LauNuts: A Knowledge Graph to identify and compare geographic regions in the European Union*

Adrian Wilke and Axel Ngonga

DICE group, Department of Computer Science, Paderborn University
adrian.wilke@uni-paderborn.de, axel.ngonga@upb.de
https://dice-research.org/

Abstract. The *Nomenclature of Territorial Units for Statistics* (NUTS) is a classification that represents countries in the European Union (EU). It is published at intervals of several years and organized in a hierarchical system where geographical areas are subdivided according to their population sizes. In addition to NUTS, there is a further subdivided hierarchy level, named *Local Administrative Units* (LAU), whose data are updated annually by EU member states. While both datasets are published by Eurostat as Excel files, an additional RDF dataset is available for NUTS up to the 2016 scheme. With this work, we provide the Linked Data community with an up-to-date Knowledge Graph in which NUTS and LAU data are linked and which contains population numbers as well as area sizes. We also publish an Open Source generator software for future released versions that will naturally arise due to changes in population numbers. These contributions can be used to enrich other datasets and allow comparisons among regions in the European Union. All resources are available at https://w3id.org/launuts.

Keywords: EU · European Union · Eurostat · Knowledge Graph · LAU · LauNuts · Linked Data · NUTS

Resource type: Knowledge Graph

DOIs: 10.5281/zenodo.7760179, 10.6084/m9.figshare.22272067.v2
Website: https://w3id.org/launuts

1 Introduction: Extension possibilities and contributions

The *Nomenclature of Territorial Units for Statistics* (NUTS) is a hierarchical system in which regions of the European Union (EU) and related states are subdivided. There is an official Resource Description Framework (RDF) dataset

* This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the project EML4U under the grant no 01IS19080B and by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) within the project OPAL under the grant no 19F2028A.

provided by Eurostat, which contains major NUTS concepts. Since some data is not included in the RDF dataset, the following possibilities for extension arise:

E.1 Extension by the finest geographical level, named *Local Administrative Units* (LAU). This level contains data of districts and municipalities and allows a more precise identification of regions.
E.2 Extension by the currently valid version (*NUTS 2021*). The published Eurostat RDF dataset is limited to data up to the *NUTS 2016* version.
E.3 Extension by URIs for different versions. The Eurostat RDF dataset focuses on the respective latest NUTS version. There are no unique URIs for obsolete versions, which would be helpful for, e.g., updating other datasets to revised NUTS versions, which are issued at intervals of three years.

With this work, we present the three following contributions to the Linked Data community:

C.1 A proposal to extend the existing Linked Data scheme by the additional LAU level as well as unique identifiers for published NUTS and LAU versions.
C.2 A Knowledge Graph (KG) generator, which can be used to build and update the KG to NUTS versions released in the future. The generator is implemented to automatically parse the file format used by Eurostat to publish new NUTS and LAU data, which are contained in Excel files.
C.3 A KG built upon the existing concepts as well as a scheme extension along with data officially published by Eurostat and links to additional entities.

The contributions can be used to enhance other KGs and scientific works which include relations to EU regions and their population. Also, tasks like Named Entity Recognition (NER) of geographical entities can be improved by including the hierarchical structure of named regions.

The remainder of this article is structured as follows: Section 2 introduces NUTS and LAU concepts and gives insights into related works. In Sec. 3, an extension of the existing *Eurostat: NUTS - Linked Open Data* dataset concepts is presented. This includes a description of the given scheme (Sec. 3.1), the added concepts of the extension (Sec. 3.2) and the data processing pipeline (Sec. 3.3). Sec. 4 lists statistics of the resulting KG. Finally, Sec. 5 provides a conclusion and an outlook towards future works.

2 Related work: Existing concepts and their usage

Related works comprise the data and schemes published by Eurostat (Sec. 2.1) and scientific works related to NUTS and LAU (Sec. 2.2).

