Exploration of Paris

Knowledge and Data Integration 2021-2022

Giuliano Andronic :	Project Leader Data Scientist
Diane Willaime :	Domain Expert Knowledge Engineer

SUMMARY

1 - Inception

2 - Informal modeling

3 - Formal modeling

4 - Data integration

5 - Conclusion

1 - INCEPTION

Defining a purpose and collecting data

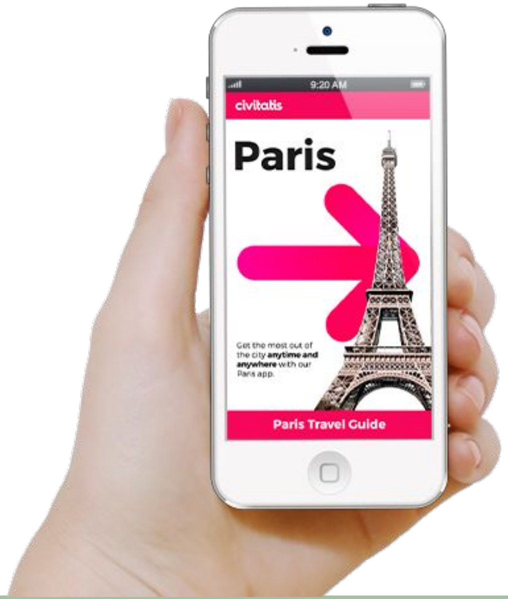
PURPOSE - Domain of Interest

- City of Paris : Activities, points of interest, …
- Targeted users : Tourists and Paris citizens

Example: You have 4 hours to spare in Paris and you don't know what to do

Ideal Goal

Input : position of the user, interests (art, music, …)
Output : list of events/activities to do


9:20 AM
civitatis
Paris
Get the most out of the city anytime and anywhere with our Paris app.
Paris Travel Guide

PURPOSE - personas and scenarios

Defining different personas and scenarios helped us to properly understand what was the kind of data we needed to satisfy the needs of the users.

Martina

She's planning a school trip to Paris for her students

Lilian

He will spend one weekend in Paris with his girlfriend

Marie

Erasmus student in Paris who loves shopping

RESOURCES

OpenData Paris : collects information about places, public transports, events and the evolution of the city

→ most of the data required for our DI project

OpenStreetMaps : data for the localization and extra places

PARIS | Data

OpenStreetMap

COMPETENCY QUESTIONS

Creation of a set of CQs representing our functional requirements.

Martina: Does a given bus stop offer shelter for me and my students? Is it available for wheelchairs?

COMMON: place, logistic information
CORE: bus stop
CONTEXTUAL: shelter, wheelchair access

COMPETENCY QUESTIONS

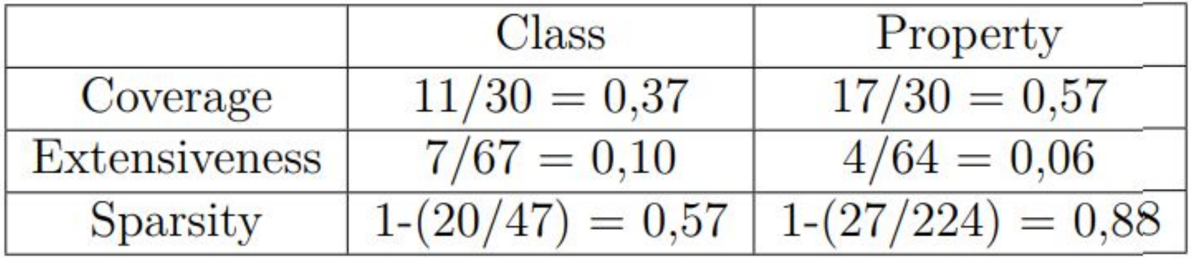
Other examples :

Tommaso: Are there any food-related activities during my stay ?

Clark : What is the nearest Coffee Shop with free wi-fi?

Lilian : Are there any romantic parks around the Eiffel Tower open at night ?

