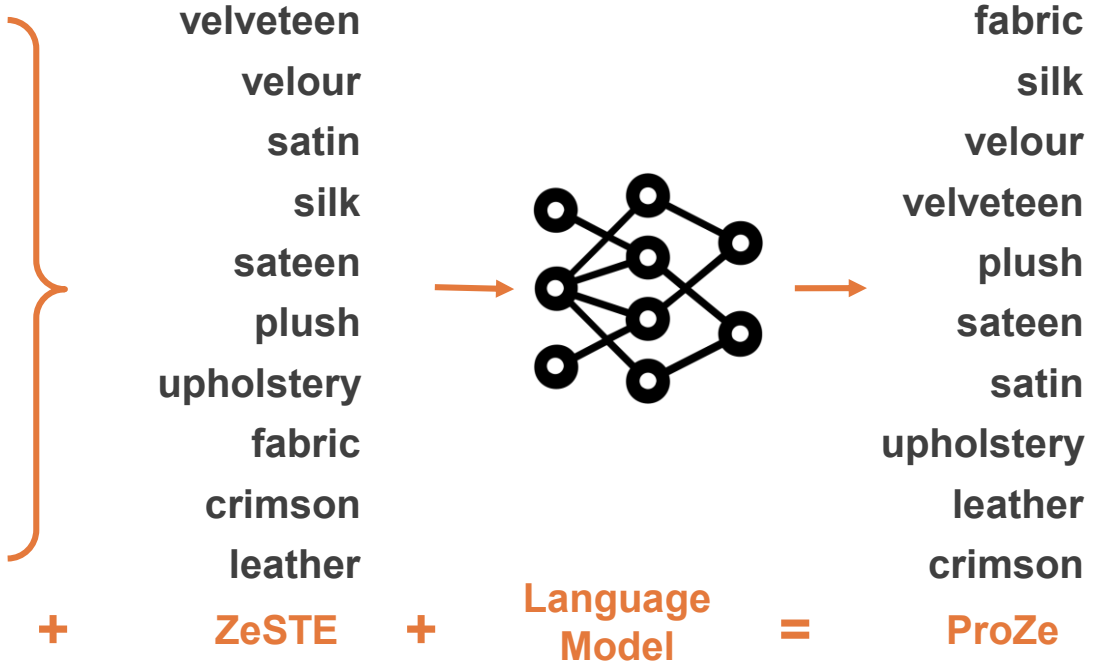
ProZe: Explainable and Prompt-guided Zero-Shot Text classification

Context

"Silk Textile"

Prompt

"[Velvet] is related to [candidate word]"



I. Harrando*, A. Reboud*, T. Schleider*, T. Ehrhart, R. Troncy

IEEE Computing special issue on Knowledge-Infused Learning for Computational Social Systems, 2022

Results

Classification	Silk Material (6 classes)	Silk Technique (7 classes)
ZeSTE	34.3%	46.9%
Entail (BART)	29.0%	64.0%
ProZe (ZeSTE + BART)	**39.0%**	**50.8%**

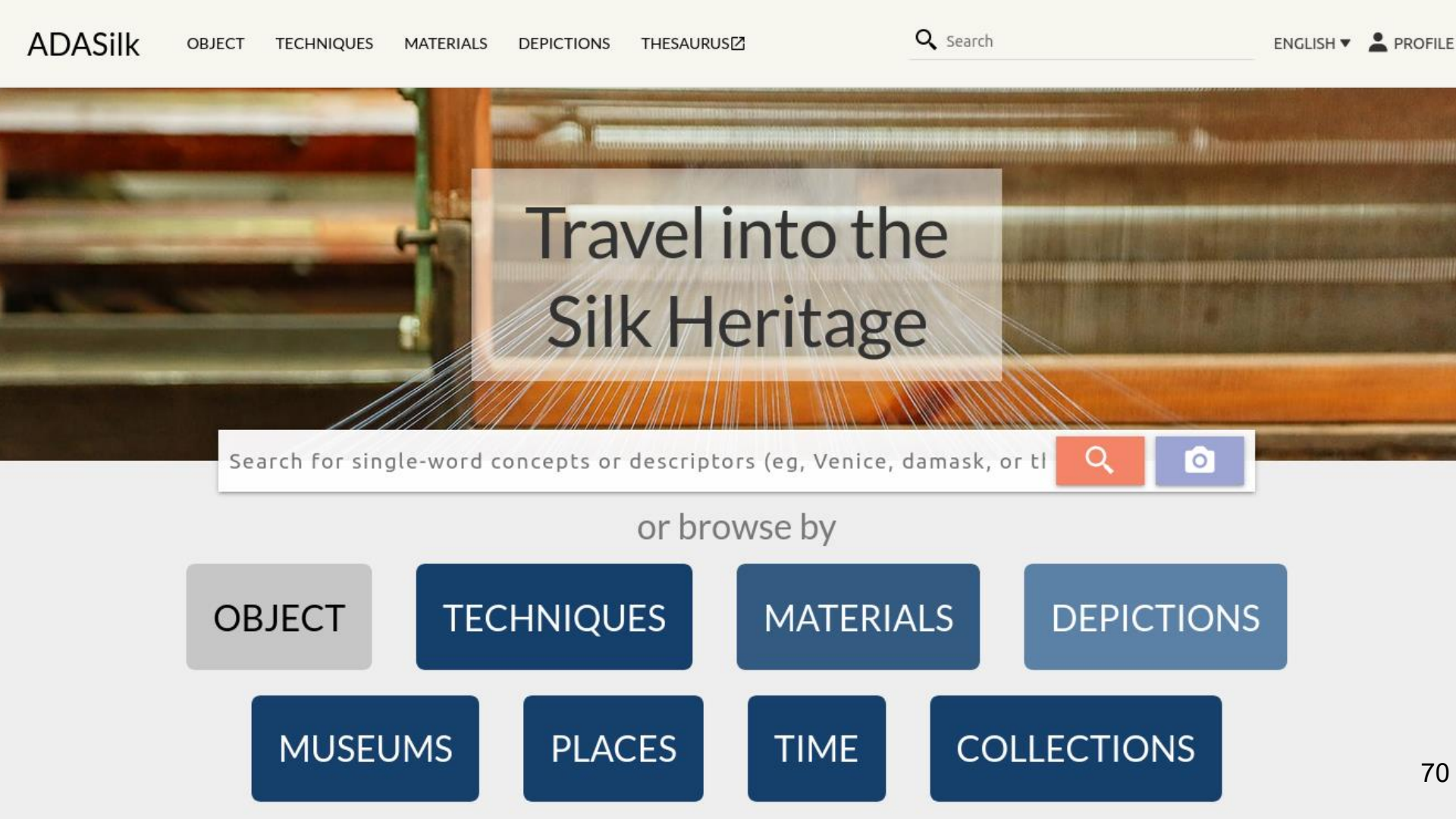
Performance comparison (accuracy) on domain-specific silk datasets

ConceptNet
Top 10 label neighborhood for
"Embroidery"

Embroidery, overstitch, running stitch, picot, stumpwork, arresene, couture, fancywork, embroider, berlin work

BART
Top 10 label neighborhood for
"Embroidery" + prompt "silk textile":

Craft artifact sewn, fabric, embroidery stitch, embroidery, detail, embroider, mending, embellishment, elaboration, filoselle


ADASilk
OBJECT
TECHNIQUES
MATERIALS
DEPICTIONS
THESAURUS
Search
ENGLISH
PROFILE
Travel into the Silk Heritage
Search for single-word concepts or descriptors (eg, Venice, damask, or tl
or browse by
OBJECT
TECHNIQUES
MATERIALS
DEPICTIONS
MUSEUMS
PLACES
TIME
COLLECTIONS

ADASilk OBJECTS TECHNIQUES MATERIALS DEPICTIONS THESAURUS Search ENGLISH PROFILE

CLEAR FILTER

Text search

Production time

Select...

Production place

Select...

Material

Select...

Technique

Select...

Depiction

Select...

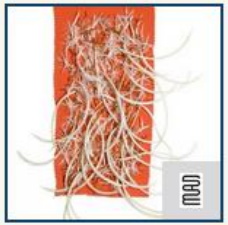
Type of object

37806 search results

Sort by Select...


Bordure pour écrans et paravents, destinés à la Salle d...

