|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sector_header | | | | | | |
|  | **NLP4StatRef: Methodological support on advanced methods for accessing, ingesting and linking textual information using semantic analysis and natural language processing** | | | | | |
| partner_flag1  partner_flag2 | Content database and Knowledge database Documentation | | | | | |
|  | | | | | |
|  | | | | | |
| Specific contract N° 000068  Under  Framework Contract N° 2018.0088, Lot 1: Methodological support | | | | | |
|  |  | evaluation_03 |  | evaluation_01 |  |  |
|  | January 2022 | | | | | |

Contents

[1 Introduction 3](#_Toc95690410)

[2 Content and Knowledge Workflow 4](#_Toc95690411)

[2.1 Data 5](#_Toc95690412)

[2.1.1 Data definition 5](#_Toc95690413)

[2.1.2 Data categorization 6](#_Toc95690414)

[2.1.3 List of processed data 7](#_Toc95690415)

[2.2 Content database 8](#_Toc95690416)

[2.2.1 Content database model 9](#_Toc95690417)

[2.2.2 How to update the content? 9](#_Toc95690418)

[2.3 Knowledge database 10](#_Toc95690419)

[2.3.1 Reused Standards 10](#_Toc95690420)

[2.3.2 Knowledge database model 14](#_Toc95690421)

[2.3.3 How to update the knowledge? 22](#_Toc95690422)

[3 How to add new data? 22](#_Toc95690423)

[3.1 Requirements and Specifications 23](#_Toc95690424)

[3.1.1 Data description 23](#_Toc95690425)

[3.1.2 Data Modelling 25](#_Toc95690426)

[3.1.3 Mapping: between data and the Content and Knowledge databases 25](#_Toc95690427)

[3.2 Data ingestion 27](#_Toc95690428)

[3.2.1 Data collection 27](#_Toc95690429)

[3.2.2 Content database population 28](#_Toc95690430)

[3.3 Knowledge instantiation 29](#_Toc95690431)

[3.3.1 Notebook creation 30](#_Toc95690432)

[3.3.2 Links about creation of SPARQL queries 30](#_Toc95690433)

[4 Articles 31](#_Toc95690434)

[4.1 Statistics Explained content & OECD Topics 31](#_Toc95690435)

[4.1.1 Requirements and Specifications 31](#_Toc95690436)

[4.1.2 Content database model 33](#_Toc95690437)

[4.1.3 Knowledge database model 33](#_Toc95690438)

[4.2 Experimental statistics 34](#_Toc95690439)

[4.2.1 Requirements and Specifications 34](#_Toc95690440)

[4.2.2 Content database model 34](#_Toc95690441)

[4.2.3 Knowledge database model 35](#_Toc95690442)

[4.3 News & Euro Indicators 35](#_Toc95690443)

[4.3.1 Requirements and Specifications 35](#_Toc95690444)

[4.3.2 Content database model 36](#_Toc95690445)

[4.3.3 Knowledge database model 36](#_Toc95690446)

[5 Statistical Data 37](#_Toc95690447)

[5.1 Statistical Data Report 37](#_Toc95690448)

[5.1.1 Requirements and Specifications 38](#_Toc95690449)

[5.1.2 Content database model 38](#_Toc95690450)

[5.1.3 Knowledge database model 38](#_Toc95690451)

[5.2 Statistical Datasets 39](#_Toc95690452)

[5.2.1 Requirements and Specifications 39](#_Toc95690453)

[5.2.2 Content database model 40](#_Toc95690454)

[5.2.3 Knowledge database model 40](#_Toc95690455)

[5.3 Statistical Reference metadata 41](#_Toc95690456)

[5.3.1 Requirements and Specifications 42](#_Toc95690457)

[5.3.2 Content database model 42](#_Toc95690458)

[5.3.3 Knowledge database model 42](#_Toc95690459)

[6 Resources 43](#_Toc95690460)

[6.1 Authority Lists ¤ Resource Information & Resource Type 43](#_Toc95690461)

[6.1.1 Requirements and Specifications 43](#_Toc95690462)

[6.1.2 Content database model 44](#_Toc95690463)

[6.1.3 Knowledge database model 45](#_Toc95690464)

[6.2 Code lists 45](#_Toc95690465)

[6.2.1 Requirements and Specifications 46](#_Toc95690466)

[6.2.2 Content database model 46](#_Toc95690467)

[6.2.3 Knowledge database model 47](#_Toc95690468)

[6.3 Eurostat Glossary 47](#_Toc95690469)

[6.3.1 Requirements and Specifications 47](#_Toc95690470)

[6.3.2 Content database model 48](#_Toc95690471)

[6.3.3 Knowledge database model 48](#_Toc95690472)

[6.4 Glossaries ¤ OECD Glossary & SE Glossary 48](#_Toc95690473)

[6.4.1 Requirements and Specifications 48](#_Toc95690474)

[6.4.2 Content database model 49](#_Toc95690475)

[6.4.3 Knowledge database model 49](#_Toc95690476)

[6.5 Taxonomies 50](#_Toc95690477)

[6.5.1 Requirements and Specifications 51](#_Toc95690478)

[6.5.2 Content database model 51](#_Toc95690479)

[6.5.3 Knowledge database model 51](#_Toc95690480)

[6.6 Terminology Lexicon 52](#_Toc95690481)

[6.6.1 Requirements and Specifications 52](#_Toc95690482)

[6.6.2 Content database model 52](#_Toc95690483)

[6.6.3 Knowledge database model 53](#_Toc95690484)

# Introduction

This document aims to provide a step-by-step approach on how to treat & load content onto both the Content database and the Knowledge database.

Section 2 is about understanding the Content and Knowledge databases and section 3 is an instruction manual on how to load content onto them.

Sections 4 to 6 are about data already available into the Content and Knowledge databases. The work done could be reused or adapted for integrating new data, new source.

# Content and Knowledge Workflow

On the one hand, the project is about making the content produced by Eurostat more accessible, and on the other hand, it is about creating bridges with external sources of information that might add value and expand the possibilities. In order to achieve this goal, 4 use cases were determined by Eurostat.

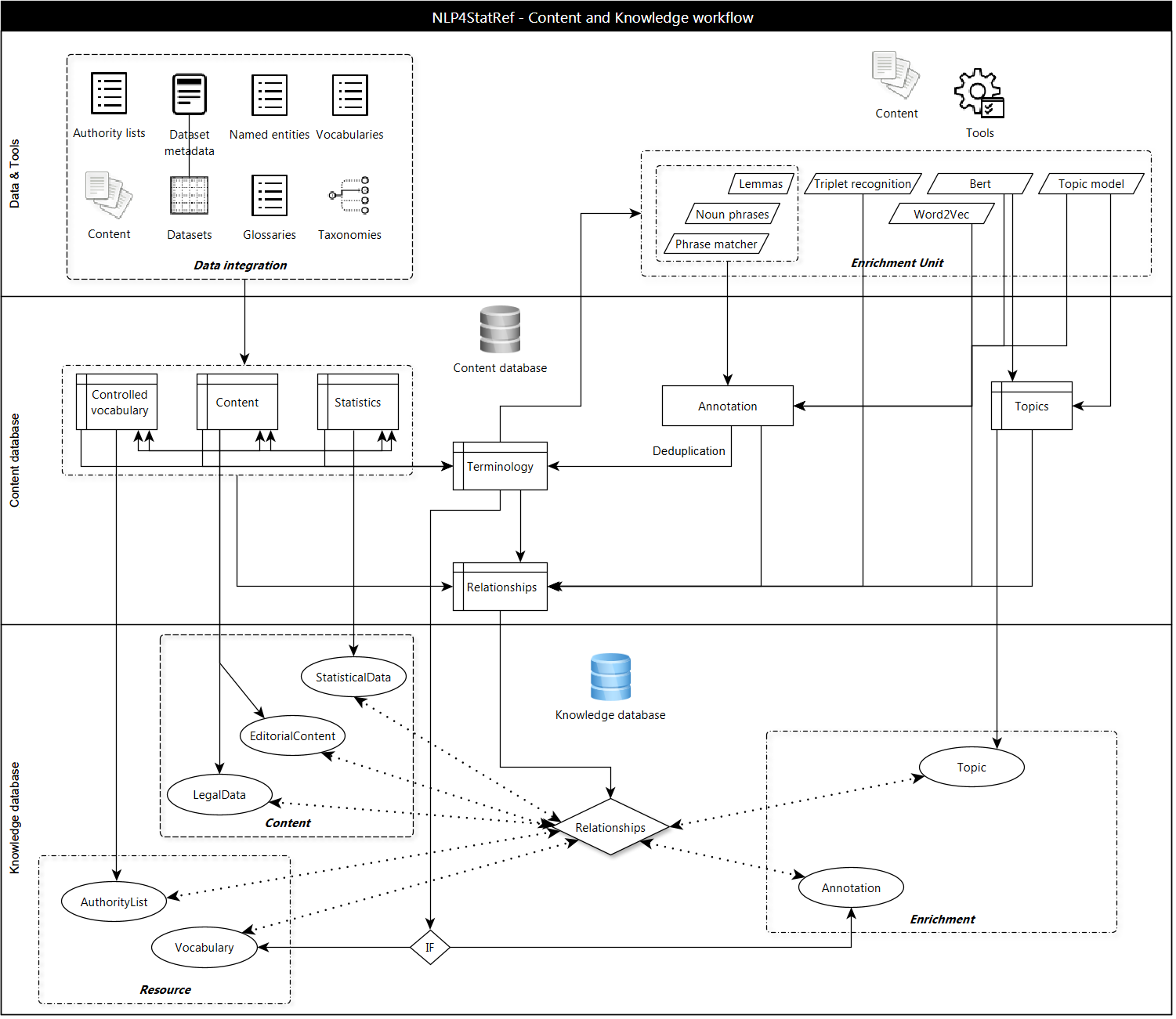
The implementation of the 4 use cases of the project and the deployment of the semantic analysis tools requires building corpora of documents and sets of concepts. This construction consists of successive stages represented in the following diagram: 

Figure 1 - Content and Knowledge workflow

This diagram shows that the heart of the Content and Knowledge databases is made of terminology and relations. The heart of the content database will be used to populate the heart of the knowledge database.

