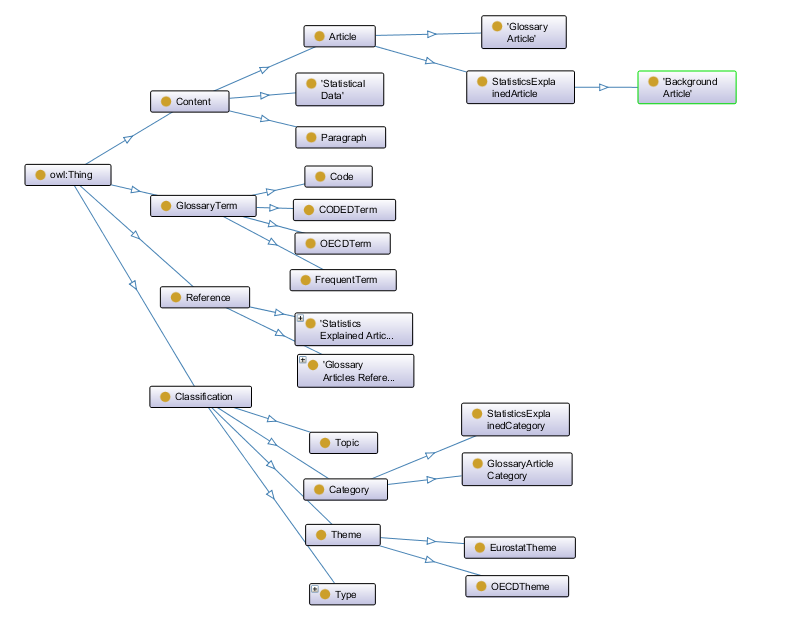
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sector_header | | | | | | |
|  | **NLP4StatRef: Methodological support on advanced methods for accessing, ingesting and linking textual information using semantic analysis and natural language processing** | | | | | |
| partner_flag1  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 |  |  |
|  | June 2022 | | | | | |

Table 1. Descriptions of the classes in the latest ontology.

