A Hospital Data Integration Project

Author:
Chiara Mazzoni
Giulia Calia
Stefano Marangoni

Knowledge & Data Integration Course
A.a. 2018/2019
Data Integration task

Input Dataset Description

The Data Integration task that our group addressed was implemented after a concrete realization of some real-world problem.

In our case, the world in question is Italy, where public health is one of the major topics of political discussion and debate.

Clinical Audit, is the practice of gathering quantitative information about clinical practice in some publicly available platform. In this way, structured information is available for citizens to be aware of the performance trend of the healthcare facilities in respect to dimensions, fluxes and degree of specialization. Furthermore, clinical audit is also a tool for the political bodies and the managerial staff to monitor the national situation of the healthcare services, in order to direct efficient interventions.

In our opinion, being informed on the level of performance and specialization of the healthcare facilities in our proximity is very important.

When people need a clinical procedure, it is often left to informal knowledge of relatives, acquaintances and word of mouth the "WHERE" would be better to go. The choice can also be guided by the family doctor or simply by the proximity of the facility, but formal quality criteria and personal preferences should and could be integrated, organically gathered and formalized so that users can be aware of the range of options in a ranked fashion and answer effectively to their health needs.

This interesting scenario resulted to already have a real-world application field, and this is PNE.

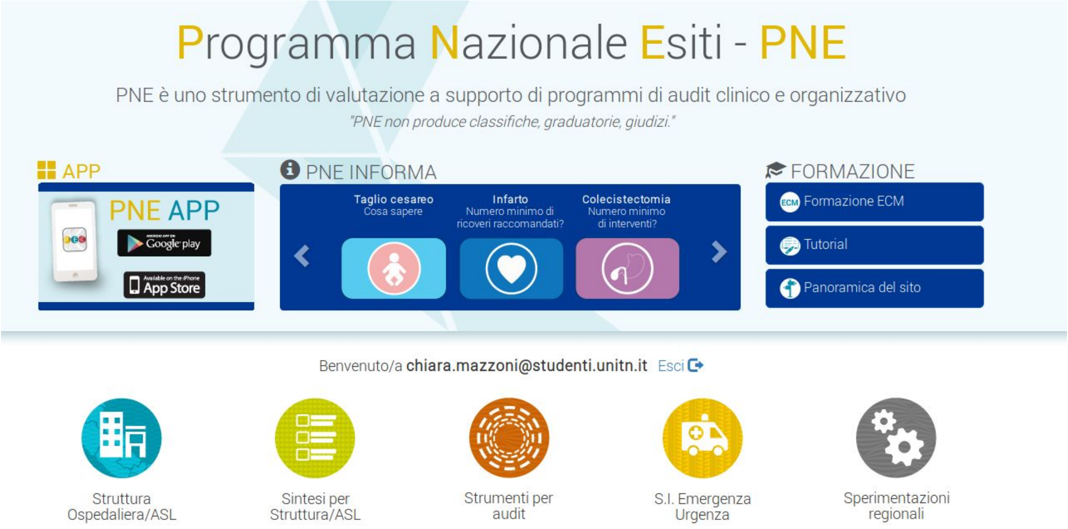
From 2013, italian national healthcare services agency (AGENAS) opened PNE platform (see PNE at this link) from which we derived the input datasets.

As the official site declaration says, PNE is a supporting tool for clinical audit, and it is not intended to produce rankings or general scores itself.

This was the starting point for us to provide the user a integrated and ready to grasp piece of information on where the best healthcare facilities are located across the national territory and for which clinical departments and procedures are evaluated the best.

For this purpose we wanted to integrate PNE recordings with Google Maps data, in order to offer specific information on proximity and performance quality of some healthcare facilities for some clinical procedure of need.

PNE platform visually offers information for a range of quality criterions (10) applied in a non-univocal way to a range of clinical procedures (16), performed across (7) different clinical departments. Non-univocal , in this case, is meant for explaining that some quality criteria, due to their generality, are applicable to more than a clinical procedure. This is not always the case, because, on the other hand, other quality criteria, only among the one we modeled, are specific to one only clinical area and procedure.



Programma Nazionale Esiti - PNE

PNE è uno strumento di valutazione a supporto di programmi di audit clinico e organizzativo

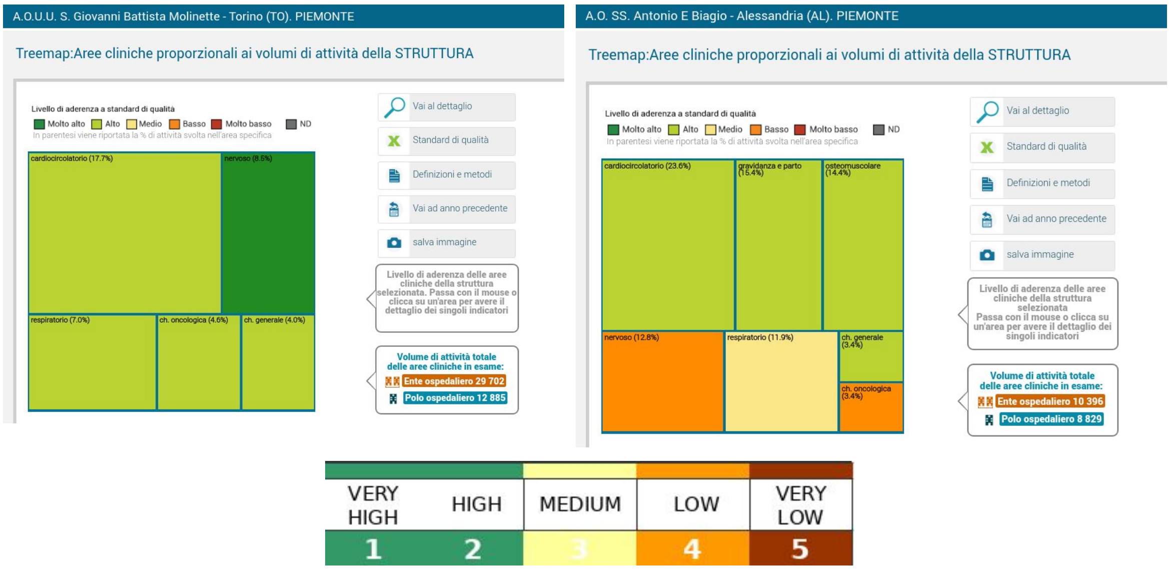
"PNE non produce classifiche, graduatorie, giudizi."