There is a space in which the concepts are managed and supervised and on which a quality check is operated. This space is enriched with the integration of external data via a controlled process: the treatments done to internal data are reproduced to external data, it eliminates duplicate concepts and it instantiates.

The populating process makes the difference between resource and annotation:

* the resources are supervised and automatically bring a relevance rate on the results;
* the annotations are enriched, the concepts put together are generating new knowledge and new relations.

The set of new tools generates new knowledge within the knowledge database.

## Data

Different types of data have been treated within the scope of the project, that can be found below:

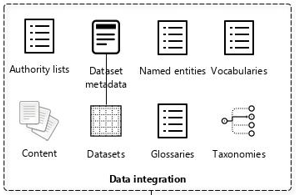


Figure 2 - Data integration

### Data definition

|  |  |
| --- | --- |
| Data | Definition |
| Authority Lists | An authority list is a controlled vocabulary used to name particular entities consistently. It is structured in separate files dedicated to different types of entity. Types are codes, proper names, metadata or terms. |
| Content | Content is for instance any type of written documents or publications: articles, book, legal information, news… |
| Dataset | A dataset is a collection of information about the same topic.  In the NLP4StatRef context, this is a statistical dataset. |
| Dataset metadata | Dataset metadata is data providing information about one or more aspects of a dataset. |
| Glossaries | Glossaries are an alphabetical list of words relating to a specific subject enriched by definitions, related content…. |
| Named entities | A named entity is a real-world object, such as a person, a location, an organization (such as the OECD, The Queen Elizabeth II, and so on). |
| Taxonomies | A taxonomy is an organization and classification of information following a pattern of specific rules and structure. |
| Vocabularies | A vocabulary is a collection of words and noun phrases of a language or a collection of concepts characterizing a specific domain. It is a resource that comprises both glossaries and taxonomies. |

### Data categorization

These data can be classified into two main categories:

* Content would involve statistical data (datasets and metadata for instance), legal data and editorial content.
* Resource will gather two subtypes: authority lists and vocabulary.

A third category of data, resulting from semantic analysis tools completes the annotation corpus: Enrichment which would involve topic modeling and annotation.

This piece of information is sum up into the diagram below:

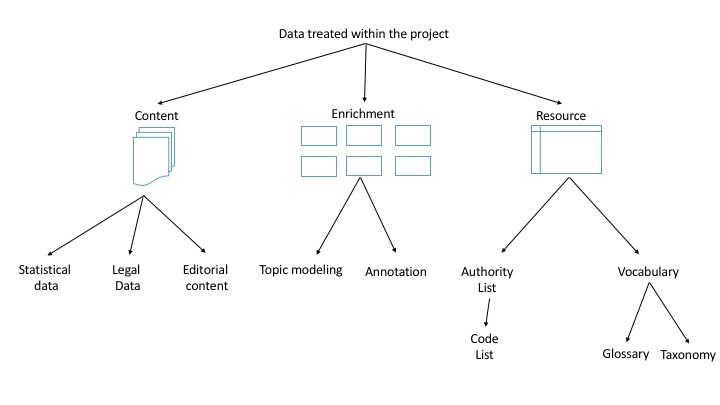


Figure 3 - Data categorization

### List of processed data

Here is a list on all data that had been analyzed for this project:

| Data | Data name | Data type | Resource type | Knowledge class |
| --- | --- | --- | --- | --- |
| Background articles  Statistics Explained website | SE Background articles | Editorial content | methodological-article | Article |
| Categories  Statistics Explained website | Eurostat Category | Vocabulary | taxonomy  concept | Taxonomy |
| Data navigation Tree  Statistics Explained website | Eurostat Database Tree | Vocabulary | taxonomy  concept | Taxonomy |
| Code lists  Ramon Metadata | Code Lists | Controlled vocabulary | code-list  code-list-item | CodeList |
| Euro Indicators | Euro Indicators | Editorial content | news | Article |
| Eurostat concept and definition | Eurostat Glossary | Vocabulary | glossary-scheme  glossary-term | Glossary |
| Eurostat databases | Statistical dataset | Statistical data | statistical-dataset | StatisticalDataset |
| Experimental Statistics | Experimental Statistics | Editorial content | statistical-article | Article |
| Family  Ramon metadata | Eurostat Family | Vocabulary | taxonomy  concept | Taxonomy |
| Glossary articles  Statistics Explained website | SE Glossary | Vocabulary | glossary-scheme  glossary-term | Glossary |
| News | News | Editorial content | news | Article |
| OECD Glossary | OECD Glossary | Vocabulary | glossary-scheme  glossary-term | Glossary |
| OECD Topic | OECD Topic | Editorial content | statistical-article | Article |
| Reference metadata  Eurostat databases | Statistical Reference metadata | Statistical data | statistical-reference-metadata | Statistical Reference metadata |
| Resource Information | Resource Information | Controlled vocabulary | named-entity-list  named-entity | AuthorityList |
| Resource Type | Resource Type | Controlled vocabulary | authority-list  entity | AuthorityList |
| Statistics Explained articles  Statistics Explained website | SE articles | Editorial content | statistical-article | Article |
| Term Extension  Ramon metadata | Eurostat Term Extension | Vocabulary | taxonomy  concept | Taxonomy |
| Terminology  Noun phrases Lexicons | Terminology Lexicon | Vocabulary | lexicon  entity | Vocabulary |
| Themes  Eurostat websites | Eurostat Theme | Vocabulary | taxonomy  concept | Taxonomy |
| Themes  OECD websites | OECD Theme | Vocabulary | taxonomy  concept | Taxonomy |

## Content database

The content database will store the data on which will be performed the treatments at a later stage of the process. This section aims at understanding its structure and how to load content on it.

### Content database model

The content database has been designed to be able to welcome all the data that has been collected within the project (scraped of downloaded). Each data has its right place, that you can find below:

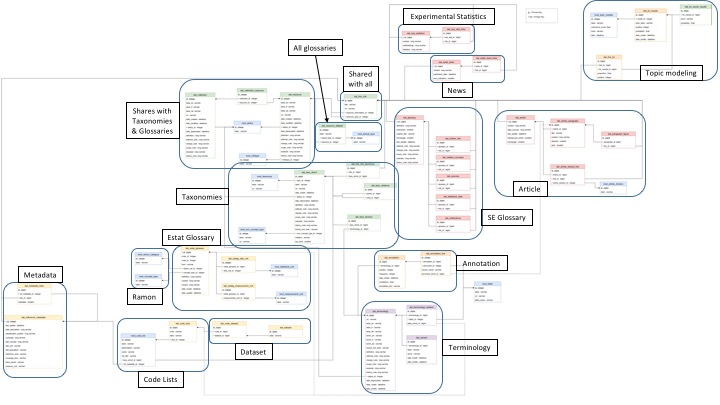


Figure 4 - Structure of the content database explained

Information has been added to make it easier for a first reader to understand which table is used to cover which data.

Plus, the articles scraped from the OECD website will appear within the Article part of the tables, as the model has been created in this purpose. In other words, these five tables will be used to structure the content from the SE articles and the articles from the OECD. The glossary on the other hand, will be covered from other tables customized specifically for this type of content.

### How to update the content?

In order to **update data**, the first thing is to see if source has changed in order to update the scraping code accordingly when the source is scrapped or to update the content database model.

If there is no change, the update is performed by executing the insertion notebook. Otherwise, the changes must be reported in the notebooks and executed them.

**Updating the model** means updating the tables described above. The tables are the structure of the content database. It has been mentioned that the OECD articles were included in the Articles part of the content database because it was thought to work like this. If new content has to be downloaded onto the content database three options are possible:

* the first option is to insert the new content on existing tables;
* the second option is to modify one or several tables (by changing the cardinality or adding a line in a table for instance);
* the third option is to create a new table that fit best the requirements of the new content.

For example, a few tables will have to be added to treat the content in French and in German.

## Knowledge database

The Knowledge database modelling reuse the proven standards, on the one hand and introduce only the classes and properties necessary to meet the needs of the use case, on the other hand.

### Reused Standards

The following standards are reused, or they have inspired the knowledge model:

| Standard | Prefix | Description |
| --- | --- | --- |
| Data Catalogue  vocabulary | dcat | DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web.  https://www.w3.org/TR/vocab-dcat-2/ |
| Data Cube  vocabulary | qb | The Data Cube vocabulary is a core foundation which supports extension vocabularies to enable publication of other aspects of statistical data flows or other multi-dimensional data sets. It is design to publish statistical data in RDF.  https://www.w3.org/TR/vocab-data-cube/ |
| Dublin core | dct | Dublin Core is a generic metadata schema for describing digital or physical resources and for establishing relationships with other resources.  https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ |
| Statistical Data and Metadata eXchange | sdmx | SDMX aims at standardising and modernising the mechanisms and processes for the exchange of statistical data and metadata among international organisations and their member countries.  https://sdmx.org/ |
| Simple Knowledge Organization System | skos | SKOS is a common data model for sharing and linking knowledge organization systems via the Web.  https://www.w3.org/TR/skos-reference/ |

The top level of the NLP4StatRef ontology has to be a DCAT profile. As the construction of the catalog is a Publisher point of view and is closely linked to the document management, we have not implemented it. However, we were careful:

* To use the recommended controlled vocabularies;
* To reuse classes in the same spirit in order to be able to easily populate the Catalog;
* To keep dct:isPartOf as a placeholder for a possible relationship between NLP4StatRef Knowledge and a Dataset, a Collection or a Master if it is implemented.

In the same way, the modeling of the statistical datasets for the project is inspired by the Data Cube vocabulary knowing that these are already structured in the Eurostat database management system.