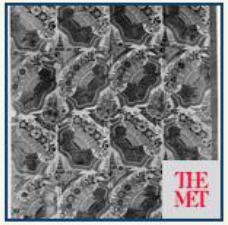

1838, Bond Street Underground Station


échantillon, Lesage, Paris, 1969, collection...

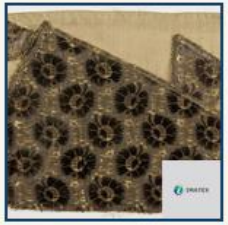

Bordure pour tenture du cabinet de repos de...


1880~, Varanasi


1870 - Myanmar


1701 / 1725, Venice


galon (obi), 1867


4143


1870 / 1899

Previous 1 2 3 4 5 6 ... 1890 1891 Next

What did we learn?

- **Many applications require more complex schemas (CIDOC-CRM, FRBR-based)**
- **KG construction from multimodal data**
- **Controlled vocabularies help partitioning the data for training classifiers and predicting missing values**
- **KGs enable explainable recommendations (entity2vec)**
- **The KG is hidden but powers explorative search interface**

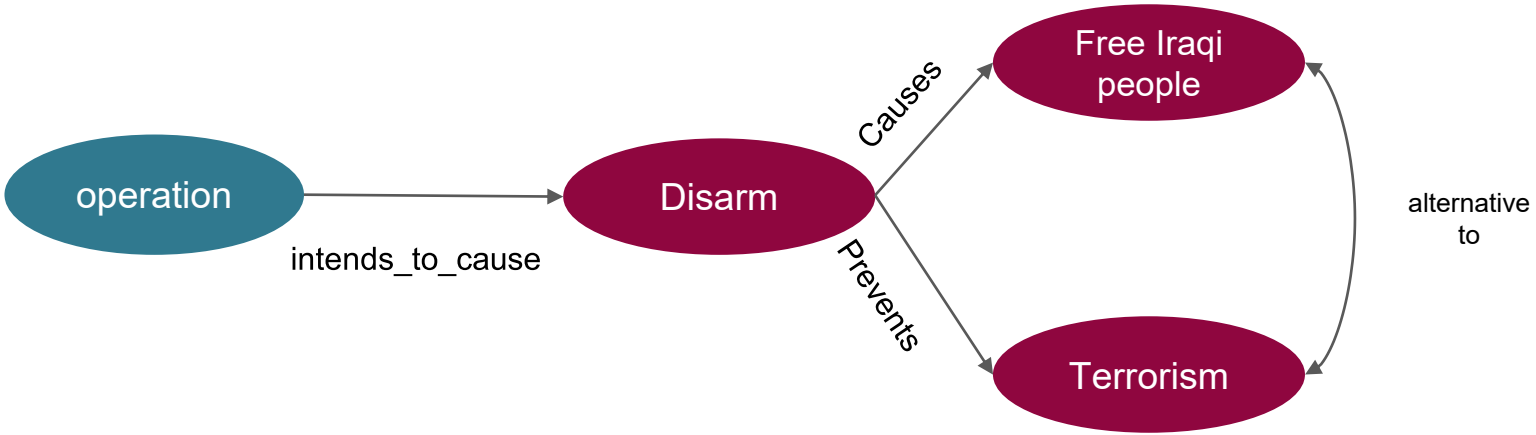
Unlocking Narratives

The role of Knowledge Graphs and AI in Story Understanding

EURECOM
Sophia Antipolis

Example

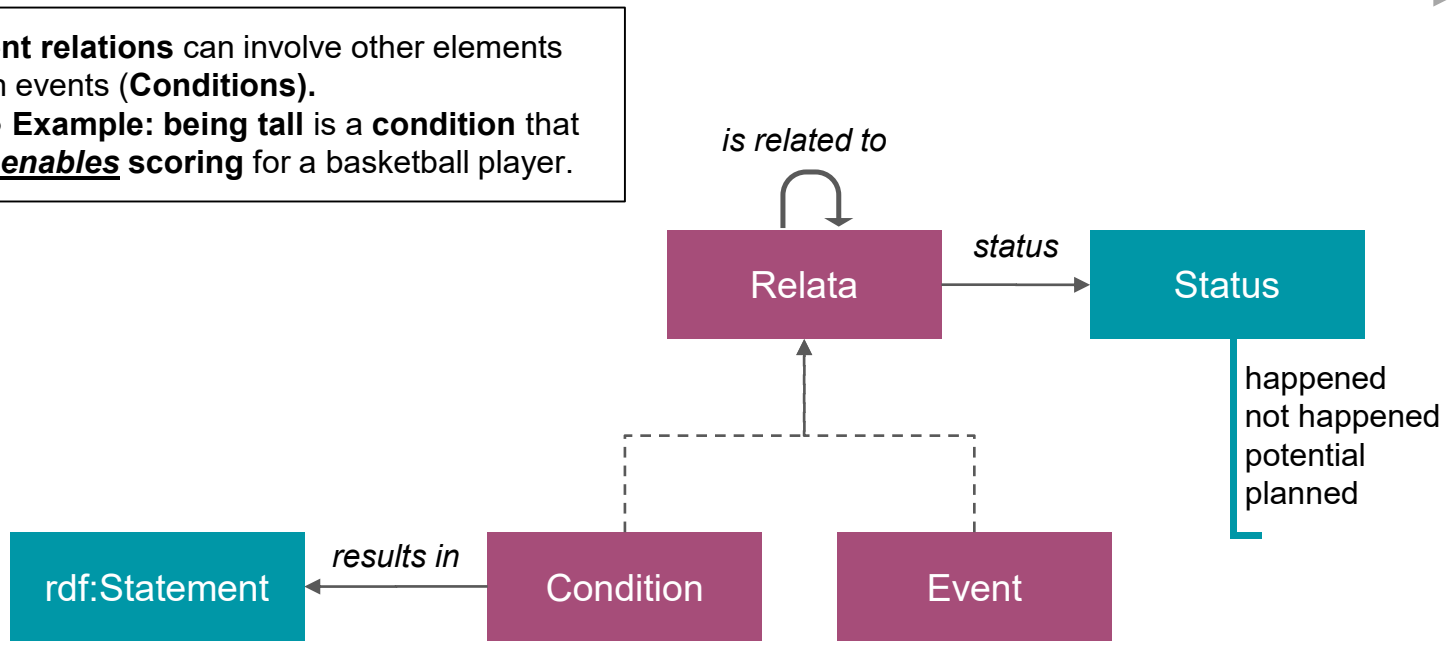
"As US claimed, the **intent** of the military operation was to disarm Iraq of weapons of mass distruction, to **end** support for terrorism and free iraqi people"



Disclaimer: this is the representation of the statement from the text, without judgement whether it is true or false.

