Immersing evolving geographic divisions in the semantic Web

Camille Bernard

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Camille Bernard

Thèse dirigée par **M. Jérôme Gensel**
codirigée par **M. Hy Dao**
et co-encadrée par **Mme Marlène Villanova-Oliver**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
et de l'**École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Immersing evolving geographic divisions in the semantic Web

Towards spatiotemporal knowledge graphs to reflect territorial dynamics over time

Thèse soutenue publiquement le **27 novembre 2019**,
devant le jury composé de :

Mme Sihem Amer-Yahia
Directrice de recherche CNRS Délégation Alpes, LIG, Université Grenoble Alpes, Présidente

Mme Nathalie Aussenac-Gilles
Directrice de Recherche CNRS Délégation Occitanie Ouest, IRIT, Université de Toulouse, Rapporteur

M. Christophe Claramunt
Professeur des Universités, Institut de Recherche de l'Ecole navale, Rapporteur

Mme Thérèse Libourel
Professeur émérite, Université de Montpellier, Examinatrice

M. Christophe Cruz
Maître de Conférences HDR, Université Bourgogne Franche-Comté, Examinateur

Mme Marlène Villanova-Oliver
Maître de Conférences HDR, Université Grenoble Alpes, Co-Encadrante de thèse

M. Jérôme Gensel
Professeur des Universités, Université Grenoble Alpes, Directeur de thèse

M. Hy Dao
Professeur titulaire, Université de Genève, Co-Directeur de thèse

La Géographie n'est autre chose que l'Histoire dans l'Espace, de même que l'Histoire est la Géographie dans le Temps.

L'Homme et la Terre – Élisée Reclus

Remerciements

Au cours de cette thèse, il y a eu bien des événements et des changements, comme certainement au cours de toutes les thèses d'ailleurs. Une période courte où tant de choses se bousculent, où les doctorants eux-mêmes sont amenés à tant changer, à voyager, faire des rencontres, franchir des obstacles tout en ne perdant pas de vue l'objectif "majeur", rendre un manuscrit et avoir par celui-ci contribué (au moins un tout petit peu) à sa discipline. Ce parcours se clôt par des échanges avec des chercheurs dont bien souvent nous lisons avec admiration les travaux depuis longtemps, c'est combien vous assurer l'honneur que me font l'ensemble des membres de mon jury par leur présence lors de cette soutenance.

Je souhaite exprimer ici en premier lieu mes plus sincères remerciements à Madame Aussenac-Gilles et à Monsieur Claramunt, tous deux rapporteurs de ces travaux. Merci à vous pour le temps consacré à l'évaluation de ces travaux, pour vos remarques et propositions, qui alimentent depuis de nouvelles réflexions. Mes sincères remerciements vont également à Madame Amer-Yahia, Madame Libourel et Monsieur Cruz examinateurs de ces travaux. Un petit mot particulier pour Madame Libourel, qui lors d'une courte discussion au cours de ma thèse, a su donner un nouvel éclairage à ma problématique.

Je remercie ensuite très chaleureusement mes trois encadrants Madame Marlène Villanova-Oliver, Monsieur Jérôme Gensel et Monsieur Hy Dao, pour leur confiance et leur soutien tout au long de ces années. Merci Jérôme pour tes lectures et relectures multiples, un peu partout dans le monde. Merci pour ton humour qui permet à tous de décompresser en réunion. Merci Hy pour tes conseils et ton enthousiasme pour ces travaux. Merci d'avoir fait le pas vers ces problématiques très informatiques. Merci beaucoup Marlène pour tous ces moments de discussion, et de co-construction de ces travaux de recherche, pour tout ce que tu m'as appris, pas uniquement en recherche. Merci pour ta rigueur scientifique, ton sens du détail mais surtout merci pour ta générosité et ton soutien que tu offres à tous les membres de l'équipe STeamer. J'ai pris beaucoup de plaisir à travailler avec vous trois et j'estime avoir eu beaucoup de chance d'être encadrée par trois chercheurs très différents mais aussi très complémentaires.

Si je remonte maintenant le temps pour en revenir au début de ce parcours universitaire, je souhaite en premier lieu remercier chaleureusement mes professeurs du DCISS (en particulier Monsieur Jean-Michel Adam, Monsieur Daniel Bardou et Monsieur Jérôme Gensel) qui, par leur démarche pédagogique, savent accompagner des étudiants en reconversion (avant l'informatique, j'ai étudié la linguistique, pas sans lien d'ailleurs avec ces travaux de thèse, car les cours qui me passionnaient à l'époque étaient alors des cours de sémantique et sémiologie).

Viennent ensuite des années en tant qu'ingénieur auprès de Madame Isabelle Salmon, Monsieur Ronan Ysebaert et Benoit Le Rubrus. Merci à vous trois pour tout ce que vous m'avez appris, j'ai maintenant quelques tic et tocs (de travail)

certainement suite à ces années, parce que vous vous engagez tellement dans ce que vous faites qu'il est difficile ensuite de ne pas en faire autant. Merci beaucoup Benoit pour toutes nos discussions et pour tes encouragements à poursuivre en thèse.

