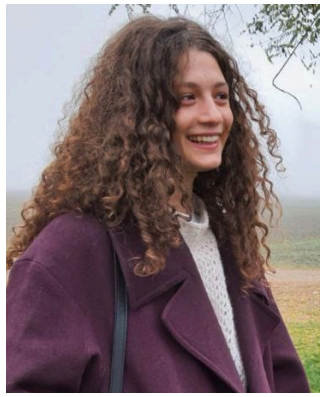
UNIVERSITY
OF TRENTO - Italy

Events in Trentino

Knowledge and Data Integration
Master in Data Science
A.Y. 2021 - 22

Speakers & overview



Anna Maria Fetz

- **Domain expert**
- **Data Scientist**

annamaria.fetz@student.unitn.it

Lucia Hrovatin

- **Domain expert**
- **Knowledge engineer**

lucia.hrovatin@studenti.unitn.it

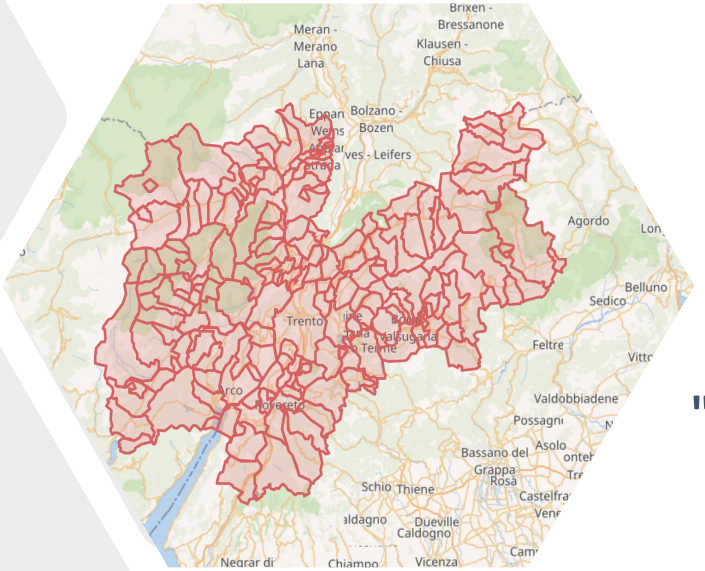
GitHub

Git repository: KDI_project

UNIVERSITY OF TRENTO - Italy

- **DoI and purpose definition**
- **Personas and CQs**
- **Inception phase**
- **Informal Modeling phase**
- **Formal Modeling phase**
- **Data integration**
- **KG exploitation**
- **Open issues**

Domain of Interest DEFINITION



2019 - 2022

"A service which helps the citizens to find events of interest in Trentino."

DoI:

events targeting **university students** in **Trento** and **Rovereto** from **2019** to **2022**.

→ Covid – 19 pandemic impact

UNIVERSITY OF TRENTO - Italy

PURPOSE FORMALIZATION

Website:

1. About **EVENTS:**

something that happens at a given place and time

Wordnet

2. acting as **collective memory**:

- **STM** (Short term memory): filters future events by location, duration, and other secondary attributes (e.g., target age, category)
- **LTM** (Long term memory): retrieves relevant information about past events

Project's RESOURCES

Different data formats were exploited:

- Scraped HTML
- CSV files
- JSON (and GEOJson)
- EXCEL spreadsheet
- Plain text

Along with lexical and knowledge resources.

PERSONAS & CQs DEFINITION

10 different personas, each defined via:

- NAME
- AGE → 18 – 30 years
- SHORT DESCRIPTION → interests
- SCENARIO → use case
- At least 4 Competency Queries (CQs)

NAME

DAVIDE

MARTINA

AGE

28 y.o.

19 y.o.

SHORT DESCRIPTION

Hiking, seminars related to its PhD

Freshman, social life

SCENARIO

He would like to retrieve hiking experiences that take place during the weekend and require medium skills. If the search will not return results, he would also be open to taking advantage of the free weekend by attending seminars which could boost his career.

Even if freshman, Martina shows excellent social skills. She has already organized an Aperitif with her new colleagues but the location is undefined yet. She plans to find the best bar by ranking them based on the last two events and the extra benefits offered during those occasions.

Inception: CQs

Actor: **ANDREA**

CQ :

Given a specific band, list all the concerts offered within Trento's municipality and their locations.

