Semantic Web 1 (2012) 1–5
IOS Press

Linked Data Representation of the Nomenclature of Territorial Units for Statistics

Editor(s): Pascal Hitzler, Kno.e.sis Center, Wright State University, Dayton, OH, USA; Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited review(s): Jesse Weaver, Rensselaer Polytechnic Institute, USA; Oscar Corcho, Universidad Politécnica de Madrid, Spain; Michael Hausenblas, DERI Galway, Ireland

Gianluca Correndo [a,*] and Nigel Shadbolt [a]
[a] *School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom.*
E-mail: [gc3,nrs]@ecs.soton.ac.uk

Abstract. The publication of public sector information (PSI) data sets has brought to the attention of the scientific community the redundant presence of location based context. At the same time it stresses the inadequacy of current Linked Data services for exploiting the semantics of such contextual dimensions for easing entity retrieval and browsing. In this paper we describe our Linked Data representation of the NUTS European statistical subdivision, created to support the e-government and public sector in publishing their data sets. The topological knowledge published in the Linked NUTS can be reused in order to enrich the geographical context of other data sets, in particular in a scenario where statistical data sets describe information that have strong ties with the territory, and therefore with its geography.

Keywords: Geographical subdivisions, linked data, RDF, statistical regions, European Union

1. Introduction

Public Sector Information (PSI) has been recognized as the single largest source of information in Europe[1], and the European Union (EU) has supported its public access and reuse since the PSI Directive of 2003 [2].

The publication of authoritative geographies by local governments can open interesting scenarios for exploiting topological knowledge in contextualising the information sources published. For example, the `geoservice` implemented within the EnAKTing project, increases the recall of the information retrieval of our `backlinking` service [4] by considering topological containment between regions. Examples of authoritative geographies already published in linked data format are the administrative geography of Great Britain [5] by the Ordnance Survey (**OS** henceforth) and the statistical geography from Office of National Statistics.

The Nomenclature of Territorial Units for Statistics [3] (NUTS from the french name of the scheme) was established by Eurostat at the beginning of 1970s, to provide a single uniform breakdown of territorial units for the production of regional statistics for the European Union. Each region at the same level is either the expression of a political will or meant to provide comparable features at statistical level (e.g. similar geographical or socio-economic requirements) in order to make comparison and analysis. The NUTS nomencla-

*Corresponding author. E-mail: gc3@ecs.soton.ac.uk.
[1] `http://ec.europa.eu/information_society/policy/psi`

ture serves different purposes in the political life of the European Union. It drives the collection, development and harmonization of statistics through the community as well as supporting a consistent analysis of the collected data. NUTS is also used for the purposes of appraising eligibility for aid from the structural funds from EU.

The current version of the NUTS nomenclature subdivides the territory of the European Union into 97 regions of level 1, 271 regions of level 2, and 1303 regions at level 3. Below that, two levels of Local Administrative Units (LAU) have been defined. The upper LAU level 1 (formerly NUTS level 4) is defined only for the following countries: Bulgaria, Cyprus, Czech Republic, Estonia, Finland, Greece, Hungary, Ireland, Latvia, Lithuania, Luxembourg, Malta, Poland, Portugal, Slovenia, Slovakia and the United Kingdom. The LAU level 2 (formerly NUTS level 5) consists of around 120.000 municipalities or equivalent units in the 27 EU Member States (as of 2007).

Since the NUTS nomenclature encodes a subdivision of a territory that is subject to frequent changes, it is expected to change accordingly. Demographical as well as political and economical indicators in fact evolve yearly making geopolitical tools suddenly obsolete. The NUTS nomenclature in fact, during the last decade, has been revised every three or four years in order to represent new member states and to update the local changes in administrative subdivisions (administrative regions can cease to exist, be split or aggregated to serve local governments' policies).

In this paper we present **Linked NUTS**, a linked data set containing the Nomenclature of Territorial Units for Statistics in RDF format. The **Linked NUTS** represents not only the geographical hierarchy of regions but also its modifications over time, an important prerequisite in order to correctly aggregate and represent statistical data sets for European countries. The creation of the **Linked NUTS** data set was motivated by the lack at the time (i.e. August 2010) of an adequate linked data representation of statistical regions covering the life time of the available data.

2. Related data sets

During the years, many data sets representing the NUTS nomenclature have been published as linked data. Not all of these data sets contain the same kind of information or provide the same kind of access to the RDF documents. The main data sets describing the NUTS are described in the following sections. A description of the proposed **Linked NUTS** will be provided in Section 3.