Au commencement de cette thèse, je prends conscience du travail de recherche remarquable accompli par Madame Christine Plumejeaud, au cours de sa thèse quelques années auparavant au LIG, STeamer. Je remercie alors Marlène et Jérôme pour m'avoir fait rencontrer Christine. S'ensuivent des échanges et des collaborations pour des articles et je te remercie Christine pour tes conseils et ta formidable énergie partagés à cette occasion. Ces travaux n'auraient pas pu autant aboutir sans ton implication et tes connaissances.

Je remercie également tous les enseignants-chercheurs de l'équipe STeamer : Paul-Annick Davoine, Danielle Ziebelin (merci Danielle pour toutes nos discussions scientifiques ou non, et pour m'avoir impliquée dans tes projets de recherche), Marlène Villanova-Oliver, Philippe Genoud, Sylvain Bouveret, Jérôme Gensel et toutes les personnes rencontrées au labo depuis le début de cette thèse. Merci aux ingénieurs, aux doctorants, aux stagiaires et mes remerciements tous particuliers à Yagmur Cinar, Lina Toro, Mahdieh Khosravi, Raffaella Balzarini, Aline Menin, Camille Cavalière, Karine Aubry, Maeva Seffar, Fatima Danash, Anna Mannai, Jacques Gautier, David Noël, Anthony Hombiat, Gabriel Lopez, Thibaut Thonet, Lies Hadjadj, Matthieu Viry, Clément Chagnaud, Adrien Dulac, Matthew Sreeves, Georgios Balikas. Un grand merci également pour leur chaleureux accueil à tous mes nouveaux collègues de l'IUT2 Grenoble départements Tech. de Co. et Info-com, et en particulier, merci pour la confiance qu'ont su me faire Didier Schwab et Benjamin Lecouteux. Un grand merci aussi à toutes les personnels techniques et administratifs du laboratoire LIG qui, par leur accompagnement, nous permettent de réaliser nos recherches sereinement. MERCI en particulier à Pascale Poulet, Michèle Hamm, Sylvianne Flammier, Francoise Jeewooth, Zilora Zouaoui, Christiane Plumere et enfin un grand merci à Christian Seguy pour m'avoir dépannée et aidée dans des installations serveur à de nombreuses reprises.

Une thèse sur la filiation me permet de placer facilement mes plus tendres remerciements à ma famille et en particulier à mes chers parents, mon grand frère, et ma petite soeur. Vous êtes là au quotidien pour moi, alors pour vous les changements se font plus continus que discrets, et c'est en ça que je souhaite vous remercier, pour toutes ces petites choses que vous faites et m'apportez depuis toujours. Je voudrais aussi remercier mes ami(e)s Julia (et p'tite patate Lucas), Cécile (et sa petite Lise), Ambre & Laura, Cynthia (et ma belle Alex, mon p'tit Darel), Miléna, Marine, André, Mihary, Anja, Ivan, Kevin. A ma belle-famille et tout particulièrement dadabe Claude et bebe Honorine, Masy, Rindra, Maeva, Josoa. Enfin, je remercie du plus profond de mon coeur Aronarivo le premier a m'avoir parlé de SIG, présent au quotidien pour me faire rire et m'aider à avancer dans mes études -tiako ianao-. A notre fille Clara Line-Tsoa, le plus Grand changement et le plus bel événement survenu au cours de cette thèse.

Résumé

Dans le Web des données ouvertes, on observe de nos jours une augmentation du volume de données provenant du secteur public, d'organismes gouvernementaux et notamment d'instituts officiels de statistique et de cartographie. Ces institutions publient des statistiques géo-codées à travers des découpages géographiques permettant aux responsables politiques de disposer d'analyses fines du territoire dont ils ont la charge. Ces découpages, construits pour les besoins de la statistique mais dérivant généralement de structures électorales ou administratives, sont nommés *Nomenclatures Statistiques Territoriales* (acronyme TSN en anglais). Les TSN codifient les unités géographiques qui, sur plusieurs niveaux d'imbrication (par exemple, en France les niveaux régional, départemental, communal, *etc.*), composent ces territoires. Or, partout dans le monde, les découpages dont ces territoires font l'objet sont fréquemment soumis à des modifications : de nom, d'affiliation, de frontières, *etc.* Ces changements sont un obstacle patent à la comparabilité des données socio-économiques au cours du temps, celle-ci n'étant possible qu'à la condition d'estimer les données dans un même découpage géographique, un processus compliqué qui finit par masquer les changements territoriaux. Dès lors, des solutions conceptuelles et opérationnelles sont nécessaires pour être en mesure de représenter et de gérer différentes versions de TSNs, ainsi que leur évolution dans le temps, dans le contexte du Web des Données Ouvertes. De tels outils permettraient en effet d'améliorer la compréhension des dynamiques territoriales, de documenter les changements territoriaux à l'origine de ruptures dans les séries statistiques et d'éviter des interprétations et manipulations erronées des données statistiques disponibles.