COMMON	CORE	CONTEXTUAL
Administrative Area	Address	Band
Audience	Artist	Concert
Duration	culturalEvent	Municipality
Event	playAction	musicEvent
Facility		
Location		
Person		
Student		

Metadata: Inception Sheet

Inception: data management

Tools:

- **Python** environment with **libraries**:
  Web scraping:

  Beautiful Soup

  Data handling:

  pandas

- REST **API**

PYROSM

OpenStreetMap PBF data parser for Python

Requests
http for humans

GitHub

Inception evaluation

Ontology vs CQs

Before parsing on scraped data and sentiment analysis on the plain text.

	Schema evaluation	
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 130 \rvert}{\lvert 46 \rvert} = \frac{\lvert 36 \rvert}{\lvert 46 \rvert} = \mathbf{0.78}$	$\frac{\lvert 121 \cap 212 \rvert}{\lvert 121 \rvert} = \frac{\lvert 87 \rvert}{\lvert 121 \rvert} = \mathbf{0.71}$
Extensiveness	$\frac{\lvert 130 - 46 \rvert}{\lvert 130 \cup 46 \rvert} = \frac{\lvert 94 \rvert}{\lvert 140 \rvert} = \mathbf{0.67}$	$\frac{\lvert 212 - 121 \rvert}{\lvert 212 \cup 121 \rvert} = \frac{\lvert 127 \rvert}{\lvert 248 \rvert} = \mathbf{0.51}$

Table 2: Summary of inception's schema evaluation

CQs vs DATA

	Data evaluation	
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 17 \rvert}{\lvert 46 \rvert} = \frac{\lvert 15 \rvert}{\lvert 46 \rvert} = 0.326 \approx \mathbf{0.33}$	$\frac{\lvert 121 \cap 98 \rvert}{\lvert 121 \rvert} = \frac{\lvert 90 \rvert}{\lvert 121 \rvert} = \mathbf{0.74}$
Sparsity	$1 - \frac{\lvert 46 \cap 17 \rvert}{\lvert 46 \cup 17 \rvert} = 1 - \frac{\lvert 15 \rvert}{\lvert 48 \rvert} = 0.687 \approx \mathbf{0.69}$	$1 - \frac{\lvert 121 \cap 98 \rvert}{\lvert 121 \cup 98 \rvert} = 1 - \frac{\lvert 90 \rvert}{\lvert 129 \rvert} = \mathbf{0.30}$

Table 3: Summary of inception's data evaluation

Purpose-driven approach while writing the CQs.

Informal Modeling phase: CQs

Actor: **ANDREA**

CQ :

Given a specific band, list all the concerts offered within Trento's municipality and their locations.

COMMON

Object	Function	Action
Duration Location Person Event	Student Administrative Area Facility	Audience

CORE

Object	Function	Action
Cultural Event Creative Work	Ticket Artist Address	Play Action

CONTEXTUAL

Object	Function	Action
Music Event	Band Concert Municipality	

E-type definition: AdministrativeArea, Duration,Location, Person, Event, Facility, Student, MusicEvent, CulturalEvent

Object Properties: PROPERTY_NAME(DOMAIN, RANGE) → **administration**(Facility, AdministrativeArea)

Data Properties: ETYPE: DATA PROPERTIES → **Event**: title, special Announcements, link, festival, edition, target Age,