#### Namespaces

NLP4StatRef ontology reuses terms from various existing specifications. Classes and properties specified in the next sections could be taken from the following namespaces:

* adms: http://www.w3.org/ns/adms#
* dcat: http://www.w3.org/ns/dcat#
* dcatap: http://data.europa.eu/r5r/
* dct: http://purl.org/dc/terms/
* euvoc: http://publications.europa.eu/ontology/euvoc#
* foaf: http://xmlns.com/foaf/0.1/
* locn: http://www.w3.org/ns/locn#
* owl: http://www.w3.org/2002/07/owl#
* odrl: http://www.w3.org/ns/odrl/2/
* prov: http://www.w3.org/ns/prov
* rdfs: http://www.w3.org/2000/01/rdf-schema#
* schema: http://schema.org/
* skos: http://www.w3.org/2004/02/skos/core#
* spdx: http://spdx.org/rdf/terms#
* xsd: http://www.w3.org/2001/XMLSchema#
* vann: http://purl.org/vocab/vann/
* voaf: http://purl.org/vocommons/voaf#
* vcard: http://www.w3.org/2006/vcard/ns#
* time: http://www.w3.org/2006/time#

We add to these namespaces, the following namespaces dedicated to NLP4StatRef:

* estat: http://nlp4statref/ontology
* estatvoc: http://nlp4statref/resource/

#### Controlled vocabularies

Properties that use controlled vocabularies are listed in the table below:

| Property URI | Used for Class | Vocabulary name | Vocabulary URI | Usage note |
| --- | --- | --- | --- | --- |
| dct:publisher | Content and Resource sub-classes | EU Vocabularies Corporate bodies Named Authority List | http://publications.europa.eu/resource/authority/corporate-body | The Corporate bodies NAL must be used for European institutions and a small set of international organizations. |
| dct:subject | Content and Resource sub-classes | NLP4StatRef Theme | http://nlp4statref/resource/taxonomy/nlp4statref-theme | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| dct:type | Content and Resource sub-classes | NLP4StatRef Resource Type | http://nlp4statref/resource/authority/resource-type | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| estat:resourceInformation | Content and Resource sub-classes | NLP4StatRef Resource Information | http://nlp4statref/resource/authority/resource-tinformation | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| estat:inCategory | Content sub-classes | NLP4StatRef Eurostat Category | http://nlp4statref/resource/taxonomy/eurostat-category | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| estat:inDatabaseTree | estat:StatisticalDataset | NLP4StatRef Eurostat Database Tree | http://nlp4statref/resource/taxonomy/eurostat-database-tree | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| estat:inFamily | estat:CodeList | NLP4StatRef Eurostat Family | http://nlp4statref/resource/taxonomy/eurostat-family | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| estat:inTermExtension | estat:Glossary | NLP4StatRef Eurostat Term Extension | http://nlp4statref/resource/taxonomy/eurostat-term-extension | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| estat:inTheme | estat:Article | NLP4StatRef Eurostat Theme | http://nlp4statref/resource/taxonomy/eurostatheme | The values to be used for this property are the URIs of the concepts in the vocabulary. |
| euvoc:status | skos:Concept | EU Vocabularies Corporate bodies Named Authority List | http://publications.europa.eu/resource/authority/concept-status | The values to be used for this property are the URIs of the concepts in the vocabulary. |

### Knowledge database model

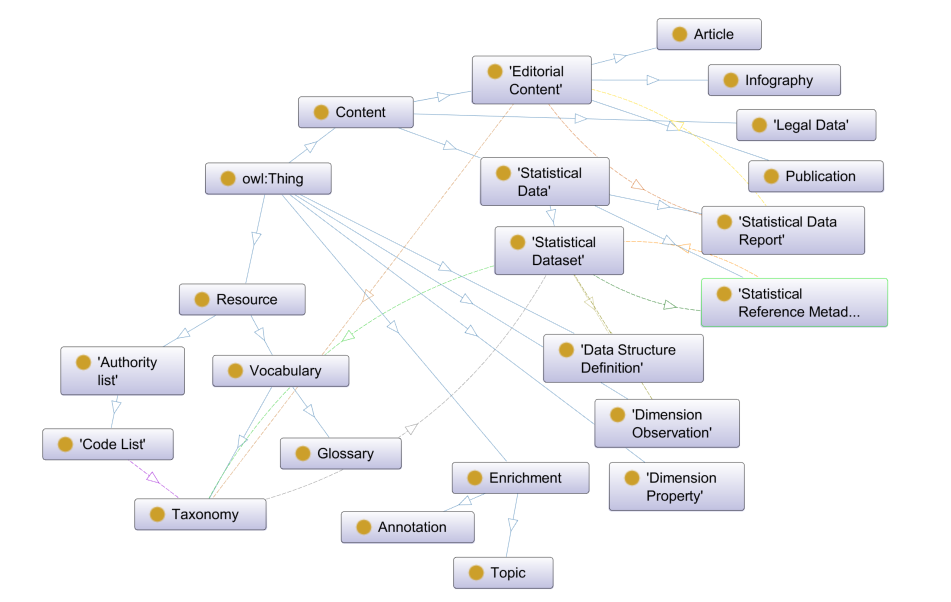


Figure 5 - NLP4StatRef Classes graph

#### NLP4StatRef Ontology ¤ Classes

The classes created to describe the relevant data to meet the use cases are the following:

| Class | Usage note | URI | Reference |
| --- | --- | --- | --- |
| Annotation | To describe with the development team. | estat:Annotation | http://nlp4statref/ontology#Annotation |
| Article | A written content, from a news to an analysis on a topic of particular interest. | estat:Article | http://nlp4statref/ontology#Article |
| Authority list | A concept collection (e.g. controlled vocabulary) in which codes, proper names, metadata or terms are defined. | estat:AuthorityList | http://nlp4statref/ontology#AuthorityList |
| Code List | A predefined, organized set of items that describe one or more statistical concepts. | estat:CodeList | http://nlp4statref/ontology#CodeList |
| Concept | An entity or unit of a vocabulary. | skos:Concept | https://www.w3.org/TR/skos-reference/#Concept |
| Content | All types of described and annotated content. | estat:Content | http://nlp4statref/ontology#Content |
| Data Structure Definition | The structure of a dataset. | estat:DataStructureDefinition | http://nlp4statref/ontology#DataStructureDefinition |
| Dimension Observation | A dimension value associated to observations of the DataSet. | estat:DimensionObservation | http://nlp4statref/ontology#DimensionObservation |
| Dimension Property | A dimension property which defines the structure of a dataset. | estat:DimensionProperty | http://nlp4statref/ontology#DimensionProperty |
| Editorial Content | All the information written and published around the statistical data, to comment on them, explain them, ... | estat:editorialContent | http://nlp4statref/ontology#EditorialContent |
| Enrichment | All enrichments from semantic analyses. | estat:Enrichment | http://nlp4statref/ontology#Enrichment |
| Glossary | An alphabetical list of terms relating to a specific subject enriched by definitions, related content…. | estat:Glossary | http://nlp4statref/ontology#Glossary |
| Infography | A content that presents a subject in a visual way: images, diagrams and short texts. | estat:Infography | http://nlp4statref/ontology#Infography |
| Legal Data | A legal document which can be legal act, case law, international agreements, and so on | estat:LegalData | http://nlp4statref/ontology#LegalData |
| Publication | A standalone and structured content published, periodically or not, like a leaflet, a book, ... | estat:Publication | http://nlp4statref/ontology#Publication |
| Resource | All types of controlled vocabularies. | estat:Resource | http://nlp4statref/ontology#Resource |
| Statistical Data | All types of described and annotated statistical content. | estat:StatisticalData | http://nlp4statref//ontology#StatisticalData |
| Statistical Data Report | A description of a statistical file (table or dataset), subject of an editorial content. | estat:StatisticalDataReport | http://nlp4statref/ontology#StatisticalDataReport |
| Statistical Dataset | A description of simple or multi-dimensional tables which represent a statistical dataset. | estat:StatisticalDataset | http://nlp4statref/ontology#StatisticalDataset |
| Statistical Reference Metadata | An explanatory document documenting statistical data. | estat:StatisticalReferenceMetadata | http://nlp4statref/ontology#StatisticalReferenceMetadata |
| Taxonomy | A classification of concepts. | estat:Taxonomy | http://nlp4statref/ontology#Taxonomy |
| Topic | To describe with the development team. | estat:Topic | http://nlp4statref/ontology#Topic |
| Vocabulary | All sets of terms listed in a lexicon, structured in a taxonomy, described in a dictionary or a glossary. | estat:Vocabulary | http://nlp4statref/ontology#Vocabulary |