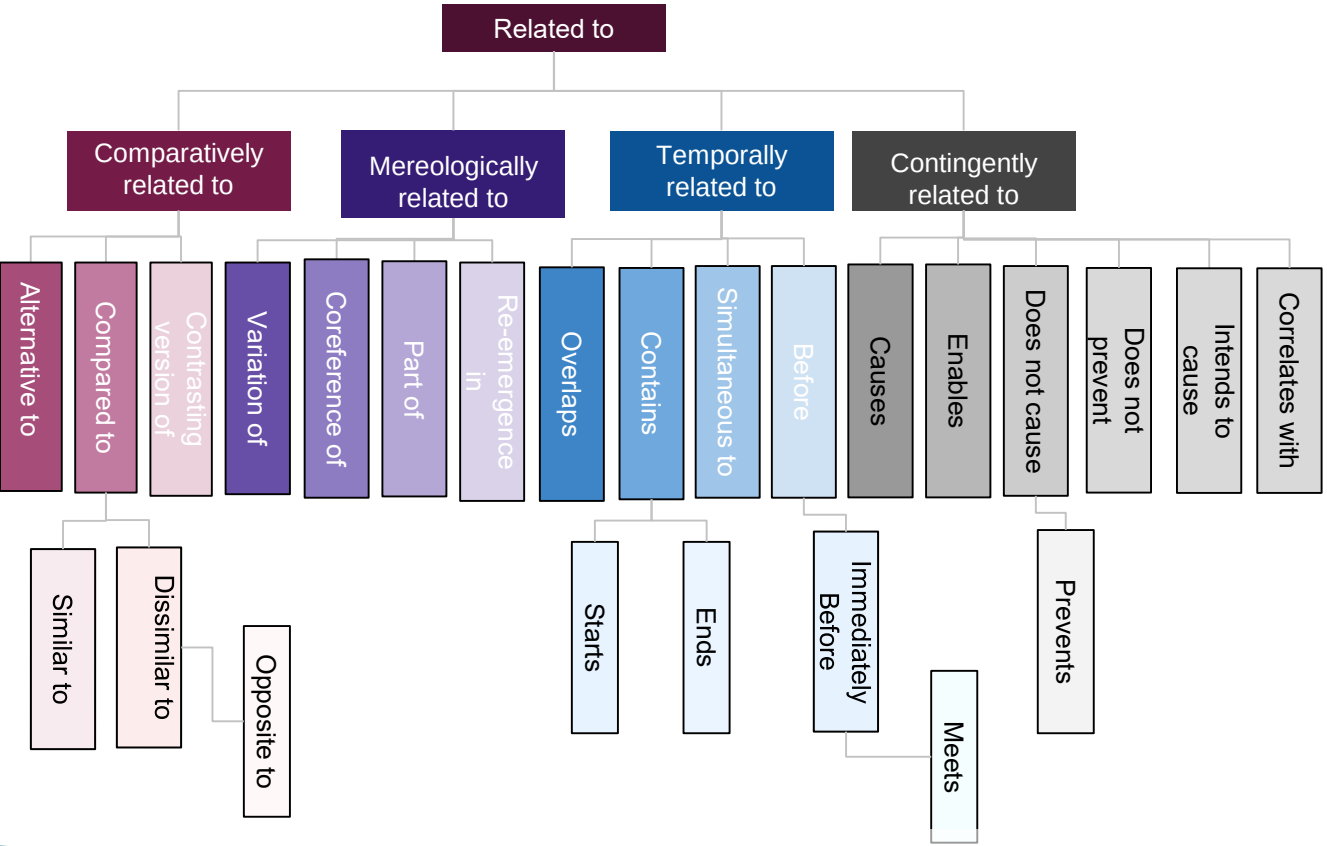
FARO: Representing Events

- **Event relations** can involve other elements than events (**Conditions).**
  - **Example: being tall** is a **condition** that ***enables*** **scoring** for a basketball player.



Y. Rebboud, P. Lisena, R. Troncy. Beyond Causality: Representing Event Relations in Knowledge Graphs. In Knowledge Engineering and Knowledge Management (EKAW), Bolzano, Italy, 2022,

FARO: Relations



- **FARO** : developed to be as much **complete** as possible.
- **Harmonizes** other data models
- **Enable reasoning**
  - **Hierarchical structure** of properties
  - **Logic constraints** (owl properties)

Y. Rebboud, P. Lisena, R. Troncy. Beyond Causality: Representing Event Relations in Knowledge Graphs. In Knowledge Engineering and Knowledge Management (EKAW), Bolzano, Italy, 2022,

Problem

- **Not existing dataset** with precise event relations
- **Our first attempt** resulted in **small** and **unbalanced** dataset.

- **Two data augmentation strategies**
  a. With GenAI
  b. With Common Sense

Target relations:
- Enabling
- Intention
- Prevention
- Direct causality

Prompt Based Data Augmentation with LLMs

Prompt(ERx) = definition(Event) + definition(ERx) + request(ER) + examples(ERx)

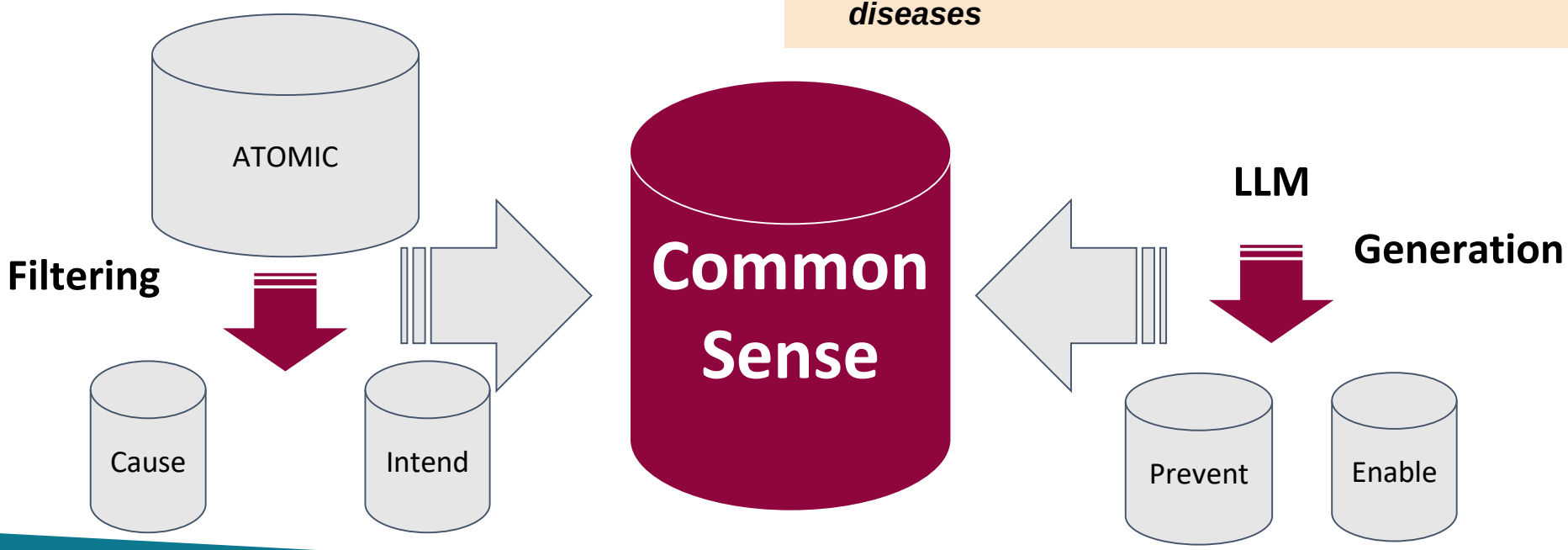
New dataset size: **2,000+ sentences**
Performance increment (F1):
Relation Classification **+27%**
Event Extraction: **+11%**

Y. Rebboud, P. Lisena, R. Troncy. Prompt-based Data Augmentation for Semantically-precise Event Relation Classification.
In Semantic Methods for Events and Stories Workshop (SEMMES), ESWC 2023, Heraklion, Greece, May 2023

Common Sense Data Augmentation

Example of Common Sense Data

exercising *regularly prevents* ***cardiovascular diseases***



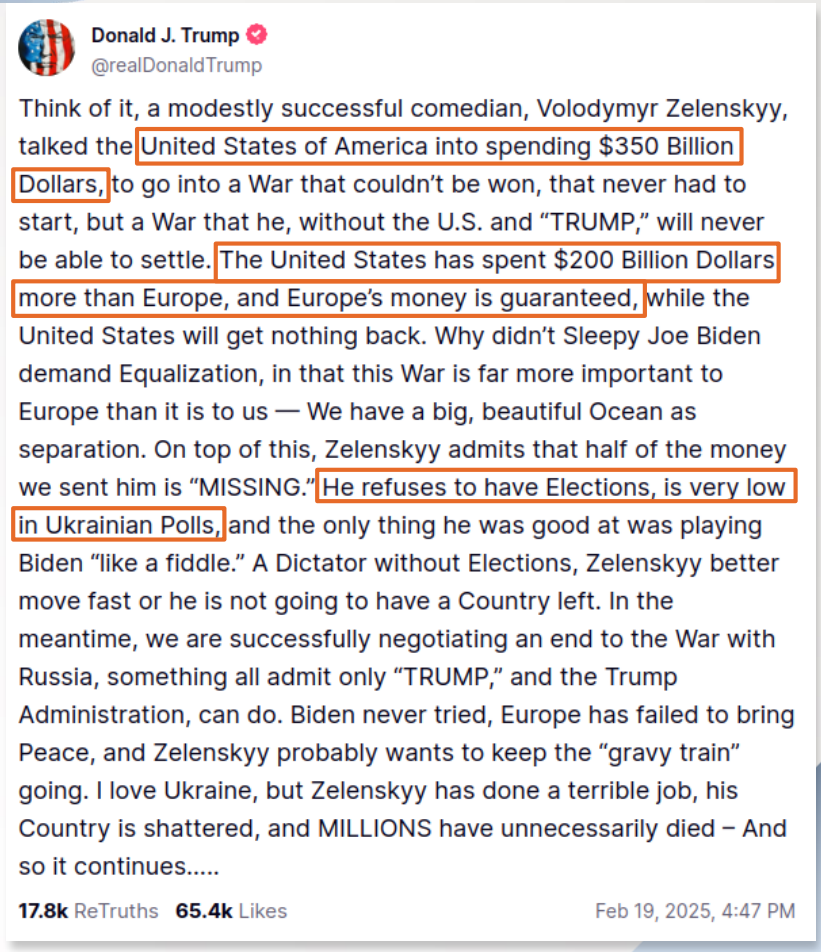
Three subtasks

Subtask	**Best performing model**	**F1 Score**	**LLM (GPT4o)**
Relation Detection *Is this sentence including a relation?*	RoBERTa-based end-to-end classifier	0.98	0.59
Relation Classification *Which relation type is in this sentence?*	RoBERTa-based end-to-end classifier	0.78	0.54
Event Extraction *What are the text token involved?*	REBEL end-to-end	0.70	0.45

