

Figure 1. The schema of the Content Database.

In the following, we describe the main tables in the Content Database, i.e. the ones which were used in the Use Cases. We omit tables which were left unused.

An important collection of texts for the project was the one of **Statistics Explained (SE) Glossary articles**. The relevant tables can be seen in the middle of Figure 1:

* Table **dat\_glossary** contains the unique IDs and the definitions of the SE Glossary articles. Their titles and links are found in the records with matching IDs in table *dat\_link\_info*, which contains all titles and links to the scraped elements from Eurostat or other internal and external sources. The values of *resource\_information\_id* for these records in this latter table, are all set to one, which corresponds to ‘Eurostat’ in table *dat\_resource*.
* Table **dat\_further\_info** is connected with *dat\_glossary* through its foreign key *glossary\_id* and points to the further information sections in the articles, through the table *dat\_link\_info*.
* Tables **dat\_related\_concepts**, **dat\_sources**, **dat\_statistical\_data** and **dat\_redirections** have similar connections and point to related content (related articles, sources, statistical data and redirections), again through table *dat\_link\_info*.
* Table **GL\_articles\_categories** is similarly linked to *dat\_glossary* and contains the (generally more than one) categories for each article.

The left side of the same figure shows the tables with the content of **Eurostat's Concepts and Definitions Database** in RAMON[[1]](#footnote-1).

* Table **dat\_estat\_glossary** contains the terms, definitions, context, remarks and RAMON categories of these summaries.
* The linked tables **dat\_estatg\_stat\_unit** and **dat\_estatg\_measurement\_unit** point to some labels of statistical and measurement units, respectively, in the look-up tables **mod\_statistical\_unit** and **mod\_measurement\_unit**.
* The RAMON categories are in the look-up table **mod\_ramon\_category**.

Finally, the left side of Figure 1 also shows some isolated tables:

* **Eurostat\_OECD\_themes** with a correspondence of Eurostat themes and sub-themes to OECD’s themes, the latter in strings separated by semicolons. If there are no corresponding OECD’s themes, the strings are empty.
* **Norm\_NPs** contains a list of 49 000 characteristic noun phrases found through semantic analysis (and then, manual filtering) of content from both Eurostat and OECD. These are the results of enrichment runs, described in section 2.3.2.
* **OECD\_Glossary** contains all content (terms, URLs, definitions, themes, related terms, related URLs, source publications) from OECD’s Glossary of Statistical Terms[[2]](#footnote-2).
* **dat\_all\_datasets** contains the full information from the parsing of the contents of Eurostat’s database from file *table\_of\_contents.xml* in the bulk download facility[[3]](#footnote-3).

In an analogous structure to the one for the SE Glossary articles, the right side of Figure 1 shows the tables with the content of the equally important collection of **SE articles**.

* Table **dat\_article** contains the context and data sources, together with a 0-1 flag *background\_article*, which indicates if the article is a background one or not. Their titles and links are found in the records with matching IDs in table *dat\_link\_info*, which, as mentioned, contains all titles and links to the scraped elements from Eurostat or other sources. The values of *resource\_information\_id* for these records in this latter table, are again all set to one, which corresponds to ‘Eurostat’ in table *dat\_resource*.
* The individual sections with their titles and contents are in table **dat\_article\_paragraph**, and are linked to table *dat\_article* through *article\_id* matching *id* in *dat\_article*. A 0-1 flag *abstract* in *dat\_article\_paragraph* indicates whether a section is an abstract or not.
* Pointers to content from paragraph figures are in the connected **dat\_paragraph\_figure** table, which contains a *link\_id* with the IDs of articles from table *dat\_link\_info*.
* Similar pointers are in table **dat\_article\_shared\_link**, which includes a key *article\_division\_id* pointing to the look-up table **mod\_article\_division** with the kind of information in the shared link (Excel file, other articles, tables, database, dedicated sections, publications, etc.).
* Table **SE\_articles\_categories** is linked to *dat\_article* and contains the (generally more than one) categories for each article.

Finally, tables **tm\_topics**, **tm\_articles\_to\_topics**, **tm\_values** and **tm\_terms** in the right side of Figure 1 contain results from some enrichment runs, namely topic modelling. The other tables shown in the right side of Figure 1 have limited information on datasets and codes used (restricted to the topic of Quality of Life) and are not discussed, because they were not used in the Use Cases.

1. <https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM&StrGroupCode=CONCEPTS&StrLanguageCode=EN> [↑](#footnote-ref-1)
2. <https://stats.oecd.org/glossary/> [↑](#footnote-ref-2)
3. <https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing> [↑](#footnote-ref-3)