EVALUATION



	Class	Property
Coverage	11/30 = 0,37	17/30 = 0,57
Extensiveness	7/67 = 0,10	4/64 = 0,06
Sparsity	1-(20/47) = 0,57	1-(27/224) = 0,88

Assess the “fitness of use of our model”

Schema level : a set of CQs VS several collected ontologies (Coverage and Extensiveness)

Data level : a set of CQs VS several collected datasets (Sparsity)

2 - INFORMAL MODELING

Selecting relevant data and ER modeling

ETYPES

Extraction of the various eTypes that will be required during the DI process, alongside their object and data properties

Martina: Does a given bus stop offer shelter for me and my students? And is it available for wheelchairs?

ETYPE: bus station
DATA PROPERTIES : shelter, wheelchair access
OBJECT PROPERTIES : location

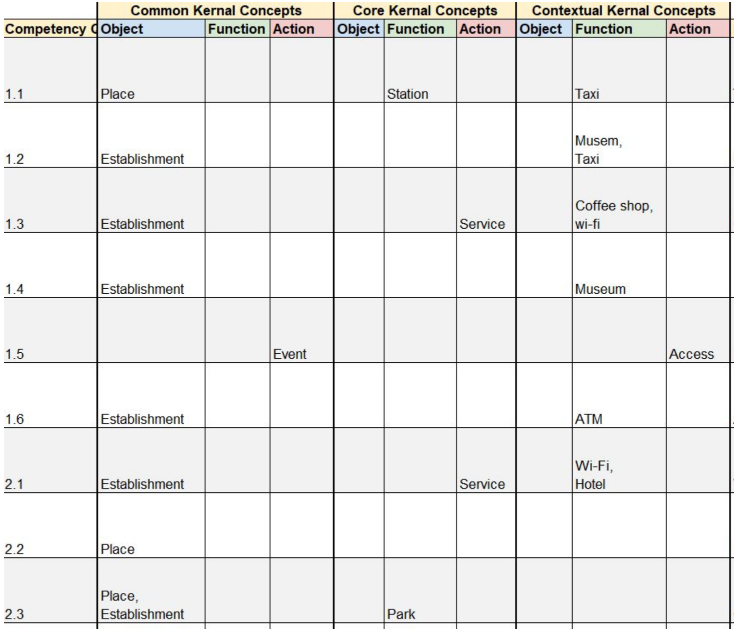
ER-MODEL - Modeling sheet

Our data must be aligned with the Foundational Teleology (reusability) :

→ Modeling Sheet

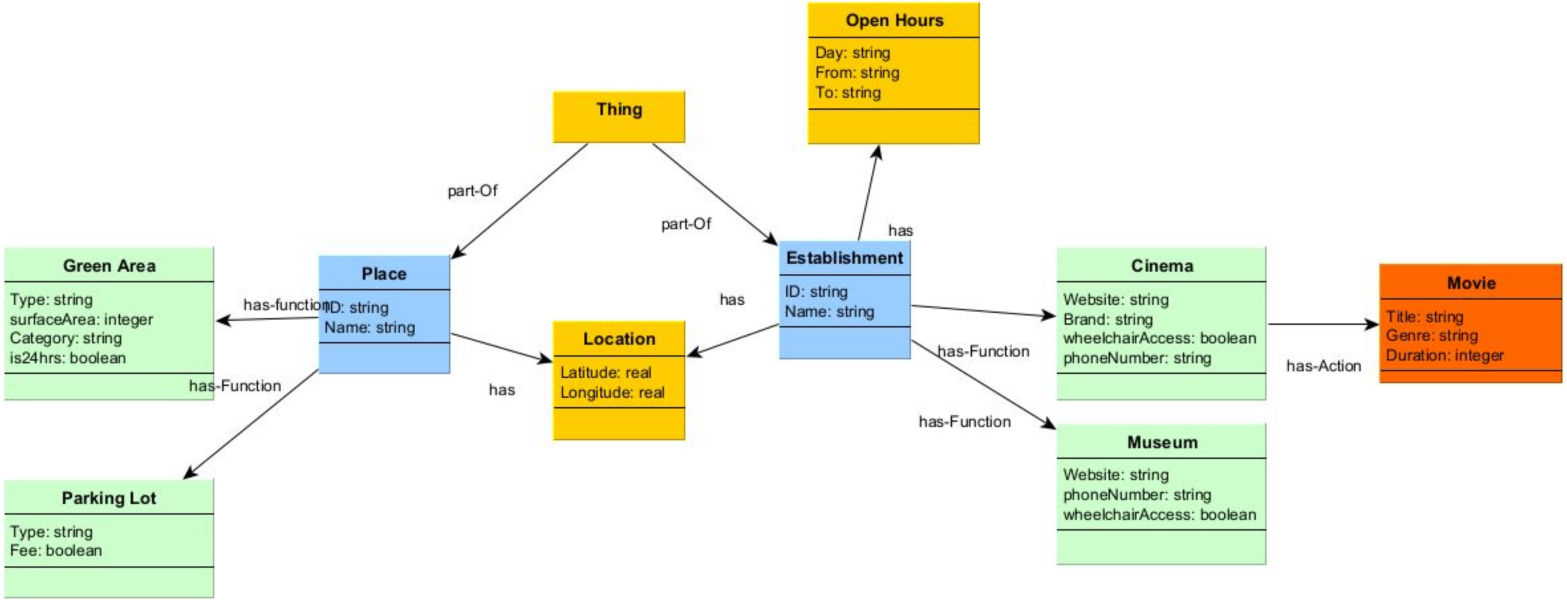
→ Grouping the kernel concepts into Objects, Functions and Activities.