Motivation

Misinformation has serious impact on societal topics

- **Nature**: Climate change denial, Australian bushfires, etc.
- **Health**: COVID “infodemic”, Vaccines, etc.
- **Politics**: US presidential elections, Brexit, Ukraine war, etc.



Donald J. Trump
@realDonaldTrump

Think of it, a modestly successful comedian, Volodymyr Zelenskyy, talked the United States of America into spending $350 Billion Dollars, to go into a War that couldn’t be won, that never had to start, but a War that he, without the U.S. and “TRUMP,” will never be able to settle. The United States has spent $200 Billion Dollars more than Europe, and Europe’s money is guaranteed, while the United States will get nothing back. Why didn’t Sleepy Joe Biden demand Equalization, in that this War is far more important to Europe than it is to us — We have a big, beautiful Ocean as separation. On top of this, Zelenskyy admits that half of the money we sent him is “MISSING.” He refuses to have Elections, is very low in Ukrainian Polls, and the only thing he was good at was playing Biden “like a fiddle.” A Dictator without Elections, Zelenskyy better move fast or he is not going to have a Country left. In the meantime, we are successfully negotiating an end to the War with Russia, something all admit only “TRUMP,” and the Trump Administration, can do. Biden never tried, Europe has failed to bring Peace, and Zelenskyy probably wants to keep the “gravy train” going. I love Ukraine, but Zelenskyy has done a terrible job, his Country is shattered, and MILLIONS have unnecessarily died – And so it continues.....

17.8k ReTruths **65.4k** Likes Feb 19, 2025, 4:47 PM

Misinformation-related factors

- **Persuasion techniques**
  - Push **agendas** and **narratives**
  - Leveraging **emotions**
  - Using **fallacies**

- **Conspiracy theories**
  - Malevolent **schemes**
  - Disproved or unproven **accusations**

Persuasion Techniques

SemEval-2024: Detection of persuasion techniques in memes

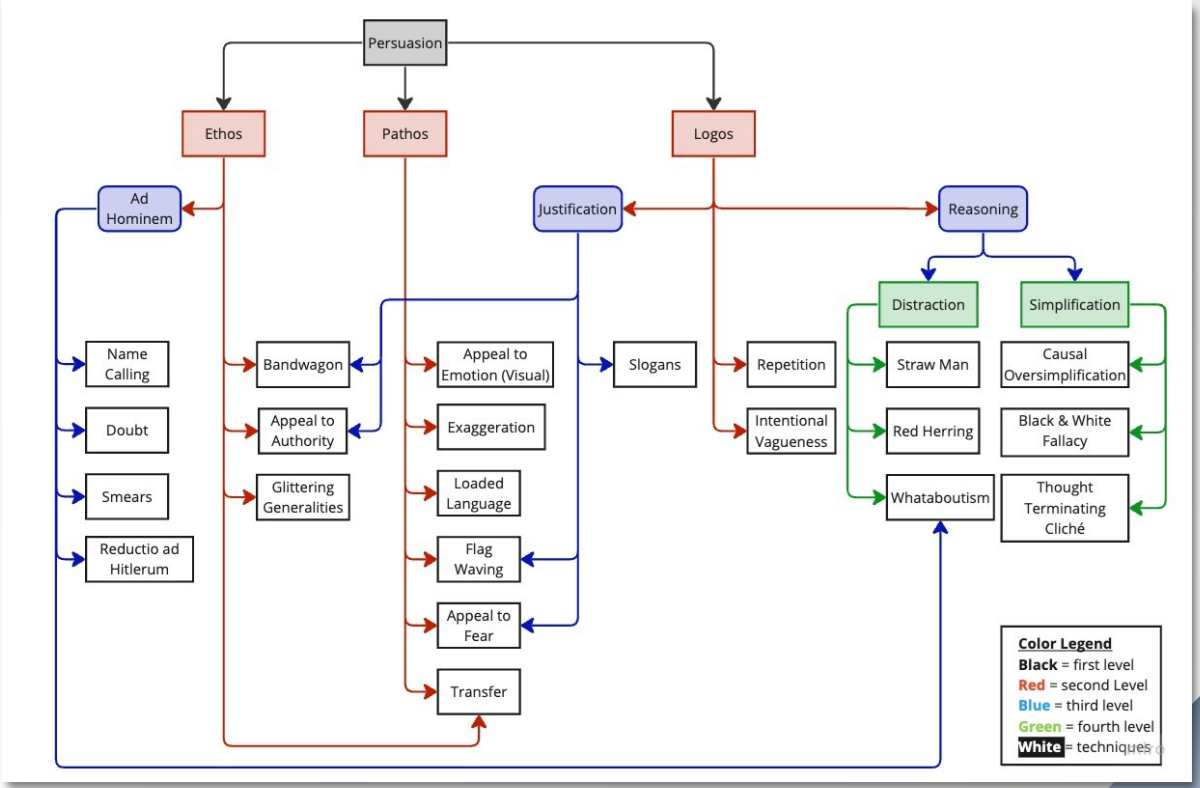
Text: "I HATE TRUMP\ n\ nMOST TERRORIST DO"

Techniques: ["Loaded Language", "Name calling/Labeling"]

SemEval 2024 - Data

- Train: 5000 memes
- Test: 1000 memes

20 different classes using a **hierarchical** structure



SemEval 2024 - Approach

- **Models:** BERT, BERT-HarMe, RoBERTa, AlBERT, DeBERTa, DistilBERT
- **Losses:** CE, BCE, Focal Loss, Hierarchical Loss (HL)
- **Data:** 2024, 2021, GPT-augmented
- **Output Classes:** 20, 28

$$\mathcal{L}^a_{BCE} = y^a \cdot log\sigma(\max(\{x^c\}_{c \in child(a)})) + (1 - y^a) \cdot log(1 - \sigma(\max(\{x^c\}_{c \in child(a)})))$$

$$\mathcal{L}_{HL} = \mathcal{L}_{BCE} + \alpha \cdot \sum_{a \in \mathcal{A}} \mathcal{L}^a_{BCE}$$

SemEval 2024 - Results

Our approach:
Ensembling of BERT-based models

"*EURECOM at SemEval-2024 Task 4: Hierarchical Loss and Model Ensembling in Detecting Persuasion Techniques*", Peskine Y., Troncy R., Papotti P., Proceedings of the 18th International Workshop on Semantic Evaluation

	Team	F1H
1	**914isthebest**	**0.752**
2	BCAmirs	0.699
3	OtterlyObsessedWithSemantics	0.697
4	TUMnlp	0.674
5	GreyBox	0.670
6	NLPNCHU	0.663
7	Puer	0.660
8	EURECOM	0.655
9	SuteAlbastre	0.652
10 / 33	UMUTeam	0.648

Code shared on github: https://github.com/D2KLab/semeval-2024-task-4

Conspiracy Theories

MediaEval: Detection of 9 COVID-related conspiracy theories in tweets

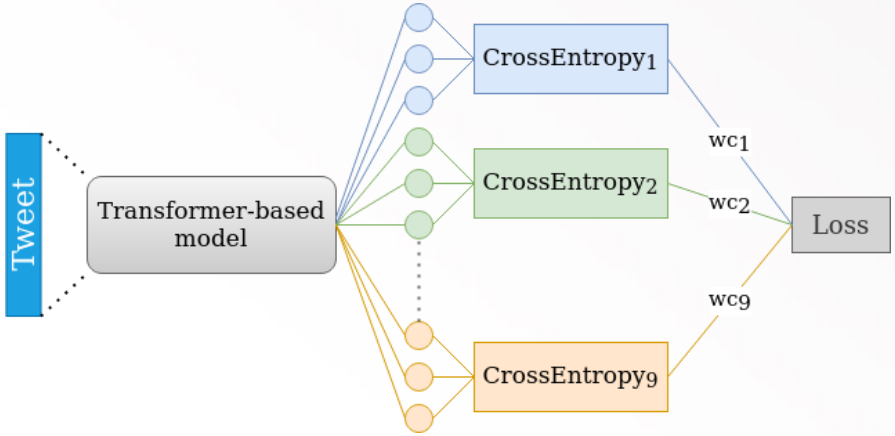
Suppressed Cures, Antivax, Fake virus, etc.