Benvenuto/a chiara.mazzoni@studenti.unitn.it Esci

Figure 1. PNE Web Platform

Accessing to the second leftmost coloured circle - Figure 1 - PNE website presents each healthcare facility as a map (Treemap) - Figure 2 - which recapitulates the clinical departments actually present in each facility and how much important they are in its activity volume.

In this map, clinical departments are color coded for their performance level (from red to dark green) in respect to official quality standards (criteria). These criteria are in detail evaluated in the Vedi Dettaglio section, with quantitative data.



A.O.U.U. S. Giovanni Battista Molinette - Torino (TO). PIEMONTE

Treemap:Aree cliniche proporzionali ai volumi di attività della STRUTTURA

Livello di aderenza delle aree cliniche della struttura selezionata. Passa con il mouse o clicca su un'area per avere il dettaglio dei singoli indicatori

Volume di attività totale delle aree cliniche in esame:
Ente ospedaliero 29 702
Polo ospedaliero 12 885

A.O. SS. Antonio E Biagio - Alessandria (AL). PIEMONTE

Treemap:Aree cliniche proporzionali ai volumi di attività della STRUTTURA

Livello di aderenza delle aree cliniche della struttura selezionata
Passa con il mouse o clicca su un'area per avere il dettaglio dei singoli indicatori

Volume di attività totale delle aree cliniche in esame:
Ente ospedaliero 10 396
Polo ospedaliero 8 829

Figure 2. PNE TreeMaps

From the PNE website, we extracted the quality data via the usage of Python web scraping libraries (requests, beautifulsoup4). We collected data, nation wide, for 982 healthcare facilities, resulting in having all or some of the quality assessment criteria, evaluated with

Very low, Low, Medium, High, Very High quality. Because not all the healthcare facilities had the totality of the assessable departments, in addition to the quality evaluation dataset we also built a department presence dataset.

For the integration of geospatial data, exact position of the healthcare facilities was retrieved via the usage of Python libraries (googlemaps ,difflib) and formally modelled with an already existing ontology specific for this purpose (geo_2007.owl), retrieved from the geospatial dataweb tutorial in Karma github documentation.

The ER Model

yEd is a standard tool to produce graphical representation of modeled data. We used it to produce an Entity Relationship Model - Figure 3 - which represented our data in its overall structure, considering the query we wanted to answer to: “Give me the best healthcare facility for receiving a specific clinical procedure, in my city/province?”

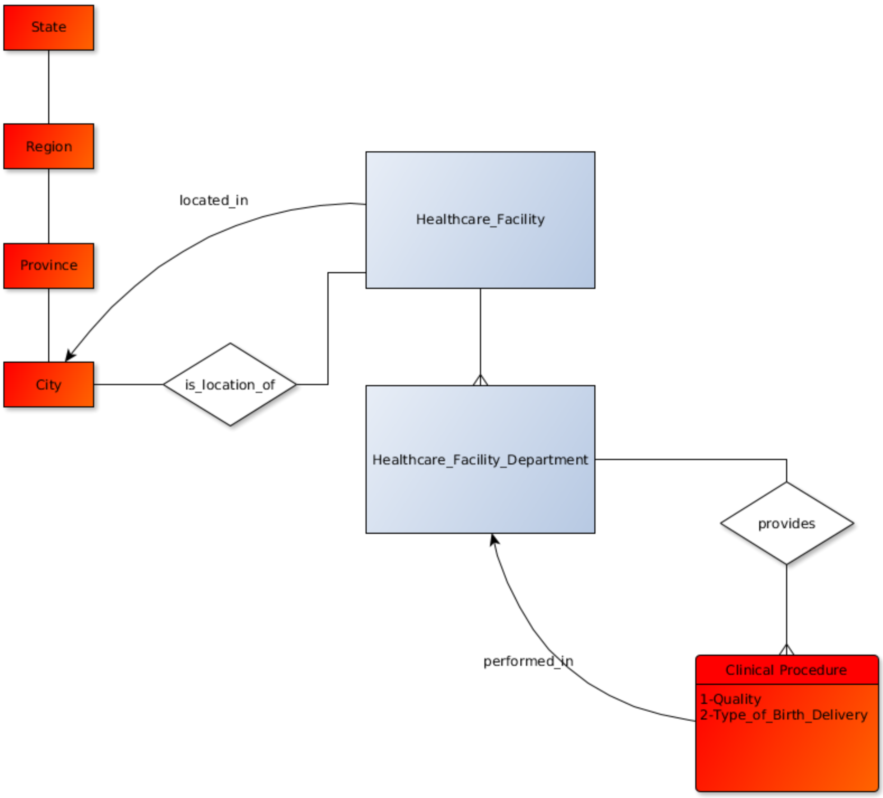


Figure 3. ER Model representing the entire structure of the data

In our model the Core entity is the Healthcare Facility, which is at the centre of our query.
Healthcare facility has a relation 1:M incoming in Healthcare Facility Department and an outgoing relationship towards city which is the lowest level Healthcare Facility location.