2.1 NUTS and LAU: Hierarchial geographical regions

The *Nomenclature of Territorial Units for Statistics* (NUTS)[1] is a geographical hierarchy of regions. For every member state of the EU and for additional

[1] `https://ec.europa.eu/eurostat/web/nuts/background`

Level	Minimum	Maximum	Example	Code 2021	Population
NUTS 0	Country level		France	FR	67.9 M
NUTS 1	3,000,000	7,000,000	Grand Est	FRF	5.5 M
NUTS 2	800,000	3,000,000	Alsace	FRF1	1.9 M
NUTS 3	150,000	800,000	Bas-Rhin	FRF11	1.1 M
LAU	-		Strasbourg	67482	0.3 M

Fig. 1. NUTS classification criteria based on population thresholds

states like the United Kingdom, respective geographical regions are sub-divided into three levels of detail. The subdivision into levels is based on thresholds of population sizes. The average population size of regions has to range between a minimum and a maximum. Fig. 1 shows the specified thresholds as well as examples of NUTS levels and regions related to Strasbourg.

With exceptions, the current NUTS scheme version is updated every 3 years. The last three versions are 2021, 2016 and 2013. With regard to the version numbers, it has to be noted that there was no large delay in releasing the versions. The naming of the scheme was changed: Up to 2016, schemes were named after the technical date of adoptions, and from 2021, it is when data becomes available. *NUTS 2016* became valid in 2018 and *NUTS 2021* has been valid since 2021. The official description of the current NUTS version was published by Eurostat [4].

The current version *NUTS 2021* comprises sub-divided regions of the 27 EU states Austria (AT), Belgium (BE), Bulgaria (BG), Croatia (HR), Cyprus (CY), Czechia (CZ), Denmark (DK), Estonia (EE), Finland (FI), France (FR), Germany (DE), Greece (GR), Hungary (HU), Ireland (IE), Italy (IT), Latvia (LV), Lithuania (LT), Luxembourg (LU), Spain (ES), Malta (MT), Netherlands (NL), Poland (PL), Portugal (PT), Romania (RO), Slovakia (SK), Slovenia (SI) and Sweden (SE), as well as the United Kingdom (UK). These country regions are sub-divided into 104 regions at the *NUTS 1* level, 283 regions at *NUTS* 2 level and 1,345 regions at *NUTS 3* level.

In addition to NUTS, there is one additional sub-divided level named *Local Administrative Units* (LAU). It consists of municipalities or equivalent units. Up to 2016, this level was sub-divided into two LAU levels. Additionally, it was named *NUTS 4* or rather *NUTS 5* up to 2003.

LAU is updated annually and in the current version (2021), it comprises the states of *NUTS 2021* (listed above) as well as additional data for Albania (AL), Iceland (IS), Liechtenstein (LI), Norway (NO), Switzerland (CH) and Turkey (TR). Related to the state of 2022-06-14, data for the following countries will also be added: Bosnia and Herzegovina (BA), Kosovo (XK), Montenegro (ME), Republic of North Macedonia (MK) and Serbia (RS). Along with the

related NUTS regions, the respective area sizes and populations are published. This data has been used in statistical and scientific works.

2.2 Usage of NUTS and LAU in statistical and scientific works

Statistical evaluations based on NUTS and LAU data were carried out in several domains. In the recent work Coronis [8], multiple public COVID-19 sources were combined with NUTS regions to compare rates of infection. The work is based on GeoVocab, which contains spatial data and was updated in 2011. In the economic domain, rental listings of Greece have been sub-divided into NUTS regions and visualized afterwards [1]. This approach could be applied to other countries and compared afterwards. Farm topology and spatial land in the German state of North Rhine-Westphalia have been combined with LAU level data [7]. It is an example of the usage of extended fine-granulated spatial data where "official statistics provide frequency tables [...] at NUTS 3 and higher level, only". Early works that focused on the UK used NUTS and the related harmonized statistics at the national level [2,3]. However, the URIs are not available anymore. In order to remain sustainably retrievable, our approach is based on a combination of open licensing of code and data, permanent identifiers via w3id.org and generator software that can parse official Eurostat data from the last 10 years and with which future releases can probably also be integrated effortlessly.

NUTS data has been combined with other data sources like postal codes, GeoNames[2] and OpenStreetMap[3] to enable users to search and retrieve information about geo entities [5]. Entities from OpenStreetMap itself have been transformed into RDF data [10].