You see this #CoronaVirus outbreak, I believe it was an intentional Bio-weapon released by these sick/mad scientists and doomsday ambassadors to test their population control theory…

MediaEval 2021 - Results

Our winning approach:

Ensembling of Covid-Twitter-BERT models



"*Detecting COVID-19-Related Conspiracy Theories in Tweets*", Peskine Y., Alfarano G., Harrando I., Papotti P., Troncy R., Multimedia Benchmark Workshop (MediaEval 2021)

	Team	MCC
1	**D2KLab**	**0.733**
2	SELAB_HCMUS	0.648
3	Deltamap	0.632
4	SELAB-HMUS-Junior	0.599
5	Upsilon Labs	0.454
6	MG-UCB	0.450
7	FakeINA	0.446
8	OTS-UEC	0.413
9/17	Delta_IITKGP	0.347

Code shared on github: https://github.com/D2KLab/mediaeval-fakenews

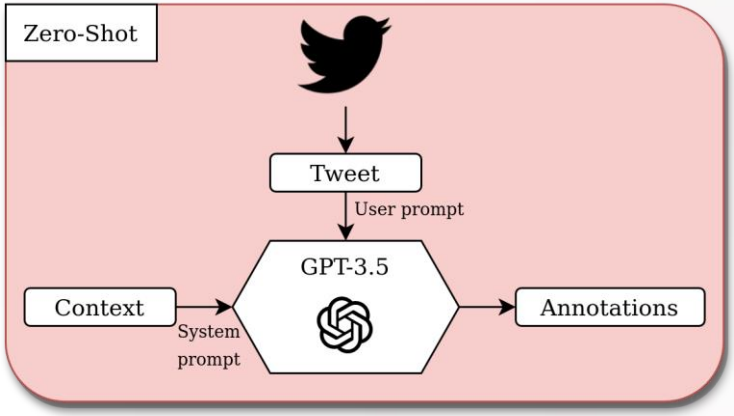
Definitions Matters

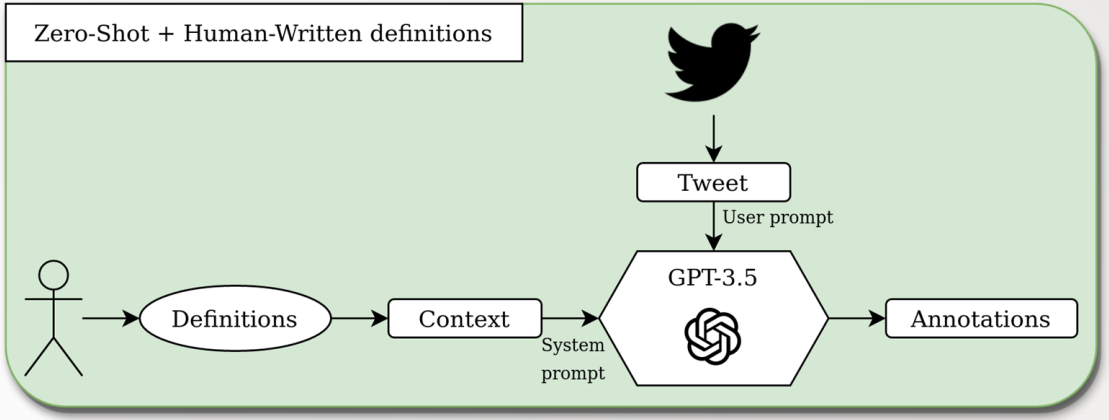
Can LLMs detect conspiracy theories?

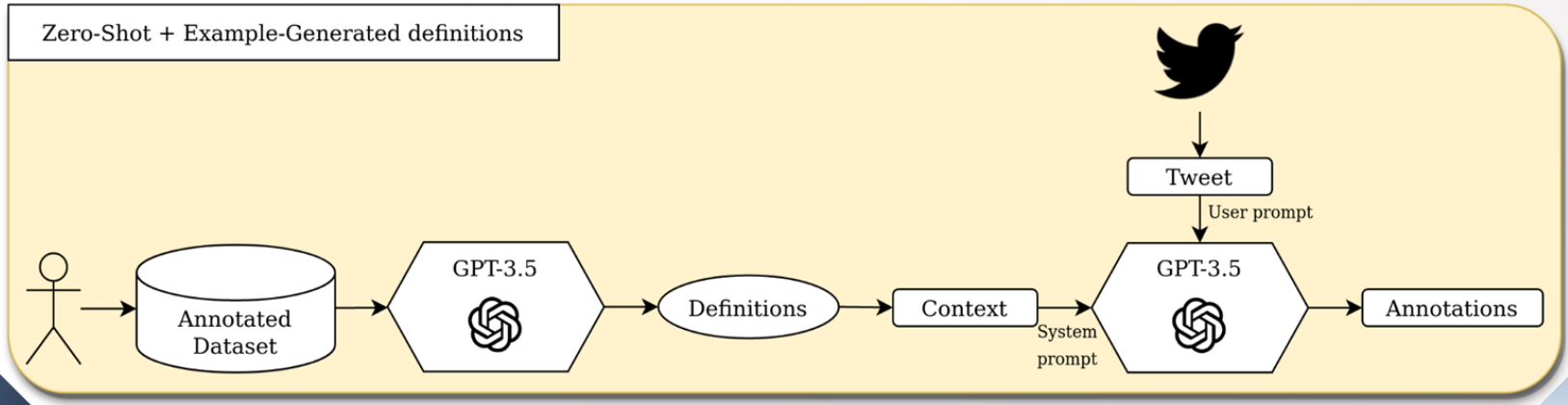
- Experiment with **GPT-3.5**
- **Zero-shot** settings
- Explore impact of different **definitions** of classes

“Definitions matter: Guiding GPT for multi-label classification”, Peskine Y., Korenčić D., Grubisic I., Papotti P., Troncy R., Rosso P., Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