Auxiliary entities, defined to filter the results of the query, are indeed City/Province/State as subclasses of Administrative District and Clinical Procedure.
City, Province, Region and State in this ER model are defined hierarchically but the cardinality of the relationships is not stated as 1:M as we could expected (State 1:M Region 1:M Province 1:M City). This was decided on purpose because, on the base of our query, the Healthcare Facility in which the user is interested in is located in only one City/Province/Region/State but having them all in individual classes allows the user to query the data on each of these hierarchical levels.
Clinical Procedure has an outgoing relationship toward Healthcare Facility Department defined as "performedIn" and an incoming relationship coming from Healthcare Facility Department defined as "provides".

In our model only Clinical Procedure has a set of structured attributes which are recapitulated in Procedure_QualityCriterion and Type_of_BirthDelivery.
Specifically, Procedure_QualityCriterion recapitulates all the Criteria attributes through which every Clinical Procedure is evaluated by.

The Ontology Search and Building

We neither had an existing ontology specifically built for modelling the data at our disposal nor a Top-level ontology would have been suitable for our purposes.
A Top-level ontology like SNOMED Clinical terms deeply and formally details everything concerning the medical field, but it would have been more exhaustive than necessary in our case.
We built an ontology from scratch, suitable to formally model our data and to subsequently map them for their application.
Modelling data at our disposal was not easy at first hand. A single piece of information extracted from the PNE web site often recapitulated a lot of other sub information, so the very first process consisted in organizing data and **deconstructing information** in order to build the appropriate relationships between the parts.
Additionally to our customized ontology we used, as previously said, a geospatial modeling ontology.

The Customized Ontology

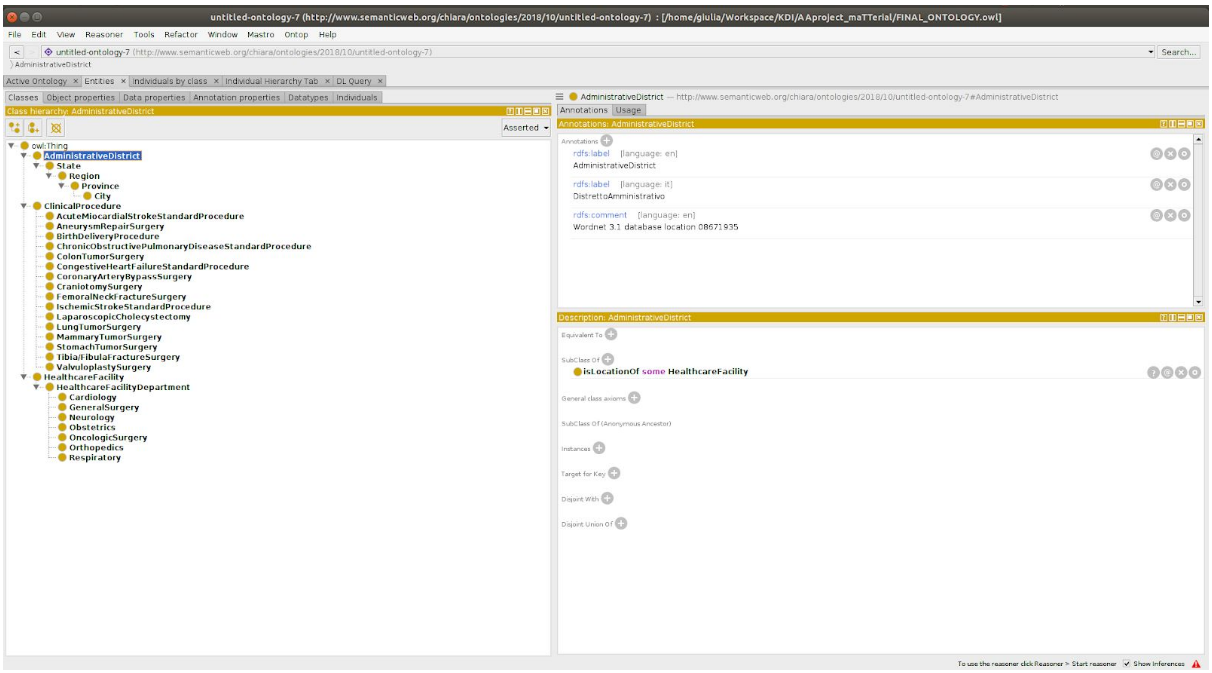
Taking into account the specificities of our data, our customized ontology includes three main classes (total class count: 31) :

- Administrative District
- Healthcare Facility
- Clinical Procedure

The class Administrative District was built to map the specific location of healthcare facilities (Healthcare Facility) nation, region and province wide plus the specific City location.
(AdministrativeDistrict <-subclassof State <-subclassof Region <-subclassof Province<-subclassof City).