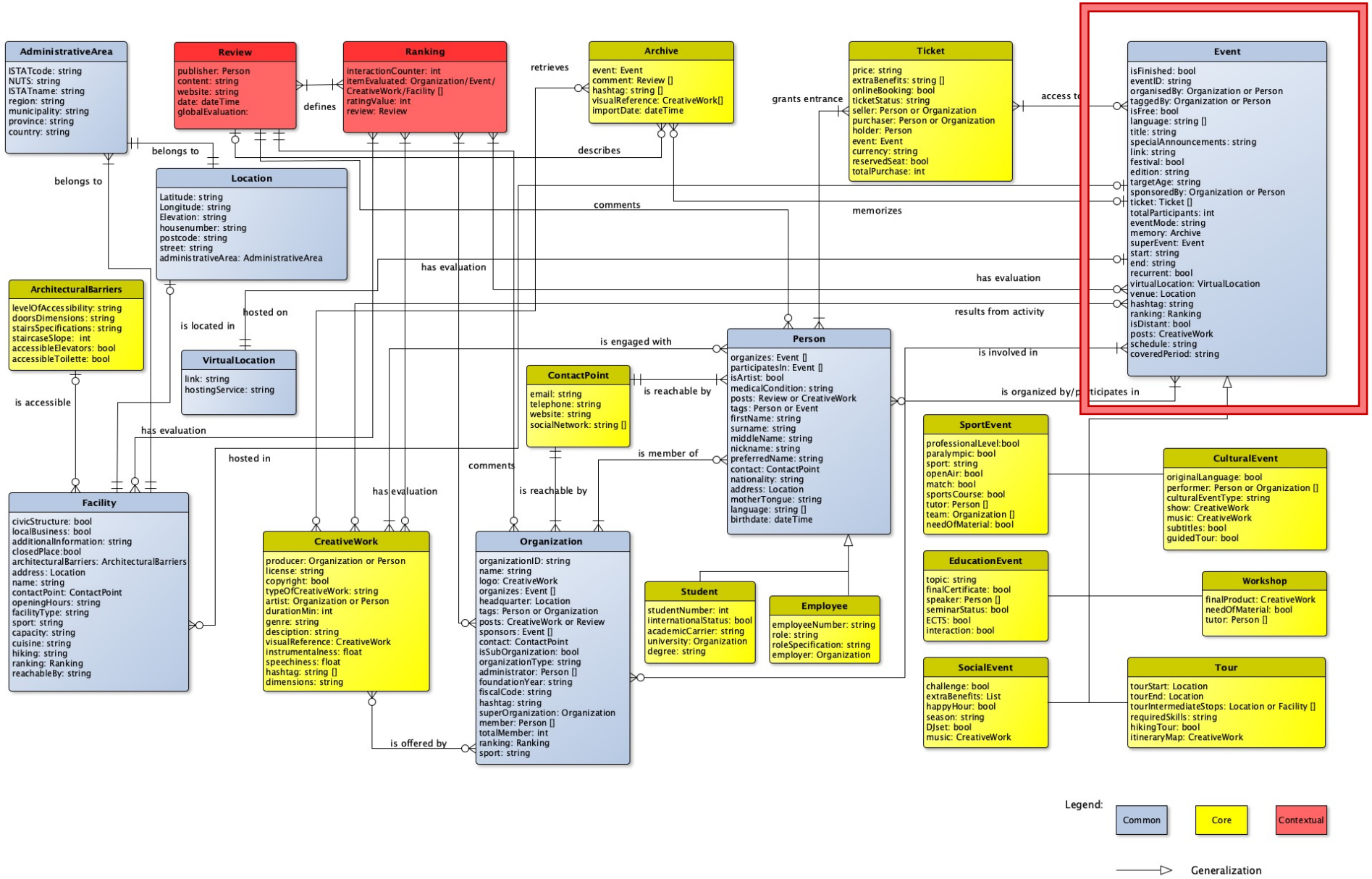
Metadata: Informal Modeling Sheet

Informal Modeling Phase: (final) ER model

Entity – Relationship model:

- Proportion: 31 – 59 – 10
- Data and purpose driven
- Reduced version due to a lack of data

Metadata: ER



Informal Modeling evaluation

Schema evaluation		
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 59 \rvert}{\lvert 46 \rvert} = \frac{\lvert 44 \rvert}{\lvert 46 \rvert} = 0.956 \approx$ **0.96**	$\frac{\lvert 121 \cap 217 \rvert}{\lvert 121 \rvert} = \frac{\lvert 115 \rvert}{\lvert 121 \rvert} =$ **0.95**
Extensiveness	$\frac{\lvert 59 - 46 \rvert}{\lvert 59 \cup 46 \rvert} = \frac{\lvert 15 \rvert}{\lvert 61 \rvert} = 0.245 \approx$ **0.25**	$\frac{\lvert 217 - 121 \rvert}{\lvert 217 \cup 121 \rvert} = \frac{\lvert 102 \rvert}{\lvert 224 \rvert} = 0.455 \approx$ **0.46**

Table 4: Summary of the evaluation on the informal modeling phase at the schema level

Consistent **reduction** of E-Types

+

different **data types** for properties

due to a lack of data

→ only 7 properties were deleted

Schema evaluation		
	Etype	**Property**
Coverage	$\frac{\lvert 46 \cap 22 \rvert}{\lvert 46 \rvert} = \frac{\lvert 22 \rvert}{\lvert 46 \rvert} = 0.478 \approx$ **0.48**	$\frac{\lvert 121 \cap 210 \rvert}{\lvert 121 \rvert} = \frac{\lvert 115 \rvert}{\lvert 121 \rvert} =$ **0.95**
Extensiveness	$\frac{\lvert 22 - 46 \rvert}{\lvert 22 \cup 46 \rvert} = \frac{\lvert 0 \rvert}{\lvert 46 \rvert} =$ **0**	$\frac{\lvert 210 - 121 \rvert}{\lvert 210 \cup 121 \rvert} = \frac{\lvert 95 \rvert}{\lvert 216 \rvert} = 0.439 \approx$ **0.44**

Table 6: Summary of the evaluation on the informal modeling phase at the schema level