There are also various visualizations of several domains, mainly published by Eurostat itself: Regions in Europe – 2022 interactive edition[4], Statistical Atlas[5], Statistics Illustrated[6], eurostat-map.js[7], NutsDorlingCartogramn[8], and Regions and Cities Illustrated[9]. To enable other EU projects to build equal works based on RDF, this work extends the existing NUTS Knowledge Graph by LAU data.

3 Extending the existing NUTS Knowledge Graph

In order to extend the existing NUTS KG, we first analyze the officially published RDF data (Sec. 3.1). Based on the scheme characteristics, we propose an extension (Sec. 3.2). In addition, we describe the single steps of the generator software (Sec. 3.3).

[2] `https://www.geonames.org/`
[3] `https://www.openstreetmap.org/`
[4] `https://ec.europa.eu/eurostat/cache/digpub/regions/`
[5] `https://ec.europa.eu/statistical-atlas/viewer/`
[6] `https://ec.europa.eu/eurostat/web/nuts/statistics-illustrated`
[7] `https://github.com/eurostat/eurostat-map.js`
[8] `https://github.com/eurostat/NutsDorlingCartogram`
[9] `https://ec.europa.eu/eurostat/cache/RCI/`

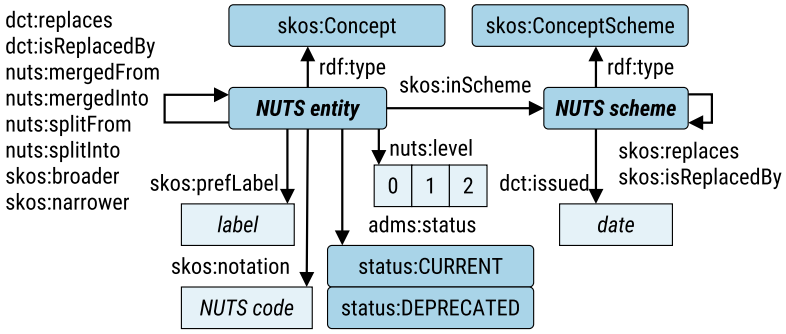

Fig. 2. Scheme of *Eurostat: NUTS - Linked Open Data*

3.1 The Eurostat Linked Open Data scheme

The Eurostat LOD scheme comprises NUTS data from country level (NUTS 0) down to NUTS 3 data. For all levels, the NUTS schemes 2016, 2013 and 2010 are included. The dataset is focused on the newest included version; changes to prior versions are described, e.g. if a region was split. Single NUTS URIs (named *NUTS entities* afterwards) are provided with the related NUTS code, NUTS scheme, label and level. Fig. 2 gives an overview and Tab. 1 lists the namespaces used in this paper.

The RDF dataset is well suited to describe the current NUTS state. However, the following disadvantages result: (a) The currently valid NUTS 2021 scheme is not included. (b) LAU-level data is not included. (c) There is no specific identifier for NUTS entities combined with related NUTS schemes. If additional data is added for a NUTS entity, e.g. population of a region, the related NUTS scheme cannot directly be addressed. (d) The NUTS level 3 is not included as literal; the data is limited to the literals 0, 1 and 2. (d) The properties *replaces* and *isReplacedBy* are part of the Dublin Core vocabulary, the RDF file erroneously uses SKOS. With regard to adding further details for regions, an extension of the scheme is necessary.

Table 1. Used prefixes, namespaces and related vocabularies

Prefix	URI
dct	http://purl.org/dc/terms/
dbo	https://dbpedia.org/ontology/
nuts	http://data.europa.eu/nuts/
owl	http://www.w3.org/2002/07/owl#
skos	http://www.w3.org/2004/02/skos/core#
status	http://publications.europa.eu/resource/authority/concept-status/