The class HealthcareFacility and its subclass HealthcareFacilityDepartment were included to map each healthcare facility to its internal healthcare facility departments and, as previously said, some healthcare facilities can have only a subset of the 7 total healthcare departments available on PNE platform.
(HealthcareFacility <-subclassof HealthcareFacilityDepartment <-subclassof {Respiratory, Orthopedics, OncologicSurgery, Obstetrics, Neurology, GeneralSurgery, Cardiology }
We did not add disjunction among the clinical departments because of the previously cited reason that the criteria used for evaluation are in some cases applied over different medical disciplines.

The class ClinicalProcedure was built to subsequently map clinical procedures quality evaluation data to each healthcare facility department specific to each Healthcare Facility - Figure 4 -.


untitled-ontology-7 (http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7) : [/home/giulia/Workspace/KDI/AAproject_maTTerial/FINAL_ONTOLOGY.owl]
File Edit View Reasoner Tools Refactor Window Mastro Ontop Help
untitled-ontology-7 (http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7)
Search...
AdministrativeDistrict
Active Ontology × Entities × Individuals by class × Individual Hierarchy Tab × DL Query ×
Classes Object properties Data properties Annotation properties Datatypes Individuals
Class hierarchy: AdministrativeDistrict
Asserted
AdministrativeDistrict — http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#AdministrativeDistrict
Annotations Usage
To use the reasoner click Reasoner > Start reasoner
Show Inferences

owl:Thing
- AdministrativeDistrict
  - State
    - Region
      - Province
        - City
- ClinicalProcedure
  - AcuteMiocardialStrokeStandardProcedure
  - AneurysmRepairSurgery
  - BirthDeliveryProcedure
  - ChronicObstructivePulmonaryDiseaseStandardProcedure
  - ColonTumorSurgery
  - CongestiveHeartFailureStandardProcedure
  - CoronaryArteryBypassSurgery
  - CraniotomySurgery
  - FemoralNeckFractureSurgery
  - IschemicStrokeStandardProcedure
  - LaparoscopicCholecystectomy
  - LungTumorSurgery
  - MammaryTumorSurgery
  - StomachTumorSurgery
  - Tibia/FibulaFractureSurgery
  - ValvuloplastySurgery
- HealthcareFacility
  - HealthcareFacilityDepartment
    - Cardiology
    - GeneralSurgery
    - Neurology
    - Obstetrics
    - OncologicSurgery
    - Orthopedics
    - Respiratory

Annotations: AdministrativeDistrict

Annotations
rdfs:label [language: en]
AdministrativeDistrict
rdfs:label [language: it]
DistrettoAmministrativo
rdfs:comment [language: en]
Wordnet 3.1 database location 08671935

Description: AdministrativeDistrict

Equivalent To
SubClass Of
isLocationOf some HealthcareFacility
General class axioms
SubClass Of (Anonymous Ancestor)
Instances
Target for Key
Disjoint With
Disjoint Union Of

Figure 4. Visualization of all the classes of the Ontology

For each clinical procedure (Clinical Procedure) we added attributes, having the role of specifically define which criteria were taken into consideration for the assessment of the Clinical Procedure, they are:
(Procedure_QualityCriterion <-subclassof {ProcedureOccurrenceCriterion, FractureProcedureBy2DaysCriterion, PTCAProcedureBy2DaysCriterion,

Under3DaysRecoveryStayCriterion, SecondaryProcedureOccurrenceBy120DaysCriterion, 135AnnualProceduresCriterion, 30DaysMortalityCriterion, 90AnnualProcedureCriterion, DeliveryComplicationOccurrenceCriterion, ProcedureWaitingTimeCriterion }, TypeOfBirthDelivery)

Procedure_QualityCriterion sub-properties are specified in their Domains and Ranges and, once mapped to the dataset, map to the previously specified quality values as string objects: Very low, Low, Medium, High, Very High.

The TypeOfBirthDelivery attribute was included for additionally defining and differentiating between Natural and Ceasarian BirthDeliveryProcedure which both have DeliveryComplicationOccurrenceCriterion assessing them.

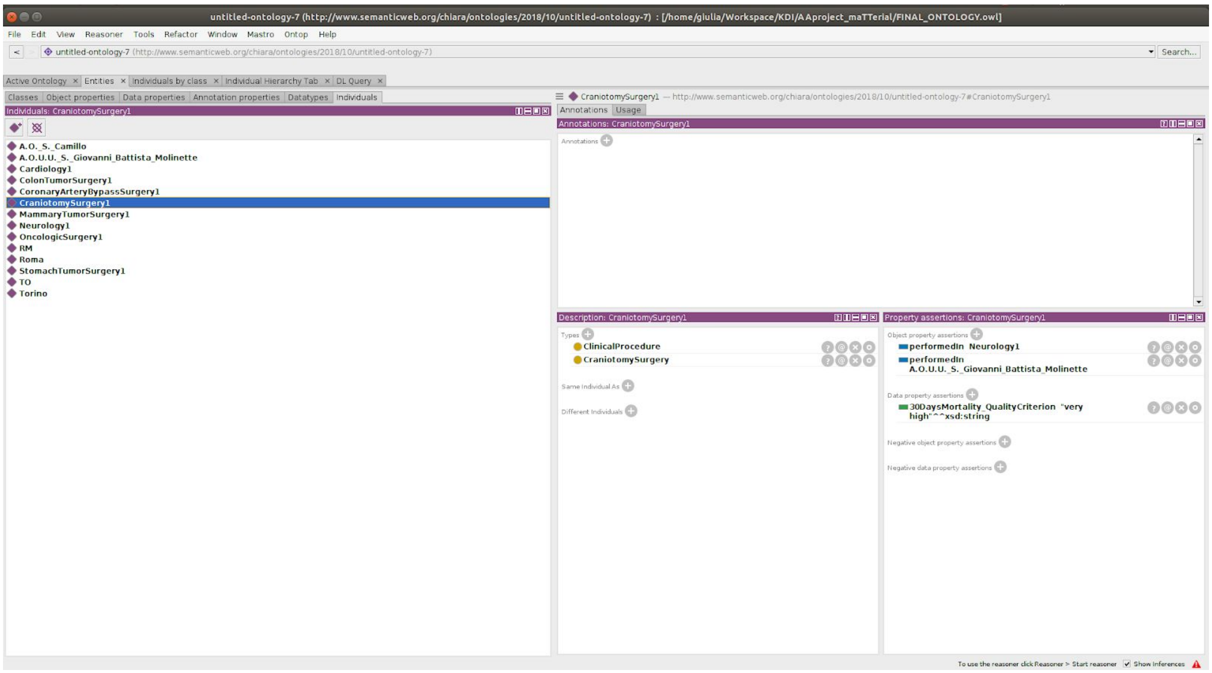
Finally, a data attribute Presence was added to correctly model the columns in our dataset reporting the presence or not of specific healthcare facility departments in the different facilities.