Dans cette thèse, nous présentons un *framework* nommé Theseus qui s'appuie sur les technologies du Web sémantique pour représenter les découpages géographiques et leurs évolutions au cours du temps sous forme de données ouvertes et liées (Linked Open Data (LOD) en anglais). Ces technologies garantissent notamment l'interopérabilité syntaxique et sémantique entre des systèmes échangeant des TSNs. Theseus est composé d'un ensemble de modules permettant la gestion du cycle de vie des TSNs dans le Web des LOD : de la modélisation des zones géographiques et de leurs changements au cours du temps, à la détection automatique des changements, jusqu'à l'exploitation de ces descriptions dans le LOD Cloud. L'ensemble des modules logiciels est articulé autour de deux ontologies nommées *TSN Ontology* et *TSN-Change Ontology*, que nous avons conçues pour une description spatiale et temporelle non ambiguë des structures géographiques et de de leurs modifications au cours du temps.

Theseus s'adresse tout d'abord aux agences statistiques, car il facilite considérablement la mise en conformité de leurs données géographiques avec les directives Open Data. De plus, les graphes de connaissances générés améliorent la compréhension des dynamiques territoriales, en fournissant aux décideurs politiques, aux techniciens, aux chercheurs et au grand public des descriptions sémantiques fines des changements territoriaux, exploitables pour des analyses fiables et traçables. L'applicabilité et la généricité de notre approche sont illustrées par des tests du framework Theseus menés sur trois TSN officielles : la Nomenclature européenne des unités territoriales statistiques (versions 1999, 2003, 2006 et 2010) de l'Institut statistique européen Eurostat ; les unités administratives de la Suisse de l'Office fédéral suisse de la statistique, décrivant les cantons, districts et communes de la Suisse en 2017 et 2018 ; l'*Australian Statistical Geography Standard*, construit par le Bureau australien de la statistique, composé de sept divisions imbriquées du territoire australien, dans les versions 2011 et 2016.

Abstract

On the Open Data Web, there is an increase of the amount of data coming from the public sector. Most of these data are created by government agencies, including Statistical and Mapping Agencies. These institutions publish geo-coded statistics through geographic divisions. These statistics are of utmost importance for policy-makers to conduce various analyzes upon the territory they are responsible for. The geographic divisions built by Statistical Agencies for purposes of data collection and restitution are called *Territorial Statistical Nomenclatures* (TSNs). They are sets of artifact areas although they usually correspond to political or administrative structures. TSNs codify the geographic areas which, on several nested levels (for instance, in the United Kingdom, the regions, districts, sub-districts levels, etc.), compose these territories. However, all around the world, these geographic divisions often change: their names, belonging or boundaries change for political or administrative reasons. Consequently, these changes are a clear obstacle to the comparability of socio-economic data over time, as this is only possible if data are estimated in the same geographical divisions, a complicated process that, in the end, hide territorial changes. Conceptual and operational solutions are needed to be able to represent and manage different versions of TSNs, as well as their evolution over time, in the Open Data Web context. By extension, these tools should improve the understanding of territorial dynamics, document the territorial changes causing breaks in the statistical series and avoid wrong interpretations and manipulations of the available statistical data.

In this thesis, we present the *Theseus Framework*. Theseus adopts Semantic Web technologies and Linked Open Data (LOD) representation for the description of the TSNs' areas, and of their changes: this guaranties among others the syntactic and semantic interoperability between systems exchanging TSN information. Theseus is composed of a set of modules to handle the whole TSN data life cycle on the LOD Web: from the modeling of geographic areas and of their changes, to the automatic detection of changes and exploitation of these descriptions on the LOD Web. All the software modules rely on two ontologies, *TSN Ontology* and *TSN-Change Ontology*, we have designed for an unambiguous description of the areas and of their changes in time and space.

This framework is intended first for the Statistical Agencies, since it considerably helps them to comply with Open Data directives, by automating the publication of Open Data representation of their geographic divisions that change over time. Second, the generated knowledge graphs enhance the understanding of territorial dynamics, providing policy-makers, technicians, researchers, general public with fine-grained semantic descriptions of territorial changes to conduct various accurate and traceable analyses. The applicability and genericity of our approach is illustrated by testing Theseus on three very different official TSNs: The European *Nomenclature of Territorial Units for Statistics* (NUTS) (versions 1999, 2003, 2006, and 2010) from the *European Eurostat Statistical Institute*; The *Switzerland Administrative Units* (SAU), from The *Swiss Federal Statistical Office*, that describes the cantons, districts and municipalities of Switzerland in 2017 and 2018; The *Australian Statistical Geography Standard* (ASGS), built by the *Australian Bureau of Statistics*, composed of seven nested divisions of the Australian territory, in versions 2011 and 2016.