#### NLP4StatRef Ontology ¤ Classes

| Class | Usage note | URI |
| --- | --- | --- |
|  |  |  |
| Classification | All types of controlled Category, Types, Topics, and others | estat:Classification |
| Category | A class that represents the Categories that EuroStat offers | estat:Category |
| GlossaryArticleCategory | A class that contains the categories for the Glossary Articles | estat:GlossaryArticleCategory |
| StatisticsExplainedCategory | A class that contains the categories for the Statistics Explained Articles | estat:StatisticsExplainedCategory |
| Topic | A class that represents the topics, provided by Quantos | estat:Topic |
| Type | A class that contains various types, statistical units, and measurement units among others | estat:Type |
| ConceptType | A class with the concept types | estat:ConceptType |
| InfoType | A class with the Information types | estat:InfoType |
| LexicalType | A class with the lexical types | estat:LexicalType |
| MeasurementType | A class with the measurement unit types | estat:MeasurementType |
| StatisticalType | A class with the statistical unit types | estat:StatisticalType |
| StatusType | A class with the status types | estat:StatusType |
| Content | All types of described and annotated content. | estat:Content |
| Article | All the information written and published around the statistical data, to comment on them and explain them. | estat:Article |
| StatisticsExplainedArticle | A class with the statistics explained articles | estat:StatisticsExplainedArticle |
| Glossary Article | A class with the Glossary Articles | estat:GlossaryArticle |
| BackgroundArticle | A class with the Statistics Explained Articles that are also Background Articles | estat:BackgroundArticle |
| Statistical Data | Abstract class for statistical content of the datasets | estat:StatisticalData |
| GlossaryTerm | All types of controlled vocabularies, dictionaries, and others. | estat:GlossaryTerm |
| Code | A class that contains the codes that the statistical datasets are related to | estat:Code |
| CODEDTerm | Glossary is an alphabetical list of terms relating to a specific subject enriched by definitions and related content. | estat:CODEDTerm |
| FrequentTerm | A class that contains the named entities existing in EuroStat Content Dataset | estat:FrequentTerm |
| OECDTerm | A class that represents the OECD data | estat:OECDTerm |
| Theme | A class that represents the Themes that EuroStat offers | estat:Theme |
| Paragraph | A class that contains the information for the paragraphs of the Statistics Explained Articles | estat:Paragraph |
| OECDTheme | A class that contains the OECD themes | estat:OECDTheme |
| EurostatTheme | A class with the Eurostat themes | estat:EurostatTheme |
| Reference | A class that represents material (resources / articles / web sources) referenced from within an instance of class content (e.g. an article) | estat:Reference |
| SEAReference | A class that contains all the subclass that indicate the type of references for the statistical explained articles | estat:SEAReference |
| GAReference | A class that contains all the subclass that indicate the type of references for the glossary articles | estat:GAReference |
| FurtherInfo | A class with the further info references for the glossary articles | estat:FurtherInfo |
| RelatedConcept | A class with the related concepts references for the glossary articles | estat:RelatedStatisticalData |
| RelatedStatisticalData | A class with the related statistical data references for the glossary articles | estat:Source |
| Source | A class with the source references for the glossary articles | estat:ExternalLink |
| ExternalLink | A class with the external references / articles / web sources | estat:ExternalLink |
| Visualization | A class with the visualization references / articles / web sources | estat:Visualization |
| Legislation | A class with the legislation references / articles / web sources | estat:Legislation |
| Methodology | A class with the methodology references / articles / web sources | estat:Methodology |
| Publication | A class with the publication references / articles / web sources | estat:Publication |
| DedicatedSection | A class with the dedicated section references | estat:DedicatedSection |
| Database | A class with references to databases | estat:Database |
| Table | A class with references to tables | estat:Table |
| OtherArticle | A class with references to other articles | estat:OtherArticle |
| Excel | A class with references to excellence resources / articles / web sources | estat:Excel |



*Figure 1. The hierarchy of classes in the ESTAT ontology.*

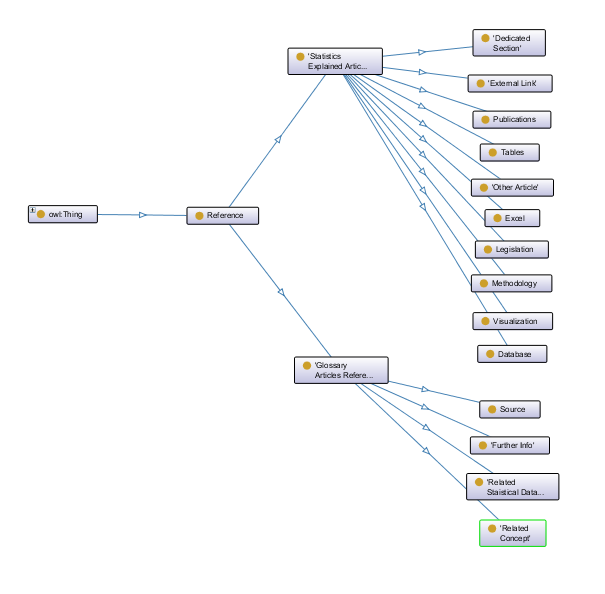


Figure 2. The Reference class.

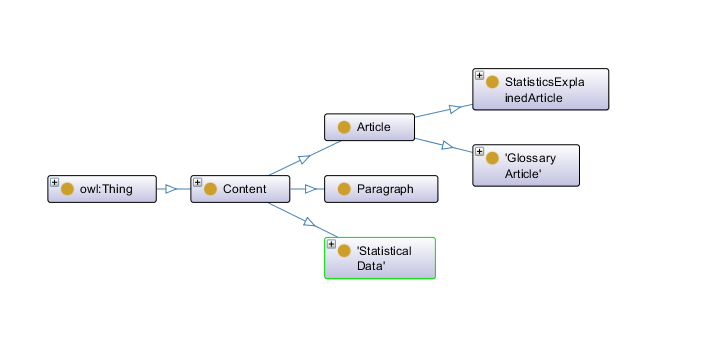


Figure 3. The Content class.

