– KnowDive Group –

Building an antenna database: data preparation, analysis and information extraction

Document Data:	Reference Persons:
December 14, 2018	A. Benoni, L. Dall'Asta, D. Tormen

Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	08.11.2018	Whole group	Document Created
1.0	19.11.2018	Whole group	Scenario description
1.1	21.11.2018	Luca & Diego	Queries added
1.2	28.11.2018	Luca & Diego	Queries adaptation to table
2.1	30.11.2018	Whole group	Huge works & RapidMiner full analysis
2.2	03.12.2018	Whole group	Report conclusions & references extraction
3.0	07.12.2018	Whole group	Minor updates
4.0	14.12.2018	Whole group	Model evaluation

1   Scenario (and Personas) description

In the Telecommunication Engineering world, the antenna design, synthesis and usage are the main goal for everyone. Basically, a TLC engineer is called to create a system which transmits a signal from one place to another. This procedure involves a non-trivial amount of knowledge and problems to be solved. The choice of the right antenna is crucial, and it is driven by several criteria to be taken into account.

When the antenna selection and purchase is the next step of a TLC project, a solution must be searched in a market, which can be both physical or *e-Commerce*. For instance, a good web platform for antennas and other electronic devices is *RS Components*[1][6]. From this web page, one could search for the word "antenna" and then select the category which comply with the scope. As an example, some of the categories are: multiband, GPS, GSM, telemetry. The main characteristics of each antenna are listed in a clear table that can be well understood without any problem. There can be found the price, the radio-frequency protocols supported, the length, the connectors used, the frequency bands and other relevant attributes, such as the name given by the producer.

However, sometimes those informations are not enough to take a good decision and a further deepening is needed. To do so, one should explore each page related to different antennas and read every data sheet before deciding what is the right choice. This procedure could be dramatically time consuming.

Our goal is to create a single data set, merging the one directly exposed by the web page we considered as example (RS Components) with another created by us with the information contained in the data sheets.

This second data set, for each antenna, contains the name, the price, the polarization, the category, the impedance and other relevant informations ignored by the starting database. Then, using this outcome, it will be possible to filter and submit different and, hopefully, more useful queries for the antenna engineer.

The utility of the output of this project is clear, since it reasonably reduces the amount of time needed by the engineer to search a solution to the antenna problem.

In the following, three different scenarios are created to better understand the working environment.

Notice that for each specific scenario explained, there have been illustrated the main keywords used by the persona for research purpose.

---

[1] https://it.rs-online.com/

1.1 Scenario 1: Expert Engineer

Arianna is an expert TLC engineer. She knows almost everything in this field, since she works everyday in antenna system organization and design. When she needs a new antenna for her project, she searches for specific attributes, such as:

- frequency;

- dimensions;

- gain;

- RF protocol;

- polarization;

- impedance;

- power;

- photograph.

1.2 Scenario 2: Student Engineer

Luca is an university student in TLC. He has to carry out a project in *Eledia Research Center*, where he has to design a TX-RX systems for satellite application. In order to do that, he searches for an antenna supporting a few features, such as:

- frequency;

- power;

- dimension.

1.3 Scenario 3: Cable Guy

Diego is a cable guy, who installs TV and other device (e.g., Wi-Fi) systems in houses. When someone calls him to place an order, he has to buy the stuff he needs for such a work. Basically, he searches for:

- application;

- connector;

- price.

2 Queries description

2.1 Action queries for expert engineer

Persona	Real world action by Persona	System action
Expert engineer	Give me all the antennas that work at 850 MHz.	A page with a list of antennas working at 850 MHz.
Expert engineer	Give me all the antennas that have a maximum dimension of 85 mm.	A page with a list of antennas having a maximum dimension of 85 mm.
Expert engineer	Give me all the antennas that have 8 dB of gain.	A page with a list of antennas having a gain of 8 dB.
Expert engineer	Give me all the antennas that support GSM protocol.	A page with a list of antennas supporting GSM protocol.
Expert engineer	Give me all the antennas that have right hand circular polarization.	A page with a list of antennas having right hand circular polarization.

