A Linked Data representation of the Nomenclature of Territorial Units for Statistics

Gianluca Correndo, Alberto Granzotto, Manuel Salvadores, Wendy Hall, and Nigel Shadbolt

Electronics and Computer Science, University of Southampton, Southampton, UK,
{gc3,ag10v,ms8,wh,nrs}@ecs.soton.ac.uk,
WWW home page: http://ecs.soton.ac.uk/people/{gc3,ag10v,ms8,wh,nrs}

Abstract. The recent 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 describes our approach for supporting the publication of geographical subdivisions in Linked Data format for supporting the e-government and public sector in publishing their data sets. The topological knowledge published can be reused in order to enrich the geographical context of other data sets, in particular we propose an exploitation scenario using statistical data sets described with the SCOVO ontology. The topological knowledge is then exploited within a service that supports the navigation and retrieval of statistical geographical entities for the EU territory. Geographical entities, in the extent of this paper, are linked data resources that describe objects that have a geographical extension. The data and services presented in this paper allows the discovery of resources that contain or are contained by a given entity URI and their representation within map widgets. We present an approach for a geography based service that helps in querying qualitative spatial relations for the EU statistical geography (proper containment so far). We also provide a rationale for publishing geographical information in Linked Data format based on our experience, within the EnAKTing project, in publishing UK PSI data.

1 Introduction

The Linked Data Initiative represents the first collaborative effort to create a *Web of Data* (WoD henceforth) of considerable scale, providing few, simple guidelines for publishing content using well established standards [2]. Such guidelines and standards are leading the way to a new paradigm of interaction between government and citizens. In order to pursue better access for citizens to information held by local as well as national public organisations, the UK government has launched[1] his portal for the publishing of Public Sector Infor-

[1] Public access to the site http://data.gov.uk has been granted the 19th of January, 2010.

mation (or PSI), adopting Linked Data tenets as future best practices. Data sets recently delivered to the public include: government expenses, NHS trusts' performances, public transportation, and a whole set of statistics about crime, mortality, census, environment, school and social indicators. Some of the data sets mentioned have been published already in Linked Data format, others have been translated within the EnAKTing project, and many others are waiting to be freed in the LOD cloud.

Such a prolific inflow of Linked Data poses new questions and challenges to the community of researchers and developers: how is it possible to integrate such different information into a meaningful schema? How is it possible to exploit the little semantics that goes a long way? How do we choreograph the publishing activity of separate organizations from the public sector? A common trait of PSI seems to be its locality: national and international public organisations are in fact mainly concerned with the collection of data about their territories, and the distribution of their resources.

In the WoD vision, links between resources from different publishers are particularly important since they are the ones that allow new data to be discovered and integrated into the current discourse. It is frequently the case that different URIs are used to refer to the same things, motivating the use of co-reference services for the resolution of instance equivalences. Knowledge of this type of relationship increases the potential for reuse since information from previously unknown sources is now accessible, and makes the problem of co-reference resolution of primary importance [8]. In any case, we can expect more and more of this linking data to be made available as the number of Linked Data publishers increases.

The publication of an authoritative geography of the UK, (its regions, counties, districts and their connections) by Ordnance Survey (the national mapping agency for Great Britain, OS henceforth) as Linked Data, has opened interesting scenarios for exploiting semantics in contextualising the information sources published on `data.gov.uk`. Aligning, in fact, the geographical dimensions present in statistical data sets to such authoritative data source it is possible to support the information retrieval task and the aggregation of statistical values by means of topological closure [5]. It is therefore important, for supporting the data publishing activity, to establish authoritative sources of geographical knowledge that could provide, not only identity, but also meaning to geographical subdivisions in UK, Europe, and worldwide.

The *Nomenclature of Territorial Units for Statistics* (NUTS henceforth) is a standard geocode scheme used by European statistical agencies for referencing regions where data was collected or aggregated. In this paper, we present our efforts in creating a reference linked data set for NUTS geographical subdivisions for supporting the data transformation, alignment, and retrieval within the European boundaries. In Section 2 we present some background information on the topic of topological representation in linked data. In Section 3 we provide a motivation on why to publish authoritative topologies for geographical subdivisions and in Section 4 we describe a linked data version of the European NUTS geography. The paper then concludes with some concluding notes in Section 5.