We stated the relationships between our entity classes via the following properties : locatedIn, performedIn, provides , isLocationOf .

Ontology Consistency Check

To investigate on the formal correctness, consistency and future mappability of our model, Protege provides the Reasoner Tool to evaluate the relationships between classes via the instantiation of individuals.

Within our model we created several instances, in particular two different healthcare facilities (A.O S. Camillo, A.O.U.U S. Giovanni Battista Molinette) -Figure 5 - located in two different cities, for which we created City and Province instances (Roma, RM, Torino, TO) providing different clinical procedures within different healthcare facility departments (CoronaryArterybypassSurgery1 in Cardiology1 for the former and MammarySurgery1, ColonTumorSurgery1,StomachTumorSurgery1 in OncologicSurgery1, CraniotomySurgery1 in Neurology1) for the latter. Additionally for A.O.U.U S. Giovanni Battista Molinette we stated a MammarySurgery1 having a negative property and negative attribute, meaning that this type of clinical procedure is not performed in this specific facility and to create an example in which, even if the OncologicSurgery department is present and performs some types of oncologic procedures (colon and stomach surgery) doesn’t provide mammary tumor surgeries.
The clinical procedures that are stated as positive properties are evaluated all by 30DaysMortality_QualityCriterion with values, “medium”,”high”, “high”,”very high”.


untitled-ontology-7 (http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7) : [/home/giulia/Workspace/KDI/AAproject_maTTerial/FINAL_ONTOLOGY.owl]
File Edit View Reasoner Tools Refactor Window Mastro Ontop Help
untitled-ontology-7 (http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7)
Search...
Active Ontology × Entities × Individuals by class × Individual Hierarchy Tab × DL Query ×
Classes Object properties Data properties Annotation properties Datatypes Individuals
CraniotomySurgery1 — http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#CraniotomySurgery1
Annotations Usage
To use the reasoner click Reasoner > Start reasoner ☑ Show Inferences

Individuals: CraniotomySurgery1

- A.O._S._Camillo
- A.O.U.U._S._Giovanni_Battista_Molinette
- Cardiology1
- ColonTumorSurgery1
- CoronaryArteryBypassSurgery1
- CraniotomySurgery1
- MammaryTumorSurgery1
- Neurology1
- OncologicSurgery1
- RM
- Roma
- StomachTumorSurgery1
- TO
- Torino

Annotations: CraniotomySurgery1

Annotations

Description: CraniotomySurgery1

Types
ClinicalProcedure
CraniotomySurgery

Same Individual As

Different Individuals

Property assertions: CraniotomySurgery1

Object property assertions
performedIn Neurology1
performedIn A.O.U.U._S._Giovanni_Battista_Molinette

Data property assertions
30DaysMortality_QualityCriterion "very high"^^xsd:string

Negative object property assertions

Negative data property assertions

Figure 5. Visualization of all the instances in the Ontology

The Geospatial Ontology

We used the geospatial ontology "geo_2007.owl" (ontology IRI http://opengis.net/gml/) suggested for modeling points, trajectories, boundaries and polygon areas.

It consists of 9 classes in total in a specific hierarchical organization:

(Feature,Geometry<-*subclassof* {Envelope,LinearRing,LineString,Point,Polygon} , KMLCustomization, SpatialReferenceSystem).

There are 2 properties :(exterior, geo:where).

There are 19 data attributes: (featurename, featuretypetag, geo:box, geo:elev, geo:floor, geo:line, geo:point, geo:polygon, geo:radius, lowerCorner, upperCorner, relationshiptag, posList, pos, lat, long). *Lat* and *long* have been used to map latitude and longitude for each specific healthcare facility retrieved in google maps; *pos* have been used to map the google maps address of the geographical coordinates.

Originally, we wanted to integrate this ontology to have direct visualization in Google Earth of the position of the facilities, once mapped in Karma. Unfortunately, Karma v. 2.2 has no geospatial data visualization mapping (as Karma v. 1.9 had), so we just added Google Maps links.

Within this ontology we instantiated A.O S. Camillo and A.O.U.U S. Giovanni Battista Molinette as Point1 and Point2 having (41.8703564 lat, 12.4555119 long) and (45.0434342 lat, 7.6690741 long) respectively.

Integration process description

The Karma integration process aimed to integrate three different datasets: two of them came from PNE platform and one from the Google retrieval of the geographical data.