The final MS was then used for the creation of the informal ER model

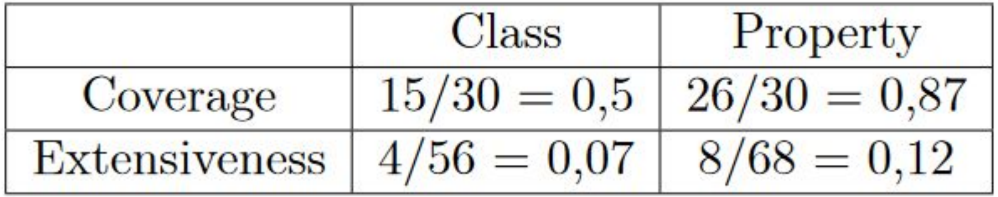


	Common Kernal Concepts			Core Kernal Concepts			Contextual Kernal Concepts		
Competency Q	Object	Function	Action	Object	Function	Action	Object	Function	Action
1.1	Place				Station			Taxi	
1.2	Establishment							Musem, Taxi	
1.3	Establishment					Service		Coffee shop, wi-fi	
1.4	Establishment							Museum	
1.5			Event						Access
1.6	Establishment							ATM	
2.1	Establishment					Service		Wi-Fi, Hotel	
2.2	Place								
2.3	Place, Establishment				Park				

ER-MODEL - partial



EVALUATION

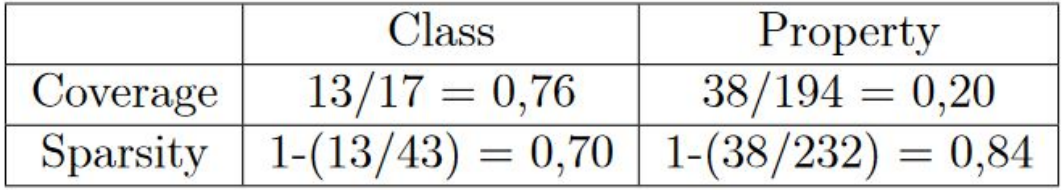


	Class	Property
Coverage	$15/30 = 0{,}5$	$26/30 = 0{,}87$
Extensiveness	$4/56 = 0{,}07$	$8/68 = 0{,}12$

Schema level : proposed informal ER model VS a set of CQs

- Coverage : if the ER model covers CQs
- Extensiveness : if the ER model properly extends CQs

EVALUATION



	Class	Property
Coverage	13/17 = 0,76	38/194 = 0,20
Sparsity	1-(13/43) = 0,70	1-(38/232) = 0,84

Data level : proposed informal ER model VS several collected datasets

- Coverage : if the ER model aligns with collected datasets
- Sparsity : if the ER model is much different from collected datasets

3 - FORMAL MODELING

Generating a shareable ETG and handling syntactic heterogeneity

ETG GENERATION

Ontology selection : to improve the reusability of our solution

Language alignment : we searched every single term used in our informal ER in the UKC Knowledge base. Most of the terms has a synonymous match, an ad-hoc definition was created for the remaining part