#### NLP4StatRef Ontology ¤ Properties

NLP4StatRef Ontology exploits the properties listed below:

| Property | URI | Range | Usage note | Card |
| --- | --- | --- | --- | --- |
| alternative | dct:alternative | rdfs:Literal | This property contains an alternative name given to an individual. This property can be repeated for parallel language versions of the description. | 0-n |
| alt Label | skos:altLabel | skos:Concept | This property contains a n alternative lexical label to an individual. This property can be repeated for parallel language versions of the description. | 0-n |
| associated Term | estat:associatedTerm | skos:Concept | This property references the term(s) associated to a concept of a lexicalized taxonomy. | 0-n |
| broader Match | skos:broaderMatch | skos:Concept | This property states a broader relation between 2 concepts in different Concept schemes. | 0-n |
| close Match | skos:closeMatch | skos:Concept | This property states a close relation between 2 concepts in different Concept schemes. | 0-n |
| code List | estat:codeList | estat:CodeList | This property gives the code list associated with a Dimension Property. | 1 |
| code List Item | estat:codeListItem | skos:Concept | This property gives the code list item associated with a Dimension Observation. | 1 |
| contributor | dct:contributor | rdfs:Literal | This property refers to the entity and person who contributed to producing the resource. | 0-n |
| created | dct:created | rdfs:Literal typed as  xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth | This property indicates the date of creation of the resource or the entity. | 0-1 |
| creator | dct:creator | foaf:Agent | This property refers to the entity responsible for producing the resource. | 0-n |
| definition | skos:definition | skos:ConceptScheme or skos:Concept | This property supplies an explanation of the intended meaning of a concept. This property can be repeated for parallel language versions of the description. | 0-n |
| deprecated | estat:deprecated | rdfs:Literal typed as  xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth | This property indicates the date of depreciation of the entity. | 0-1 |
| description | dct:description | rdfs:Literal | This property contains a free-text account of the individual. This property can be repeated for parallel language versions of the description. | 0-n |
| dimension Property | estat:dimensionProperty | estat:DimensionProperty | This property indicates a Dimension Property expected on the Statistical Dataset. | 1 |
| exact Match | skos:exactMatch | skos:Concept | This property indicates a similar meaning between 2 concepts in different Concept schemes. | 0-n |
| example | skos:example | skos:ConceptScheme or skos:Concept | This property supplies an example of the use of a concept. This property can be repeated for parallel language versions of the description. | 0-n |
| has For Reference Metadata | estat:hasForReferenceMetadata | estat:StatisticalReferenceMetadata | This property refers to the Statistical Reference metadata description of the Statistical Dataset | 0-1 |
| history Note | skos:historyNote | skos:ConceptScheme or skos:Concept | This property describes significant changes to the meaning or the form of a concept. This property can be repeated for parallel language versions of the description. | 0-n |
| in Category | estat:inCategory | skos:Concept | This property refers to the concept of the Eurostat Category Taxonomy associated to the article. | 0-n |
| in Database Tree | estat:inDataBaseTree | skos:Concept | This property refers to the concept of the Eurostat Database Tree Taxonomy associated to the Statistical Dataset. | 0-n |
| in Family | estat:inFamily | skos:Concept | This property refers to the concept of the Eurostat Family Taxonomy associated to the Code List. | 1 |
| in Scheme | skos:inScheme | estat:AuthorityList or estat:Glossary or skos:ConceptSceme | This property relates a concept to a concept scheme in which it is included. |  |
| in Taxonomy | estat:inTaxonomy | skos:Concept | This property refers to a concept of a taxonomy. | 0-n |
| in Term Extension | estat:inTermExtension | skos:Concept | This property refers to the concept of the Eurostat Term Extension Taxonomy associated to the Eurostat Glossary. | 0-n |
| in Theme | estat:inTheme | skos:Concept | This property refers to the concept of the Eurostat Theme Taxonomy. | 0-n |
| is Part Of | dct:isPartOf |  | Placeholder to confirm. |  |
| issued | dct:issued | rdfs:Literal typed as  xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth | This property contains the date of formal publication. | 0-1 |
| license | dct:license | rdfs:Resource | This property refers to the license under which the Catalogue can be used or reused. | 0-1 |
| member Of Category | estat:memberOfCategory | estat:StatisticalDataReport or estat:Article | This property indicates an individual that reference the concept of the Category taxonomy. |  |
| member Of Database Tree | estat:memberOfDatabaseTree | estat:StatisticalDataset | This property indicates the Statistical Dataset associated to the concept of the Eurostat Database Tree Taxonomy. | 0-n |
| member Of Family | estat:memberOfFamily | estat:CodeList | This property indicates the Code Lists associated to the concept of the Eurostat Family Taxonomy. | 0-n |
| member Of Taxonomy | estat:memberOfTaxonomy | rdfs:Resource | This property indicates an individual that reference the concept of the taxonomy. | 0-n |
| member Of Term Extension | estat:memberOfTermExtension | estat:Glossary | This property indicates the terms associated to the concept of the Eurostat Term Extension Taxonomy. | 0-n |
| modified | dct:modified | rdfs:Literal typed as  xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth | This property contains the most recent date of modification or update. | 0-1 |
| narrow Match | skos:narrowMatch | skos:Concept | This property states a narrower relation between 2 concepts in different Concept schemes. | 0-n |
| notation | skos:notation | rdfs:Literal | This property contains a string that is an identifier or a classification code. | 0-1 |
| prefLabel | skos:prefLabel | rdfs:Literal | This property contains a preferred label of the individual. This property can be repeated for parallel language versions of the label. | 1-n |
| publisher | dct:publisher | skos:Concept | This property refers to an entity (organisation) responsible for making the data available. | 0-1 |
| reference Metadata Of | estat:referenceMetadataOf | estat:StatisticalDataset | This property indicates the Statistical Dataset described by the Statistical Reference Metadata. | 1 |
| related | skos:related | skos:Concept | This property enables the representation of associative (non-hierarchical) links between concepts within a same Concept scheme. | 0n |
| related Content | estat:relatedContent | rdfs:Resource | This property indicates a content that references a content or a concept. | 0-n |
| related Editorial Content | estat:relatedEditoria Content | estat:EditorialContent  and subclasses | This property indicates an editorial content that references a content or a concept. | 0-n |
| related Legal Data | estat:relatedLegalData | estat:LegalData | This property indicates a legal data that references a content or a concept. | 0-n |
| related Match | skos:relatedMatch | skos:Concept | This property states an associative mapping link between 2 concepts in different Concept schemes. | 0-n |
| related Statistical Data | estat:relatedStatisticalData | estat:StatisticalData  and subclasses | This property indicates a statistical data that references a content or a concept. | 0-n |
| report Data | estat:reportData | estat:StatisticalDataReport | This property gives the URI of the Statistical Data report subject of a Statistical article or a Background Article. | 0-1 |
| report Source | estat:reportSource | estat:Article | This property indicates the Editorial Content which comments the Statistical Data Report. | 0-1 |
| resource Information | estat:resourceInformation | skos:Concept | This property refers to the source of information. The controlled vocabulary is https://nlp4statref/resource/authority/resource-information. | 1 |
| rights | dct:rights | rdfs:Resource | This property refers to a statement that specifies rights associated with the resource. | 0-1 |
| scopeNote | skos:scopeNote | skos:ConceptScheme or skos:Concept | This property supplies information about the intended meaning of a concept. This property can be repeated for parallel language versions of the description. | 0-n |
| source | dct:source | estat:Glossary or skos:Concept | This property links to the original metadata that was used in creating the resource. | 0-1 |
| source Of Information | estat:sourceOfInformation | skos:Concept | This property gives the references of the original source of information. | 0-n |
| status | euvoc:status | skos:Concept | This property refers to a status of the concept in the present time span. | 0-1 |
| structure | estat:structure | estat:DataStructureDefinition | This property indicates the structure to which the statistical Dataset conforms. | 1 |
| subject | dct:subject | skos:Concept | This property refers to the classification http://nlp4statref/resource/taxonomy/nlp4statref-theme. | 0-n |
| title | dct:title | rdfs:Literal | This property contains a name given to an individual. This property can be repeated for parallel language versions of the description. | 0-n |
| type | dct:type | skos:Concept | This property refers to a type of resource. The controlled vocabulary is http://nlp4statref/resource/authority/resource-type. | 1 |
| type | rdf:type | rdfs:Class | This property indicates the class of which the individual is an instance. | 0-1 |

### How to update the knowledge?

In order to **update knowledge**, if there is no change in the Content database model, the update is performed by executing the insertion notebook. Otherwise, the changes must be reported in the instantiation notebook and executed it.

**Updating the model** means updating the ontology and the instantiation notebook. This is the same process as the one to do for a new data flow. [Section 3]

# How to add new data?

The diagram below is showing the path used to implement new data. The first step is to identify which data might be worth the trouble. Then, it is important to describe it first, as it will be a huge help to create the model and, the mapping, in order to scrap it properly and grab the relevant information.

Once the information is scraped, it needs to be inserted within the content database, and in order to do so, each information needs to find its right place within an existing table (the tables corresponding to the information extracted are also specified in the mappings). The information can then be inserted into Virtuoso.

Knowledge instantiation adds a semantic model to the data, which includes a formal classification with classes, subclasses, relationships and instances on one hand, and rules for interpreting the data, on the other. The formal classification is described by the NLP4StatRef Ontology [section 2.3.2] and the rules, based on classes and properties, are described by the mappings [section 3.1.3].

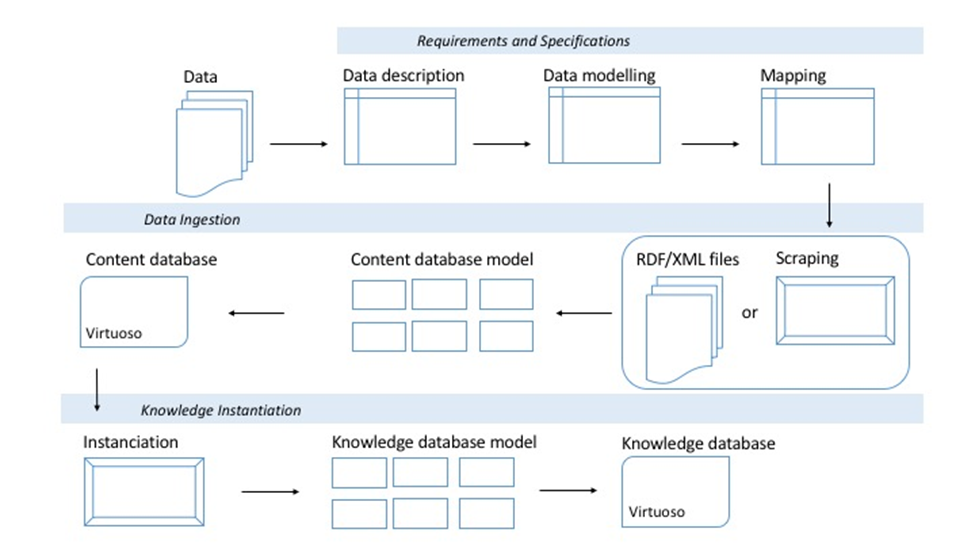


Figure 6 - How to add new data

## Requirements and Specifications

The process of adding new data must follow a list of steps that are detailed in the diagram above. The phase called requirements and specifications is the first phase of the steps. In this phase, the core data is being broke down to be analyzed and to see how the data can be inserted into the model.