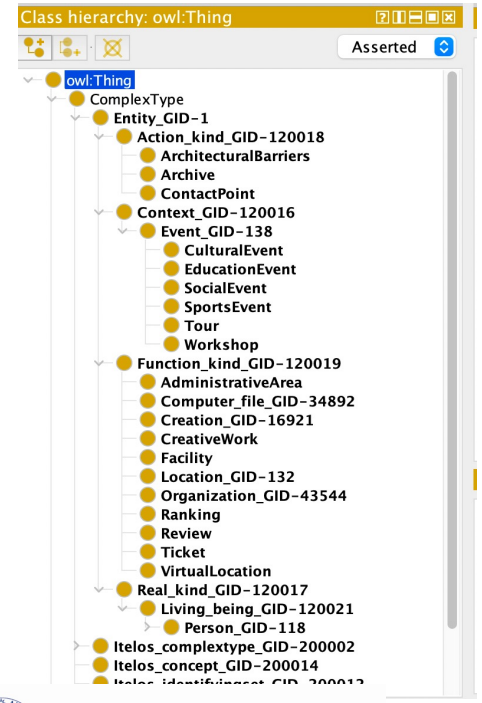
Formal Modeling Phase: ETG generation

Tools:

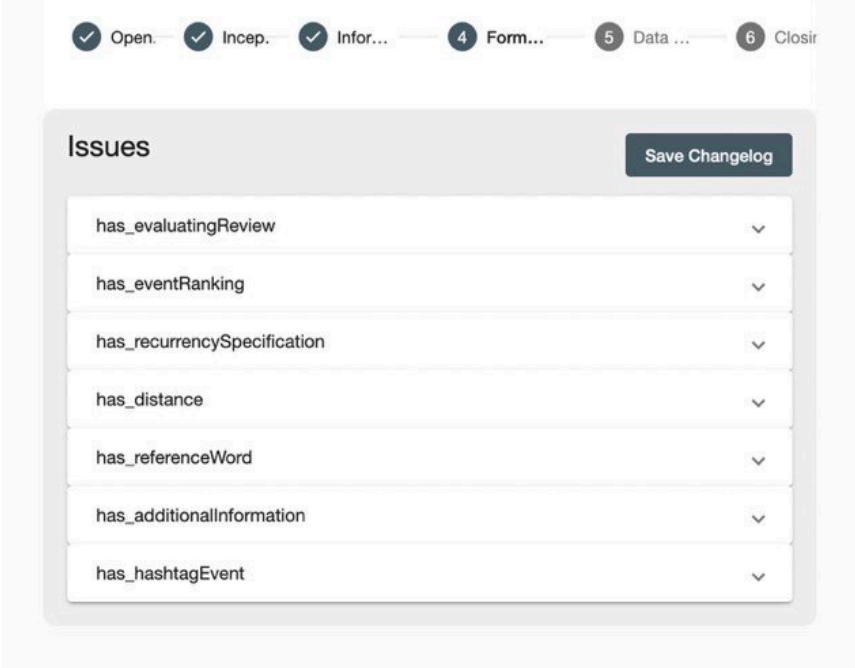
KOS and **Protégé**

Procedure:

1. ETG on Protégé



2. Linguistic alignment and issues resolution on KOS



3. Final ETG

Metadata: ETG

Formal Modeling Phase:
Data management - syntactic heterogeneity

1. Formats according to the ER:
   - string
   - integer/float
   - boolean
   - dateTime

2. STANDARDS whenever possible:
   - ISO 639-1 for language codes
   - ISO 8601 for timestamps
   - WGS 84 / UTM zone 32N for geocoordinates

3. Variables aligned via **has_** labels

4. Sentiment analysis on plain text (reviews) with Python library **NLTK**

5. Georeferencing via geocoder **Nominatim**

```
event = {"has_eventID": string,
        "has_title": string,
        "has_type":string,
        "has_mode": string, # offline, online or blended
        "has_cost": {
            "has_ticketID":string,
            "has_event": Event,
            "has_freeEntrance": bool,
            "has_onlineBooking": bool,
            "has_extraBenefits": string, # all extras included in the ticket
            "has_totalPurchase": int,
            "has_price": string, #either price or url to the purchasing web service
            "has_currency": string,
            "has_purchaser": string
                    },
        "has_link": string,  # link to the main page of the event
        "has_targetAge": string,
        "has_edition": int,
        "has_festivalStatus": bool,
        "has_language": [],
        "has_start": dateTime, # starting date of the event
        "has_end": dateTime,# ending date of the event
        "has_recurrency": bool, # recurrent event or not
        "has_schedule": string, # define the pace (weekly, daily, monthly, yearly)
        "has_organizer": string,
        "has_specialAnnouncements": string,
        "has_description" :string,
        "has_terminated": bool,
        "has_hashtag": [],
        "has_distance": bool,
        "has_transportMode": [],
        "has_venue": string,
        "has_virtualLocation":string,
        "has_superEvent":string
        }
```

Formal Modeling evaluation

Cue validity measure is the most suited one, but a SW down prevented its computation.