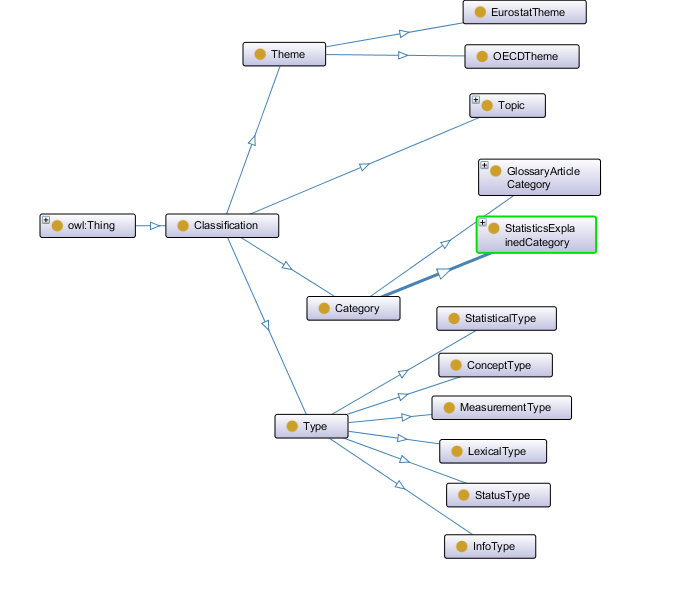


Figure 4. The Classification class.

Table 2. Descriptions of the object properties in the latest ontology.

#### NLP4StatRef Ontology ¤ Object Properties

| Property | URI | Domain | Range | Usage |
| --- | --- | --- | --- | --- |
| hasClassification | estat:hasClassification | estat:Content | estat:Classification | A property that shows if an article from the Editorial Content has any Classification |
| hasCategory | estat:hasCategory | estat:Content | estat:Category | A property that relates an article from the Editorial Content with its Categories |
| hasCategoryOfGlossaryArticle | estat:hasCategoryOfGlossaryArticle | estat:GlossaryArticles | estat: GlossaryArticleCategory | A property that relates an article from the glossary category with its category(-ies) |
| hasCategoryOfStatisticExplainedArticle | estat:hasCategoryOfStatisticExplainedArticle | estat:StatisticsExplainedArticles | estat:StatisticsExplainedCategory | A property that relates an article from the statistics explained category with its category(-ies) |
| hasTopic | estat:hasTopic | estat:Content | estat:Topic | A property that relates an article from the Editorial Content with its Topics |
| hasType | estat:hasType | estat:Content | estat:Type | Some articles from the Editorial Content may have some information about their lexical type, status type, and others |
| hasGlossaryTerm | estat:hasGlossaryTerm | estat:Content | estat:GlossaryTerm | A property that shows if an article from the Editorial Content has any Glossary Term |
| hasCode | estat:hasCode | estat:StatisticalData | estat:Code | A property that relates a dataset to its code(-s). Basically is the labels of the Statistical Data |
| hasFrequenTerm | estat:hasNamedEntity | estat:Content | estat:FrequentTerme | A property that relates an article from the Editorial Content with the named entities that it contains |
| hasCODEDTerm | estat:hasCODEDTerm | estat:Content OR estat:Classification | estat:CODEDTerm | A property that relates entities from the Editorial Content with the CODED Terms |
| hasOECDTerm | estat:hasOECDTerm | estat:Content | estat:OECDTerm | A property that relates entities from the Editorial Content with the OECD Terms |
| hasTheme | estat:hasTheme | estat:Content | estat:Theme | A property that relates the Editorial Content entities with their Theme(-s) |
| hasReference | estat:hasReference | estat:Content OR estat:CODEDTerm OR estat:OECDTerm | estat:Reference | A property that relates articles or terms with references material (e.g. databases, articles, legislation, etc.) |
| relatedTerm | estat:relatedTerm | estat:GlossaryTerm | estat:GlossaryTerm | A property that relaetes glossary terms to another glossary term |
| hasSubTheme | hasSubTheme | estat:EurostatTheme | estat:EurostatTheme | A property that relates a OECD Entity with the EuroStat sub Theme(-s) |
| relatedTheme | estat:relatedTheme | estat:EurostatTheme | estat:OECDTheme | A property that relates a OECD Entity with the EuroStat Theme(-s) |
| hasParagraph | estat:hasParagraph | estat:StatisticsExplainedArticles | estat:Paragraph | A property that relates the Statistics Explained Articles with its paragraphs |
| hasOECDTheme | estat:hasOECDTheme | estat:OECDTerm | estat:OECDTheme | A property that relates the OECD themes with the OECD terms |
| hasURI | estat:hasURI | estat:Reference | estat:Content | A property that relates referenced resources with an entity from the content class |

