Textos para Discussão

TD-IEA n. 42/2020

LUPA'S DATA MODEL FOR DATA WAREHOUSES AND DYNAMIC REPORTS[1]

Daniel Kiyoyudi Komesu[2]

Abril de 2021

[1]Registrado no CCTC: 40/2020.

[2]Economista, Assessor Técnico do Instituto de Economia Agrícola.

ABSTRACT: This text documents the building process of the data model, exposing the software tools and concepts used in the work. The text aims to bring more transparency to society about LUPA data structure and the data model used in the dynamic report publicly available on IEA's website. The goal is to instigate and help researchers to use LUPA as subject or data source for their research papers.

Key-words: data model, business intelligence, LUPA, data warehouse.

1 – INTRODUCTION

On the first quarter of 2020 the Power BI dynamic report of was put on-line at *Instituto de Economia Agrícola* (IEA) website. I was responsible for building the data model used in that report and I will document in this text the build process and the rationale behind the work.

Designing a data model[3] for Business Intelligence (BI) software requires a deep process of introspection into the data architecture being modeled. Datasets with large number of variables, several many-to-many cardinality relationships, and hierarchical levels can be hard to model.

Levantamento Censitário das Unidades de Produção Agropecuária do Estado de São Paulo (LUPA) is a census survey of São Paulo state's rural properties (2009; 2019). This is an example of a dataset with a large quantity of variables that can be filtered and correlated in complex combinations. Thus, planning the data model for Business Intelligence (BI) reports was essential to deliver a quality product.

This text aims to bring more transparency to society about LUPA's data structure and the data model used in the dynamic report (available on-line)[4]. The text documents the building process of the data model, exposing the software tools and concepts used in the work.

My final goal is to instigate and help researchers to use LUPA as a subject or a data source for their research papers. I think that exposing its data structure is the best way to reach these goals.

2 – TOOLS AND REFERENCES

To process data, I used Python programming language[5]. All the data cleaning and regularization work was possible thanks to the use of the Pandas package[6]. Microsoft Excel spreadsheet software was not an option as there is a limitation in the number of rows that can be loaded, poor performance issues and lack of data processing reproducibility.

Although Power BI software offers data cleaning and transformation tools – through Data Analysis Expressions (DAX)[7] and M formula language[8], I wanted to avoid vendor lock in. So, all processed data outputs were in Comma Separated Values (CSV) file format. In case the institution decides to change the BI software, the data can be easily imported in other software tools. All the logic and calculated measures can be easily replicated as well.

---

[3]For a more detailed description of what is a data model see Wikipedia contributors (2020a).
[4]The Power BI report is available at: http://www.iea.agricultura.sp.gov.br/out/bilupa.php
[5]Visit Python Software Foundation website at https://www.python.org/ for more information.
[6]Visit Pandas website at https://pandas.pydata.org/ for more information.
[7]Visit Microsoft's DAX documentation at https://docs.microsoft.com/pt-br/dax/ for more information.
[8]Visit Microsoft's M formula language documentation at https://docs.microsoft.com/en-us/powerquery-m/ for more information.

In the data warehouse field, several types of database schema are used for storing data. A common choice is the star schema, illustrated by Figure 1.