Definitions Matters



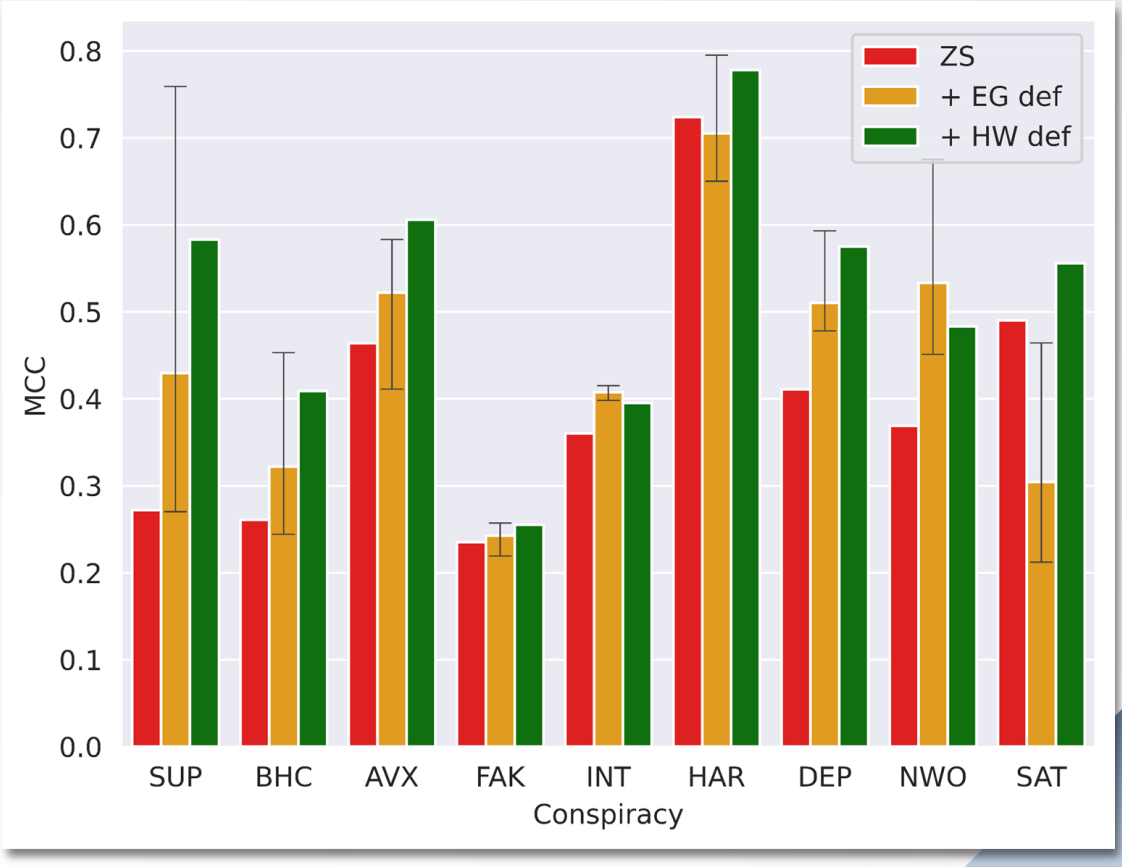




Classification Results

- **Definitions** improve classification performance
- **Human-Written** (HW) perform best on average
- Best **Example Generated** (EG) definitions outperform their HW counterparts

Suppressed Cures (SUP), Behaviour Control (BHC), Anti-Vaccination (AVX), Fake Virus (FAK), Intentional Pandemic (INT), Harmful Radiation (HAR), Depopulation (DEP), New World Order (NWO), Satanism (SAT)



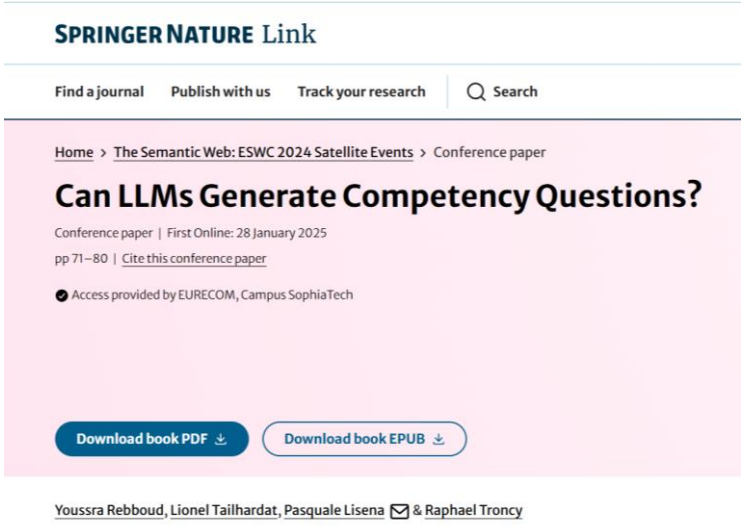
Code shared on github: https://github.com/dkorenci/gpt-def-zeroshot

What did we learn?

- **LLMs can extract information from text but under-perform for Sequence Labeling task (regardless of your prompt engineering effort)**
- **Fine-Tuning SLMs for specialized tasks is effective (ensemble!)**
- **LLMs are good at generating synthetic data to fill-in the blanks (class imbalance, data distribution)**
- **Common Sense Knowledge Graphs are great seeds for LLMs to generate useful data**

What’s next?

LLM for Knowledge Engineering



SPRINGER NATURE Link

Find a journal | Publish with us | Track your research | Search

Home > The Semantic Web: ESWC 2024 Satellite Events > Conference paper

Can LLMs Generate Competency Questions?

Conference paper | First Online: 28 January 2025

pp 71–80 | Cite this conference paper

Access provided by EURECOM, Campus SophiaTech

Download book PDF | Download book EPUB

Youssra Rebboud, Lionel Tailhardat, Pasquale Lisena & Raphael Troncy

https://doi.org/10.1007/978-3-031-78952-6_7

Benchmarking LLM-based Ontology Conceptualization: A Proposal

Youssra Rebboud[4,*], Pasquale Lisena[1], Lionel Tailhardat[1,2] and Raphael Troncy[1]

[1]EURECOM, Sophia Antipolis, France
[2]Orange, Paris, France

Abstract

This study presents a benchmark proposal designed to enhance knowledge engineering tasks through the use of large language models (LLMs). As LLMs become increasingly pivotal in knowledge extraction and modeling, it is crucial to evaluate and improve their performance. Building on prior work aiming at reverse generating competency questions (CQs) from existing ontologies, we introduce a benchmark focused on specific knowledge modeling tasks including ontology documentation, ontology generation, and query generation. In addition, we propose a baseline evaluation framework that applies various techniques, such as semantic comparison, ontology evaluation criteria, and structural comparison, using both existing ground truth datasets and newly proposed ontologies with corresponding CQs and documentation. This rigorous evaluation aims to provide a deeper understanding of LLM capabilities and contribute to their optimization in knowledge engineering applications.

Keywords

Benchmark Proposal, Knowledge Engineering, Knowledge Representation, Large Language Models

1. Introduction

The knowledge engineering and semantic web communities are increasingly experimenting with Large Language Models (LLMs) to build ontologies and knowledge graphs. Key tasks being explored include: creating views on heterogeneous data lakes [1], RDF triples and SPARQL query generation [2], named entity recognition and relation extraction [3], RML mappings creation [4] or schema and ontology matching [5, 6, 7]. Hence, we observe that the various stages of the knowledge engineering process are revisited in the era of LLMs (e.g. LOT [8]). However, their systematic usage need to be further assessed as the results greatly vary depending on the underlying LLM being used and other factors.

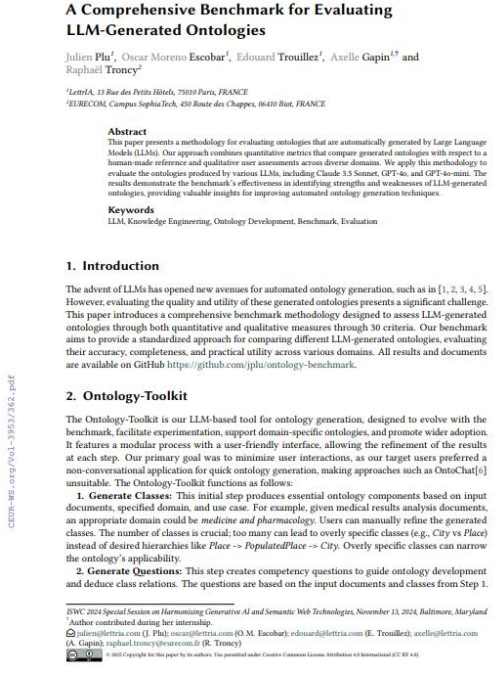
In previous work [9], we have evaluated six LLMs using zero- and few-shot approaches with three prompting strategies, inputting either classes alone, classes with properties, or a schema summary.[1] These configurations were tested across five ontologies to assess the LLMs' ability to reverse generate Competency Questions (CQs). These ontologies were precisely selected because expert made competency questions having lead to their conceptualization were made available. We observed that while providing competency question examples generally improved performance for this task, in some cases, adding more detailed information from certain ontologies unexpectedly reduced LLM effectiveness. This highlights the need to further investigate the characteristics of the ontologies that impact the accuracy of LLM responses and vice versa.