2.1. Eurostat (FUB)

The Freie Universität Berlin has republished a portion of the geographical data provided by the Eurostat in structured format as Linked Data[2]. The data set contains a list of countries, groups of countries like the EU, and regions of Europe (one of the latest versions of the NUTS). The entities are resolvable via content negotiation in RDF/XML format and a SPARQL endpoint is provided. There is no clear ontological commitment of the data, containment relations are expressed only between a region and its country of belonging despite the various levels of containment that are present in the NUTS nomenclature. Furthermore regions' descriptions provide no indication of the time validity of the statistical observation.

2.2. NUTS-RDF

GeoVocab.org has published the latest version of the latest NUTS nomenclature as linked data[3] in which are included countries like Norway and Turkey, not present in previous versions. Note that there is only one version of the NUTS nomenclature and no temporal validity is provided for the represented regions.

The entities are resolvable via content negotiation in RDF/XML and Turtle format but no API or SPARQL endpoint is provided. A data dump is provided for each version of the data set (two versions are provided so far).

This data set mainly provides three kind of information: topological relationships with other regions (not only NUTS), alignment towards other data sets, and shape information. The vocabulary used to describe the topology is the *NeoGeo Spatial Ontology* whose semantics is based on RCC8 [1] while the regions' shapes are provided in a range of formats (e.g. KML and RDF).

[2]Metadata available at: `http://thedatahub.org/en/dataset/fu-berlin-eurostat`

[3]Metadata available at: `http://thedatahub.org/en/dataset/nuts-geovocab`

2.3. Eurostat NUTS dump

The Eurostat provides itself an RDF dump of the 2008 version of the NUTS[4]. Unfortunately the URIs for the single regions are hash based and the download of the whole file which defines the whole nomenclature is required every time an entity is resolved. In this dump only the hierarchical structure and the labels are provided.

2.4. EIONET

The European Environment Information and Observation Network (EIONET for short) provides a linked data representation of the NUTS nomenclature[5].

The entities are resolvable via content negotiation only, providing documents in RDF/XML and Turtle format. The represented entities provide only topological information and labels but no temporal validity is expressed in the data set.

3. Linked NUTS data set

Within the research activities of the EnAKTing project[6] many data sets has been published as linked data and in all of them, geographical entities were used. In order to extend to the whole EU the same support provided by the OS ontology and the services developed we published the **Linked NUTS** data set[7].

In **Linked NUTS**, every region in the nomenclature is available as an HTTP resolvable URI only. The URIs from the data set follow the following regex pattern:

`http://nuts.psi.enakting.org/id/.+`

From the main page it is possible to access the links to the Eurostat pages which contain the data used to generate the linked data representation. The Eurostat releases the data using an open license with attribution (same license for the boundaries served by the `geoservice`). The Linked NUTS releases the data following an Open Data Commons Attribution license.

By using the content negotiation mechanism, user agents can specify the format of preference and retrieve a document in either RDF/XML or Turtle. When asking for a representation of an NUTS region, clients are redirected via HTTP 303 to a document describing the resource identified by the initial URI; for example:

`http://nuts.psi.enakting.org/id/UKG32`

describes the NUTS 3 region of Solihull, in the West Midlands. When we request the Turtle document describing such entity (e.g. `curl -L -G -H "Accept: application/x-turtle" http://nuts.psi.enakting.org/id/UKG32`) we are redirected to the URL of the Turtle document (i.e. `http://nuts.psi.enakting.org/id/UKG32/ttl`) which contains the RDF document describing the entity (see Figure 1). RDF/XML documents' URIs will be postfixed with `/rdf` and HTML documents with `/doc`.

The Linked NUTS data set contains a total of 2068 regions of different levels, many of which span different versions of the nomenclature. In fact we can count: $1,342$ regions for the first version; $1,436$ the second; $1,779$ for the third and fourth version. We can count therefore many regions that are valid for more than one version. The definition of the NUTS regions, except the alignment towards Freebase, counts $17,700$ triples, $13,362$ of which employ an object property (`rdf:type` included) and only $4,338$ triples employ a datatype property.

The version of NUTS currently covered by this data set does not include the newest version, released the 1^{st} of January 2012. Since the updates to the nomenclature are infrequent any change to the data set is decided on a per request basis.