New ER model = severe changes in matching of:

- Ontologies (wide scope) and ETG *
- CQs (project's scope) and ETG

	Schema evaluation	
	Etype	**Property**
Coverage of Ont	$\frac{\lvert 130 \cap 22 \rvert}{\lvert 130 \rvert} = \frac{\lvert 18 \rvert}{\lvert 130 \rvert} = \mathbf{0.14}$ *	$\frac{\lvert 212 \cap 210 \rvert}{\lvert 212 \rvert} = \frac{\lvert 141 \rvert}{\lvert 212 \rvert} = 0.665 \approx \mathbf{0.67}$
Coverage of CQ	$\frac{\lvert 46 \cap 22 \rvert}{\lvert 46 \rvert} = \frac{\lvert 21 \rvert}{\lvert 46 \rvert} = 0.456 \approx \mathbf{0.46}$	$\frac{\lvert 121 \cap 210 \rvert}{\lvert 121 \rvert} = \frac{\lvert 119 \rvert}{\lvert 121 \rvert} = 0.983 \approx \mathbf{0.98}$
Sparsity on Ont	$1 - \frac{\lvert 130 \cap 22 \rvert}{\lvert 130 \cup 22 \rvert} = 1 - \frac{\lvert 18 \rvert}{\lvert 134 \rvert} = \mathbf{0.87}$	$1 - \frac{\lvert 212 \cap 210 \rvert}{\lvert 212 \cup 210 \rvert} = 1 - \frac{\lvert 141 \rvert}{\lvert 281 \rvert} = \mathbf{0.50}$
Sparsity on CQ	$1 - \frac{\lvert 46 \cap 22 \rvert}{\lvert 46 \cup 22 \rvert} = 1 - \frac{\lvert 21 \rvert}{\lvert 47 \rvert} = 0.554 \approx \mathbf{0.55}$	$1 - \frac{\lvert 121 \cap 210 \rvert}{\lvert 121 \cup 210 \rvert} = 1 - \frac{\lvert 119 \rvert}{\lvert 212 \rvert} = \mathbf{0.43}$
Extensiveness on Ont	$\frac{\lvert 22 - 130 \rvert}{\lvert 130 \cup 22 \rvert} = \frac{\lvert 3 \rvert}{\lvert 134 \rvert} = 0.022 \approx \mathbf{0.02}$	$\frac{\lvert 210 - 212 \rvert}{\lvert 212 \cup 210 \rvert} = \frac{\lvert 69 \rvert}{\lvert 281 \rvert} = 0.245 \approx \mathbf{0.25}$
Extensiveness on CQ	$\frac{\lvert 22 - 46 \rvert}{\lvert 46 \cup 22 \rvert} = \frac{\lvert 1 \rvert}{\lvert 47 \rvert} = \mathbf{0.02}$	$\frac{\lvert 210 - 121 \rvert}{\lvert 121 \cup 210 \rvert} = \frac{\lvert 91 \rvert}{\lvert 212 \rvert} = 0.429 \approx \mathbf{0.43}$