The creation of the dataset was done from scratch, populating it with data extracted recursively from separate PNE platform pages. We decided to implement the extraction of data with Python because the process was a bit more complicated to implement in Rapidminer when some platform pages created exceptions.

The first dataset contains the information about the presence of the clinical departments in the healthcare facilities: in other words, a more general idea on how the facilities are organized and which macro clinical areas are covered. This information on its own does not give an exhaustive idea on WHERE the user should go for his needs, but more generally it assesses the overall quality of the clinical performance of the health services.

The second dataset from PNE covers the detailed evaluation of each single clinical procedure taken into consideration, leaving a blank space if the specific procedure is not performed within the clinical department.

The third dataset was built from scratch for gathering geospatial coordinates for each healthcare facility.

Karma Web Tool allowed us to import the datasets and start building the very first version of our healthcare facilities database, containing only italian healthcare facilities data at the moment.

First of all, directly in Karma, we created columns for including unique identifiers (URI) of the healthcare facilities, in order to have both the customized and the geospatial ontology mapping univocally to the same healthcare facility.

As a good practice, a label column was also included for each URI column and for other minor issues Py-transformations were used, essentially for technical purposes. Given that the clinical records datasets were built from scratch by us starting from non structured data extracted from json files of the PNE website, data cleaning and preparation was massive before the integration in Karma.

For instance, we had some issues with the names of some facilities which were not appropriately reported, such that the retrieval of their geographical coordinates was problematic. This required the usage of *difflib* python library for the maximal string match with the healthcare facilities names located in a separate file, retrieved from the web.

The most delicate part of the Karma integration was the mapping of the clinical procedures with theirs procedure quality criterion/a.

The PNE datasets were particularly coherent within themselves, so we had attributes and properties mapping in bulk to the PNE datasets.

Within Karma integration we realized that a data attribute indicating the presence of the healthcare facility was missing so we added it to our ontology and we mapped the clinical departments columns with the already cited data property *Presence*.

Karma allows to create more copies of the same class if more than one column contains the same kind of information, but at the same time the entity would have a different property mapping it to other classes. This was the case for the class BirthDeliveryProcedure which has three different criteria evaluating it. Given that the type of birth delivery procedure is an information modeled in the ontology as a data property (TypeOfDelivery, ranging from "Natural" to "Caesarian"), for mapping the same class with different Procedure_QualityCriterion we created BirthDeliveryProcedure1 and BirthDeliveryProcedure2, with the latter being the "Natural" kind of birth delivery.

The main goal of our integration was to exploit the Google Earth visualization of the mapping of the geospatial points within the third dataset. Unfortunately this functionality was available in the previous version of Karma but not in the present one. We opted, instead, for including a direct link to google.maps visualization creating a dedicated column and mapping it under the **geo:pos** data property which had only the *string* data type restriction, so perfectly adhering to the purpose.

Output Dataset and queries description

For the sake of visualization and exploration of the output database from Karma integration, we opted for the web tool GraphDB.

GraphDB interface allows the user to check how the integration has occurred, visually inspecting the overall and hierarchical structure - Figure 6 - of data and having a grasp of either the number of connections created between the classes and the number of times these properties are called for connecting those classes.

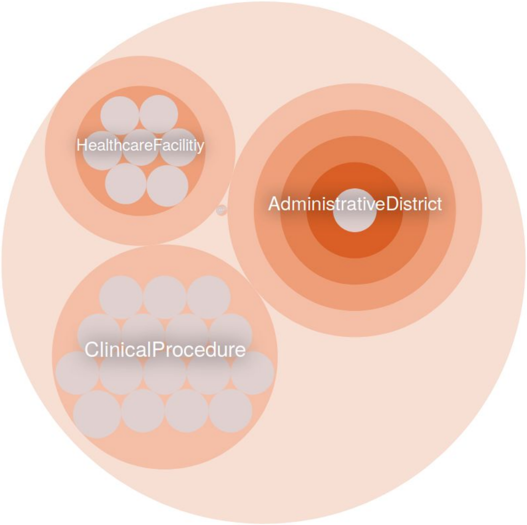


Figure 6. GraphDB representation of the class hierarchy

Karma integration produces a graph in turtle format (.ttl) which can be imported in GraphDB and queried.

Our output dataset is the result of the integration of departments presence, clinical procedures quality and localization of facilities so this are the information available to be queried.
In a ideal future scenario in which more healthcare facilities would be included in the performance evaluation and more procedures would be taken into consideration for developing metrics for quality evaluation, the list of quality criteria would be expanded and possibly reorganized in a structured way and in more general macro categories.
At present, PNE records use clinical procedure quality criteria that specifically refer to numerical numbers (FractureProcedureBy**2**DaysCriterion, PTCAProcedureBy**2**DaysCriterion, Under**3**DaysRecoveryStayCriterion, SecondaryProcedureOccurrenceBy**120**DaysCriterion, **135**AnnualProceduresCriterion, **30**DaysMortalityCriterion, **90**AnnualProcedureCriterion) that have been taken as standard quality thresholds.
At this 'infancy' phase of our database the user can ask about all the facilities performing a restricted selection of procedures, and additionally filter for a specific level of quality.
The user can ask about how many facilities are located above a certain latitude (we could take as a reference the latitude of Rome, which divides the northern regions from the southern regions) and draws some conclusions from the number of results of the query.
As previously mentioned, we instantiated a limited number of entities in the ontology so the first results of certain specific queries are these nodes and only these will present some specifics.
For other queries less comprehensive ids will be the result of the query.
Here we report a selection of the queries we ran for exploring the database.