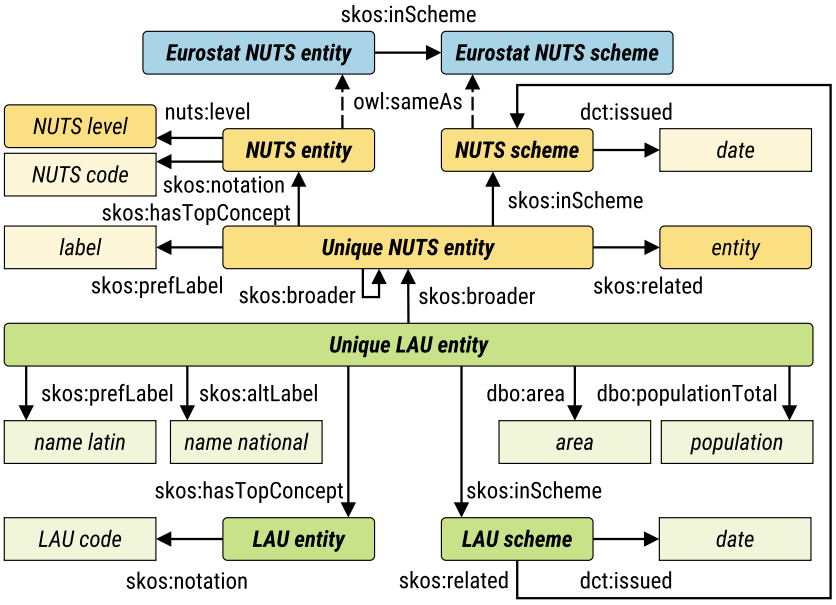

Fig. 3. Extension of Eurostat scheme with LAU data

3.2 Extension of the Eurostat scheme

In order to uniquely address a NUTS entity, we introduce a combination of a *NUTS entity* and a related *NUTS scheme*. This combination is named *Unique NUTS entity* and is shown in Fig. 3. The figure also shows existing Eurostat concepts in blue, while all data generated in our approach is colored yellow and green. Additional concepts from Fig. 2 (e.g. the NUTS label) remain valid but are not additionally visualized. A *Unique NUTS entity* has a label (the name of the respective region in English) and can be related to other entities, e.g. the region URI in Wikipedia, Wikidata or DBpedia. The NUTS hierarchy is represented by *skos:broader* properties between pairs of *Unique NUTS entities*. The inverse narrower direction can easily be inferred and is not explicitly modelled to keep the amount of data to generate low.

In addition, we introduce the same NUTS concepts for LAU-level data. A *Unique LAU entity* is related to both a *LAU entity* with a code and a *LAU scheme* representing the issued year. In addition, we add the respective area and population sizes and use *skos:prefLabel* for Latin names and *skos:altLabel* for names using non-Latin characters. Fig. 4 shows the symmetric design of the scheme and the single parts of URIs, which allow directly addressing NUTS and LAU codes of individual years. LAU entities can be listed by traversing scheme paths (e.g. using SPARQL) and be directly addressed by URIs. In addition to this scheme extension, we processed published data and built a KG.

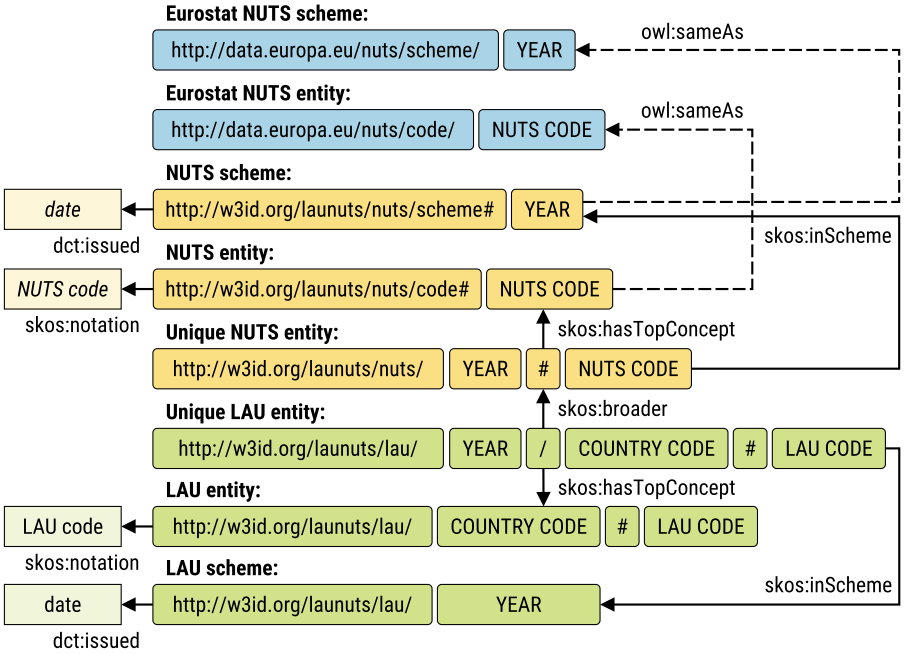

Fig. 4. Extension of Eurostat NUTS URIs with LAU and unique identifiers

3.3 Data analysis and processing

The LauNuts approach was developed in several iterations following the Linked Data life cycle [6]. Actions such as *manual revision* and *quality analysis* towards the final KG generation are included implicitly in every stage of the workflow.