Contents

List of Publications

The following papers were published as part of this thesis:

— **Bernard, C., Villanova-Oliver, M., Gensel, J., and Le Rubrus, B.** (2017a) Spatio-Temporal evolutive Data Infrastructure: a Spatial Data Infrastructure for managing data flows of Territorial Statistical Information, International Journal of Digital Earth, 10(3):257–283,. ISSN: 1753-8947, 1753-8955. DOI-URL https://www.tandfonline.com/doi/full/10.1080/17538947.2016.122200.

— **Bernard, C., Villanova-Oliver, M., Gensel, J., and Dao, H.** (2017b) TSN et TSN-change : ontologies pour représenter l'évolution des découpages territoriaux statistiques. Conférence internationale de géomatique Sageo, Novembre 2017, Rouen, France.

— **Bernard, C., Villanova-Oliver, M., Gensel, J., and Dao, H.** (2018a) Modeling changes in territorial partitions over time: Ontologies tsn and tsn-change. ISBN: 978-1-4503-5191-1. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC'18, pages 866–875. ACM. DOI-URL http://doi.acm.org/10.1145/3167132.3167227.

— **Bernard, C., Villanova-Oliver, M., Gensel, J., and Dao, H.** (2018b) Ontologies pour représenter l'évolution des découpages territoriaux statistiques. Revue Internationale de Géomatique, 28 4 (2018) 409-437. DOI-URL: https://doi.org/10.3166/rig.2019.00069.

— **Bernard, C., Plumejeaud-Perreau, C., Villanova-Oliver, M., Gensel, J., and Dao, H.** (2018c) An ontology-based algorithm for managing the evolution of multi-level territorial partitions. ISBN: 978-1-4503-5889-7. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL'18, pages 456–459, New York, NY, USA. ACM. DOI-URL http://doi.acm.org/10.1145/3274895.3274944.

— **Bernard, C., Plumejeaud-Perreau, C., Villanova-Oliver, M., Gensel, J., and Dao, H.** (will be published in 2020) Semantic graphs to reflect the evolution of geographic subdivisions. Handbook of Big Geospatial Data edited by Martin Werner and Yao-Yi Chiang, Springer.

Acronyms

ASGS Australian Statistical Geography Standard
CLC Corine Land Cover
COSP Change Of Support Problem
ESMS Euro-SDMX Metadata Structure
GUI Graphical User Interface
LAU Local Administrative Units
LOD Linked Open Data
MAUP Modifiable areal Unit Problem
MUA Morphological Urban Areas
NSA National Statistical Agency
NUTS Nomenclature of territorial units for statistics
OGC Open Geospatial Consortium
QB RDF Data Cube
R2RML Relational Databases to RDF Mapping Language
RDF Resource Description Framework
SA Statistical Agency
SAU Switzerland Administrative Units
SDMX Statistical Data and Metadata eXchange
SMA Semantic Matching Algorithm
SPARQL SPARQL Protocol and RDF Query Language
TSI Territorial Statistical Information
TSN Territorial Statistical Nomenclature
TU Territorial Unit
UMZ Urban Morphological Zones
URI Uniform Resource Identifier
VCS Version Control System
W3C World Wide Web Consortium

List of Figures

List of Tables

List of Listing Codes

Chapter 1

Introduction

1.1 Context

Directives or laws are enacted all around the world to open up data to citizens. Publishing data on the Open Data Web is the most common way to achieve this open data movement. Thus, public institutions in the world are facing the challenge of publishing data on the Open Data Web, on behalf of governments or other political organizations. For instance, the *Open Data Directive*[1] in Europe sets up a legal framework to make public sector information widely accessible and reusable. Indeed, according to the European Commission, by allowing public sector data to be re-used one can foster the participation of citizens in political and social life and increase the transparency of public policies. Similarly, in Brazil, the *Information Access Law*[2] has been applied by public authorities since it came into force on 16 May 2012. Then, in 7 years, Brazil has experienced a considerable improvement in the volume of information accessible to population. As a consequence, the volume of data coming from the public sector is growing rapidly. Today, more and more political union, or state (or region) in the world has an Open Data Portal in order to centralize data and make it accessible to citizens (*e.g.*, the opendata.swiss portal[3], the U.S. Government's open data portal[4], the European Union Data Portal[5], etc.).