In this paper, we propose to develop a benchmark to systematically compare the performance of LLMs for knowledge engineering tasks, specifically focusing on the stages of specification, conceptualization, and validation of an ontology. The core of our proposal is to leverage ontologies that have been published alongside a set of CQs and have been evaluated through the corresponding authoring tests expressed in SPARQL. The goal of the LLM will typically be to understand user intents expressed in natural language

ISWC 2024 Special Session on Harmonising Generative AI and Semantic Web Technologies, November 13, 2024, Baltimore, Maryland
*Corresponding author.
youssra.rebboud@eurecom.fr (Y. Rebboud); pasquale.lisena@eurecom.fr (P. Lisena); lionel.Tailhardat@orange.com (L. Tailhardat); raphael.troncy@eurecom.fr (R. Troncy)
0000-0003-3507-5646 (Y. Rebboud); 0000-0003-3094-5585 (P. Lisena); 0000-0001-5887-899X (L. Tailhardat); 0000-0003-0457-1436 (R. Troncy)

[1]The code is available at https://github.com/D2KLab/llm4ke

https://ceur-ws.org/Vol-3953/365.pdf



A Comprehensive Benchmark for Evaluating LLM-Generated Ontologies

Julien Plu[1], Oscar Moreno Escobar[1], Edouard Trouillez[1], Axelle Gapin[1,*] and Raphaël Troncy[2]

[1]LettrIA, 13 Rue des Petits Hôtels, 75010 Paris, FRANCE
[2]EURECOM, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, FRANCE

Abstract

This paper presents a methodology for evaluating ontologies that are automatically generated by Large Language Models (LLMs). Our approach combines quantitative metrics that compare generated ontologies with respect to a human-made reference and qualitative user assessments across diverse domains. We apply this methodology to evaluate the ontologies produced by various LLMs, including Claude 3.5 Sonnet, GPT-4o, and GPT-4o-mini. The results demonstrate the benchmark's effectiveness in identifying strengths and weaknesses of LLM-generated ontologies, providing valuable insights for improving automated ontology generation techniques.

Keywords

LLM, Knowledge Engineering, Ontology Development, Benchmark, Evaluation

1. Introduction

The advent of LLMs has opened new avenues for automated ontology generation, such as in [1, 2, 3, 4, 5]. However, evaluating the quality and utility of these generated ontologies presents a significant challenge. This paper introduces a comprehensive benchmark methodology designed to assess LLM-generated ontologies through both quantitative and qualitative measures through 30 criteria. Our benchmark aims to provide a standardized approach for comparing different LLM-generated ontologies, evaluating their accuracy, completeness, and practical utility across various domains. All results and documents are available on GitHub https://github.com/jplu/ontology-benchmark.

2. Ontology-Toolkit

The Ontology-Toolkit is our LLM-based tool for ontology generation, designed to evolve with the benchmark, facilitate experimentation, support domain-specific ontologies, and promote wider adoption. It features a modular process with a user-friendly interface, allowing the refinement of the results at each step. Our primary goal was to minimize user interactions, as our target users preferred a non-conversational application for quick ontology generation, making approaches such as OntoChat[6] unsuitable. The Ontology-Toolkit functions as follows:

1. Generate Classes: This initial step produces essential ontology components based on input documents, specified domain, and use case. For example, given medical results analysis documents, an appropriate domain could be *medicine and pharmacology*. Users can manually refine the generated classes. The number of classes is crucial; too many can lead to overly specific classes (e.g., *City* vs *Place*) instead of desired hierarchies like *Place -> PopulatedPlace -> City*. Overly specific classes can narrow the ontology's applicability.

2. Generate Questions: This step creates competency questions to guide ontology development and deduce class relations. The questions are based on the input documents and classes from Step 1.

ISWC 2024 Special Session on Harmonising Generative AI and Semantic Web Technologies, November 13, 2024, Baltimore, Maryland
*Author contributed during her internship.
julien@lettria.com (J. Plu); oscar@lettria.com (O.M. Escobar); edouard@lettria.com (E. Trouillez); axelle@lettria.com (A. Gapin); raphael.troncy@eurecom.fr (R. Troncy)

https://ceur-ws.org/Vol-3953/362.pdf

EURECOM Sophia Antipolis

LLMs are starving for KGs

- **List of crawlers visiting our KGs (1 month):**
  - OpenAI (OpenAI-Searc, ChatGPT-User, **GPTBot**)
  - Bytedance (Bytespider)
  - Apple (Applebot)
  - Meta AI (MetaExternal)
  - Anthropic (**ClaudeBot**)
  - Microsoft (BingBot, LinkedInBot)
  - DuckDuckGo (DuckAssist)
  - CommonCrawl (CCBot)
  - Amazon (**Amazonbot**)
  - Perplexity (Perplexity-User, Perplexitybot)
- **Visiting mostly:** /describe/, /fct/, /sparql /resource

data.doremus.org: 3,430,585 requests

memad.eurecom.fr: 1,831,538 requests

data.cimple.eu: 1,280,429 requests

explorer.cimple.eu: 676,140 requests

vocab.odeuropa.eu: 303,161 requests

asrael.eurecom.fr: 241,378 requests

data.silknow.org: 103,334 requests

skosmos.silknow.org: 68,660 requests

ada.silknow.org: 12,402 requests

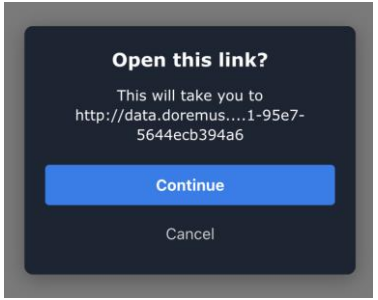
data.odeuropa.eu: 9,688 requests

std.eurecom.fr: 217 requests

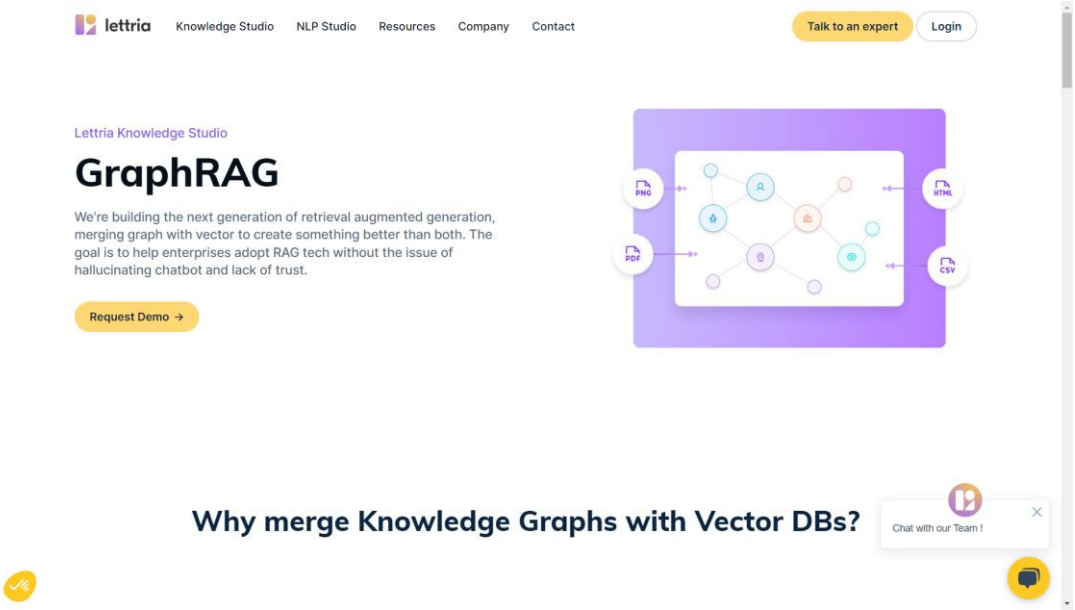
data.edf.eurecom.fr: 117 requests

yang.eurecom.fr: 48 requests

sirene.eurecom.fr: 17 requests



Graph RAG and Path RAG



https://www.lettria.com/features/graphrag

From Local to Global: A Graph RAG Approach to Query-Focused Summarization

Darren Edge[1†] Ha Trinh[1†] Newman Cheng[2] Joshua Bradley[2] Alex Chao[3]
Apurva Mody[3] Steven Truitt[2]
Jonathan Larson[1]

[1]Microsoft Research
[2]Microsoft Strategic Missions and Technologies
[3]Microsoft Office of the CTO

{daedge,trinhha,newmancheng,joshbradley,achao,moapurva,steventruitt,jolarso}
@microsoft.com