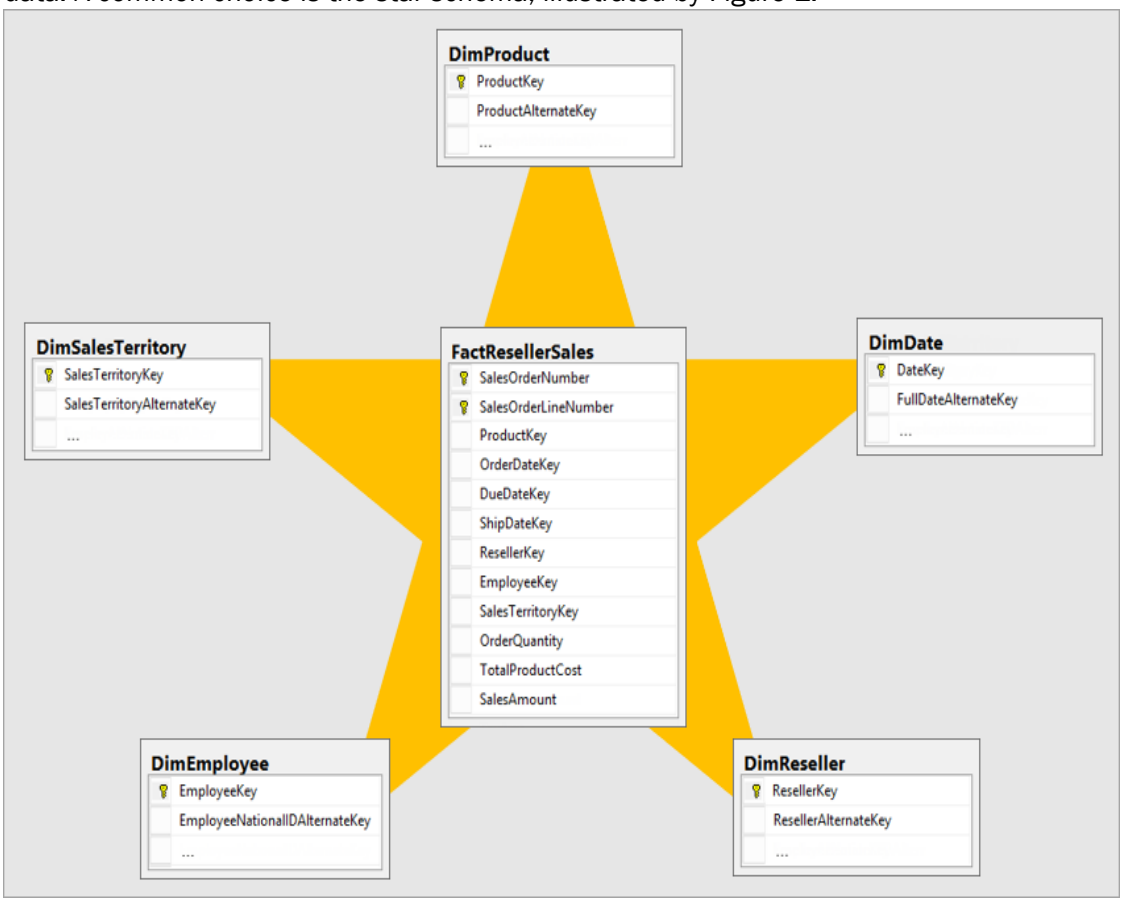

Figure 1 – Star schema.
Source: MICROSOFT POWER BI DOCS (2019).

There are many other schemas commonly used in multidimensional analysis models, some examples are snowflake schema, star cluster schema, galaxy schema etc. Başaran (2005) compared various database schema used in data warehouses, pointing pros and cons for each one.

3 – THE DATA MODEL

The first step to build a data model is to know what problem we want to solve and the desired features of the model.

While designing the model, I wanted the following features:

- Flexible and generic (will work with future LUPA versions)
- Maintainable and extensible in the long run
- Visualization of correlations between multiple variables
- Fast performance

The next step is to understand the data structure. I start by identifying dimensions and facts in the data. LUPA uses UPA (*Unidade de Produção Agropecuária* - Agricultural Production Unit) as its primary unit, so I it as my starting point.

I named this fact table in the model as "InfoUPA". This table is built by joining all information which has a one-to-one relationship with one UPA. For example, one UPA has only one value for the number of its owners[9]. On the other hand, the total area of a property looks like in the same situation. But as this information can be derived from the sum of the land occupation, I left it out from the fact table.

After determining the fact table, I started to look at the dimensions to create tables that will filter the data. Turns out LUPA presents lots of variables. To avoid wide tables, I used row modeling[10] to obtain narrow tables, which offers better performance for data-bases and BI software and allows easy filtering by attributes.

In addition to better performance, row modeled tables allow flexibility for adding and dropping attributes. This feature is ideal in the LUPA data model, as it is likely that the set of attributes will change in next censuses. Without the necessity of change in tables structures, the model can be easily extended, and this reduces the need of future reworks.

LUPA survey groups variables in 9 topics (listed in Annex 1). Two topics (machines & equipment, improvements & facilities) were combined because of its data structure's similarities. Then, 8 additional fact tables were created.