2.2 Action queries for student engineer

Persona	Real world action by Persona	System action
Student engineer	Give me all the antennas that work at 850 MHz.	A page with a list of antennas working at 850 MHz.
Student engineer	Give me all the antennas that support 10 W power.	A page with a list of antennas supporting 10 W power.
Student engineer	Give me all the antennas that have a maximum dimension of 85 mm.	A page with a list of antennas having a maximum dimension of 85 mm.

2.3 Action queries for cable guy

Persona	Real world action by Persona	System action
Cable guy	Give me all the antennas that have a TV-SAT application.	A page with a list of antennas having a TV-SAT application.
Cable guy	Give me all the antennas that have SMA Male connector.	A page with a list of antennas having SMA Male connector.
Cable guy	Give me all the antennas that have a price range from 10 euros to 50 euros.	A page with a list of antennas having a price range from 10 euros to 50 euros.

2.4 Generalized queries

Persona	Generalized query	Expected answer
Arianna	Give me all the polarizations supported by the antenna Siretta Oscar 20X.	A list of the polarizations of the specific antenna.
Diego	Give me the photo of the antenna Siretta Oscar 20X.	The URL linking to the corresponding image.
Luca	Give me the connectors used by the antenna Siretta Oscar 20X.	A list of the connectors of the specific antenna.
Arianna	Give me all the RF protocols used by the antenna Multiband EAD.	A list of the RF protocols of the specific antenna.
Arianna	Give me the RF protocol used by the frequency 900 MHz.	A single output with the corresponding RF protocol.
Luca	Give me all the frequencies used by the antenna Siretta Oscar 20 X.	A list of the frequencies of the specific antenna.
Arianna	Give me all the frequencies used by the RF protocol Wi-Fi.	A list of the frequencies used by the specific RF protocol.
Arianna	Give me all the antennas that have the polarization linear.	A list of the antennas using the specific polarization.
Diego	Give me the antenna that have the following photo (URL).	A single output with the corresponding antenna.
Arianna	Give me all the antennas that have U.FL connector.	A list of the antennas using the specific connector.
Arianna	Give me all the antennas that have UTMS RF protocol.	A list of the antennas using the specific RF protocol.
Luca	Give me all the antennas that have 1575.42 MHz frequency.	A list of the antennas working at the specific frequency.