2 Background

Many of the PSI data sets published so far can be plotted within a spatial and temporal dimension, in other words, all data can be linked together by its spatial and temporal indexes. Within this context, the need to provide linked entities for such dimensions and the means for supporting reasoning is of key importance. This is unsurprising, the representation and reasoning of spatial and temporal entities have always been considered to be an important part of common-sense reasoning in Artificial Intelligence. In this section, we will mainly focus on qualitative spatial representation and reasoning. There are two major approaches to qualitative spatial representation - point based and region based [4]. Region based approaches, such as Topology [6] which describe relationships between spatial regions are more intuitive than point based approaches. The commonly known approaches for formalizing topological properties of spatial regions are based on work from Whitehead [12] and Clarke [3] who axiomatized mereotopologies (a theory that combines mereology and topology) using a single primitive relation and binary connectivity relationships.

By using these primitive relations, other relations can be defined. The Region Connection Calculus (RCC8) proposed by Randell, Cui and Cohn [10] defines a set of jointly exhaustive and pairwise disjoint relations DC, EC, PO, EQ, TPP, NTPP, TPPi an NTPPi, as illustrated in Figure 1, and is the most well-known approach in the domain. Since the RCC Calculus is expressed in first-order predicate calculus, a wide range of theorem provers can be used for reasoning. For instance, Given a fixed vocabulary of relations, Ri, given R1(x,y) and R2(y,z), one can answer questions about the possible relations (from the set Ri) that can hold between x and z by looking up the composition table [7]. Although general first-order theorem proving is too inefficient to be useful for many purposes [9], it is relatively simple to implement and particularly useful in our case for reasoning are the geographic location relationships.

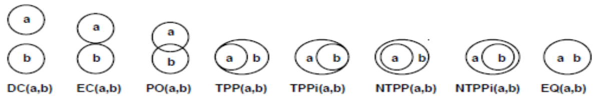

Fig. 1. RCC Eight Jointly Exhaustive and Pariswise Disjoint Relations

Within the Linked Data context, there are several services providing resolvable URIs for geographic locations. Geo-names[2] for example, is a community based service that provides geographical representation of geographical entities covering all countries worldwide and manages eight million URIs for geographical resources. Freebase[3] maintains a collaborative knowledge base of more than

[2] `http://www.geonames.org` last accessed 27/11/2010

[3] `http://www.freebase.org` last accessed 27/11/2010

24 thousands administrative geographic entries worldwide. As a further example, the British national mapping agency, Ordnance Survey, maintains a continuously updated linked data set of the topography of Great Britain and recently released a new version that include topological information at the level of postcodes.

3 Rational for Topological Knowledge Publishing

The Linked Data principles [2] promote a Web of Data whose architecture is inherently decentralised, relying on data already published (when available) in order to give semantics and context to new data. The growth the WoD has experienced over recent years relies on the simplicity of publishing and linking data. However, up to now a semantically coherent orchestration of data publishing is still a mirage. Nevertheless, relying purely on data linkage for the discovery and browsing of linked data resources would lead to a serious knot to untie in the near future. The use of ontologies and powerful ontology languages in publishing Linked Data will be an effort that must be justified against a scenario where such explicit semantics are rarely exploited.

In publishing UK Public Sector Information, but the experience is generalizable, we have identified an issue concerning data accessibility and navigability that addresses in particular the missing exploitation of semantics (in this case about qualitative spatial description of geographical entities). In this paper we present a solution to overcome such issue that soundly enhance data retrieval and browsing when geographical dimensions are involved.

The issue is about the usage of geographical entities for contextualising local information (i.e. information that are related to a particular geographical location, for example the population of a region, the MPs of a constituency, or various statistical data based on territory). In publishing this kind of information, we provided alignments of our data (at least for the geographical dimensions represented in the data) to authoritative knowledge bases using co-reference systems [8]. The problem we have to deal with originates from the fact that, since the public sector information published was originated by different sectors of UK government, the kind of spatial classifications used were highly heterogeneous, ranging from local parishes to counties and up to European regions (e.g. South East of England). The different granularities used to classify the data means, in Linked Data terms, that related information sources link to different URIs. Some data may be in fact relevant for constituencies, while others may use a different granularity (by county for example), and the URI of a county is obviously different from the set of URIs of all its constituencies. Available knowledge bases about the geographical or administrative subdivision of a territory can be exploited to cover such gap in data granularity.