The downside of row modeling is the creation of many-to-many relationships. Alt-hough it is possible to connect two fact tables directly in Power BI and use the columns to filter the data, this would be very space inefficient. Instead, for each one of the 8 newly created fact tables, I normalized the data and created a new dimension table. I ended up converting these fact tables into associative entities[11] (also known as bridge tables) to connect the UPA facts table with the dimension tables.

The table for techniques added another level of details to the model because it refers to information about crops and techniques combined. This table must support being filtered by crop, technique and UPA at the same time. This case was treated differently because of this idiosyncrasy. If related directly to the dimensions, this would create a cir-cular relationship problem[12]. Because the bridge and InfoUPA tables must filter each other (bidirectional relationship) to make multiple correlations between dimension, an uncareful designed relationship would create an ambiguity problem (Figure 2).

[9]As the report does not need information of each individual owner of a property, only the quantity, the number of owners per property is enough.
[10]A table where facts about each entity is stored in multiple rows rather than multiple columns. See Wikipedia contributors (2020b).
[11]See Wikipedia contributors (2020c) for detailed information about association entities.
[12]See Truong (2009) for more information about circular relationship problems in database design.

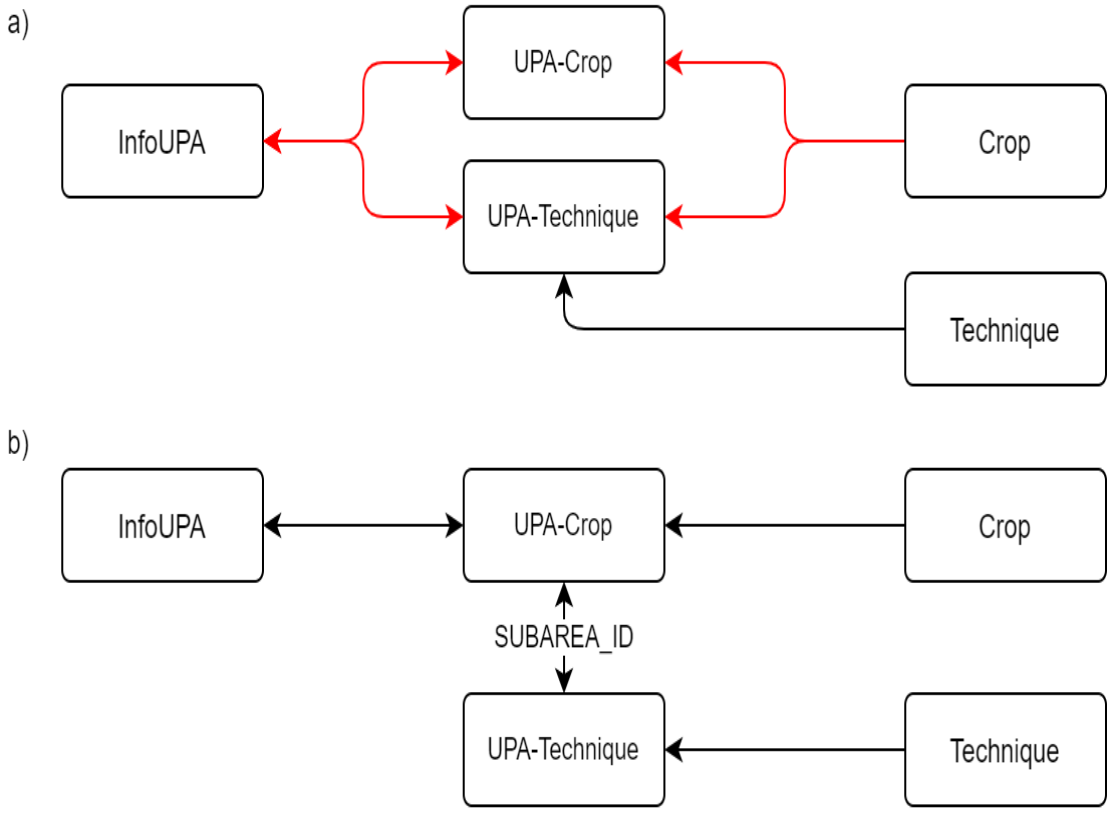

Figure 2 – Circular relationship problem (a) and the design implemented to solve it (b).
Source: The author.

The workaround for this problem was to create an identifier for each crop in a property, named SUBAREA_ID (Figure 2b).