Data Integration phase: ENTITY MATCHING

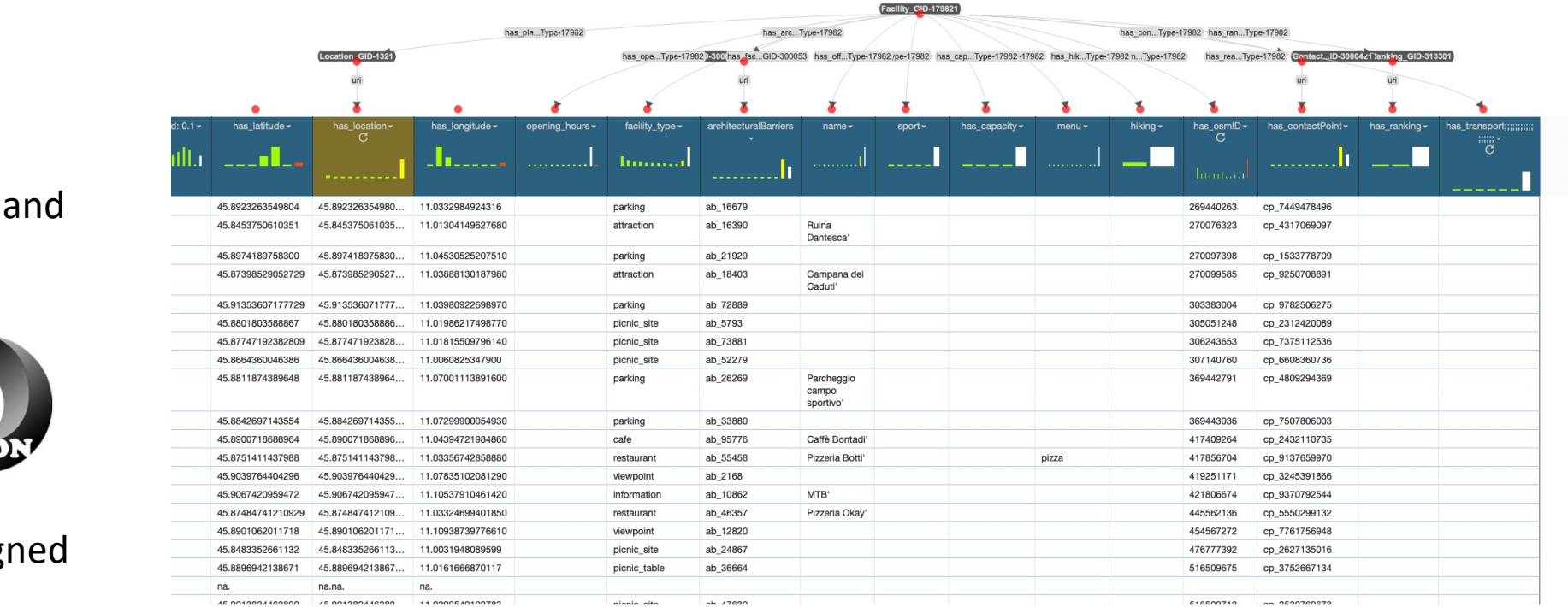
Tool: Karmalinker

Procedure:

One RDF file per E-Type

DATA: dataset cleaned and aligned

KNOWLEDGE: ETG aligned



has_latitude	has_location	has_longitude	opening_hours	facility_type	architecturalBarriers	name	sport	has_capacity	menu	hiking	has_osmID	has_contactPoint	has_ranking	has_transport
45.8923263549804	45.892326354980...	11.0332984924316		parking	ab_16679						269440263	cp_7449478496		
45.8453750610351	45.845375061035...	11.01304149627680		attraction	ab_16390	Ruina Dantesca'					270076323	cp_4317069097		
45.8974189758300	45.897418975830...	11.04530525207510		parking	ab_21929						270097398	cp_1533778709		
45.87398529052729	45.873985290527...	11.03888130187980		attraction	ab_18403	Campana dei Caduti'					270099585	cp_9250708891		
45.91353607177729	45.913536071777...	11.03980922698970		parking	ab_72889						303383004	cp_9782506275		
45.8801803588867	45.880180358886...	11.01986217498770		picnic_site	ab_5793						305051248	cp_2312420089		
45.87747192382809	45.877471923828...	11.01815509796140		picnic_site	ab_73881						306243653	cp_7375112536		
45.8664360046386	45.866436004638...	11.0060825347900		picnic_site	ab_52279						307140760	cp_6608360736		
45.8811874389648	45.881187438964...	11.07001113891600		parking	ab_26269	Parcheggio campo sportivo'					369442791	cp_4809294369		
45.8842697143554	45.884269714355...	11.07299900054930		parking	ab_33880						369443036	cp_7507806003		
45.8900718688964	45.890071868896...	11.04394721984860		cafe	ab_95776	Caffè Bontadi'					417409264	cp_2432110735		
45.8751411437988	45.875141143798...	11.03356742858880		restaurant	ab_55458	Pizzeria Botti'			pizza		417856704	cp_9137659970		
45.9039764404296	45.903976440429...	11.07835102081290		viewpoint	ab_2168						419251171	cp_3245391866		
45.9067420959472	45.906742095947...	11.10537910461420		information	ab_10862	MTB'					421806674	cp_9370792544		
45.87484741210929	45.874847412109...	11.03324699401850		restaurant	ab_46357	Pizzeria Okay'					445562136	cp_5550299132		
45.8901062011718	45.890106201171...	11.10938739776610		viewpoint	ab_12820						454567272	cp_7761756948		
45.8483352661132	45.848335266113...	11.0031948089599		picnic_site	ab_24867						476777392	cp_2627135016		
45.8896942138671	45.889694213867...	11.0161666870117		picnic_table	ab_36664						516509675	cp_3752667134		
na.	na.na.	na.												