### Data description

Before integrating data from an internal or external source, this source has to be described. This description covers both the structure of the source and the content itself.

Description of the source structure is useful:

* For preparing the data collection, especially if it will be done by scrapping;
* To identify the Content type and Knowledge type;
* To specify Content and Knowledge workflow.

For the NLP4StatRef project, this was done in an Excel spreadsheet called NLP4StatRef-ResourceDescription.

In the following tab, on the first column can be found the title of the columns of the Resource Description spreadsheet, and on the second one, a brief explanation.

|  |  |
| --- | --- |
| Column name | Usage note |
| Work date | The last work done on the line has to figure here |
| Owner | The task owner |
| WorkStatus | Indicate if the work is to be started, done or in progress |
| WorkAction | Here to indicate if the work is created or modified |
| EditorialNote | Notes for the different users |
| ChangeNote | Things that have changed |
| ResourceName | The name of the resource that is going to be described in the line |
| Description | A quick description of the data that is in the page |
| ContentType | The content type is in this case either a document, vocabulary or statistical data |
| URL HomePage | The URL of the homepage |
| HomePage Content | A quick description of what we can find in the homepage |
| ContentUnit | What is the nature of the information in the page? Is it an article or a glossary entry? |
| Unique Identifier | The unique identifier is a way to identify the data that is unique for each entry |
| Identification Reference | These are the elements which make it possible to identify the information |
| Citation Reference | These are the elements used to cite information |
| Metadata | This is for the metadata that might be in the page |
| Named Entities | This is for the named entities that might be in the page |
| Content Description | This is where the content of the pages will be described  How the page is built, how many times can one component can be present, what kind of information is also present, and so on. Example: Related concepts [Text, Url] 0-n  In the described content, it is possible – but not necessary [0-n] - to find a link [Text, Url] to a Related Concept. |
| External Resources | Here has to be indicated in which part of the page a link to external content can be found |
| Associated Content | Here has to be indicated in which part of the page a PDF to downloaded (or a spreadsheet) for example can be found |
| Comments | Any comments the task owner or a team member seem appropriate can figure here |

### Data Modelling

Data modeling is a process of describing the structure, associations, relationships, and constraints of available data. It is used to establish standards and define data management rules.

This modelling is first done with a mind mapping tool, then the formalization is created in the ontology.

For each dala flow, a description of the content is made, centered on the data available and relevant to the project. This data is organized into 3 families: metadata, core content and available relationships.

When access to data is not homogeneous according to the features offered, the information is saved for future developments.

This description

* prefigures the structure of the Content database;
* makes it possible to compare flows in order to identify areas of overlap or even possible reuses of an existing workflow.

A second branch of the description is focused on the knowledge description. Content information is transformed into objects and properties which will then be formalized in the ontology.

### Mapping: between data and the Content and Knowledge databases

A mapping is a specification used to describe how:

* integrate data in the Content database;
* Instantiate the Knowledge database from the Content database.

The 1st goal of the mapping is to "link" the element of the processed source with the relevant element of the database:

* Content database: path of the Table\Column;
* Knowledge database: class and/or property.

In addition to this "linking", this mapping describes all the treatments and controls that must be performed during the insertion:

* Treatments are inserted into the processes of population and instantiation;
* Controls are either be embedded in the insertion process or turned into constraints.

For the NLP4StatRef project, this was done in an Excel spreadsheet called NLP4StatRef-[Data name]-Mapping, consisting of a sheet for the Content database [CD-Data name] and a sheet for the Knowledge database [KD-Data name]:

| Content database mapping [CD-Data name] | |
| --- | --- |
| Column name | Usage note | |
| Work date | The last work done on the line has to figure here | |
| Work status | Indicate if the work is to be started, done or in progress | |
| Note | Any note that may be needed | |
| Resource page | The page and the type of resource described | |
| Resource data | The name of the data that will be the subject of the line | |
| Description | The nature of the data named in the column before | |
| Cardinality | The cardinality of the data | |
| Content database | In which table of the content database this data can be located | |
| Treatment | If there is any treatment to be done to the data, it will be indicated here | |
| Controls | If there is any control to be done to the data, it will be indicated here | |
| Comments | Any comment necessary can appear here | |

| Knowledge database mapping [KD-Data name] | |
| --- | --- |
| Column name | Usage note | |
| Work date | The last work done on the line has to figure here | |
| Work status | Indicate if the work is to be started, done or in progress | |
| Note | Any note that may be needed | |
| Resource page | The page and the type of resource described | |
| Content data | The path of the Content database element which is mapped | |
| Domain [Subject] | The class name of the subject | |
| Property [Predicat] | The property name of the predicat | |
| Range [Object] | The class name or the data property of the object | |
| M | R | O | Property opt.: **M**andatory | **R**ecommended | **O**ptional | |
| Content | URI of the object when this is a fixed value | |
| Resource | URI of the controlled vocabulary in which the object is defined | |
| Comments | Any comment necessary can appear here | |

## Data ingestion

The data ingestion is the phase in which the data gets ready to be included within the Content Database. It includes the scraping, the design of the tables and the use of Virtuoso.

### Data collection

According to data and to data access features available, the data collection could be done by:

* Scrapping of data;
* Download of data;
* Data manually built.

For the NLP4StatRef project, all these methods were used and the integration of the results in the Content database is paired and documented:

| Data name | Format | Collection method | Reference |
| --- | --- | --- | --- |
| Background articles  Statistics Explained website | Web pages | Scraping | Section 4.1.1.2 |
| Categories  Statistics Explained website | Web pages | SKOSification | Section 6.5 |
| Data navigation Tree  Statistics Explained website | Excel | SKOSification | Section 6.5 |
| Code lists  Ramon Metadata | Dic | Bulk download | Section 6.2 |
| Euro Indicators | Web page + Pdf | Copy/Paste | Section 4.3.1.2 |
| Eurostat concept and definition | Csv | Download | Section 6.3 |
| Eurostat databases | Tsv | Bulk download | Section 5.2 |
| Experimental Statistics | Web page | Scraping | Section 4.2 |
| Family  Ramon metadata | Web pages | SKOSification | Section 6.5 |
| Glossary articles  Statistics Explained website | Web page | Scraping | Section 6.4.1.2 |
| News | Web page | Scraping | Section 4.3.1.1 |
| OECD Glossary | Web page | Scraping | Section 6.4.1.1 |
| OECD Topics | Web page | Scraping | Section 4.1.1.3 |
| Reference metadata  Eurostat databases | Web page | Scraping | Section 5.3 |
| Resource Information | RDF | SKOSification | Section 6.1.1.1 |
| Resource Type | RDF | SKOSification | Section 6.1.1.2 |
| Statistics Explained articles  Statistics Explained website | Web page | Scraping | Section 4.1.1.1 |
| Term Extension  Ramon metadata | Web pages | SKOSification | Section 6.5 |
| Terminology  Noun phrases Lexicons | Web pages | SKOSification | Section 6.5 |
| Themes  Eurostat websites | Web pages | SKOSification | Section 6.5 |
| Themes  OECD websites | Web pages | SKOSification | Section 6.5 |

The guidelines for populating the content database are described in the [How to load data into the Content database](#_How_to_load) section. The specific rules by data, if applicable, are describe mappings.

### Content database population

Here are all the files that will allow to create the core structure of the data and populate tables:

|  |  |
| --- | --- |
| Content type | Notebooks and SQL files |
| CodeList and datasets | Concatenate\_dic\_files.ipynb |
| Estat\_codelist\_dataset.sql |
| Estat\_codelist\_label\_data.sql |
| Estat\_dataset\_code\_data.sql |
| Estat\_dataset\_label\_data.sql |
| Estat\_dictionnary\_code\_data\_batch1.sql |
| Estat\_dictionnary\_code\_data\_batch2.sql |
| Estat\_dictionnary\_code\_data\_batch3.sql |
| Estat\_dictionnary\_code\_data\_batch4.sql |
| Estat\_dictionnary\_code\_data\_batch5.sql |
| Estat\_dictionnary\_code\_data\_time\_addition.sql |
| Eurostat glossary | Estat13k\_glossary\_data.sql |
| Estat13k\_glossary.sql |
| Estat13k\_modalities\_data.sql |
| Estat13k\_modalities.sql |
| Estat13k\_stat\_and\_measurement\_unit\_data.sql |
| Statistics explained | Cdb\_articles\_v1.sql |
| Cdb\_global\_se\_data\_v1.sql |
| Cdb\_glossary\_v2.sql |
| Cdb\_link\_info\_v1.sql |
| Cdb\_resources\_mod\_data.sql |
| Estat\_news\_exp\_stat.sql |
| Taxonomies | Estat\_taxonomies.sql |
| Terminology | Estat\_terminology\_annotation.sql |
| Topic model | Estat\_topic\_model.sql |

When the Content database is setup, the following scripts allow to populate data within the tables:

* Cdb\_global\_v2.sql
* Cdb\_insert\_link\_info.ipynb
* Cdb\_insert.ipynb

## Knowledge instantiation

Before instantiating the KD, Python must be installed and connected to Virtuoso.

### Notebook creation

Instantiation is a two steps process

1. Initialize the model
2. Create the SPARQL queries

These 2 steps are based on the mapping [section 3.1.3 - Knowledge database mapping [KD-Data name]].