We first *explored* data sources and discovered, inter alia, the officially published sources for NUTS[10], LAU[11] and Linked Open Data[12]. The majority of the data is provided as Excel files. NUTS data is currently available as 7 Excel files for the schemes of the years 2021, 2016, 2013, 2010, 2006, 2003, 1999 and 1995 with 31 sheets in total; for LAU, there are 14 Excel files with 495 sheets for the years from 2010 to 2021. The RDF file related to Linked Open Data contains 20,001 triples.

The *extraction* started with sighting the data. Simply opening some of the Excel files was not possible for the following reasons: Google Sheets ("file is too large to preview"), LibreOffice Calc ("the maximum number of columns per sheet was exceeded") and Apache POI ("OutOfMemoryError: Java heap space"). We finally installed the following extraction queue, explicitly stated here as it could be interesting for other developers working on the topic: (1)

[10] `https://ec.europa.eu/eurostat/web/nuts/history`
[11] `https://ec.europa.eu/eurostat/web/nuts/local-administrative-units`
[12] `https://ec.europa.eu/eurostat/web/nuts/linked-open-data`

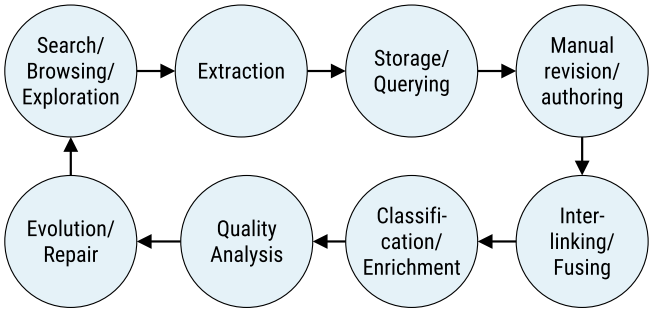

Fig. 5. The Linked Data life cycle

Converting XLS files to XLSX using LibreOffice (7.3.7.2). (2) Converting XLSX files to CSV using ssconvert/Gnumeric (1.12.51). (3) Extracting single sheet names using in2csv/csvkit (1.0.7). (4) Renaming CSV files. The additionally provided Eurostat RDF file could be read using Apache Jena (4.6.1) without any problems.

For further *querying*, the stored CSV and RDF data were used as a cache. To ensure reproducibility, even if the Excel source files are updated in the future, the pre-processed data are published on an FTP server[13].

The *manual revision* of the data started with data analysis of the RDF source. In order to reuse existing Semantic Web concepts, we created a scheme from the available RDF data (see Fig. 2). The scheme is extensible, and details of the most important nuts are provided. However, the predicates used in the RDF file *replaces* and *isReplacedBy* are not part of the used vocabulary SKOS, but DCT (see Tab. 1). Regarding the predicate *nuts:level* and related literals, the NUTS levels 0, 1 and 2 are included, and level 3 is not included. Additionally, the RDF data is limited to the NUTS schemes 2016, 2013, and 2010. The provided Excel files contain values that must be handled individually and partially cleaned. The LAU Excel files provide LAU codes, related NUTS codes, names in Latin characters, names in national (non-Latin) characters, and area and population data. The values of Latin names are sometimes duplicates of non-Latin names. In other cases, there are no Latin names given. Furthermore, some row headings describing the same concepts are named differently in single files. An example of required cleaning is the code FR7, which occurs twice in NUTS 2013. In addition, the LAU 2021 file contains a sheet with 1 million rows, where each contains a cell with a value 0. Overall, the data was evaluated to be usable with additional cleaning.