3.1. Vocabularies used

From the Turtle document shown in Figure 1 we can see that every *NUTS region* has associated: a code (the one assigned by the NUTS nomenclature), a temporal validity, topological information, an alignment towards the linked data cloud, and one or more links to a shape.

Topological relationships between NUTS regions are described using the **OS spatial** ontology[8] whose semantics is based on region connection calculus RCC8 [1]. At the time of creation of the Linked NUTS data set, the **OS** spatial ontology was the most used vocabulary for describing topological relationships. It is in fact used not only to describe the administrative and

[4] Metadata available at: `http://thedatahub.org/en/dataset/eurostat-rdf`

[5] Available at `http://rdfdata.eionet.europa.eu/page/eea/countries/EIONET`

[6] `http://www.enakting.org`

[7] Metadata available at: `http://thedatahub.org/en/dataset/linked-nuts`

[8] `http://www.ordnancesurvey.co.uk/oswebsite/ontology/`

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix fb: <http://rdf.freebase.com/ns/> .
@prefix os: <http://data.ordnancesurvey.co.uk/ontology/spatialrelations/> .
@prefix ward: <http://statistics.data.gov.uk/id/electoral-ward/>.
@prefix nuts: <http://nuts.psi.enakting.org/id/>.
@prefix enakting: <http://nuts.psi.enakting.org/def/>.

nuts:UKG32 a enakting:NUTSRegion; rdfs:label "Solihull";
    enakting:code "UKG32"^^enakting:NUTS3Code ;
    enakting:validity nuts:v1 , nuts:v2 , nuts:v3 , nuts:v4 ;
    owl:sameAs fb:guid.9202a8c04000641f80000000002f23f8 ;
    os:containedBy nuts:UKG3 ;
    os:contains ward:00CTFT , ward:00CTFU , ... ;
    enakting:shapeGeoJson
        <http://geoservice.psi.enakting.org/nuts/polygon?nuts_id=UKG32&format=json> ;
    enakting:shapeKML
        <http://geoservice.psi.enakting.org/nuts/polygon?nuts_id=UKG32&format=kml> .
```

Fig. 1. Turtle serialization for the region of Solihull (UKG32)

electoral subdivisions in UK, but the statistical subdivisions too.

Moreover, since `http://data.gov.uk` already provides the URIs for the further two levels of the British statistical geography, the level 3 NUTS regions for the UK contain one or more LAU level 1 regions from such source. For example, in Figure 1, the region of Solihull contains a number of LAU regions already published by the UK government (e.g. Bickenhill as `ward:00CTFT`, Blythe as `ward:00CTFU` etc.).

One dimension although is not represented in the **OS** spatial ontology, the temporal extent of a given geographical subdivision. Dublin Core provides the means for defining temporal validities of documents although the way to encode time spans make use of Literals and it is not based on any framework. In order to describe temporal validity for the NUTS regions, an entity has been created for each one of the version of the NUTS, starting from the *gentlemen's agreement* of 1999.

Every NUTS region can be valid in one or more versions of the nomenclature, and every version of the nomenclature is represented as an instance of an **OWL Time Interval** (for example `nuts:v1` in Figure 1) and the property `enakting:validity` links every region to the version of the nomenclature valid for the region code. When a new version of the nomenclature is released, only the definition of the last temporal interval is updated by setting a precise date for the interval end. For all the codes that remain unchanged in the new version, only a new triple is added for sanctioning the extended validity of the code (for example in Figure 1 the region of Solihull is valid for all the four versions of the nomenclature). Otherwise the code ceases to be valid, superseded by one or more new regions.

OWL Time ontology has been chosen because it provides a light framework, but with a well founded semantics, for describing temporal entities. Since a *version*, in the data set, is represented as a temporal interval it seemed natural to adopt such ontology.

In order to represent the reorganization of regions in the nomenclature, we defined two additional properties to describe operations over the nomenclature. The transformations of regions after each reorganization of the nomenclature are represented with two properties: `enakting:merge`, and `enakting:split`. A **merge** is the operation that erase two or more regions from the nomenclature to replace them with a new region, and **split** is the inverse operation.

Linked NUTS provides also the shape files for the fourth version of the NUTS nomenclature (i.e. the one valid from 2009 until 2012) which was the latest version at the time of the data set creation. It does so by linking a NUTS region with the `geoservice`[9] developed within EnAKTing project [4]. The shape files are available for all the EU regions mentioned in the NUTS version 4, from the country level (i.e. NUTS 0) to the most granular (i.e. NUTS level 3).

[9] `http://geoservice.psi.enakting.org/`

The `geoservice` provides shape files in two formats, KML and GeoJSON, and the user can decide what format retrieve by using content negotiation or by referring explicitly to the desired document.

3.2. Linked NUTS alignment via `owl:sameAs`

The latest version of the Linked NUTS has been aligned towards the Linked Data cloud, and the alignments have been deployed via the `sameas`[10] service. The alignment has been performed using the Google Refine[11] tool against the Freebase data set and manually checked for errors. The initial alignment provided by Freebase reconciliation service left 230 regions out of $1,351$ not aligned, further 350 regions have been manually aligned using Freebase search functionalities. In total, the alignment counts $1,106$ matches (roughly 82% of all regions). Both alignments, automatic and manual, have been manually checked for correctness.

The choice of aligning the data set to Freebase only is due to the fact that Google Refine provides only this data set as target for reconciliation. However, this is not a limitation since Freebase is only the entry point to the linked data cloud. In fact, via the `sameas` service, which computes and manages the equivalence bundles for aligned resources, it is possible to retrieve equivalent entities from many different domains. For example, resolving the following URL `http://sameas.org/text?uri=http://nuts.psi.enakting.org/id/UKG32` we are able to retrieve almost 60 equivalent entities:

```
<http://metoffice.dataincubator.org/areas/...
<http://dbpedia.org/resource/Solihull>
<http://os.rkbexplorer.com/id/osr700000000...
<http://data.ordnancesurvey.co.uk/id/70000...
<http://dbpedialite.org/things/637106#id>
<http://linkedgeodata.org/triplify/node209...
<http://mortality.psi.enakting.org/id/Soli...
<http://mpii.de/yago/resource/Solihull>
<http://nuts.psi.enakting.org/id/UKG32>
<http://openlylocal.com/id/councils/375>
<http://rdf.freebase.com/ns/m.02z8_s>
<http://statistics.data.gov.uk/id/local-au...
<http://sw.opencyc.org/concept/Mx4rwg_eAMI...
<http://sws.geonames.org/2637546/>
<http://transport.data.gov.uk/id/local-aut...
<http://umbel.org/umbel/ne/wikipedia/Solih...
...
```

[10] `http://sameas.org`

[11] `http://code.google.com/p/google-refine/`

Table 1
Average and standard deviation for equivalent URIs distribution

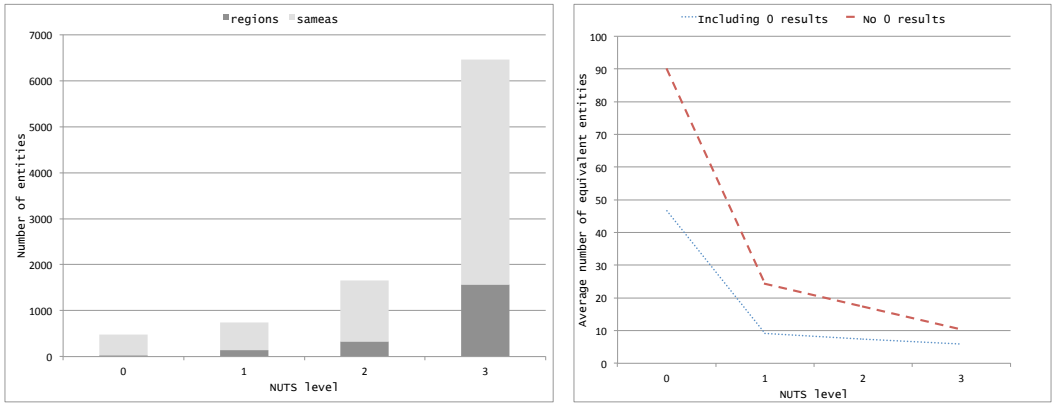
level	μ_l	σ_l
0	89.86	49.06
1	24.27	17.46
2	17.33	12.80
3	10.36	10.03

Once the data set has been aligned to other entities in the Linked Data cloud, it is interesting to see how regions of different size are represented in other data sets. We gathered therefore instance equivalence information using the `sameas` service and run some analysis.

It is worth to note that the `sameas` service considers as "equivalent" entities that are related by properties other than `owl:sameAs`, but whose semantics are close to it (e.g. `skos:exactMatch`, `skos:closeMatch`, `umbel:isLike`, etc.). In particular, for DBpedia, redirects from one resource URI to another (i.e. `dbprop:redirect`) are considered to be equivalence statements. It is then up to the data consumer to decide if the semantics provided by the `sameas` service are good enough, or whether it instead requires a further step of data filtering to take into account only certain types of equivalences.

Figure 2(a) reports the number of regions per level (dark grey bars) and the number of equivalent URIs for those regions (light grey bars). Despite the growing trend illustrated by Figure 2(a), the average number of equivalent URIs decreases as we proceed to lower levels of the subdivision (i.e. smaller regions) as depicted in Figure 2(b). This tells us that the regions belonging to level 0 of the NUTS subdivision (i.e. countries) are, not surprisingly, the most represented in the Linked Data cloud. Figure 2(b) shows the distribution of the average number of equivalent URIs for a region in two cases: considering the case where there are no equivalents as 0, and considering only those regions that have at least one equivalent region (dropping therefore the 0 cases). In both cases the trend vaguely resembles a logarithmic distribution where the bigger the region is, the more equivalent URIs can be retrieved for integrating information, and if we compute then the mean and standard deviation of this distribution we can clearly see an inverse trend happening for the standard deviation (see Table 1).

Considering that the integration task usually includes a phase where equivalent regions' URIs are resolved in order to join pieces of information from the returned documents, we can consider the distributions



(a) Number of equivalent URIs by NUTS level (b) Average number of equivalent URIs by NUTS level

Fig. 2. NUTS entity equivalence statistics