The main actors in this process are (National) Statistical Agencies (NSA or SA) and Mapping Agencies which create and disseminate official statistics and geographic information (such as the administrative or electoral boundaries), on behalf of their government, in order to monitor their jurisdictions. For the (N)SAs the Open Data challenge is all the more important that their statistics cover a multitude of themes (demographic, economic or environmental, ...), and evolve in time and space, *i.e.*, they are multi-dimensional. Official statistics measure diverse socio-economic or natural phenomena that occur and evolve on these jurisdictions (*e.g.*, demography). The expression *geo-coded statistics* is used to designate such official territorial statistics, meaning that data refer to a territorial reference system, using alphanumerical codes of *geographic areas* (*e.g.*, numbers assigned to each area) Eurostat [2001]. All these *geographic areas* are organized by (N)SAs into what is called

1. https://ec.europa.eu/digital-single-market/en/open-data
2. (Law number 12.527 at http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm
3. https://opendata.swiss/
4. https://www.data.gov/
5. https://data.europa.eu/euodp/en/home?

a *Territorial Statistical Nomenclature* (TSN). A *Territorial Statistical Nomenclature* is a set of artifact geographic areas (also called Territorial Units (TU)) built by (N)SAs to observe a given territory at several geographic subdivision levels (*e.g.*, regions, districts, sub-districts levels). Numerous TSNs exist throughout the world. A *Territory*, in this thesis report, refers to the largest geographic area of the TSN, (that bounds all the others subdivisions), a portion of space (on Earth) well delimited. One example of TSN is the Eurostat [6] nomenclature, called the *Nomenclature of Territorial Units for Statistics* (NUTS) [7], that provides four nested divisions of the European Union (EU) territory, for the collection of EU regional statistics: Level 0 corresponds to the division of the European territory into its State members, Level 1 splits each of the States obtained at Level 0 into major Regions, Level 2 divides each major Region of the Level 1 into basic Regions, and Level 3 into small Regions. The concept of TSN is a central topic in this thesis.

Geo-coded statistics are of utmost importance for policy-makers to conduct various analyses, in time and within the space of their jurisdiction. For instance, using data available at two or more periods of time, they can observe the evolution of the unemployment rate in a given administrative region. As a matter of fact, the analysis of territories through the evolution of statistical series over time allows stakeholders to be aware of the impact of past policies, to understand territory at present time, and to better grasp the future [Bernard et al., 2017]. Thus, there is a strong demand from governments, organizations and researchers regarding time-series of official territorial statistics.

However, even if statistical data are available for several points in time, they are often not comparable through time due to changes in concepts (*e.g.*, definition of unemployment), in acquisition methods or in territorial units. This latter aspect is the focus of the thesis. Past geo-coded data cannot be compared to more recent data if the geographic areas observed have changed in the meantime *i.e.*, data collected in different versions of a TSN are not directly comparable because the observed geographic areas are potentially not the same areas anymore. Then, territorial changes lead to broken time-series and are source of misinterpretations of statistics, and statistical biases when not properly documented. Territorial changes are very frequent in Europe (for instance in France, in 2016, administrative regions have been merged into greater regions) or in the U.S.A., as a result of a well-known process called *gerrymandering* or, more broadly *redistricting*. They lead to broken statistical time-series because data are collected for and from areas that have changed. This problem, well known as the *Change of Support Problem* (COSP) [Openshaw and Taylor, 1979; Gotway Crawford and Young, 2005; Howenstine, 1993], describes the phenomenon where data collected in different zonings or versions of a zoning are not comparable due to potential differences between the geographic areas used as supports for the collected data.

As a result of territorial changes, TSNs also change over time: statistical agencies usually update their geographic areas every 1 to 10 years in order to reflect the real

6. The European Statistical Agency that provides official statistics to the NSAs of the European Union member states

7. `http://ec.europa.eu/eurostat/web/nuts/overview`

world evolution over time. However, the codes of the geographic areas themselves do not necessarily change, even if the areas have changed their name or merged with a neighbor. Then, the main drawback of using such codes for areas lies in their lack of consistency in time and space, as they may designate a region that has changed over time in its boundaries or/and name.

To address this problem, statistical services often transfer former statistical data into the latest version of the TSN. Hence, statistical data sets do not contain traces of territorial changes. However, this non-evolving view hampers a good understanding of the territory life itself. Indeed, changes of the areas are not meaningless because they are decided or/and voted by an authority pursuing some objective. Thereby, solutions for representing different versions of the geographic divisions, and their evolution on the Open Data Web are to be proposed in order to enhance the understanding of the way territories evolve (*i.e.*, which means in this report, understand the way the territories are organized into several geographic areas with neighborhood, governance and genealogy relationships, and understand the way these relationships evolve over time) providing statisticians, researchers, citizens with descriptions to comprehend the motivations and the impact of changes on geo-coded data (on electoral results for instance). In fact, providing an explicit representation of territorial changes through times is a prerequisite to a reliable analysis of time-series of statistical data. Therefore, it is crucial to keep and enrich such information about territorial changes with metadata and other resources available on the Web that may contribute to explain the changes (*e.g.*, societal reasons, historical events).

Going back to the open data challenge for the SAs, it should be noted that there are different degrees of data openness, depending on the data format chosen by institutions. While this format determines what can be made with data available on the Web and how they can be linked with other resources available on the Web.