To connect the InfoUPA table to bridge tables I created an identifier code from the concatenation of three other identifier codes: LUPA version code, municipality code and UPA code within the municipality. The resulting code, named UPA_ID, uniquely identifies an UPA across LUPA versions and municipalities. This step was necessary (to simplify joins and) because BI software in general does not support complex joins, like composite primary key joins.

A simplified conceptual diagram of the data model is shown in Figure 3 (complete version in Annex 1).

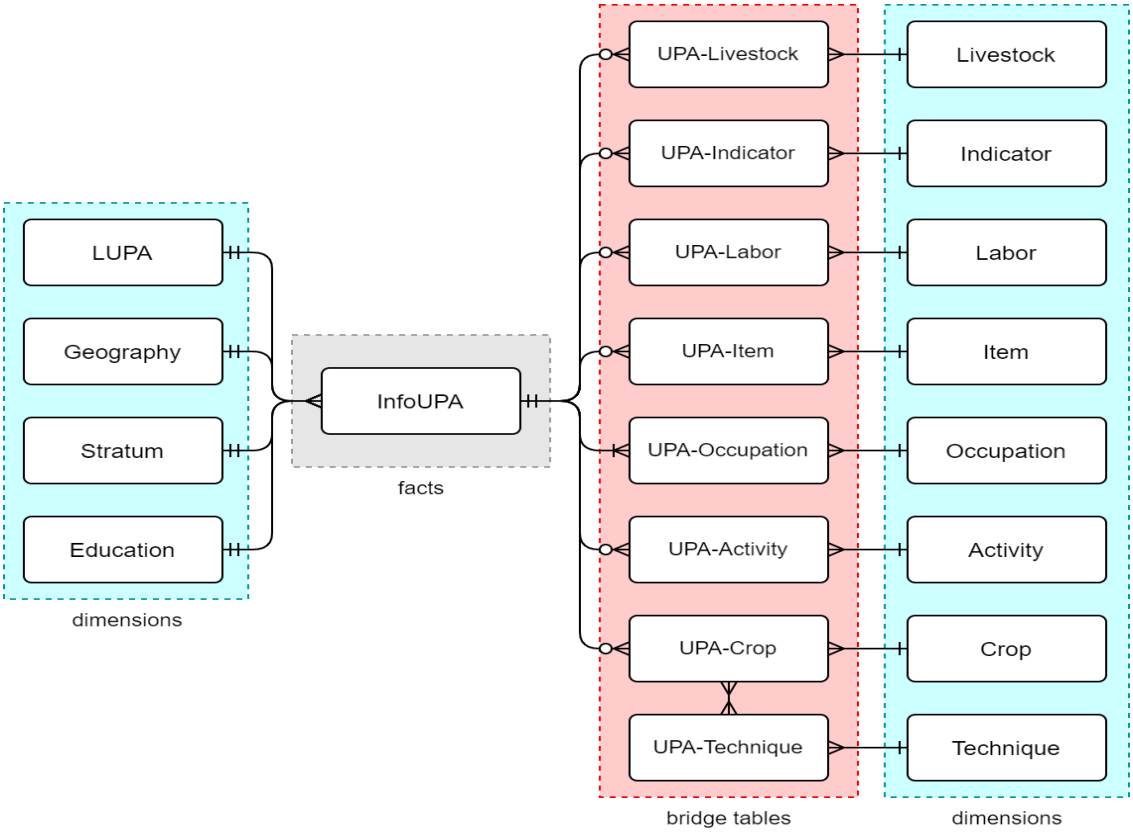

Figure 3 – Simplified Entity Relationship Diagram of LUPA data model.
Source: The author.

4 – CONCLUDING REMARKS

The data model described here is used in the Power BI report available at the institution's website. The challenges and experiences gained in the process were valuable to the institution and I think that these lessons will be appreciated by the readers as well. Also, the public scrutiny of the process, pointing issues or errors, can give directions for improvement not seen by the institution's staff.

RECOMMENDED READING

Ballard, C.; Herreman, et al. **Data Modeling Techniques for Data Warehousing**. San Jose, California: International Business Machines (IBM) Corporation: International Technical Support Organization, fev 1998.

REFERENCES

BAŞARAN, B. P. **A comparison of data warehouse design models**. Ankara: Atilim University, jan 2005.