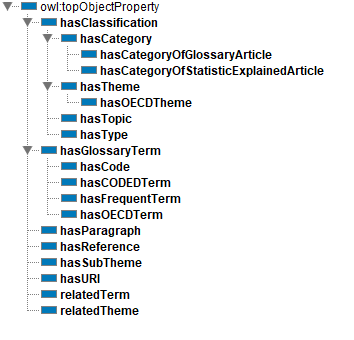


Figure 5. The hierarchy of the object type relations in the new ESTAT ontology.

Table 3. Descriptions of the datatype properties in the latest ontology.

#### NLP4StatRef Ontology ¤ DataType

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

| Property | URI | Domain | Range | Usage |
| --- | --- | --- | --- | --- |
| content | estat:content | estat:Article or estat:Paragraph | xsd:string | This property relates the articles from the editorial content with their Content (i.e., their Abstract or a small Description) |
| context | estat:context | estat:Article OR estat:OECDTerm | xsd:string | A property that relates the Articles from the Editorial Content with their context (a small description of the context that the article is applicable to) |
| dataSource | estat:dataSource | estat:Article | xsd:string | A property that connects the articles from the Editorial Content with their Data Source(-s) (which are strings) |
| dateCreated | estat:dateCreated | estat:Content OR estat:GlossaryTerm | xsd:dateTimeStamp | A property that relates a glossary entity with its creation date |
| dateUpdated | estat:dateUpdated | estat:Content OR estat:GlossaryTerm | xsd:dateTimeStamp | A property that relates a glossary entity with the date that was updated |
| databasePath | estat:databasePath | estat:StatisticalData | xsd:string | A property that relates the datasets with the path of its names |
| definition | estat:definition | estat:CODEDTerm OR estat:OECDTerm | xsd:string | A property that relates a glossary entity with its definition |
| fileDescription | estat:fileDescription | estat:StatisticalData | xsd:string | A property that relates a dataset with its file description |
| id | estat:id | estat:Content OR estat:GlossaryTerm OR  estat:StatisticalData | xsd:integer | A property that indicates the ID of the glossary or the content entity |
| keyword | estat:keyword | estat:Topic | xsd:string | A property that relates the an instance from the class Topic with its keyword(-s) |
| term | estat:term | estat:GlossaryTerm | xsd:string | A property that indicates the label of the dataset and the label of the Vocabulary entity |
| level | estat:level | estat:StatisticalData | xsd:string | A property that indicates the depth that the dataset is in the Statistical Data tree |
| paragraph | estat:paragraph | estat:Paragraph | xsd:string | A property that relates a paragraph from the paragraph class with its context |
| remark | estat:remark | estat:CODEDTerm | xsd:string | A property that relates a glossary entity with its remark |
| sourcePublication | estat:sourcePublication | estat:OECDTerm | xsd:string | A property that relates the OECD enities with their source Publication |
| title | estat:title | estat:Content OR estat:GlossaryTerm | xsd:string | This property relates the articles from the editorial content with their Title |
| fileLink | estat:fileLink | estat:StatisticalData | xsd:anyURI | A property that indicates the link of a dataset |
| hasURL | estat:hasURL | estat:Reference | xsd:anyURI | A property that contains the URL(s) of a referenced resource |