Taking as an example some of the PSI data sets published within EnAKTing, we adopted the Ordnance Survey administrative ontology in order to provide context to our data items (i.e. SCOVO items instances[4] and local governmental

[4] `http://purl.org/NET/scovo`

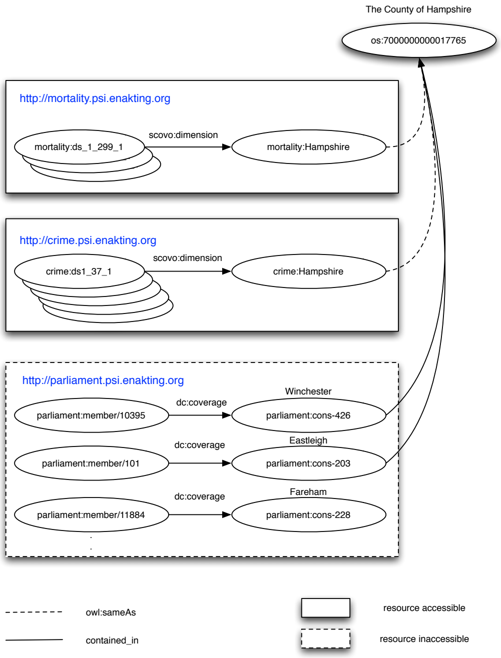

Fig. 2. Resource irretrievable via geographical gap

data). The SCOVO ontology allows us to describe statistical data as a collection of *Items* where each item describes a statistical value (i.e. a single cell in a multidimensional table) along with all the dimensions that characterise it. In the case of UK PSI statistics, many data sets collected were related to geographical regions (counties, districts, etc.)

In this case, users who wished to discover useful information about their own region (e.g. the County of Hampshire, top Figure 2) would start their searching activity by browsing one of its available URIs. The OS URI for such geographical entity would be `os:7000000000017765`[5], but any equivalent URI provided by a co-reference system will provide the same results as will be described in the following. Using a backlinking service [11] for resolving the entities linking to the given URI for Hampshire, we are able to retrieve links to mortality statistics (`mortality:ds1_299_[1...3]`[6]) and crime statistics (`crime:ds1_37_[1...11]`[7]). In Figure 2 those URIs are contained in boxes labelled as *"accessible"*, meaning that those URIs are retrievable following back already existent arcs. Those SCOVO data sets' items address in fact Hampshire county as

[5] `PREFIX os:<http://data.ordnancesurvey.co.uk>`

[6] `PREFIX mortality:<http://mortality.psi.enakting.org/id/>`

[7] `PREFIX crime:<http://crime.psi.enakting.org/id/>`

one of their dimensions. What is missing is the further data collected that reports valuable information about regions contained in Hampshire. In particular, within the EnAKTing project, we published linked data about the singular constituencies too. In detail we published, for each of constituency, an historical record of the MP in charge for that constituency, his/her voting records and expenses. In Figure 2 those resources are contained in dotted boxes labelled as *"inaccessible"*, meaning that they cannot be retrieved with the existent knowledge.

The aim of publishing authoritative topologies (of administrative geographies as well as statistical ones) is to cover such representational gaps, allowing therefore citizens to retrieve information resources relevant to their region of interest. Moreover, the integration of different geographical knowledge bases could lead to the possible reuse of available information sources in contexts different from the one that originated them.

4 Linked Representation of NUTS regions

The Nomenclature of Territorial Units for Statistics [1] (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 nomenclature 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).

4.1 Linked Data representations of NUTS

The hierarchical nature of the NUTS nomenclature can be easily described with the Ordnance Survey ontology[8] whose semantics is based on region connection calculus RCC8 [10]. One dimension although is not represented in such 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 is based on 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. Each *NUTS version* is an instance of OWL time instance class and each *NUTS region* belongs to at least one *NUTS version*.