MARTINS, V. A., et al. **LEVANTAMENTO CENSITÁRIO POR UNIDADES DE PRODUÇÃO AGROPECUÁRIA 2016/17**. São Paulo: SAA: IEA: CDRS, 2019. Available at: <http://www.cdrs.sp.gov.br/projetolupa/estudos_lupa/IE-LUPA-2016-2017.pdf>. Access on: 2020-08-24 18:58 UTC.

MICROSOFT POWER BI DOCS. **Understand star schema and the importance for Power BI**. Redmond, Washington: Microsoft Corporation, set 2019. Available at: <https://docs.microsoft.com/en-us/power-bi/guidance/star-schema>. Access on: 2020-08-24 18:50 UTC.

MICROSOFT POWER BI DOCS. **Many-to-many relationship guidance**. Redmond, Washington: Microsoft Corporation, fev 2020. Available at: <https://docs.microsoft.com/en-us/power-bi/guidance/relationships-many-to-many>. Access on: 2020-08-24 18:53 UTC.

SÃO PAULO (Estado). Secretaria de Agricultura e Abastecimento do Estado de São Paulo. Instituto de Economia Agrícola. Coordenadoria de Desenvolvimento Rural Sustentável. **Projeto LUPA 2007/08**: Censo Agropecuário do Estado de São Paulo. São Paulo: SAA: IEA: CDRS, 2009.

SÃO PAULO (Estado). Secretaria de Agricultura e Abastecimento do Estado de São Paulo. Instituto de Economia Agrícola. Coordenadoria de Desenvolvimento Rural Sustentável. **Projeto LUPA 2016/17**: Censo Agropecuário do Estado de São Paulo. São Paulo: SAA: IEA: CDRS, 2019.

TRUONG, PHAM DINH. **Prevent Circular References in Database Design**. CodeProject, 2 ago 2009. Available at: <https://www.codeproject.com/Articles/38655/Prevent-Circular-References-in-Database-Design>. Access on: 2020-08-31 15:51 UTC.

WIKIPEDIA CONTRIBUTORS. **Data model**. Wikipedia, The Free Encyclopedia: 20 jul 2020a. Available at: <https://en.wikipedia.org/w/index.php?title=Entity%E2%80%93attribute%E2%80%93value_model&oldid=968692062>. Access on: 2020-08-24 19:20 UTC.

WIKIPEDIA CONTRIBUTORS. **Entity–attribute–value model**. Wikipedia, The Free Encyclopedia: 11 ago 2020b. Available at: <https://en.wikipedia.org/w/index.php?title=Data_model&oldid=972313763>. Access on: 2020-08-24 19:15 UTC.

WIKIPEDIA CONTRIBUTORS. **Associative entity**. Wikipedia, The Free Encyclopedia: 11 mai 2020c. Available at: <https://en.wikipedia.org/w/index.php?title=Associative_entity&oldid=956130476>. Access on: 2020-08-25 19:37 UTC.

ANNEX 1 - COMPLETE LUPA'S DATA MODEL

This annex presents a brief description of LUPA's data structure and the full data model described in the text. See Annex 1 in Martins et al. (2019) to get a detailed description of LUPA's attributes and variables.

Unidade de Produção Agropecuária (UPA) - Agricultural Production Unit - is the census' primary unit. In LUPA dataset, each property can have different combinations of crops and techniques. At property level, there are many attributes and variables of various types, from binary to scalar and categorical (exclusive and non-exclusive). The total area of an UPA is equal to the summation of its land occupation.

The following topics are present in LUPA dataset:

- **livestock**: livestock type and its quantity.
- **crops**: crop type and its total area in the property.
- **techniques**: area for each combination of technique and culture.
- **labor**: labor type and its quantity in the property.
- **land occupation**: area of each land occupation in the property (the summation equals the property's total area).
- **rural economic activities**: categorical non-exclusive indicator if the property has other economic activities.
- **socioeconomic indicators**: categorical non-exclusive indicator if the property has some characteristic.
- **machines and equipment**: type and quantity of machines and equipment in the property.
- **improvements and facilities**: type of improvements and facilities and its quantity in the property.