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Figure 6. The hierarchy of the data type relations in the new ESTAT ontology.

Moreover, the NLP4StatRef-Ontology-datasets-v2.3.owl file contains the hierarchy of the Eurostat datasets that exist in the Content Database. This is explained below.

**The datasets ontology**

To explain the principles followed, we note that the Eurostat Database consists of a navigation tree for accessing the various categories of statistical datasets of Eurostat organized in various ways, such as organized by Themes, on EU policy, or on cross cutting topics. Many tables are present in more than one node in the tree. However, when the same table is present in several places, the same code is used to represent it. This led us to model the Eurostat Database as follows.

Each non-terminal node of the tree is modelled as an OWL class and is a subclass of the previous node of the tree. E.g. the “Air Transport” node of the “Tables by Themes” major subtree of the Database is represented as follows:

estat:t\_avia rdf:type owl:Class.

estat:t\_avia rdfs:subClassOf estat:t\_transp.

where t\_avia is the code of this node, and t\_transp is the code for the parent node “Transport”, which itself has “Tables by Themes” as its parent node. The top-level class of this part of the ontology is estat:StatisticalData.

Furthermore, each non-terminal node of the database tree is also represented as an instance of the class estat:StatisticalData, in order to keep information related to the node, information that is represented using properties that have estat:StatisticalData as range. Notice that this dualism (being both a class and an instance) is allowed in OWL2 and is called **punning[[1]](#footnote-2)**. Such a treatment of non-terminal nodes allows SPARQL queries to return both terminal and non-terminal nodes of the database, when a query is addressed to the instances of the class estat:StatisticalData. The “Air Transport” node is represented as follows:

estat:t\_avia rdf:type estat:StatisticalData.

estat:t\_avia rdfs:label "Air transport"^^xsd:string.

estatdata:t\_avia estat:level "2"^^xsd:string.

estat:t\_avia estat:databasePath "Tables by themes; Transport; Air transport"^^xsd:string.

estat:t\_avia estat:hasCode estatdata:t\_avia.

Notice, that for each instance of Class estat:StatisticalData a human-readable text is kept, along with the level of the node in the navigation tree and the path of the nodes up to the root of the tree, that consists of all the labels of the ancestor nodes. Finally, for each node (either terminal or non-terminal) an instance of the estat:Code is also created to host the textual representation of the code of the node. The instance of the estat:StatisticalData class is linked to this latter Code instance through the estat:hasCode object property.

estatdata:t\_avia rdf:type estat:Code.

estatdata:t\_avia estat:term "t\_avia"^^xsd:string.

The following figures represent some part of the navigation tree as a class hierarchy, using either the codes of the nodes or their textual description.



*Figure 7. Part of the latest datasets ontology in the Knowledge Database – textual descriptions.*



*Figure 8. Part of the latest datasets ontology in the Knowledge Database – node codes.*

Terminal nodes of the navigation tree are represented only as instances of estat:StatisticalData class and they represent statistical tables. For example, the “Air transport of passengers by airport and type of transport (monthly data)” table is represented as follows:

estatdata:ttr00017 rdf:type estat:t\_avia.

estatdata:ttr00017 rdfs:label "Air transport of passengers by airport and type of transport monthly data"^^xsd:string.

estatdata:ttr00017 estat:level "3"^^xsd:string.

estatdata:ttr00017 estat:fileLink "https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?sort=1&file=data%2Fttr00017.tsv.gz"^^xsd:anyURI.

estatdata:ttr00017 estat:databasePath "Tables by themes; Transport; Air transport; Air transport of passengers by airport and type of transport monthly data"^^xsd:string.

estatdata:ttr00017 estat:hasCode estatdata:ttr00017.