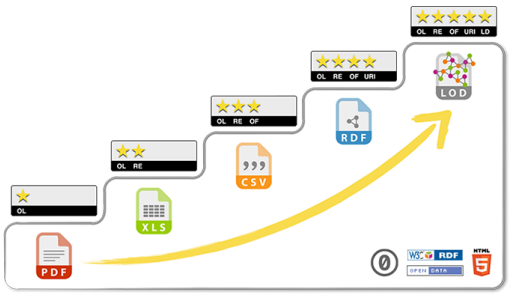


Figure 1.1 – Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data [Hausenblas, 2012].

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a deployment scheme for Open Data (Figure 1.1) that starts from, at least, the publication of data as PDF data on the Web, then as structured data (*e.g.*, within an Excel file instead of an image scan of a table, so that a computer program may

extract the value of each cell). The next step is to make data available but in a non-proprietary format (*e.g.*, comma-separated values instead of Excel). Data reach 4 stars when each of them (*e.g.*, each cell of a table) may be identified uniquely using a *Unique Resource Identifier* (URI), so that people can point at data using their URI. The Open data process ends when linking each URI with each other *i.e.*, each cell data being identified uniquely by a URI is connected to other data on the Web using a *link* (also identified by a URI). Indeed, the more data are linked, the more users can discover new facts, by going from one node of the distributed Web graph database to another. Such a linking of two resources on the Web is called a *triple* (**subject-predicate-object**) (Figure 1.2). Most of the time, the *Resource Description Framework* (RDF) standard syntax is used to write these triples, called RDF triples. For instance, from the data resource "Helsinki" identified by the URI `http://dbpedia.org/resource/Helsinki` on the LOD Web, users discover people born in Helsinki by visiting linked nodes.

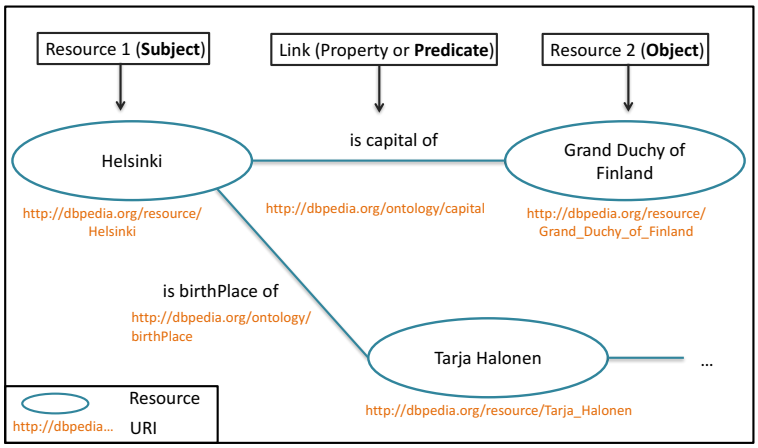


Figure 1.2 – Example of Linked Resources on the Web using URI and the RDF syntax for their identification and representation.

For the (N)SAs, the stakes behind adopting the LOD approach and technologies are crucial because the more contextualized data are (linked to historical, environmental information, etc.), and linked to each other, in time and space, the more analyzes carried out on the territories using these data will be multi-criteria, relevant and reusable by the policies and researchers. As a matter of fact, the more statistical data are linked together (two statistical indicators [8] that describe the same geographic area could be linked together for instance on the spatial dimension of data), the more analysts may explore correlations, causalities and understand the territory under analysis. So far, (N)SAs share their statistics at the data set level. Then, users have to download the whole data set in order to perform their analysis,

8. According to Eurostat (Source: `https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Statistical_indicator`) "*A statistical indicator is the representation of statistical data for a specified time, place or any other relevant characteristic [...]. It is a summary measure related to a key issue or phenomenon and derived from a series of observed facts. Indicators can be used to reveal relative positions or show positive or negative change.*.

even if they are interested in only one indicator or one value of the file. Then, (N)SAs are far from disseminating atomic data from which one may automate the process, re-used, or infer new fact from it. Similarly, to address a specific problem, analysts have to download multiple isolated data sets, each of them resolving only a part of the problem.