3   Dataset(s) extraction/generation and Dataset(s) description

Starting from the scenario described in Sec. 1, we produced our *Entity-Relation Diagram*, in order to model our environment. The ER diagram obtained is shown in Fig. 1 ([8]).
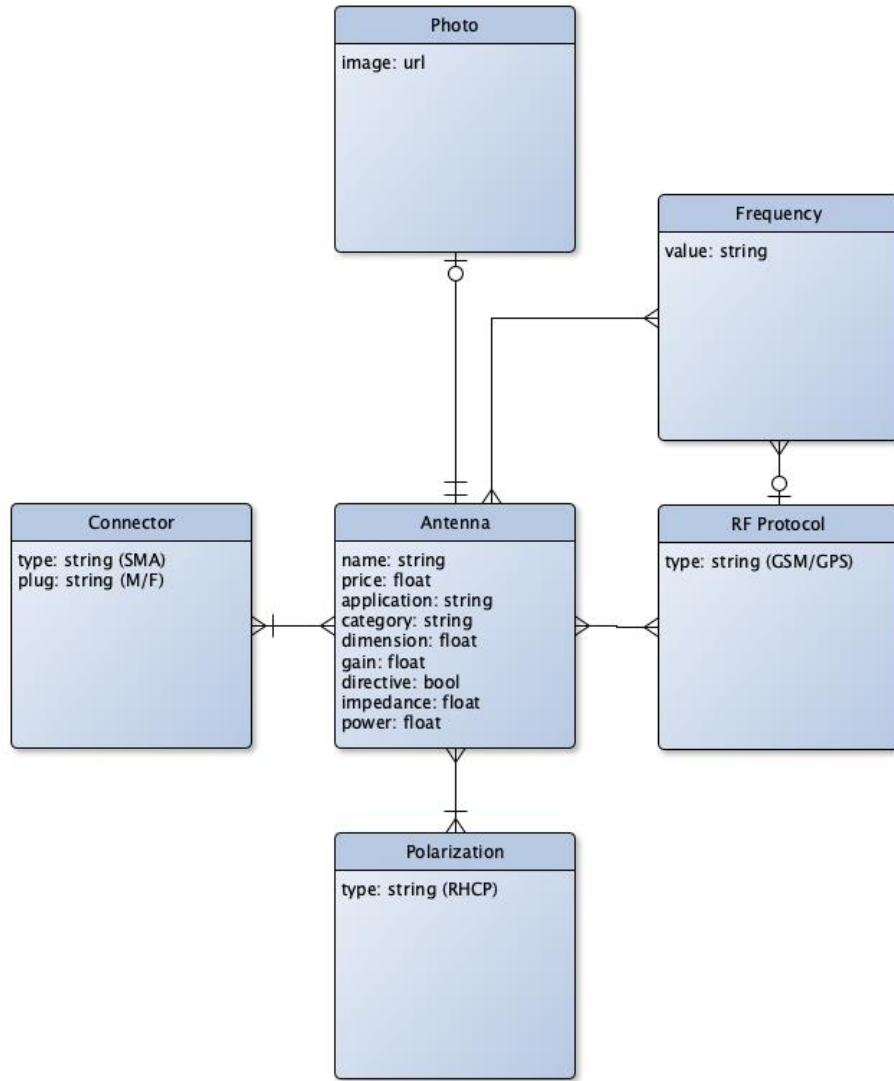

Figure 1: *Entity-Relation Diagram* for the antenna environment.

As said in Sec. 1, two different procedures have been followed in order to create the two datasets that have been used. To sum up: a first dataset has been obtained by the direct reformulation of the e-Commerce web site [6] table after the "*Antenna*" keyword search (see Fig. 2 for detailed explanation), then a second dataset has been obtained entering in each antenna description and reading further the data sheet (see Fig. 3 for detailed explanation).

(a) Category selection

(b) Row selection

Figure 2: **First dataset generation** - After the research, a category must be chosen and then a row must be selected.

(a) Datasheet retrieving

(b) Information adding

Figure 3: **Second dataset generation** - After an antenna has been selected, its datasheet must be read.

A complete diagram of the procedure can be seen in Fig. 4 ([8]).

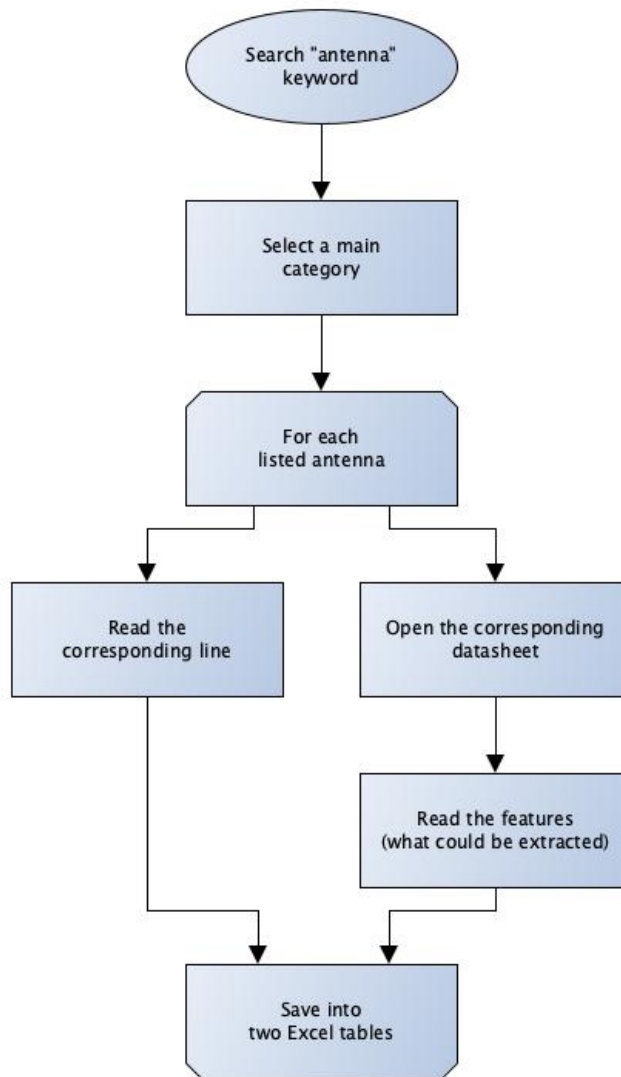Figure 4: **Datasets generation** - The procedure.