Every *NUTS region* has a code (the one assigned by the NUTS nomenclature), a label, and a temporal validity. Additionally, two or more regions can be merged into another region or vice versa, one region can be split into two or more regions, due to a reorganization in the nomenclature. Geographical containment information are represented using the OS ontology topological properties, one region can contains one or more other regions creating topologies that can be queried afterwards using a geoservice like the one provided by the EnAKTing project[9].

Every NUTS region is available as resolvable URI at the following address: `http://nuts.psi.enakting.org/id/{NUTScode}` (e.g. `http://nuts.psi.enakting.org/id/UKG32` describe the NUTS 3 region of Solihull). Moreover, since the data set `http://statistics.data.gov.uk` provides URIs for the further two levels of the statistical geography, the level 3 NUTS regions for the UK contains one or more LAU level 1 regions from such source (see Figure 3). For example, the URI for the *Inner London - East* NUTS level 3 region (`http://nuts.psi.enakting.org/id/UKI12`) contains a number of LAU regions whose linked data has been already published by the UK government. All the NUTS regions are aligned to entities in the linked data cloud and available via the sameAs service. In particular the NUTS regions within the UK are aligned to the LAU regions, as already mentioned, in `http://statistics.data.gov.uk`, which contains finer grain subdivisions for statistical purposes, and to administrative regions in `http://data.ordnancesurvey.co.uk` (see Figure 3).

To support the user's experience in browsing and discovery of new resources in the WoD, we have developed a geographical service for querying the UK territory structure. Knowledge about geographical containment is coupled with the instance equivalence knowledge provided by the `http://sameas.org` service and exploited to link information that is contextually related because of their spatial dimension. Recent extensions to the geoservice allow also to provide shape files for plotting regions' borders on a map and create in this way mashups on the fly.

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

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

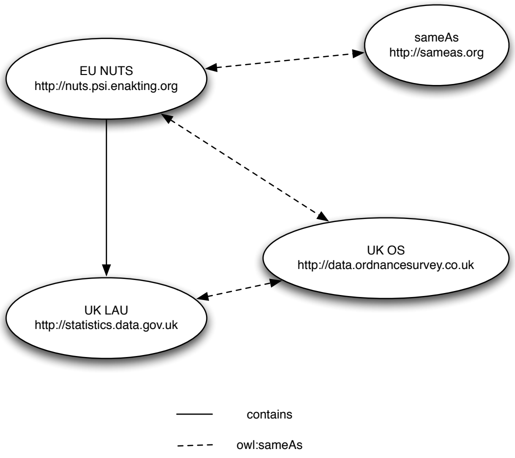
sameAs
http://sameas.org
EU NUTS
http://nuts.psi.enakting.org
UK OS
http://data.ordnancesurvey.co.uk
UK LAU
http://statistics.data.gov.uk
contains
owl:sameAs

Fig. 3. NUTS alignment to UK geographies

An example of such service can be seen in Figure 4 where the polygons of the NUTS region of first level for the UK are rendered in different colours[10]. This service is also integrated with the `http://sameas.org` service in order to bridge the boundaries of data publishing and increase the reuse of such information. Exploiting instance equivalence axioms from data publishers allow us in fact to retrieve and reuse the topological knowledge as well as geometrical information from authoritative sources (the linked data version of NUTS has been created using Eurostat documentation) no matter the starting data set.

5 Conclusions

We have presented in this paper a linked data version of the NUTS statistical geography and a service that helps users in browsing geographical resources within the boundaries of EU. One of the novelties of the geoservice that supports the NUTS data set is the use of a co-reference system (`http://sameas.org`) to extend the containments from one geographic data set to others where such containments are not so rich or complete. Due to the particular nature of the knowledge provided (i.e. closure of geographical containment properties), there is the possibility of overwhelming the user with information when asking about top level features (e.g. `UK`). In order to cope with this eventuality, the service has the capability to limit the results by depth levels. Therefore, when asked about all the entities contained in a top level feature such as `England` at the first level of depth, the service will return only: `East Midlands`, `Northern Ireland`, `East of England`, `Wales`, `West Midlands`, `South East England`, `South West`,