Using the LOD paradigm, each statistical data set available on the Web may be identified by a URI representing a resource, while each indicator, each data and metadata composing the data set as well: everything being a resource node in the distributed Web graph database. Thus, data sets as LOD are no longer isolated instances, they are immersed in the Web as graph(s). They are all interconnected, and at a finer granularity, the indicators and the statistical values can be interconnected as well. Furthermore, LOD technologies foster syntactic interoperability, and most of all semantic interoperability between systems. Indeed, for each data published on the LOD Web, it is necessary to define which set of things – real-world objects, events, situations or abstract notion – it belongs to, by linking it to a concept defining this set. As part of LOD, *ontologies* are documents that formally define these concepts and their relations [Berners-Lee et al., 2001]. They help both people and machines to communicate, supporting the share of semantics and not only syntax [Maedche and Staab, 2001]. Hence, using RDF triples, the syntax of data is homogenized, data can be transferred from one system to another, and because of the explicit semantic, systems "understand" data they receive and can determine the appropriate process to be applied (*e.g.*, data tagged as geospatial data may automatically be displayed on a map). Thus, the term **Semantic Web** is also used to denote data sets, ontologies and technologies on the LOD Web. More recently the term **knowledge graphs** emerges to denote graph on the LOD Web containing both data sets and ontologies, bearing formal semantics, which can be used to interpret data and infer new facts [Ontotext, 2018]. A kowledge graph can be envisaged as a network of several data sets and ontologies which are relevant to a specific domain, and on which one can apply a reasoner in order to derive new knowledge [Ehrlinger and Wöß, 2016]. Even if the term is borrowed from a commercial data graph, the *Google Knowledge Graph*, it is now used to denote also available open graphs, such as DBpedia, YAGO, and Freebase [Paulheim, 2017].

Thus, there are many benefits for (N)SAs in using LOD technologies for their statistics:
– users may navigate from one data set to another;
– data and metadata are interlinked;
– systems "understand" the data they receive and can determine the appropriate process to apply to data;
– data are put in context, as they are linked to other resources on the LOD Web;
– each statistical indicators and value become addressable;
– using a network of ontologies and data sets, a knowledge graph representing the dynamics of the territories over time may be constructed, combining several statistical data sets from various disciplines (*e.g.*, environment, socio-economy, ethology, transports), at different instants in time, observed in different meshes and versions of these meshes. Also historical and political information could be mobilized in order to explain the changes over time, such as the changes in boundaries over time.

From this knowledge graph one can build intelligent tools for the restitution of these very disparate data which, taken as a whole, may help in understanding the complexity of the change phenomenon, and the cascading effect of changes through all levels of the European territory, for instance. These tools might also be able to infer new data (such as the estimation of unemployment values after a redistricting), using a cross-sectoral approach for accurate estimation.

(N)SAs have just started publishing their data on the LOD Web to benefit from all the technologies and ontologies around this new data format. Various initiatives to disseminate Linked Statistical Data emerge throughout the world. For instance, the *Aragón Statistical Office Open Data Portal* provides LOD statistics on municipalities of the Aragon region of Spain [9]; the Italian *Istat Linked Open Data Portal* [10] and the *e-Stat Japanese Portal* [11] disseminate statistical LOD for the National SAs of Italia and Japan; the *European Union (EU) Open Data Portal* [12] gives access to (L)OD published by EU institutions such as Eurostat that provides official statistics on the European Union territory.

The W3C *RDF Data Cube* ontology (QB) is widely used to describe these LOD statistics in a way that is compatible with the *Statistical Data and Metadata eXchange* (SDMX), an ISO standard for exchanging and sharing statistical data and metadata among organizations [Cyganiak and Reynolds, 2014]. Using the QB vocabulary, one can publish statistical observations and a set of dimensions that define what the observation applies to: time, gender and geographic areas, for instance. However, the QB ontology does not provide the necessary vocabulary to achieve the description of the geographic areas and of the TSN these areas belong to. TSN levels and TUs have to be described elsewhere, using other ontologies than QB.

The (N)SAs, for their part, often create their own ontology for the description of statistical areas (*e.g.*, the *Territorio Ontology* [13] of the *Italian National Institute for Statistics*, the *Geography Ontology* [14] of the Scottish Government), which results into a counterproductive proliferation of non-aligned vocabularies. Consequently, there is no semantic interoperability between TSN data set nowadays.

The spatiotemporal model of Plumejeaud et al. [2011] (created few years ago by researchers of our STeamer LIG group) offers a way to represent hierarchies of geographic divisions used for statistical purposes. The geographic areas that compose the TSN are described over time, in relation to their ancestors or descendants. The constructed lineages of areas provide policy-makers with information about the changes of the territory observed over time. However, this conceptual model still needs to be immersed in the LOD world if ones want to address the NSAs today's challenge with regard to dissemination of data. This still requires, as noticed by the *Spatial Data on the Web Working Group* [15](a group that gathers members from the

9. http://opendata.aragon.es/
10. http://datiopen.istat.it/index.php?language=eng
11. http://data.e-stat.go.jp/lodw/en
12. http://data.europa.eu/euodp/home
13. http://datiopen.istat.it/odi/ontologia/territorio/
14. http://statistics.gov.scot/vocabularies/
15. https://www.w3.org/2015/spatial/wiki/Main_Page

World Wide Web Consortium (W3C) and the Open Geospatial Consortium (OGC), "*a significant change of emphasis from traditional Spatial Data Infrastructures (SDI) by adopting a Linked Data approach.* " [Tandy et al., 2017].