in Table 1 as an indicator of how manageable the integration process could be. For countries in fact (see Table 1, level 0), the standard deviation is very high, indicative of the fact that the data is very sparse and therefore difficult to estimate. For smaller regions the distribution of the equivalent entities converge to a normal distribution with a standard deviation considerably smaller than the previous case. Assuming a normal distribution we can then estimate the expected amount of data to collect from the WoD, making the data collection phase more predictable. We can then conclude that the more fine grained the region is, the more predictable the data collection for that region will be.

4. Conclusions

We have presented in this paper a linked data version of the NUTS statistical geography and how it is linked to external geographical entities in the cloud and to internal services that enrich the description of the regions. An important aspect represented in this data set, which differentiates it from other NUTS representations, is the temporal extent of geographical subdivisions which has changed frequently during the years. New entities can be defined, old ones can be abolished or change status, and this is true for many kind of geography. For example, in the UK administrative geography, Southampton, once part of Hampshire, became a *Unitary Authority* on the 1^{st} of April 1997. Since then, Southampton has been administratively detached from the county of Hampshire (i.e. not contained any more), although being still part of it as a *ceremonial county*.

Versioning of information resources is an important aspect in linked data community and it is even more important when publishing Public Sector Information, whose content and validity must be put into context. Within the EnAKTing project we experienced how dynamic the geographical information in governmental data sets can be. We therefore considered beneficial for the community to mint reference URIs for all of the NUTS regions, current and past, providing enough information to reconcile and exploit European statistical data sets over a considerable span of time.

In case of British governmental data, the use of coreference systems allowed us to exploit the knowledge created in one organization (the OS administrative ontology) in different, and potentially novel, data collections, overlapping a qualitative spatial dimension that was not present before [4]. The creation of services like the `sameas`, the `geoservice`, and the `backlinking` service enable us to reuse local knowledge (topological containment for example) in different contexts. Posing at the same time many questions about the quality management of the knowledge and the entity alignments used; what kind of classification to use, and which alignments to trust.

Acknowledgements

This work was supported by the EnAKTing project funded by the Engineering and Physical Sciences Research Council under contract EP/G008493/1.

References

[1] D. A. Randell, Z. Cui, A. G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, San Mateo, California, USA, 1992 , pages 165–176.

[2] European Union. Directive 2003/98/EC on the re-use of public sector information. *Official Journal of the European Union*, 31 December 2003.

[3] Eurostat. Regions in the European Union. Nomenclature of territorial units for statistics NUTS 2006 /eu-27. In *Eurostat, methodologies and working papers*, 2011.

[4] G. Correndo, M. Salvadores, Y. Yang, N. Gibbins, and N. Shadbolt. Geographical service: a compass for the web of data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2010)*, Raleigh, North Carolina, USA, April 27, 2010, CEUR Workshop Proceedings, ISSN 1613-0073.

[5] J. Goodwin, C. Dolbear, G. Hart. Geographical linked data: The administrative geography of Great Britain on the semantic web. In *Transaction in GIS*, 12 (1) (2009), pages 19–30.