A third information source has been created. This contains information about the application of each registered antenna. For the purpose of the third dataset creation, we added it manually identifying the application of each antenna using our experience.

Notice that this part is the most innovative one, since it creates information instead of retrieving it.

4 Dataset(s) cleaning, merging and analysis description

After the creation of the three datasets (described before), we now reach the merging step. Then, we needed to clean the obtained result, in order to create a final dataset with the ordered and complete data. As a final part, we did some statistics and graphical representation to better comprehend our work.

The usage of *RapidMiner* [5] and its integration with *Python* [4, 2] has been introduced for process purposes.

In the following, the whole procedure is explained, while the graphic process of *RapidMiner* is shown in Fig. 5.

4.1 Merging procedure

Before to explain this part, it is important to understand the three datasets we used.

Antenna 1 It contains the datasheet information.

Antenna 2 It contains the e-commerce table information.

Antenna 3 It contains the application information.

Since the *Antenna 3* dataset has been obtained starting from *Antenna 2* dataset, a first merge has been carried out considering those two. Then, the obtained table has been merged with the *Antenna 1* dataset. The merging procedure described until now results in a raw dataset, containing duplicate and unsorted attributes.

Notice that the joining technique we used is the *left* one: this means that the main importance is given to the first input table, while the second table has secondary importance.

4.2 Cleaning procedure

From the raw results obtained from the previous procedure, we needed a smart way to clean our table. For instance, the duplicate columns remaining come from both datasheet and e-commerce table information. More specifically, the duplicate columns are:

1. *Frequency & Banda di frequenza*;

2. *Gain & Guadagno*;

3. *Antenna length & Lunghezza antenna.*

Since a TLC engineer (Arianna person) trusts more the datasheet information than the commercial standard table, we created a custom *Python* script, using the *Pandas* module, which processes the dataset in the following way:

- For each row in duplicate column:
    - If the first[2] column does not have data:
        * Take the second[3] column data

It is clear that a "hierarchical merging" has been performed, subsequently we give a greater priority to the first column (created from the datasheet). Notice that if a value is present in both the columns, than we take into account only the first one; otherwise, if both values are missing, we check it as a missing value.