[†]These authors contributed equally to this work

Abstract

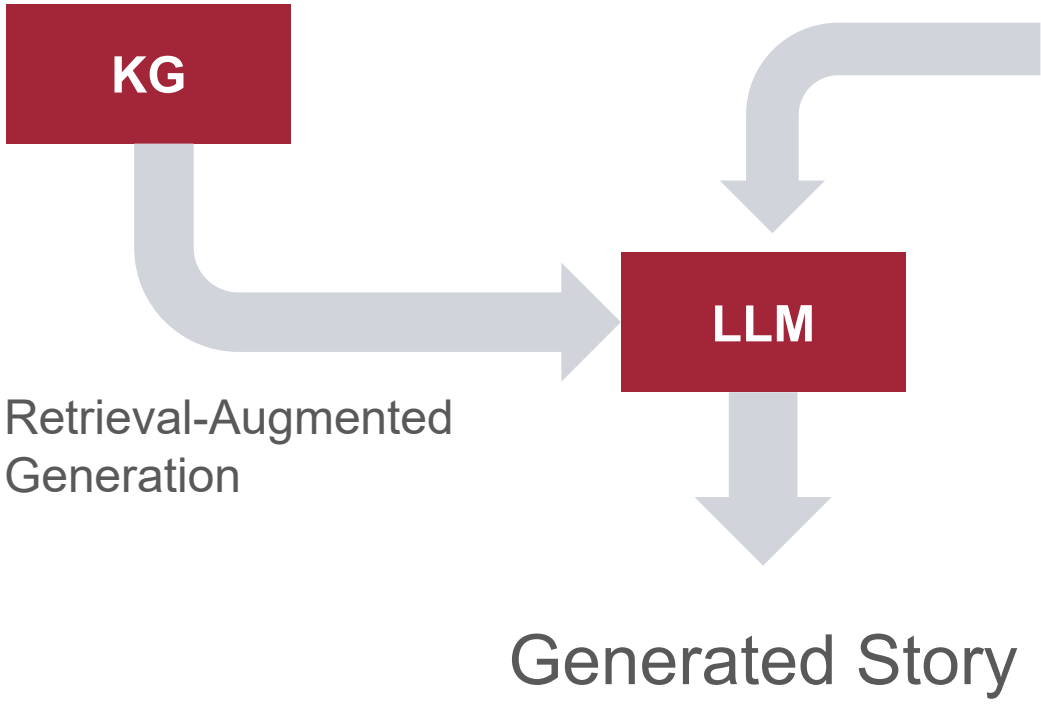
The use of retrieval-augmented generation (RAG) to retrieve relevant information from an external knowledge source enables large language models (LLMs) to answer questions over private and/or previously unseen document collections. However, RAG fails on global questions directed at an entire text corpus, such as "What are the main themes in the dataset?", since this is inherently a query-focused summarization (QFS) task, rather than an explicit retrieval task. Prior QFS methods, meanwhile, fail to scale to the quantities of text indexed by typical RAG systems. To combine the strengths of these contrasting methods, we propose a Graph RAG approach to question answering over private text corpora that scales with both the generality of user questions and the quantity of source text to be indexed. Our approach uses an LLM to build a graph-based text index in two stages: first to derive an entity knowledge graph from the source documents, then to pre-generate community summaries for all groups of closely-related entities. Given a question, each community summary is used to generate a partial response, before all partial responses are again summarized in a final response to the user. For a class of global sensemaking questions over datasets in the 1 million token range, we show that Graph RAG leads to substantial improvements over a naïve RAG baseline for both the comprehensiveness and diversity of generated answers. An open-source, Python-based implementation of both global and local Graph RAG approaches is forthcoming at https://aka.ms/graphrag.

1 Introduction

Human endeavors across a range of domains rely on our ability to read and reason about large collections of documents, often reaching conclusions that go beyond anything stated in the source texts themselves. With the emergence of large language models (LLMs), we are already witnessing attempts to automate human-like sensemaking in complex domains like scientific discovery (Microsoft, 2023) and intelligence analysis (Ranade and Joshi, 2023), where sensemaking is defined as

Preprint. Under review.

Interactive storytelling



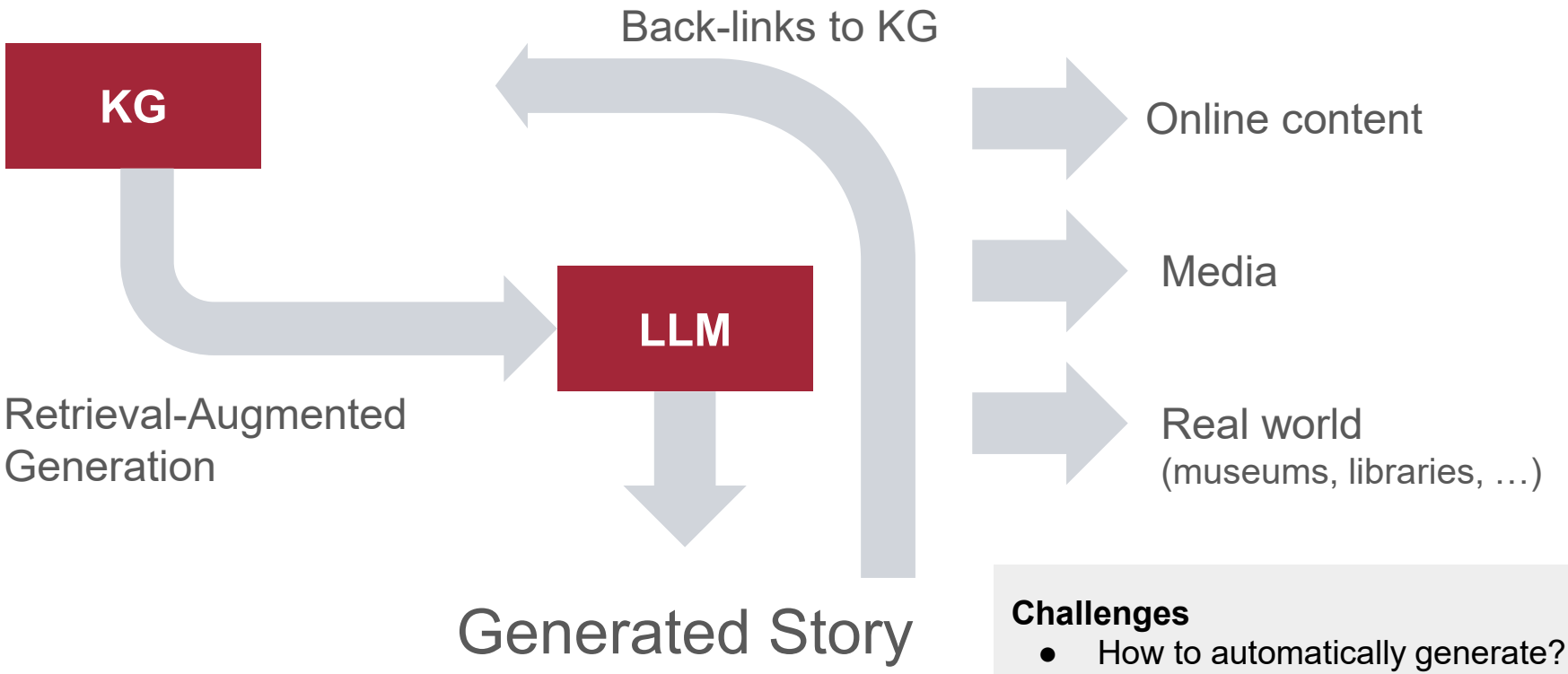
HUMAN-COMPUTER INTERACTION

Goal of the narrative
- Education
- Memorisation

Personalisation on the storyteller
- Style
- Metaphors

Personalisation on the Listener
- Cultural gap
- Generational gap
- Preferences

Enriched storytelling



Challenges

- How to automatically generate?
- How to guide and control this generation?

Let’s discuss!

Credits: Pasquale Lisena, Thibault Ehrhart, Giuseppe Rizzo, Vuk Milicic, Ahmad Assaf, Alison Reboud, Amine Dadoun, Benjamin Klotz, Emilie Palagi, Enrico Palumbo, Fanfu Wei, Florent Moriconi, Ghislain Atemezing, Houda Khrouf, Ismail Harrando, Jixiong Liu, José Luis Redondo Garcia, Julien Plu, Lionel Tailhardat, Thomas Schleider, Youri Peskine, Youssra Rebboud