Table A.1.1 – Dimensions in the data model

Dimension	Cardinality
Rural economic activities	9
Livestock	27
Crops	214
Tecniques	9
Socioeconomic indicators	32
Labor	14
Machines and equipment / improvements and facilities	80
Land occupation	8

Source: The author.

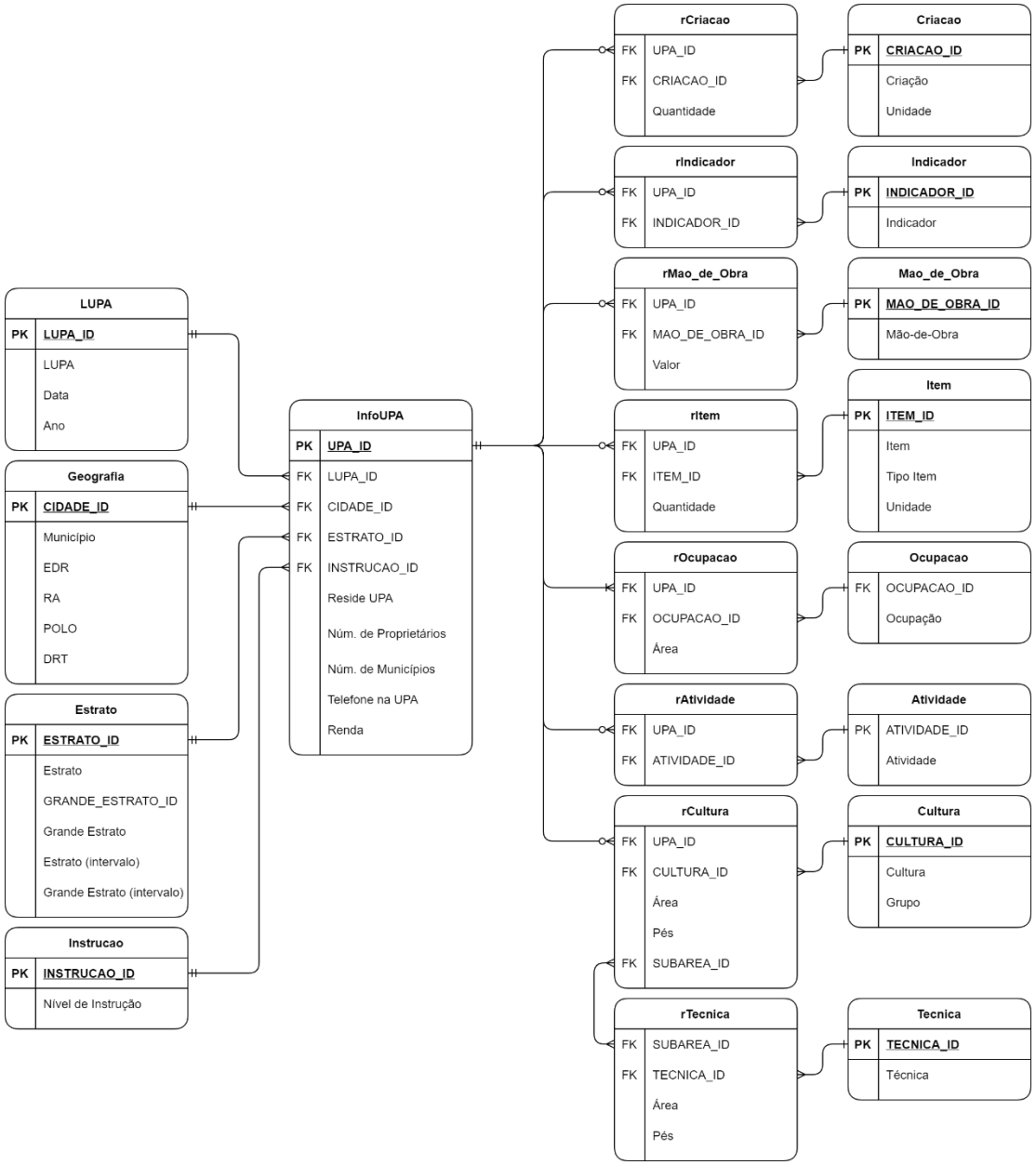

Figure A.1.1 – Complete LUPA data model.
Source: The author.

Recebido em 31/08/2020.Liberado para publicação em 09/11/2020.