---

[2]Most important column, the one obtained from the datasheet.
[3]Less important column, the one obtained from the commercial table.
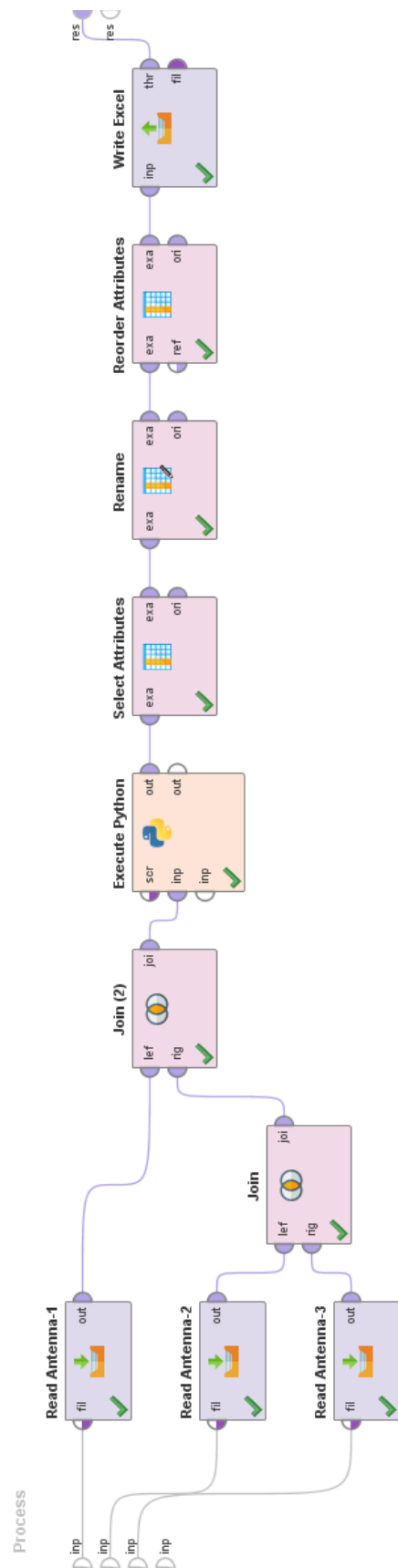
Figure 5: *Dataset cleaning and merging* - The whole procedure consisting of a cascade of existing operators and a custom Python script.

4.3 Analysis

The results obtained by the previous analysis are briefly analyzed here.

Firstly, from a rapid view, it can be seen that the majority of the considered antennas belong to multiband and telemetry category (see Fig. 6).
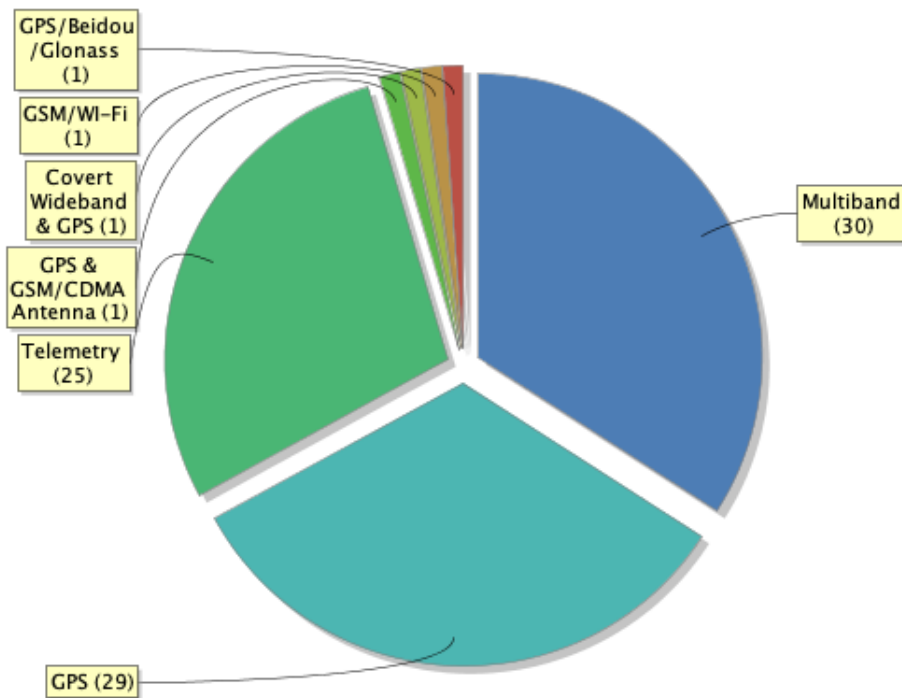
Figure 6: **Analysis** - Considered antennas: distribution of the categories.