We *interlinked* the generated data with two data sources. First, the official Eurostat NUTS URIs have been reused. Second, as a proof of concept, we also created links to Wikipedia URIs[14] representing regions at NUTS levels 0

[13] `https://hobbitdata.informatik.uni-leipzig.de/LauNuts/sources/`

[14] `https://en.wikipedia.org/w/index.php?title=First-level_NUTS_of_the_European_Union&oldid=1126125069`

Table 2. Knowledge Graph sizes

	0	NUTS-1	NUTS-2	NUTS-3	LAU	Area	Population	Linked	Triples
2016	28	132	309	1,376	89,284	54,313	45,570	0	568,396
2021	37	162	371	1,551	98,891	98,825	90,074	151	707,816
2021b	65	294	680	2,927	188,175	153,138	135,644	151	1,181,549
All	177	829	1,992	9,122	1,591,703	1,151,106	1,266,290	151	8,039,437

and 1. Therefore, we processed JSON data retrieved using the Wikipedia API and parsed the embedded Markdown code. The Wikipedia URIs can be used to create additional *skos:relatedTo* links to Wikidata and DBpedia as these KGs are also linked to existing Wikipedia URIs.

Additional steps of the Linked Data life cycle are integrated into the used workflow. The *classification* of entities is built in as the overall data integration is based on the used RDF schemes. The *quality analysis* and *evolution* were conducted by several iterations during development and comparing official numbers about the data and concrete values with actually created entities in the KG. The development started in 2019 as part of the OPAL research project and has been used to access geo labels for Question Answering (QA)[9].

4 Results: Open Software and Knowledge Graph

This work provides three main contributions as listed in Sec. 1. The first contribution (C.1) is the scheme extension described in Sec. 3.2, which allows the integration of LAU data versions, which are updated annually. The scheme makes extensive use of common RDF vocabularies. Additionally, created URIs use permanent identifiers of w3id.org to be available in the future.

The generator software (C.2) is published as Open Source (*GNU AGPLv3* license) on GitHub[15]. This enables extensions or reuse of the code in other projects. It is designed to extract NUTS and LAU data in the format of published Eurostat data of the last 10 years; therefore, it is probably possible to effortlessly process data published in the future. The software is parameterized to process only single steps (e.g. data extraction or KG building) and subsets of available data (e.g. only specified NUTS or LAU versions or single country data).

Generated Knowledge Graphs (C.3) contain up to 7 NUTS versions (from 1999 to 2021) and 12 LAU versions (from 2010 to 2021) with labels, area, and population sizes. The current version is named *LauNuts2021b* to be referenced unambiguously and comprises the NUTS schemes 2021 and 2016 as well as LAU data from 2021 and 2020. In addition, entities of the NUTS levels 0 and 1 and the *NUTS 2021* scheme are linked to Wikipedia URIs. As new LAU versions are published annually, new KG versions are expected to be generated in the future. The KG is published under the *CC BY 4.0 International* license on FTP[16],

[15] `https://github.com/dice-group/launuts`
[16] `https://hobbitdata.informatik.uni-leipzig.de/LauNuts/`

Zenodo and Figshare and therefore is accessible by respective Document Object Identifiers (DOI). Tab. 2 shows an overview of contained entities and literals in the KG and sub-graphs for 2021 and 2016. The KG in version *LauNuts2021b* contains 1,181,549 triples.

5 Outlook and Conclusion

5.1 Possibilities for future work

Extending the KG with postal codes could enable a more precise linking to other KGs. Postal codes[17] are available for the NUTS schemes 2021, 2016, 2013, and 2010. For example, for NUTS 2021, there are lists for 35 countries. Mappings between geodata[18] and NUTS as well as LAU codes are available in different file formats and scales. An extension with geodata would enable the identification of geographic regions for given points of interest.

NUTS and LAU codes could complete mappings in well-known Knowledge Graphs independently from the LauNuts KG. In Wikidata, there is the property P605[19] which represents NUTS links. It is already used for entities, e.g. for Alsace[20]. In DBpedia, there is the property nutsCode[21]. It is used, e.g., for Cornwall[22]. The property is listed as an equivalent property to the Wikidata property P605[23].