The following notebooks instantiate the Knowledge database:

To be completed after updating the knowledge database and synchronizing it with the baselined ontology.

|  |  |
| --- | --- |
|  | Notebook name |
|  |  |
|  |  |

### Links about creation of SPARQL queries

| How to | Link |
| --- | --- |
| Cheat sheet SPARQL | http://www.iro.umontreal.ca/~lapalme/ift6281/sparql-1\_1-cheat-sheet.pdf |
| Simple queries | https://www.w3.org/2001/sw/DataAccess/rq23/examples.html |
| Advanced queries | https://csiro-enviro-informatics.github.io/info-engineering/tutorials/tutorial-intro-to-rdf-and-owl.html#5-part-3-query-the-rdf-data-using-sparql-in-topbraid-15-20mins |
| Python package for SPARQL queries | https://rdflib.dev/sparqlwrapper/ |
| Update the knowledge graph | https://stackoverflow.com/questions/42814427/how-to-insert-new-instance-into-owl-ontology-using-sparql/42830041 |

# Articles

Articles are one of the 3 editorial contents integrated in the corpus and described by the ontology.

The 2 others – Infography and Publication - were also created since they are referenced in the processed content. However, being only referenced and not integrated into the corpus, these 2 types are not described even if they are already provided for by the ontology.

## Statistics Explained content & OECD Topics

The article subsection comprises SE statistical article, SE background article and the OECD Topics. They are all made of text, datasets and links to internal and external websites. Their difference is that SE statistical article are about statistics produced by Eurostat and SE background article are about methodology or providing contextual elements, and the OECD topics are about statistics produced by the OECD, about Quality of Life.

### Requirements and Specifications

#### SE articles

<https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Category:Statistical_article>

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ ContentTypes > Statistical articles |
| Data model | PDF: NLP4StatRef-SE-Article-Model |
| Mapping | Excel: NLP4StatRef-SE-Article-Mapping |
| Data collection | Notebook: article\_spider.py[[1]](#footnote-1) |
| Resource Type | http://nlp4statref/resource/authority/resource-type#statistical-article |

SE articles are articles written about statistics produced by Eurostat. They are made of text, links to other pages to internal and external websites as well as datasets.

The SE Background and the SE articles have a similar structure but they differ in a number of ways, such as their content, as explained above. While the SE articles are about statistics produced by Eurostat, the SE Background articles are about methodology. The core information contain within the article is thus different. As they shared a same structure, they are identical for the ontology but distinguished by their Resource Type [Section 6.1.1.2].

Within almost each SE articles, there is a spreadsheet that gathers all the tables and graphs present in the article. This spreadsheet is called Source data for tables and graphs. A link downloads instantly in one spreadsheet all the data present in the article.

These links were interesting and thus have been treated as an object in themselves. [Section 5.1 Statistical Data Report] They are also present within the SE Background Articles.

#### SE Background articles

<https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Category:Background_article>

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ ContentTypes > Background articles |
| Data model | PDF: NLP4Stat-Ref-SE-Background-Article-Model |
| Mapping | Excel: NLP4StatRef-SE-Background-Article-Mapping |
| Data collection | Notebook: background\_article\_spider.py |
| Resource Type | http://nlp4statref/resource/authority/resource-type#methodological-article |

The SE Background articles are articles written about methodology, they are complementing the SE articles.

Just like for the SE articles, there is almost always a spreadsheet is called Source data for tables and graphs. The link downloads instantly in one spreadsheet all the data present in the article.

These links were interested and thus have been treated as an object in themselves. [Section 5.1 Statistical Data Report]

#### OECD Topic

<https://www.oecdbetterlifeindex.org/topics/housing/>

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ ContentTypes > OECD Topics |
| Model | PDF: NLP4StatRef-OECD-Topic-Model |
| Mapping | Excel: NLP4StatRef-OECD-Topic-Mapping |
| Data collection | Notebook: OECD\_Topics.ipynb |
| Resource Type | http://nlp4statref/resource/authority/resource-type#statistical-article |

### Content database model

Articles are modelized into 5 tables connected with dat\_link\_info:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_article | Table for the generic information of the article |
| dat\_article\_paragraph | Table for information regarding a paragraph |
| dat\_paragraph\_figure | Tables for the figures |
| dat\_article\_shared\_links | Tables for shared links |
| mod\_article\_division | Table for the name of the different related content |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

The article model is simple and not very restrictive in order to take into account all the forms of articles. Most of the knowledge acquired relates to the relations between the article and its annotations and its relations.

#### NLP4StatRef Ontology ¤ Article

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:Article | dct:type | dct:issued | dct:isPartOf |
| estat:resourceInformation | dct:publisher | estat:inCategory |
| dct:source | dct:modified | estat:inTheme |
|  |  | estat:relatedEditorialContent |
|  |  | estat:relatedLegalData |
|  |  | estat:relatedStatisticalData |
|  |  | estat:reportData |

## Experimental statistics

<https://ec.europa.eu/eurostat/web/experimental-statistics/>

Experimental statistics are statistics that have not yet reached full maturity as they use new data sources and methods. The Experimental Statistics are all divided in distinctive parts: one part is about presenting the data to the user, one part is about methodology, one part is about linking similar content, one part is about getting feedback.

As it is an experimental content, a dedicated Content database model has been created. An evolution could be to move this content to the Article model created for articles.

### Requirements and Specifications

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ ContentTypes > Topics |
| Model | PDF: NLP4StatRef-Experimental-Statistics-Model |
| Mapping | Excel: NLP4StatRef-Experimental-Statistics-Mapping |
| Data collection | Notebook: JNB#1 Experimental Statistics.ipynb |
| Resource Type | http://nlp4statref/resource/authority/resource-type#statistical-article |

Like for the SE articles and SE Background articles, there is always and at least one spreadsheet is called Access the statistics. The link downloads instantly in one spreadsheet all the data present in the article.

These links were interested and thus have been treated as an object in themselves. [Section 5.1 Statistical Data Report]

For some Experimental statistics, statistics are statistical database. [Section 5.2]

### Content database model

The Experimental Statistics are modelized into 2 tables connected with dat\_link\_info:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_exp\_statistics | Tables for the different paragraphs |
| dat\_exp\_stat\_links | Tables for the links within these paragraphs |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

Experimental statistics are estat:Article [section 4.1.3.1].

The instantiation is from the knowledge perspective the same as the one done for SE articles.

## News & Euro Indicators

There are two type of content that are similar: News and Euro Indicators. They are at the same location in the Eurostat website. The Euro Indicators have not been uploaded in the content database because they are PDF documents and it was not a part of the project. However, the content of the PDF document is very close to the News article, that is why the design of the model has been thought out to include them very easily.

A dedicated Content database model has been created. An evolution could be to move this content to the Article model created for articles.

### Requirements and Specifications

#### Eurostat News

<https://ec.europa.eu/eurostat/web/main/news/whats-new>

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ ContentTypes > Page of What’s New |
| Data model | PDF: NLP4StatRef-News-Model |
| Mapping | Excel: NLP4StatRef-News-Mapping |
| Data collection | Notebook: news\_spider.py  cdb\_estat\_news\_v1.sql |
| Resource Type | http://nlp4statref/resource/authority/resource-type#news |

#### Euro Indicators

<https://ec.europa.eu/eurostat/web/main/news/euro-indicators>

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ ContentTypes > Euro Indicators |
| Model | PDF: NLP4StatRef-Euro-Indicators-Model |
| Mapping | Excel: NLP4StatRef- Euro-Indicators -Mapping |
| Resource Type | http://nlp4statref/resource/authority/resource-type#news |

Euro indicators provide general economic information on the euro area, European Union and individual Member States. They are produced by Eurostat.

They have not been collected because treating PDF was not part of the project. In order to collect these data, an OCR has to be used. However, all the material is ready in case this project goes further, there is no need to adapt the tables of the Content Database in order to integrate the Euro Indicators as they fit in the existing ones as can be seen below.

### Content database model

The News article are modelized into 2 tables, connected to the table dat\_link\_info:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_estat\_news | Table for the generic content |
| dat\_estat\_news\_links | Table for the links in the News article |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

News and Euro Indicators are estat:Article [section 4.1.3.1].

# Statistical Data

Statistical data modeling not only describes datasets but also all related aspects, such as access, description and other metadata information. It also facilitates the linking of datasets with other content and resources. This modelling outlines a Statistical Content ecosystem summarized in the following schema:

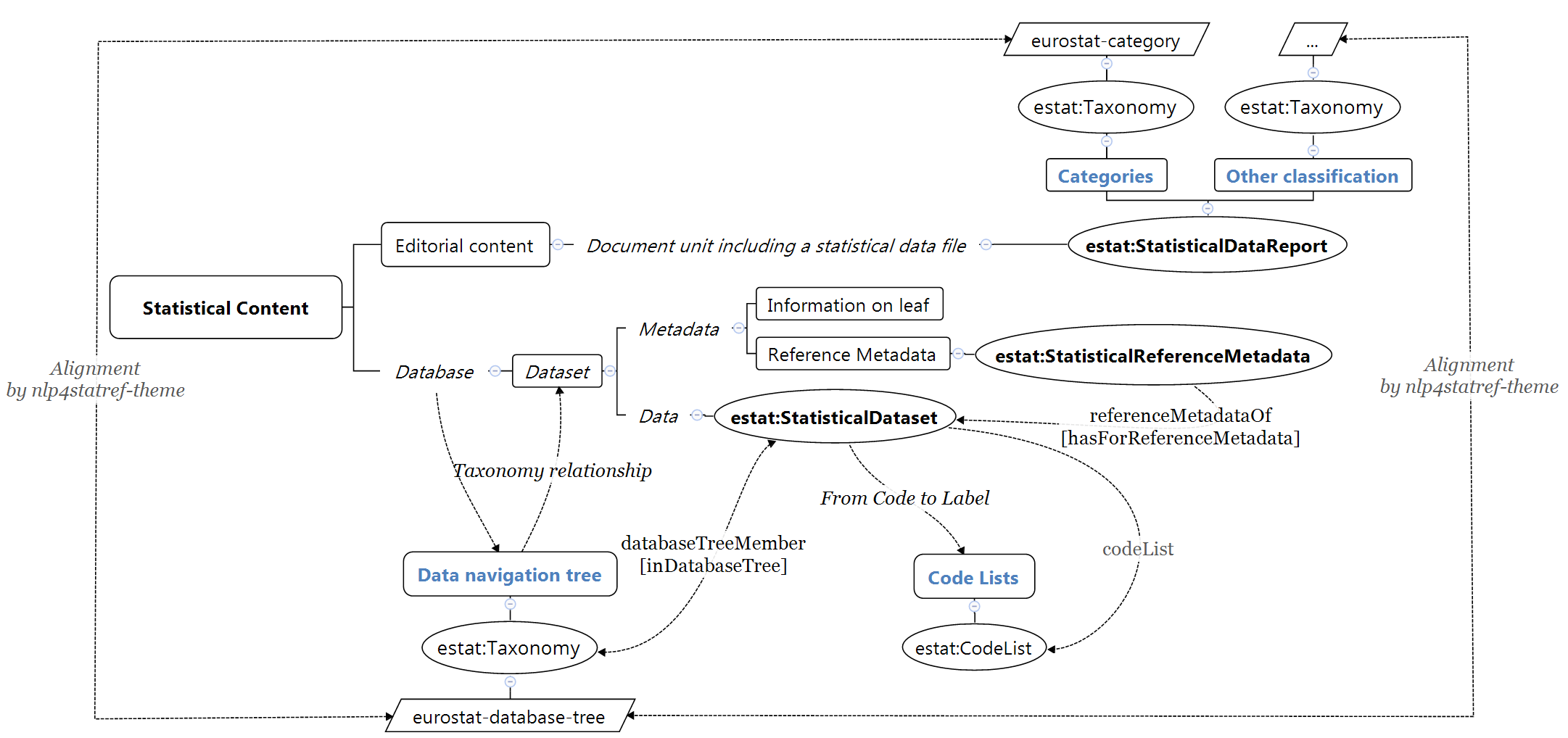


Figure 7 - The Statistical Content ecosystem.

This ecosystem is mainly based on the following data:

|  |  |
| --- | --- |
| Statistical data classes | Contextualization resources |
| Statistical Data Report [Section 5.1] | Data navigation tree [section 6.5] |
| Statistical Datasets [Section 5.2] | Code Lists [section 6.2] |
| Statistical Reference metadata [Section 5.3] | Dedicated classifications like Categories in the Statistics Explained website [section 6.5] |

## Statistical Data Report

Within statistical articles, there is sometimes a link to download spreadsheet that gathers all the tables and graphs present in the article. These spreadsheets have been treated as an object in themselves.

In NLP4StatRef corpus, they are present within the Experimental statistics, SE Articles or SE Background articles.

### Requirements and Specifications

|  |  |
| --- | --- |
| Data description | No dedicated description as these data are included in other content |
| Data model | PDF: NLP4StatRef-Data-Report-Model |
| Mapping | Statistical Data Reports inherit properties from the document they come from. Therefore, their mapping is included in the content of the document mapping in a dedicated sheet named KD-DataReport  Excel: NLP4StatRef-Experimental-Statistics-Mapping  Excel: NLP4Stat-Ref-SE-Background-Article-Mapping  Excel: NLP4StatRef-SE-Statistical-Article-Mapping |
| Data collection | Statistical Data Reports are collected during the main content collection. |
| CD population | Statistical Data Reports are integrated in the same process as that carried out for the main content. |
| Resource type | http://nlp4statref/resource/authority/resource-type#statistical-data-report |

### Content database model

Statistical Data reports are modelized into 3 tables:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_link\_info | Table for the statistical data report identification and metadata |
| dat\_article\_shared\_link | Table for the link between the article and its statistical data report |
| mod\_article\_division | id='1' Label='Excel' |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

Statistical Data Report inherits of properties from its main document but it has a life of its own.

#### NLP4StatRef Ontology ¤ StatisticalDataReport

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:StatisticalDataReport | dct:modified | dct:description | dct:isPartOf |
| dct:publisher |  |  |
| dct:source |  |  |
| dct:title |  |  |
| dct:type |  |  |
| estat:inCategory |  |  |
| estat:reportData |  |  |
| estat:resourceInformation |  |  |

## Statistical Datasets

<https://ec.europa.eu/eurostat/web/main/data/database>

Eurostat data presented in simple or multi-dimensional tables.

It is not a question here of redefining the model of this content type which is already standardized but of modelling the useful information to respond to the use cases of the project.

### Requirements and Specifications

|  |  |  |
| --- | --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ ContentTypes > Database | |
| Data model | PDF: NLP4StatRef-Datasets-TSV-Model | |
| Mapping | Excel: NLP4StatRef-Datasets-TSV-Mapping | |
| Data collection | | Bulk download |
| Resource Type | | http://nlp4statref/resource/authority/resource-type#statistical-dataset |

This integration was done using SQL commands following the mapping.

An update of the content database and its regular execution will require an evolution in particular to collect the data directly at the source and not by using scraping tool or download feature. This development will improve the modelling of the Content and Knowledge databases and will enrich knowledge and annotations.

### Content database model

Depending on the format of the dataset (table, database, Excel spreadsheet, tsv file,… ) and the access mode chosen (Data navigation tree icons, Download, Bulk down load), the content is more or less rich.

The minimum is described as follows: Columns 2-n ; Rows 2-n

* 1st column: codes or labels or both
* Other columns: figures
* 1st row: header
* Other rows: dataset values

Datasets, especially when the 1st column only contains codes, are associated with Code Lists allowing the codes to be transformed into labels and vice versa.

Statistical datasets are modelized into 2 tables:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_dataset | Table for the Dataset file |
| dat\_code\_dataset | Table for the Dataset items: |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

In order to get the label associated to each code, dat\_code\_dataset is linked to dat\_code\_dico of Code lists content type.

### Knowledge database model

#### NLP4StatRef Ontology ¤ StatisticalDataset

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:StatisticalDataset | dct:issued | dct:description | dct:isPartOf |
| dct:publisher | dct:modified |  |
| dct:title | dct:subject |  |
| dct:type | estat:hasForReferenceMetadata |  |
| estat:resourceInformation | estat:inDatabaseTree |  |
| estat:structure |  |  |

#### NLP4StatRef Ontology ¤ DataStructureDefinition

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:DataStructureDefinition | estat:dimensionProperty |  |  |

#### NLP4StatRef Ontology ¤ DimensionProperty

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:DimensionProperty | estat:codeList |  |  |

#### NLP4StatRef Ontology ¤ DimensionObservation

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:DimensionObservation | estat:codeListItem |  |  |
| estat:dataSet |  |  |
| estat:dimensionProperty |  |  |

## Statistical Reference metadata

<https://ec.europa.eu/eurostat/cache/metadata/en/ilc_esms.htm>

Reference metadata describe statistical concepts and methodologies used for the collection and generation of data, they also provide information on data quality. They can be found by clicking on the “M” at the end of each node of the database tree.

It is not a question here of redefining the model of this content type which is already standardized but of modelling the useful information to respond to the use cases of the project.

### Requirements and Specifications

|  |  |
| --- | --- |
| Data model | PDF: Work in progress |
| Mapping | Excel: Work in progress |
| Data collection | Download: Copy-Paste |
| Resource Type | http://nlp4statref/resource/authority/resource-type#statistical-reference-metadata |

### Content database model

Reference metadata are modelized into 2 tables:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_reference\_metadata | Tables about the paragraphs |
| dat\_metadata\_links | Table about the links present in the page |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

Work in progress

#### NLP4StatRef Ontology ¤ estat:StatisticalReferenceMetadata

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:StatisticalReferenceMetadata |  |  |  |
|  |  |  |
|  |  |  |

# Resources

Resources modelling describes concepts - *code, metadata, proper name, term* - within a collection – *classification, definition, file, glossary, lexicon, taxonomy, terminology*. Two main family of resources are considering in the ontology:

* Authority lists

An authority list is a controlled vocabulary used to name particular entities consistently. It is structured in separate files dedicated to different types of entity. Types are codes, proper names, metadata or terms.

* Vocabulary

A vocabulary is a more or less organized, structured set of concepts which are more or less defined or described.

Resources currently available are:

|  |  |
| --- | --- |
| AuthorityList | Vocabulary |
| CodeLists [section 6.2] | Eurostat Glossary [Section 6.3] |
| Resource Information [Section 6.1] | OECD Glossary [Section 6.4.1.1] |
| Resource Type [Section 6.1] | SE Glossary [Section 6.4.1.2] |
|  | Taxonomies [Section 6.5] |
|  | Terminology Lexicon [Section 6.6] |

## Authority Lists ¤ Resource Information & Resource Type

For the need of NLP4StatRef project, 2 dedicated resources were built. They have been designed to be aligned with different existing resources.

These resources are built in an Excel spreadsheet, structured to be automatically transformed into a SKOS resource.

### Requirements and Specifications

#### Resource Information

|  |  |
| --- | --- |
| Data resource | Excel: NLP4StatRef\_Resource > ResourceInformation  RDF: NLP4StatRef\_ResourceInformation |
| Data model | PDF: NLP4StatRef-ResourceInformation-Model |
| Mapping | Excel: NLP4StatRef-ResourceInformation-Mapping |
| Resource Type | https://nlp4statref/resource/authority/resource-type#named-entity-list |

Resource Information is a named entities authority list that references all the organizations whose data is integrated into the Content and Knowledge databases.

#### Resource Type

|  |  |
| --- | --- |
| Data resource | Excel: NLP4StatRef\_Resource > ResourceType  RDF: NLP4StatRef\_ResourceType |
| Data model | PDF: NLP4StatRef-ResourceType-Model |
| Mapping | Excel: NLP4StatRef-ResourceType-Mapping |
| Resource Type | https://nlp4statref/resource/authority/resource-type#authority-list |

Resource Information is an authority list that references all types of content, data and resources integrated into the Content and Knowledge databases.