What emerged from our statistical studies is in the following:

1. the category influences the power distribution of the antenna (see Fig. 7);

2. the category and the price seem to be linked (see Fig. 8).
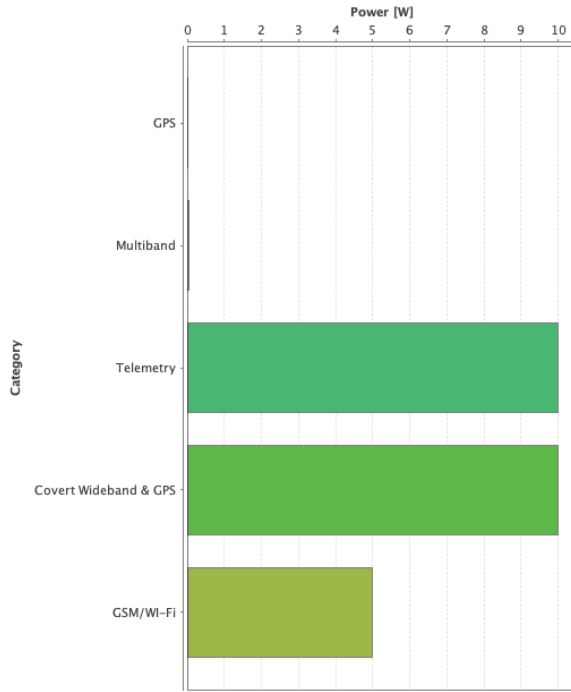
(a) Category vs. minimum power
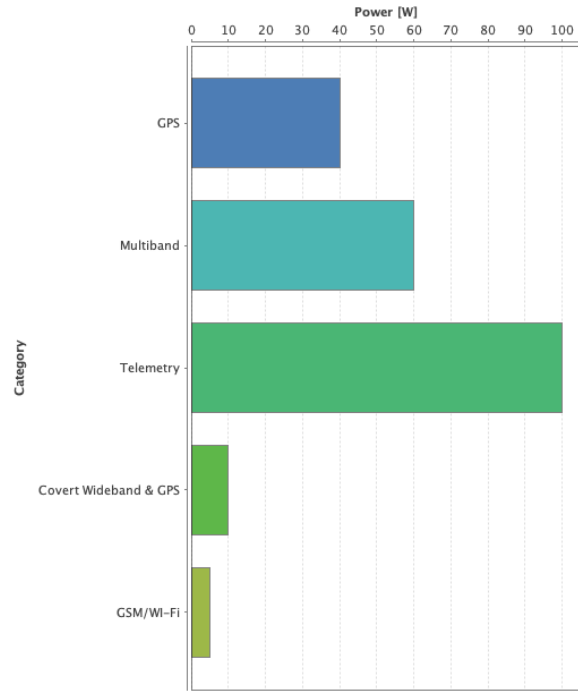
(b) Category vs. maximum power

Figure 7: **Analysis** - Category vs.: (a) minimum working power and (b) maximum working power.

It is clear from Fig. 7 that the most power consuming category of the list considered is the *telemetry*, while the *GSM-WiFi* seems to be the lightest category in terms of average power usage.
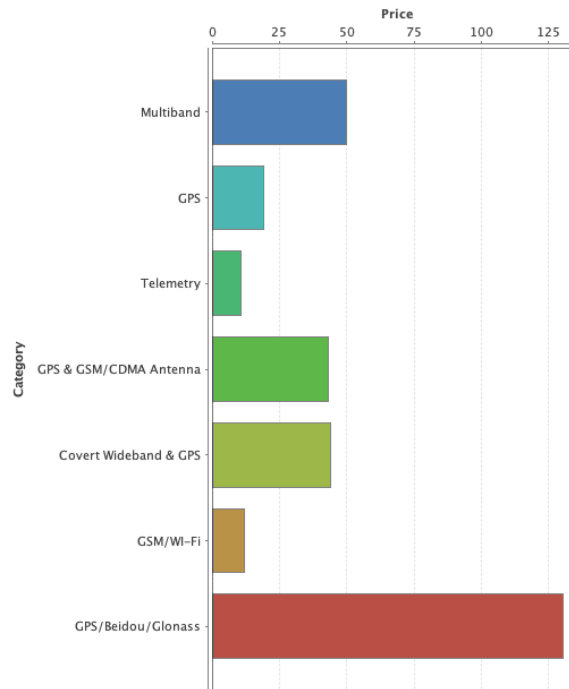
Figure 8: **Analysis** - Category vs. mode price.