The generated entities could be completely linked to Wikipedia URLs. For *NUTS 2021*, the levels 0 and 1 have already linked in this work as a proof of concept. Pages in the Wikipedia category *Nomenclature of Territorial Units for Statistics*[24] contain tables with NUTS codes and linked Wikipedia pages. These mappings could then be utilized for KG linking, as Wikipedia pages are also linked from Wikidata and DBpedia.

5.2 Conclusion

With this work, we extended the existing Eurostat KG with the suggestions listed in Sec. 1: We added (E.1) LAU data, (E.2) the current *NUTS 2021* version, and (E.3) URIs for different NUTS and LAU versions.

The KG can be utilized for tasks such as Named Entity Recognition or entity disambiguation by using the provided literals and geographical hierarchy. Other

[17] `https://ec.europa.eu/eurostat/web/nuts/correspondence-tables/postcodes-and-nuts`
[18] `https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units`
[19] `https://www.wikidata.org/wiki/Property:P605`
[20] `https://www.wikidata.org/wiki/Q1142`
[21] `https://dbpedia.org/property/nutsCode`
[22] `http://dbpedia.org/resource/Cornwall`
[23] `http://mappings.dbpedia.org/index.php/OntologyProperty:NutsCode`
[24] `https://en.wikipedia.org/wiki/Category:Nomenclature_of_Territorial_Units_for_Statistics`

use cases are updates of outdated data to the newest NUTS and LAU versions or comparisons of EU regions based on population numbers.

The provided LauNuts KG and the generator software are available with open licensing and are ready to use for upcoming research projects related to EU regions and on the national level.

References

1. Boutsioukis, G., Fasianos, A., Petrohilos-Andrianos, Y.: The spatial distribution of short-term rental listings in greece: a regional graphic. Regional Studies, Regional Science **6**(1), 455–459 (2019). https://doi.org/10.1080/21681376.2019.1660210
2. Correndo, G., Granzotto, A., Salvadores, M., Hall, W., Shadbolt, N.: A Linked Data representation of the Nomenclature of Territorial Units for Statistics. In: Auer, S. et al. (ed.) Proceedings of the Workshop on Linked Data in the Future Internet. vol. 700 (2010), `http://ceur-ws.org/Vol-700/Paper1.pdf`
3. Correndo, G., Shadbolt, N.: Linked Nomenclature of Territorial Units for Statistics. Semantic Web **4**(3), 251–256 (2013). https://doi.org/10.3233/SW-2012-0079
4. European Commission, Eurostat: Statistical regions in the European Union and partner countries : NUTS and statistical regions 2021 : 2022 edition. Publications Office of the European Union (2022). https://doi.org/10.2785/321792
5. Neumaier, S., Savenkov, V., Polleres, A.: Geo-Semantic Labelling of Open Data. In: Fensel, A. et al. (ed.) SEMANTiCS 2018. Procedia Computer Science, vol. 137, pp. 9–20. Elsevier (2018). https://doi.org/10.1016/j.procs.2018.09.002
6. Ngonga Ngomo, A., Auer, S., Lehmann, J., Zaveri, A.: Introduction to Linked Data and Its Lifecycle on the Web. In: Koubarakis, M. et al. (ed.) Reasoning Web. Lecture Notes in Computer Science, vol. 8714, pp. 1–99. Springer (2014). https://doi.org/10.1007/978-3-319-10587-1_1
7. Pahmeyer, C., Schäfer, D., Kuhn, T., Britz, W.: Data on a synthetic farm population of the German federal state of North Rhine-Westphalia. Data in Brief **36**, 107007 (2021). https://doi.org/https://doi.org/10.1016/j.dib.2021.107007
8. Santipantakis, G.M., Vouros, G.A., Doulkeridis, C.: Coronis: Towards integrated and open COVID-19 data. In: Velegrakis, Y. et al. (ed.) EDBT 2021. pp. 686–689. OpenProceedings.org (2021). https://doi.org/10.5441/002/edbt.2021.84
9. Schmidt, M.: A Question Answering (QA) System for the Data Catalog Vocabulary (DCAT). Bachelor's thesis, Paderborn University (2020), `https://github.com/projekt-opal/dcat-qa/blob/thesis/thesis.pdf`
10. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: LinkedGeoData: A core for a web of spatial open data. Semantic Web **3**(4), 333–354 (2012). https://doi.org/10.3233/SW-2011-0052