→ tool : KOS platform

Schema alignment : we had to align our model over the Foundational Teleology. This step requires a lot of focus due to the large amount of entities and properties defined in our model

→ tool : Protégé

schema.org

K

protégé

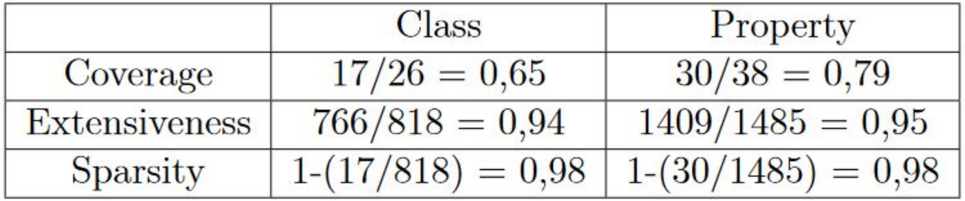
DATA MANAGEMENT

Development of code to prepare the data for the final integration with the ontology.

- Remove unnecessary fields from the datasets
- Translate and assign a standard english name to necessary fields
- Fix the format of some of the data

Node.JS was used for the creation and execution of the code and JSON format was used for the final datasets.

EVALUATION



	Class	Property
Coverage	17/26 = 0,65	30/38 = 0,79
Extensiveness	766/818 = 0,94	1409/1485 = 0,95
Sparsity	1-(17/818) = 0,98	1-(30/1485) = 0,98

- *Coverage :* if the reference ontologies covers the proposed ETG
- *Extensiveness :* if the reference ontologies extends the proposed ETG
- *Sparsity :* if the reference ontologies is different from the proposed ETG

4 - DATA INTEGRATION

Building and populating the final ETG

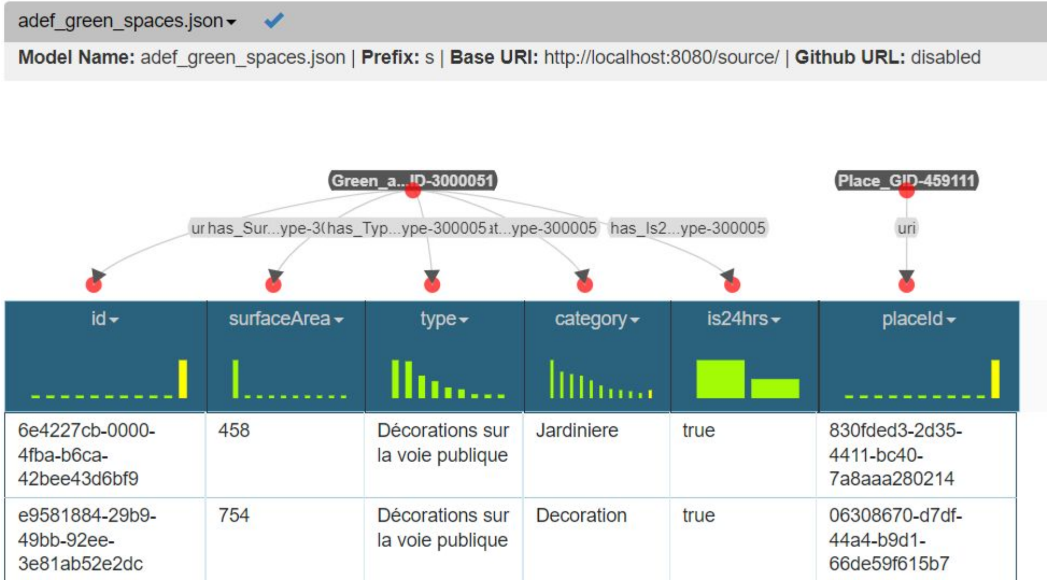
DATA MANAGEMENT & ENTITY MATCHING

No specific issue of semantic heterogeneity

No need for entity matching

Entity alignment over opening hours

Tool used : *Karmalinker*



adef_green_spaces.json

Model Name: adef_green_spaces.json | **Prefix:** s | **Base URI:** http://localhost:8080/source/ | **Github URL:** disabled

id	surfaceArea	type	category	is24hrs	placeId
6e4227cb-0000-4fba-b6ca-42bee43d6bf9	458	Décorations sur la voie publique	Jardiniere	true	830fded3-2d35-4411-bc40-7a8aaa280214
e9581884-29b9-49bb-92ee-3e81ab52e2dc	754	Décorations sur la voie publique	Decoration	true	06308670-d7df-44a4-b9d1-66de59f615b7