It is clear from Fig. 8 that the *Glonass* category is the most expensive one, with a mode (maximum of the histogram) that exceeds the 125 euros. Another good point is that common consumer-antennas (such as *Wi-Fi* and *telemetry*) are low cost products, with a price mode of around 10 euros.

5 Related ontology proposal

Since for our purpose and subject there is no ontology already developed, we created our new one. In order to do so, we started from the *Entity-Relationship* model shown in Fig. 1. The proposed approach takes inspiration from the one illustrated in [9].
We prepared the ontology following a standard approach, while starting from a top level ontology found in [7]:

1. *create the classes and subclasses examples where needed*:

    (a) Mind product

        i. Real product
            A. Photo
        ii. Rules
            A. RF Protocol
                • GPRS protocol
                • GPS protocol
                • GSM protocol
                • WiFi protocol

    (b) Physical object

        i. Artifact
            A. Antenna
                • GPS antenna
                • GSM antenna
                • telemetry antenna
            B. Connector
                • SMA connector
                • type N connector
                • UFL connector

    (c) Property

        i. Frequency
        ii. Polarization
            A. circular polarization
                • left hand circular polarization
                • right hand circular polarization
            B. linear polarization
                • horizontal polarization
                • vertical polarization

2. *create the data properties following the attributes of the ER model*:

    (a) application

(b) category

(c) dimension

(d) directive

(e) gain

(f) image

(g) impedance

(h) name

(i) plug

(j) power

(k) price

(l) type

(m) value

3. *create the object properties following the relationships of the ER model*:

(a) has connector

(b) is connector of

(c) has frequency

(d) is frequency of

(e) has photo

(f) is photo of

(g) has polarization

(h) is polarization of

(i) has protocol

(j) is protocol of

4. *create individuals and test the results.*

The procedure just illustrated led us to a complete ontology for our purpose. A graphical representation is shown in Fig. 9, created with *Protégé* [3] and represented with an online tool [1].

Instead of creating a single class table for each one identified in the ontology, we managed the classification through a *universe* table reordering the dataset. The procedure of sorting has been carried out in the *Cleaning and merging* phase in Sec. 4, where the columns have been sorted using the entity attributes as keys.

Figure 9: *Ontology proposal* - Developed environment for antenna ontology.

6 Related datasets proposal and integration example

The project illustrated before has been carried out using around a hundred antennas. However, it is possible to integrate the three datasets with other individuals, or creating new datasets instead.
The whole procedure works with the structure described before, so there is no difference with the obtained results if the entire structure is followed.

Basically, every e-commerce platform can be used as integration dataset, while searching for the word *Antenna*. A brief example of websites which could be useful for further integration and development is provided in the following.

- *RS Components* (`https://it.rs-online.com/`)

- *Digi-Key* (`https://www.digikey.it/`)

- *Tessco* (`https://www.tessco.com/`)

- *Banggood* (`https://www.banggood.com/`)

- *Distrelec* (`https://www.distrelec.it/`)

- *Amazon* (`https://www.amazon.com/`)

From these websites, a new scraping method can be implemented and tested, in order to upgrade our work with an automatic procedure for data retrieval.

7   Final considerations and open issues

7.1   Model evaluation

In order to evaluate and validate our work, we used the standard method proposed, concerning the coverage and the flexibility of the classes and the properties. We considered only the *Competency Questions - Model* part, since all the others are not feasible with our topic.

In Tab. 1 are shown the obtained values for our model[4] compared with ideal values for each field.