### Content database model

Resource Information and others named entities lists in a SKOS file are modelized into the following tables:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_resource | Table for Concept |
| dat\_resource\_altLabel | Table for Lexical Label |
| mod\_lexical\_type | Table for the type of Lexical Label |
| mod\_info\_type | Table for the Scheme |
| dat\_collection | Table for Collection |
| dat\_collection\_resource | Table for the link between a concept and a collection |
| mod\_status | Table of the content status |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

#### NLP4StatRef Ontology ¤ AuthorityList

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:AuthorityList | rdf:type | dct:description | dct:contributor |
| dct:type | dct:modified | dct:created |
| estat:resourceInformation | dct:publisher | dct:creator |
|  | dct:title | dct:isPartOf |
|  |  | dct:issued |
|  |  | dct:license |
|  |  | dct:rights |

Each item of the Authority List is an individual of skos:Concept:

|  |  |
| --- | --- |
| Resource Type | http://nlp4statref/resource/authority/resource-type#named-entity |
| Resource Type | http://nlp4statref/resource/authority/resource-type#concept |
| Mandatory properties | dct:type  estat:resourceInformation  skos:inScheme  skos:prefLabel  euvoc:status |

## Code lists

<https://ec.europa.eu/eurostat/data/metadata/code-lists>

Code lists are predefined, organized sets of items that describe one or more statistical concepts. They are the building blocks for defining statistical indicators. Code lists are used to build multi-dimensional tables.

For the Content and Knowledge databases, a Code list is a file containing Codes/Label pairs. This means 2 different objects to consider:

1. The code list file
2. Each Code/Label pair as a Code list item.

### Requirements and Specifications

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ Resources > Code List |
| Data model | PDF: NLP4StatRef-Code-Lists-Model |
| Mapping | Excel: NLP4StatRef-Code-Lists-Mapping |
| Data collection | Bulk download |
| Resource Type | http://nlp4statref/resource/authority/resource-type#code-list |

There are 2 methods for collect Eurostat Codes lists:

* Scrap the website;
* Download data with the bulk download.

Data are not exactly the same, some information are not available in the bulk download. As the main need for the project is to have all the code list items, the first integration was done using the bulk download.

This integration was done using SQL commands following the mapping.

An update of the content database could be done after scrapping the website. The Content database model is structured for this update.

### Content database model

Code lists are modelized into 2 tables:

|  |  |
| --- | --- |
| Table name | Usage note |
| mod\_code\_list | Table for the Code List file |
| dat\_code\_dico | Table for the Code List items |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

#### NLP4StatRef Ontology ¤ CodeList

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:CodeList | dct:type | dct:publisher | dct:alternative |
| estat:resourceInformation |  | dct:description |
| rdf:type |  | dct:isPartOf |
| skos:notation |  | dct:title |
|  |  | estat:inFamily |

Each pair Code/Label is an individual of skos:Concept:

|  |  |
| --- | --- |
| Resource Type | http://nlp4statref/resource/authority/resource-type#code-list-item  or  http://nlp4statref/resource/authority/resource-type#code-label-item |
| Mandatory properties | dct:type  estat:resourceInformation  skos:inScheme  skos:notation  skos:prefLabel |

## Eurostat Glossary

<https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GLOSSARY&StrNom=CODED2&StrLanguageCode=EN>

Eurostat concept & definition represents 12361 concepts with defining elements about statistical terms that can be found in any statistical work.

### Requirements and Specifications

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ Resources > Eurostat Glossary |
| Data model | PDF: NLP4StatRef-Eurostat-Glossary-Model |
| Mapping | Excel: NLP4StatRef-Eurostat-Glossary-Mapping |
| Data collection | Bulk download |
| Resource Type | https://nlp4statref/resource/authority/resource-type#glossary |

### Content database model

|  |  |
| --- | --- |
| Table name | Usage note |
| Dat\_estat\_glossary | Table for the concept |
| Dat\_estatg\_stat\_unit | Table for the link between term and the statistical unit |
| Dat\_estatg\_measurement\_unit | Table for the link between term and the measurement unit |
| Mod\_satistical\_unit | Table for the statistical unit |
| Mod\_measurement\_unit | Table for the measurement unit |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

Work in progress

## Glossaries ¤ OECD Glossary & SE Glossary

The glossaries of the OECD and of Eurostat's Statistics Explained website have a similar and classic structure. Because of this, they are modeled in the same way in the Content and Knowledge databases.

### Requirements and Specifications

#### OECD Glossary

<https://stats.oecd.org/glossary/>

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ Resources > OECD Glossary |
| Data model | PDF: NLP4StatRef-OECD-Glossary-Model |
| Mapping | Excel: NLP4StatRef-OECD- Glossary -Mapping |

#### SE Glossary

<https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Thematic_glossaries>

|  |  |
| --- | --- |
| Data description | Excel: NLP4StatRef-ResourceDescription  ⮊ Resources > SE Glossary |
| Data model | PDF: NLP4StatRef-SE-Glossary-Model |
| Mapping | Excel: NLP4StatRef-SE-Glossary-Mapping |
| Data collection | Notebook: glossary\_spider.py  cdb\_glossary\_v1.1sql |

### Content database model

Glossaries are modelized into 6 tables, as follows:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_glossary | Table for the concept |
| dat\_further\_info | Table for the further information paragraph |
| dat\_related\_concepts | Table for the related concept paragraph |
| dat\_sources | Table for the sources paragraph |
| dat\_statistical\_data | Table for the statistical data paragraph |
| dat\_redirections | Table for the redirection paragraph |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

#### NLP4StatRef Ontology ¤ Glossary

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:Glossary | rdf:type | dct:publisher | dct:alternative |
| dct:type |  | dct:description |
| dct:title |  | dct:isPartOf |
| estat:resourceInformation |  |  |
| dct:source |  |  |

Each term is an individual of skos:Concept:

|  |  |
| --- | --- |
| Resource Type | http://nlp4statref/resource/authority/resource-type#glossary-term |
| Mandatory properties | dct:type  estat:resourceInformation  skos:inScheme  skos:prefLabel  skos:definition |

## Taxonomies

Classifications available in Eurostat and OECD websites have been created in the form of lexicalized taxonomies in order to enrich the annotations on the one hand and to allow classification alignments on the other:

| Classification | Taxonomy name | URI |
| --- | --- | --- |
| Categories  Statistics Explained website | eurostat-category | http://nlp4statref/resource/taxonomy/eurostat-category |
| Data navigation Tree  Statistics Explained website | eurostat-database-tree | http://nlp4statref/resource/taxonomy/eurostat-database-tree |
| Family  Ramon metadata | eurostat-family | http://nlp4statref/resource/taxonomy/eurostat-family |
| Term Extension  Ramon metadata | eurostat-term-extension | http://nlp4statref/resource/taxonomy/eurostat-term-extension |
| Themes  Eurostat websites | eurostat-theme | http://nlp4statref/resource/taxonomy/eurostat-theme |
| Themes  OECD websites | oecd-theme | http://nlp4statref/resource/taxonomy/oecd-theme |

### Requirements and Specifications

|  |  |
| --- | --- |
| Data model | PDF: NLP4StatRef-Taxonomy-Model |
| Mapping | Excel: NLP4StatRef-Taxonomy -Mapping |
| Resource Type | https://nlp4statref/resource/authority/resource-type#lexicon |

### Content database model

Taxonomies are modelized into a main table associate to 5 tables:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_taxo\_elmnt | Table for the description of the concept |
| dat\_link\_info\_taxonomy | Table for the links between the taxonomy and other content type |
| dat\_taxo\_relations | Table for the taxonomy tree |
| dat\_taxo\_termino | Table for the links between concept and term for the lexicalization of the taxonomy |
| mod\_srm\_concept\_type | Table for the type of the concept: Structuration | Terminology |
| mod\_taxonomy | Table for the taxonomy tree |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

#### NLP4StatRef Ontology ¤ Taxonomy

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:Taxonomy | dct:type | dct:description | dct:created |
| dct:title | dct:modified | dct:isPartOf |
| estat:resourceInformation | dct:publisher |  |
| rdf:type |  |  |

Each term is an individual of skos:Concept:

|  |  |
| --- | --- |
| Resource Type | http://nlp4statref/resource/authority/resource-type#concept |
| Mandatory properties | dct:type  estat:resourceInformation  euvoc:status  skos:broader  skos:inScheme  skos:prefLabel |

## Terminology Lexicon

A lexicon is a list of terms that can be enriched with lexical or syntactic variants. A Terminology lexicon has been done for the annotation need.

### Requirements and Specifications

|  |  |
| --- | --- |
| Data resource | Excel: NLP4StatRef\_Termino-V2  RDF: NLP4StatRef\_Termino |
| Data model | PDF: NLP4StatRef-Lexicon-Model |
| Mapping | Excel: NLP4StatRef-Lexicon-Mapping |
| Resource Type | https://nlp4statref/resource/authority/resource-type#lexicon |

### Content database model

The 2 main tables for Lexicon are:

|  |  |
| --- | --- |
| Table name | Usage note |
| dat\_terminology | Table for the description of the term |
| dat\_variant | Table for the variants of the term |

These tables are described in section 2.2.1. Content database model and the correspondence between the data collected and the above tables is described in the mapping.

### Knowledge database model

#### NLP4StatRef Ontology ¤ Vocabulary

| Class URI | Mandatory prop. | Recommended prop. | Optional prop. |
| --- | --- | --- | --- |
| estat:Vocabulary | rdf:type |  | dct:created |
| dct:type |  | dct:description |
| estat:resourceInformation |  | dct:isPartOf |
|  |  | dct:issued |
|  |  | dct:modified |
|  |  | dct:publisher |
|  |  | dct:title |

Each item of lexicon is an individual of skos:Concept:

|  |  |
| --- | --- |
| Resource Type | http://nlp4statref/resource/authority/resource-type# entity |
| Mandatory properties | dct:type  estat:resourceInformation  skos:inScheme  skos:prefLabel |

1. <https://github.com/eurostat/NLP4Stat/blob/testing/Content%20Database/CDB%20content/Scrapper/Spiders/article_spider.py> [↑](#footnote-ref-1)