This terminal node is an instance of its parent node in the tree (t\_avia), which is a class. In addition to the non-terminal nodes, terminal nodes have one more property, which is the link (estat:fileLink datatype property) to the table. Similarly to the non-terminal nodes, there is a also a corresponding estat:Code instance.

estatdata:ttr00017 rdf:type estat:Code.

estatdata:ttr00017 estat:term "ttr00017"^^xsd:string.

Notice that when the same table is accessible from two (or more) different paths of the navigation tree, then the corresponding instance belongs to two (or more) classes at the same time, something that is common in knowledge graphs. E.g. the following represents the table “Air passenger transport between main airports in each reporting country and partner reporting countries” that is reachable both by the “Database by themes” and “Cross cutting topics” subtrees.

estatdata:avia\_paoac a estat:Code,

estat:avia\_pao, estat:cli\_dri\_tran;

rdfs:label "Air passenger transport between main airports in each reporting country and partner reporting countries";

estat:databasePath "Cross cutting topics; Climate change; Drivers; Transport; Air passenger transport between main airports in each reporting country and partner reporting countries", "Database by themes; Transport; Air transport; Air transport measurement passengers; Overview of the air passenger transport by country and airports; Air passenger transport between main airports in each reporting country and partner reporting countries";

estat:fileLink "https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?sort=1&file=data%2Favia\_paoac.tsv.gz"^^<http://www.w3.org/2001/XMLSchema#anyURI>;

estat:hasCode estatdata:avia\_paoac;

estat:level "4", "5";

estat:term "avia\_paoac" .

**Alignment with existing vocabularies**

A recommended mapping from the NLP4StatRef ontology to DCAT is presented in the following table. These mappings are axiomatized using the predicates rdfs:subClassOf, rdfs:subPropertyOf, owl:equivalentClass, owl:equivalentProperty, skos:closeMatch. The table depicts the axiomatisations of the upper-level classes and properties, which are inherited to the subclasses and subproperties, taking into account the semantics of RDFS/OWL. The external vocabularies considered are:

* **DCMI Metadata Terms (dc:** [**http://purl.org/dc/terms/,**](http://purl.org/dc/terms/,) **dcmi:** [**https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/dcmitype/**](https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/dcmitype/)**):** Dublin Core (DC) is an improved digital cataloging system for making search engines much more accurate and efficient. The schema for Dublin Core has many terms for describing resources such as web pages and media like video and images. It also has data regarding physical objects. The main objective is to create a powerful and accommodating catalog involving all web objects. It can be used for better search engine optimization. The metadata generated from this can be used for quick description of web resources and for combining metadata from different standards.
* **SKOS - Simple Knowledge Organization System (skos:** [**http://www.w3.org/2004/02/skos/core**](http://www.w3.org/2004/02/skos/core)**#)**: SKOS provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. Because SKOS is based on RDF, these representations are machine-readable and can be exchanged between software applications and published on the World Wide Web.
* **DCAT - Data Catalog Vocabulary (dcat:** [**http://www.w3.org/ns/dcat**](http://www.w3.org/ns/dcat)**#):** DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. It enables a publisher to describe datasets and data services in a catalog using a standard model and vocabulary that facilitates the consumption and aggregation of metadata from multiple catalogs. This can increase the discoverability of datasets and data services.
* **Schema.org (sdo:** [**http://schema.org/**](http://schema.org/)**)**: Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond. The vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model.
* **FOAF Vocabulary (foaf:** [**http://xmlns.com/foaf/spec/)**](http://xmlns.com/foaf/spec/))**:** FOAF is a descriptive vocabulary expressed using RDF and OWL. It is a machine-readable ontology describing persons, their activities and their relations to other people and objects.
* **RDF Data Cube Vocabulary (qb:** [**http://purl.org/linked-data/cube**](http://purl.org/linked-data/cube)**#)**: The Data Cube vocabulary is focused purely on the publication of multi-dimensional data on the web. It enables such information to be represented using the RDF standard and published following the principles of linked data. The vocabulary is based upon the approach used by the SDMX ISO standard for statistical data exchange. The cube model is very general and so the Data Cube vocabulary can be used for other data sets such as survey data, spreadsheets and OLAP data cubes.