[10] `http://geoservice.psi.enakting.org/geob?uri=http://nuts.psi.enakting.org/id/UK`

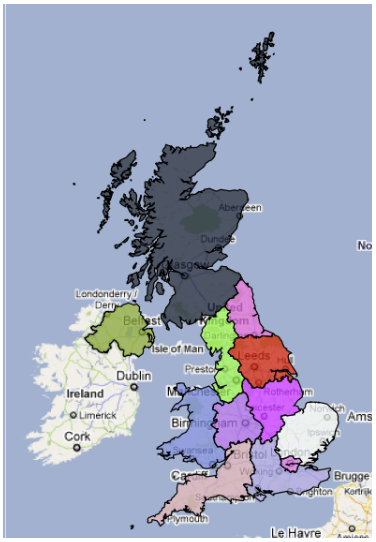

Fig. 4. United Kingdom NUTS level 1 regions

`Scotland`, `Yorkshire & the Humber`, `North West`, `Wales`, `London` (different from `the City of London`).

Another important aspect tackled in this work is the temporal extent of geographical subdivisions. The version of NUTS geography will change shortly (a review is due in 2011) and has changed frequently during the years due to a number of different causes. 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 hot topic in Linked Data community and it is even more important when publishing Public Sector Information, whose content and validity must be put into context.

The research work reported here tackles an important aspect of Linked Data, the exploitation of explicit semantic content for enhancing resource retrieval and browsability. The choice to tackle geographical knowledge rather than some other data facet is mainly due to the analysis of the available data sources, their structure and the available knowledge exploitable for a better integration of the available information.

The use of co-reference systems allowed us to exploit the knowledge created in one organization (Eurostat and Ordnance Survey in this case) in different, and potentially novel, data collections, overlapping a qualitative spatial dimension that was not present before. Such reuse of knowledge is potentially innovative but poses many questions about the management of the quality of the knowledge and the entity alignments used. The presence, integration, and comparison of different geographical knowledge bases can be beneficial for the maintenance and discovery of entity alignments of good quality.

Another interesting aspect related to the use of co-reference services integrated with an additional knowledge source is the ability to exploit the data semantics in order to change the navigability of the datasets. Such change in the navigability is clear when new arcs are provided within the same data set (e.g. between *dbpedia* resource where they were not linked before) or between resources belonging to different data sets [5].

6 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. Regions in the european union. Nomenclature of territorial units for statistics NUTS 2006 /eu-27. Technical report, Eurostat.
2. T. Berners-Lee. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.
3. B. L. Clarke. A calculus of individuals based on “connection”. *Notre Dame J. Formal Logic*, 22(3):204–218, 1981.
4. A. G. Cohn, B. Bennett, J. Gooday, and N. M. Gotts. Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, 1(3):275–316, 1997.
5. G. Correndo, M. Salvadores, Y. Yang, N. Gibbins, and N. Shadbolt. Geographical service: a compass for the web of data. In *Linked Data on the Web (LDOW2010)*, April 2010.
6. M. J. Egenhofer. A formal definition of binary topological relationships. In 3^{rd} *International Conference, on Foundations of Data Organization and Algorithms (FODO)*, pages 457–472, New York, NY, USA, 1989. Springer-Verlag New York, Inc.
7. C. Freksa. Temporal reasoning based on semi-intervals. *Artif. Intell.*, 54(1-2):199–227, 1992.
8. H. Glaser, A. Jaffri, and I. Millard. Managing co-reference on the semantic web. In *WWW2009 Workshop: Linked Data on the Web (LDOW2009)*, April 2009.
9. Grzegorczyk. Undecidability of some topological theories. *Fundamenta Mathematicae*, 38:137–152, 1951.
10. A. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. *In B. Nebel, W. Swartout, and C. Rich, editors, Principles of Knowledge Representation and Reasoning*, 1992.
11. M. Salvadores, G. Correndo, M. Szomszor, Y. Yang, N. Gibbins, I. Millard, H. Glaser, and N. Shadbolt. Domain-specific backlinking services in the web of data. In *Web Intelligence*, September 2010.
12. A. N. Whitehead. *Process and Reality.* The MacMillan Company, New York, NY, USA, 1929.