- Give me the facilities located in province of Rome, which provides some procedures with their quality (return each procedure with the value of quality).

```
PREFIX hc:<http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#>
PREFIX gml: <http://www.opengis.net/gml/>
select * where {
        ?fac  hc:locatedIn hc:RM.
        ?fac  hc:provides ?procedure.
         ?procedure hc:Procedure_QualityCriterion ?quality.
}
```

- Give me the google maps link for the facility A.O S.Camillo

```
PREFIX gml: <http://www.opengis.net/gml/>
select * where {
   gml:A.O S.Camillo  gml:pos  ?link
}
```

- Give me the facilities in province of Turin, which provides procedures with a high quality

```
PREFIX hc:<http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#>
PREFIX gml: <http://www.opengis.net/gml/>

select * where {
        ?fac  hc:locatedIn  hc:TO.
         ?fac  hc:provides  ?procedure.
         ?procedure hc:Procedure_QualityCriterion ?quality.
        FILTER (?quality='high')
}
```

- Give me the department (and corresponding facility) whose procedures have high quality level

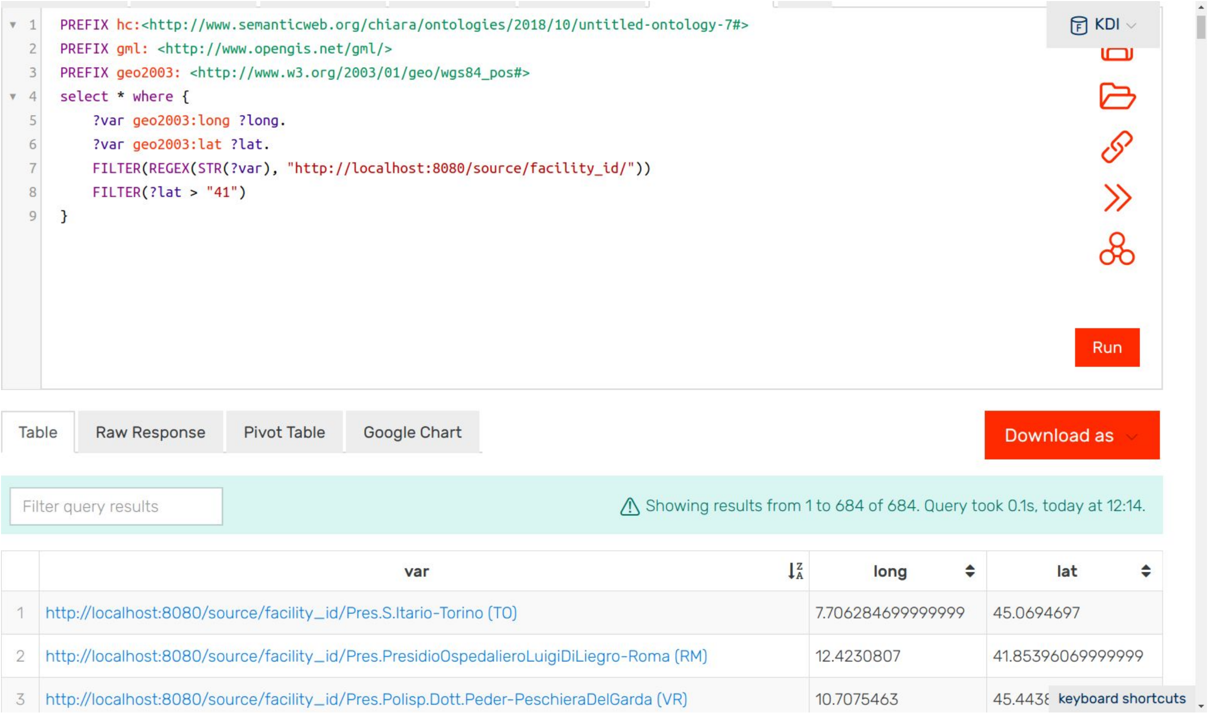
```
PREFIX hc:<http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#>
select * where {
?proc ?criterion "high" .
 ?proc hc:performedIn ?department
}
```

- Give me the facility whose ProcedureWaitingTime_QualityCriterion (mapping the Tibia/FibulaFractureSurgery) has a low quality

```
PREFIX hc:<http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#>
select * where {
        ?fac hc:ProcedureWaitingTime_QualityCriterion ?qual .
        ?fac  ?crit  ?qual
        FILTER (?qual = 'low')
}
```

- Give me the facilities whose latitude is higher than 41 (latitude of Rome) - Figure 7 -.

```
PREFIX hc:<http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#>
PREFIX gml: <http://www.opengis.net/gml/>
PREFIX geo2003: <http://www.w3.org/2003/01/geo/wgs84_pos#>
select * where {
?var geo2003:lat ?lat.
FILTER(REGEX(STR(?var), "http://localhost:8080/source/facility_id/"))
FILTER(?lat > "41")
```



```
PREFIX hc:<http://www.semanticweb.org/chiara/ontologies/2018/10/untitled-ontology-7#>
PREFIX gml: <http://www.opengis.net/gml/>
PREFIX geo2003: <http://www.w3.org/2003/01/geo/wgs84_pos#>
select * where {
    ?var geo2003:long ?long.
    ?var geo2003:lat ?lat.
    FILTER(REGEX(STR(?var), "http://localhost:8080/source/facility_id/"))
    FILTER(?lat > "41")
}
```

Showing results from 1 to 684 of 684. Query took 0.1s, today at 12:14.

	var	long	lat
1	http://localhost:8080/source/facility_id/Pres.S.Itario-Torino (TO)	7.706284699999999	45.0694697
2	http://localhost:8080/source/facility_id/Pres.PresidioOspedalieroLuigiDiLiegro-Roma (RM)	12.4230807	41.85396069999999
3	http://localhost:8080/source/facility_id/Pres.Polisp.Dott.Peder-PeschieraDelGarda (VR)	10.7075463	45.4438

Figure 7. Results of the most complex query above

Input/Output Dataset comparative analysis description

Our data integration task was inspired by noticing limitations in browsing the PNE website and mobile app.