| **NLP4StatRef element** | **Alignment axiomatisation** | **External element** |
| --- | --- | --- |
| estat:Classification | rdfs:subClassOf | skos:Concept |
| estat:MeasurementType | rdfs:subClassOf | qb:AttributeProperty |
| estat:StatisticalType | rdfs:subClassOf | qb:AttributeProperty |
| estat:Content | rdfs:subClassOf | dcat:Resource |
| estat:Article | rdfs:subClassOf | dcmi:Text  foaf:Document  sdo:Article |
| estat:StatisticalData | rdfs:subClassOf | dcat:Dataset  sdo:Dataset  dcmi:Dataset |
| estat:GlossaryTerm | rdfs:subClassOf | skos:Concept |
| estat:Paragraph | rdfs:subClassOf | dcmi:Text |
| estat:Reference | rdfs:subClassOf | skos:Concept  dcat:Resource |
| estat:ExternalLink | rdfs:subClassOf | sdo:WebPage |
| estat:Legislation | owl:equivalentClass | sdo:Legislation |
| estat:hasClassification | rdfs:subPropertyOf | skos:related |
| estat:hasCategory | rdfs:subPropertyOf | dc:subject |
| estat:hasTopic | rdfs:subPropertyOf | dc:subject |
| estat:hasType | owl:equivalentProperty | dc:type |
| estat:hasGlossaryTerm | rdfs:subPropertyOf | dc:subject |
| estat:hasTheme | owl:equivalentProperty | dcat:theme |
| estat:hasReference | owl:equivalentProperty | dc:references |
| estat:hasParagraph | rdfs:subPropertyOf | dc:hasPart |
| estat:content | rdfs:subPropertyOf | dc:abstract |
| estat:context | rdfs:subPropertyOf | dc:description |
| estat:dateCreated | rdfs:subPropertyOf owl:equivalentProperty | dc:date  dc:issued |
| estat:dateUpdated | owl:equivalentProperty | dc:modified |
| rdfs:subPropertyOf | dc:date |
| estat:definition | owl:equivalentProperty | skos:definition |
| estat:fileDescription | rdfs:subPropertyOf | dc:description |
| estat:id | rdfs:subPropertyOf | dc:identifier  dcat:identifier |
| estat:keyword | owl:equivalentProperty | dcat:keyword |
| estat:remark | rdfs:subPropertyOf | skos:note |
| estat:title | owl:equivalentProperty | dc:title |
| estat:hasURL | rdfs:subPropertyOf | dcat:accessURL  sdo:url |
| estat:hasURI | rdfs:subPropertyOf | sdo:sameAs |
| estat:fileLink | rdfs:subPropertyOf | dcat:downloadURL |
| estat:dataSource | rdfs:subPropertyOf | sdo:backstory |
| estat:sourcePublication | rdfs:subPropertyOf | sdo:isBasedOn |
| estat:hasSubTheme | rdfs:subPropertyOf | skos:narrower |
| estat:relatedTerm | rdfs:subPropertyOf | dc:relation  skos:related |
| estat:relatedTheme | rdfs:subPropertyOf | dc:relation  skos:related |

1. [https://www.w3.org/TR/owl2-new-features/#F12:\_Punning](https://www.w3.org/TR/owl2-new-features/%23F12:_Punning) [↑](#footnote-ref-2)