EVALUATION

Evaluation on **data level**. We need to check if :

- The CQs in inception phase can be answered by our EG *(evaluation based on practical applications)*
- Our collected datasets are sufficiently used and the dataset schema is aligned to ETG properties *(sparsity)*

5 - CONCLUSION

OPEN ISSUES

Datasets exploitation

Most of the datasets were in French.
Difficult to translate some concepts (ex : arrondissements)
Lost semantic meaning

User's localisation

Needed an additional algorithm able to calculate distances within GraphDB

Waterfall model

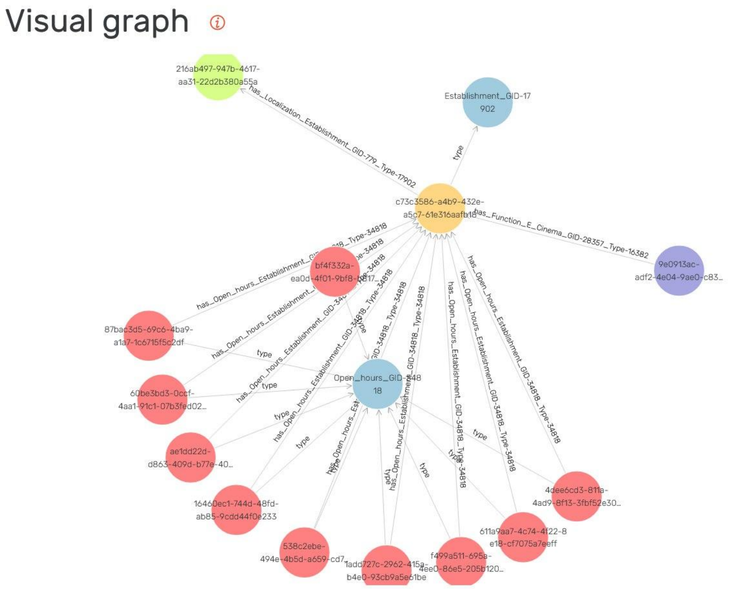
DI process : iterative process where each phase is based on the output of the previous one

OUTCOME EXPLOITATION

Total number of eTypes : 24
Total number of properties : 24 object properties and 91 data properties

Possibles exploitations :

- Family trip to Paris
- Shopping afternoon
- Business trip
- ...


Visual graph
216ab497-947b-4617-aa31-22d2b380a55a
Establishment_GID-17902
has_Localization_Establishment_GID-779_Type-17902
type
c73c3586-a4b9-432e-a5c7-61e316aafb18
has_Function_E_Cinema_GID-28357_Type-16382
9e0913ac-adf2-4e04-9ae0-c83...
has_Open_hours_Establishment_GID-34818_Type-34818
bf4f332a-ea0d-4f01-9bf8-b817...
87bac3d5-69c6-4ba9-a1a7-1c6715f5c2df
60be3bd3-0ccf-4aa1-91c1-07b3fed02...
ae1dd22d-d863-409d-b77e-40...
16460ec1-744d-48fd-ab85-9cdd44f0e233
538c2ebe-494e-4b5d-a659-cd7...
1add727c-2962-415a-b4e0-93cb9a5e61be
f499a511-695a-4ee0-86e5-205b120...
611a9aa7-4c74-4f22-8e18-cf7075a7eeff
4dee6cd3-811a-4ad9-8f13-3fbf52e30...
Open_hours_GID-34818
type

Thank you for your attention

Let’s do the demonstration !

References

Github repository : https://github.com/g1nus/KDI

Schema.org : https://schema.org/docs/schemas.html

OpenData Paris : https://opendata.paris.fr/pages/home/

OpenStreetMaps : https://www.openstreetmap.org/

SparQL tutorial : https://www.stardog.com/tutorials/sparql

KDI website : https://unitn-kdi-2021.github.io/unitn-kdi-2021-website/