PNE declares that comparative purposes among italian regions are not pursued or implemented within the platform, that it was not designed to draw any general conclusion on the national situations and across regions. Because these aspects are not purposely shown or discussed within PNE, this leaves space for some satellite application, namely ours.

Throughout the design process of our modelling and integration of data, we focused on the interpretability, readability, understandability of the information that the database would provide to the user.

As previously mentioned, we thought of our ideal application scenario as the informal context of a citizen willing to have an idea on the quality of some clinical procedures, performed in some healthcare facility within the geographical neighbourhood of his choice.

In this regard PNE can evaluate the clinical procedure with :

- criteria reporting the percentage of cases that fall within certain quality ranges; ranges that identify high+/medium/low+ quality;
- criteria reporting the waiting time in days for a procedure to be delivered.

PNE input dataset, along the color legend for implying quality levels, included also some quantitative metrics supporting the significance of the platform statements. This includes the p.value for the percentage reported which is adjusted for the most prominent confounding factors, whose influence has been inferred by multivariate regression models.

The adjustment of the percentage is used as a refinement of the record, which in this way can be compared across different healthcare facilities throughout all the territory even if, statistically it is calculated that they are exposed to different degrees of confounding factors like heterogeneity of the population in respect to age, gender, disease severity, chronic comorbidities.
Additionally PNE reports also the national median for each particular criterion.
These statistical metrics refers to a number of cases taken into consideration, so they are calculated on the volume of activity of the specific department under investigation and, in our opinion, the correlation between the activity volume and the quality of performance should be investigated more in detail, in order to draw some conclusions on facilities with comparable activity volume, resources and personnel.

The quantitative metrics (properties) just described were excluded from the data integration task, because we thought these metrics could not be the starting point of a user query. In the modelling phase we focused on the more understandable and immediate qualitative metric for the procedures but now we do not exclude the value of a more complete assessment of performance, backed with statistics, that it could be implemented in the future.
Drawing differences between the original dataset (PNE platform) and the output dataset we report:
34 classes for the input dataset
33 classes for the output model dataset (not mapped class:Region)
24 properties for the input dataset
20 properties for the output model dataset (not mapped properties: activity volume, adjusted percentage,national median percentage)
18 class questions (queries)
14 property questions (queries)

We do think that, in the era of digital visual art and entertainment, visual information have more opportunities for the effective perception and understanding of the information by the user.
For this purpose we appreciated the role of colours and treemaps within PNE, but for our purposes, not having a comprehensive view of all the structures organized for quality levels could have been considered a shortcoming.
We thought of visualizing data with a geographical map or a more sophisticated version of it, being a thematic map. Our basic solution for this purpose is implemented within Google Fusion Tables.
The web application allows to import datasets, merge them and mapping geographical coordinates for a map visualization.
Within this visualization it is possible to filter output dataset results for each column present in the dataset, in our case, the kind of procedure and quality levels, which here have been decoded in numerical values from 1 to 5 (from 'very low' to 'very high') and NA is for not

available data, interpretable as ‘ photographic negatives’ of the information, namely on where some kind of procedures are not performed compared to others.
The Google Fusion Tables visualization map is available at this link.

DB generation proposal

On one hand we generated an output dataset that integrates all the data at our disposal, on the other we also imagined the fashion in which data could be organized into, for best exploration and answers.
The Database in question, entering the clinical procedure and a particular province or city in a dedicated search bar , would return a list, sorted by clinical procedure quality, provided by a list of nearest facilities, whose positions can be visualized directly in a map.
Tabs with different quality levels would be available to filter, mix and match the results allowing the user to have a sense of the range of results possible.
We can imagine the extension of the model originally built on italian data to different countries. This would require a more structured way to organize and standardize quality criteria, which indeed could vary from country to country. At the end of a possible restructuring phase, the database would fit also data coming from non-italian healthcare facilities and could be available and useful for a non-italian user. Given the expansion of the HDI model to different countries, also the country of origin should be entered by the user.

Final considerations and open issues

To tackle the data integration task for this project we wanted to approach a real-life problem, something that we personally felt to deserve some attention.
The identification of an unresolved issue, something which creates discomfort or uncertainty, often leads to the visualization of some unexplored space, ready to be expanded with the power of digital resources.
Flexibility and expandability of the database has not been demonstrated yet and open issues exist on the organization of new procedure quality criteria, especially if coming from different healthcare systems.
Nevertheless we hope that clinical audit will be increasingly implemented to be more and more user friendly and effectively reachable by the user, which is not the case at this moment.
Before this project, we did not have idea of the PNE existence, probably both PNE platform and mobile app are not sufficiently used at present and have limited resonance in citizens lives. We also searched for similar projects and discovered this website, basically implementing the same idea.
We can finally say that our hope is in a future in which the citizen is guided and supported in important choices also by these kind of integrated data platforms.