In Tab. 2-4 a more specific evaluation is carried out for each topic of interest.

Value type	Ideal value	Obtained value
Class coverage	0.8	1.0
Class flexibility	0.2	0.0
Property coverage	0.8	-0.5
Property flexibility	0.5	0.6

Table 1: **Model evaluation** - CQ-Model cover and flexibility.

It is clear from the values in 1 that our model has a full class coverage, since we started from the action-queries to obtain the ER diagram (the result is a logical consequence). This result is supported also by the class flexibility value. What we obtained for the property case is nearly an ideal scenario. However, it is clear that something is wrong with the definition of the property coverage metric.

---

[4]Since we did not realized a model (both formal and informal), we relied on the ER diagram.

Schema level	Answer
Does the model including cycles in the class hierarchy?	No
Does the Model uses any polysemous terms for its class or property name?	No
Is Multiple Domain / Range defined for any property?	Yes, but with no conflict
Does any class have more than one direct parent class?	No
Does the Model include multiple classes which have same meaning?	No
Is the class Hierarchy over specified?	No
Does the model use *isA* as a object Property or relation?	No
Does the model have any leaf class for which there is no relation with the rest of the model?	No
Did you use miscellaneous or others as one of the class name?	No
Does the model have any chain of Inheritance in class hierarchy?	Yes
Do all properties have explicit domain and range declarations?	Yes
Does the model have any classes or properties which are not used?	No
Are a collection of elements included as a group in a number of class/attribute?	Yes

Table 2: **Model evaluation** - Schema level evaluation.

Linguistic level	Answer
Does all elements of the model (i.e. class and property) have human readable annotations?	No
Do all elements of the model follow the same naming convention?	Yes

Table 3: **Model evaluation** - Linguistic level evaluation.

Metadata level	Answer
Is provenance information (Creator, Version, Date) available for the final protege model?	Yes
Is provenance information available for any property or class which is taken from some reference standard or ontology?	No

Table 4: **Model evaluation** - Metadata level evaluation.

7.2   Final considerations

As the project converged at its final point, we can highlight some considerations and issues related both to the implementation and the results.

During the dataset creation and merging, we saw that a good amount of fields has been left blank: this problem is not fixable/correctable at the state of the art. The main reason for this is the lack of information which is not easily retrievable.

Another consideration is the convergence point we reached: the real queries have been illustrated but not implemented. So this will be the topic of another work.

The last emerged point is related to the modularity of the results: each process can be easily used and integrated in other environments.

Concluding, we think that this project has been a good source of inspiration for us, since we never worked with databases, queries and data integration.

We found it interesting and useful as it increased our knowledge and awareness.

We also hope that this project would be useful for other people, since we encountered many difficulties along the starting of our TLC engineering career.

7.3   Future works and updates

As said in Sec. 6, this work has been carried out using around a hundred entries. This can be defined as a *pilot project*, since other works can be integrated and improved.

As an example, a taxonomy of applications can be created and integrated, enlarging the vision of our ontology proposal and also the starting dataset.

Another example is explained in Sec. 6, where a scraping procedure is suggested.

References

[1] *Graphical ontology editor*, `http://owlgred.lumii.lv/`.

[2] *Pandas, the python data analysis library*, `https://pandas.pydata.org/`.

[3] *Protégé*, https://protege.stanford.edu/.

[4] *Python*, `https://www.python.org/`.

[5] *Rapidminer documentation*, `https://docs.rapidminer.com/`.

[6] *Rs components*, `https://it.rs-online.com/web/`.

[7] *Wordnet, a lexical database for english*, `https://wordnet.princeton.edu/`.

[8] *yed - graph editor*, `https://www.yworks.com/products/yed/`.

[9] Pasapitch Chujai, Nittaya Kerdprasop, and Kittisak Kerdprasop, *On transforming the er model to ontology using protégé owl tool*, International Journal of Computer Theory and Engineering **6** (2014), 484–489.