Entity matching

Data Integration evaluation

- QUANTITATIVE

HIGH generalization
or
PURPOSE-DRIVEN
approach

Data evaluation		
	Etype	**Property**
Sparsity	$1 - \frac{\lvert 22 \cap 18 \rvert}{\lvert 22 \cup 18 \rvert} = 1 - \frac{\lvert 18 \rvert}{\lvert 22 \rvert} = \mathbf{0.18}$	$1 - \frac{\lvert 210 \cap 204 \rvert}{\lvert 210 \cup 204 \rvert} = 1 - \frac{\lvert 186 \rvert}{\lvert 228 \rvert} = 0.185 \approx \mathbf{0.19}$

- QUALITATIVE dimensions:
  - **Consistency:** one-to-one in Domain/Range specification with wordy names as consequence.
  - **Accuracy**: hierarchically not over-specified thanks to the ER adjustments .
  - **Completeness:** complete model, with the exception of has_record and has_memory

Knowledge Graph Exploitation

Tools:

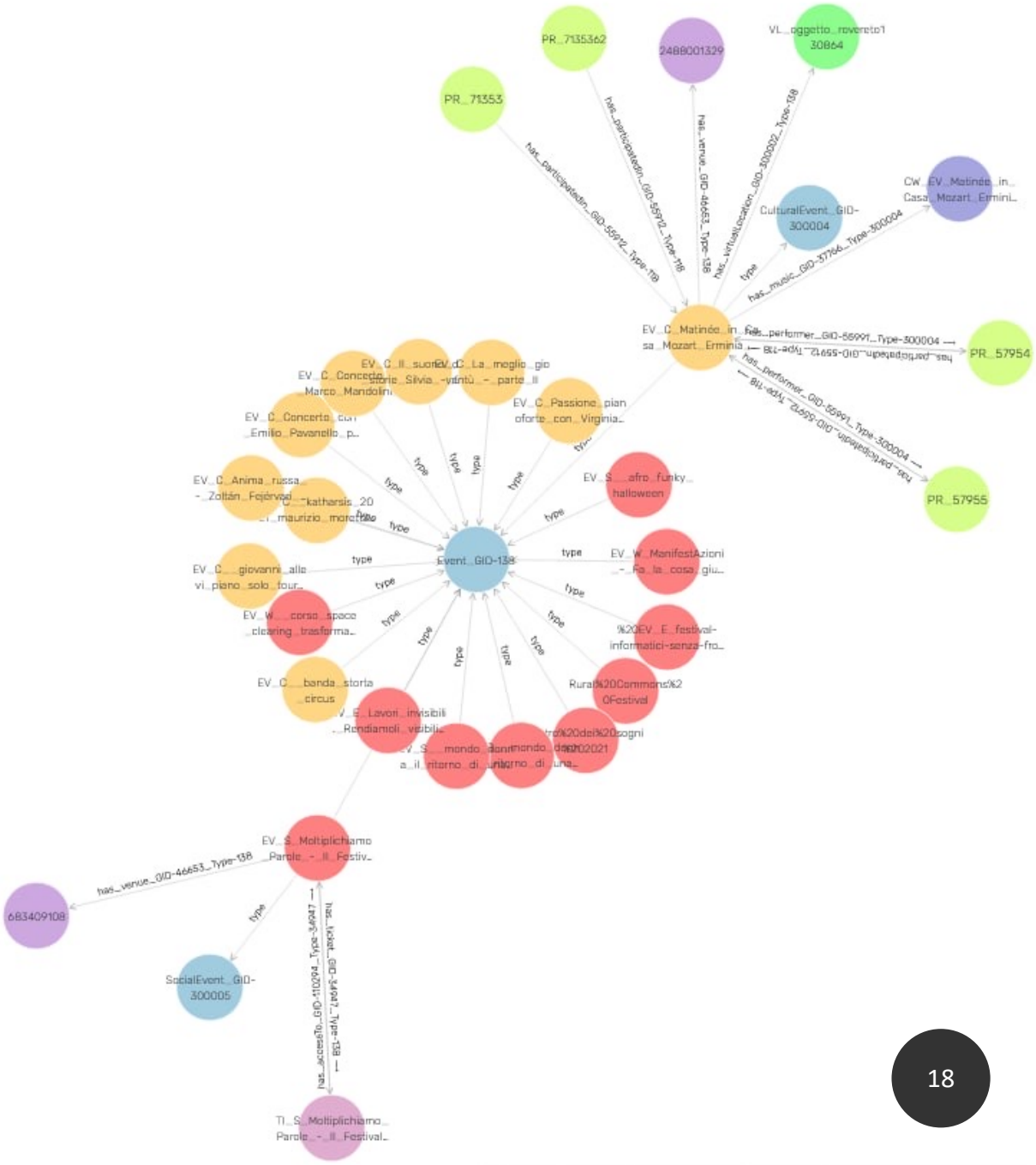
GraphDB

EXPLOITATION:

- **MANUAL** query

```
PREFIX  foaf:  <http://knowdive.disi.unitn.it/etype#>

CONSTRUCT WHERE {
    ?s a foaf:Event_GID-138 .
}LIMIT 100
```



Knowledge Graph Exploitation (SPARQL)

EXPLOITATION:

- **SPARQL** query – **COMMON CQ**

```
# Return all university student associations in Trento that organize events
PREFIX  foaf:  <http://knowdive.disi.unitn.it/etype#>

SELECT DISTINCT ?organization ?word ?headq ?event
WHERE {
 ?s a foaf:Organization_GID-43544;
    foaf:has_organizer_GID-108535_Type-43544 ?event ;
    foaf:has_officialDenomination_GID-34017_Type-43544 ?organization;
    foaf:has_headquarter_GID-19132_Type-43544 ?headq ;
    foaf:has_referenceWord_GID-300126_Type-43544 ?word;
    FILTER (contains(?word, 'studenti'))
```

	organization	word
1	"ESN - Associazione Erasmus Student Network"	"studenti internazionali', 'studenti', 'università'"
2	"JETN - Junior Enterprise Trento"	"studenti', 'università', 'enterprise', 'marketing'"

Knowledge Graph Exploitation (SPARQL)

EXPLOITATION:

- **SPARQL** query – **CORE CQ**

```
# Get all the events targeting university students and weekly offered in the bars of the city center
PREFIX  foaf:  <http://knowdive.disi.unitn.it/etype#>
SELECT  ?target ?title ?pace ?venue ?name
WHERE {
    ?s a foaf:SocialEvent_GID-300005;
    foaf:has_title_it_GID-300033_Type-138 ?title ;
    foaf:has_targetAge_GID-300054_Type-138 ?target ;
    foaf:has_recurrencySpecification_GID-300123_Type-138 ?pace ;
    foaf:has_venue_GID-46653_Type-138 ?venue ;
    FILTER (?target= '18-30' && ?pace= "weekly")

    ?venue foaf:has_officialName_GID-34017_Type-17982 ?name
}
```

	target	title	pace	venue	name
1	"18-30"	"Mercolegin"	"weekly"	http://localhost:8080/source/969087852	"Domo"

Knowledge Graph Exploitation (SPARQL)

EXPLOITATION:

- SPARQL query – **CONTEXTUAL CQ**

```
# Knowing the title of the Event, retrieve its evaluation and reviews.
PREFIX  foaf:  <http://knowdive.disi.unitn.it/etype#>

SELECT  ?event ?ranking ?review ?title ?content
WHERE {
    ?event a foaf:Workshop_GID-4602;
    foaf:has_eventRanking_GID-300124_Type-138 ?rank ;
    foaf:has_title_it_GID-300033_Type-138 ?title ;
    FILTER (?title = " corso andquotil colloquio di selezioneandquot")
    ?rank foaf:has_evaluatingReview_GID-300127_Type-31330 ?review .
    ?review foaf:has_content_GID-35359_Type-300000 ?content .
```

title	content
" corso andquotil colloquio di selezione andquot"	"Ho effettuato già parecchi corsi e la mia valutazione è pienamente positival Corsi per tutti i gusti e talvolta gratuiti. Il prezzo pagato comunque per tutti i corsi effettuati è irrisorio in confronto a quello che sono riuscito ad imparare. Se avete un obiettivo, anche imprenditoriale, è il luogo giusto per poter imparare nuove skills!"

UNIVERSITY OF TRENTO - Italy

OPEN ISSUES

1. Not realistic identifiers' definition

$$Ticket_{ID} = \text{TI\_E\_Abitare\_la\_speranza}$$

2. Approximation and specificity of some E-Types due to :
   - provincial, Italian or European **regulations**
   - low granularity/quality of data

3. Events' **duration** and **location**

DURATION: *bounded temporal interval*?
→ Events' such as festivals or courses last longer and have sub-events with *per-se* existence.

LOCATION: *point in the space?*
→ Should online events be spatially located?
→ Venue vs Location

UNIVERSITY
OF TRENTO - Italy

Thanks for the attention!