Regarding TSN spatial data, even if some agencies published their TSN as LOD (*i.e.*, using the RDF syntax), in most cases, TSNs data are available online as **ESRI ®shapefiles** *i.e.*, an open format for geospatial data. If we refer to the Tim Berner-Lee deployment scheme for Open Data (see Figure 1.1), in the field of TSNs, such a deployment reaches level 3. TSN data are, most of the time, not yet available as Linked (Open) Data.

1.2 Problematic

We focus through this thesis report on the spatial dimension of geo-coded statistics, and we try to immerse TSNs on the LOD Web by taking into account the heterogeneity of existing TSNs and their evolving nature. Our work is a continuation of the previous work from Plumejeaud [2011] and extend the proposed model and algorithm for TSNs.

Since no ontology in the LOD Web is generic enough to enable the description of any TSN in the world, semantic interoperability of TSNs is not yet achieved though it would foster the exchange of information among SAs in the world, the comparison of these statistical areas in the world and the processing of these geo-coded statistics. Similarly, no ontology on the LOD Web enables the description of TSN evolution and changes over time while, by describing TSN changes, one can meet stakeholders' needs for tools to:
a) geo-visualize changes in boundaries over time and **understand the way territories evolve** through neighborhood, governance and genealogy relationships [Plumejeaud et al., 2011], and also through information on the historical, societal or legal contexts when the change occurred;
b) automatically **estimate the values of socio-economic indicators** in a new geographic division, using a program able to determine the operation (*e.g.*, aggregation, disaggregation) to be performed on data according to the nature of the territorial change [Goodchild et al., 1980; Flowerdew, 1991];
c) **simulate the evolution of the territory** to observe the effect of a redistricting. For instance, the fusion of two municipalities into one may change the average income per capita, then may impact the budget subventions but, in turn, should reduce the cost of waste collection and treatment;
d) compare two territories each other, especially their **evolution over time**, in order to examine the relevance of following a similar trajectory (in terms of recomposition, extension, etc.) or not.

The fundamental question we address in this thesis report is: How to immerse in the LOD Web any evolving hierarchy of geographic divisions, used for statistical purpose? How to provide statisticians and researchers with semantic descriptions of changes and life lines of evolving hierarchies and geo-spatial entities, paying attention to report on these changes on each of the element composing the hierarchy and on the cascading effects of changes? Secondly, how to interlink these TSNs as LOD and their change descriptions with other data on the LOD Web, towards spatiotemporal knowledge graphs composed of geo-spatial, statistical and historical data, so that experts, but also citizens can explore and exploit these knowledge graphs, cross multiple data to conduct various analyses on many countries, or regions and understand, compare and predict the evolution of these regions?

Building such a knowledge graph implies to overcome four main challenges:

(1) to reduce the lack of semantic interoperability between systems for TSNs exchange.

(2) to provide a description of TSNs evolution over time in a way that it helps understanding the reason for the changes, and assists statisticians in the operation of transferring statistical data from one TSN version to another.

(3) to take into account the vertical/hierarchical dimension of TSNs (*i.e.*, made of several embedded geographic divisions) in their evolutions.

(4) to populate automatically such a descriptive model of TSNs and of their evolution over time, by preserving its genericity, handling the fact that the characteristics of the TUs identity may vary depending on countries and on the quality of the input TSN data sets.

1.3 Contributions

To meet these challenges, we adopt a descriptive approach for the dynamics of the territories, strongly defended within the community of geo information science [Claramunt and Thériault, 1995; Wachowicz, 2003; Del Mondo et al., 2013; Harbelot et al., 2015]. As far as we know, this is the first time that this approach is used in the context of statistics in order to describe the processes that rule the evolution of areas and of the structures these areas belong to. Thus, we extend these descriptive approaches for the dynamics of the territories by applying them to complex, and multi-level territorial structures. We try to describe the evolution of hierarchical links in the structure over time. Our model is multiscalar and allows to zoom in, zoom out to visualize in a global way the main changes of the structure from one version to another, but also all the sub-changes, including the change of name of the smallest unit of the structure, all elements being interconnected in time and space. We focus on the links between the elements of a TSN (hierarchical links, spatiotemporal links, filiation links) and the automatic publication of their description on the LOD Web.

Most of the time in statistics, data sets do not contain traces of territorial changes, *i.e.* the evolution of the areas are lost. While, in our approach, according to Grasland and Madelin [2006], we no longer consider the *Change of Support Problem* as a "Problem" but as a "Potential" that provides information about data and territories. Rather than erasing each former version of a TSN, we preserve it together with the new one. Our challenge is then to automatically find:
– *What* or,
– *Who* (legal authority) is responsible for the change in boundaries,
– *Where*, and
– *When* changes have occurred on the territory,
– *Why* (because of a reform?) the change occurs and above all,
– *How* entities change (*i.e.*, the nature of the change(s) *e.g.*, a fusion of TUs).

Thus, we try to automatically identify, describe and contextualize change events occurring between TSN versions.

Figure 1.3 below presents, in a simplified way, the kind of RDF graph we intend to obtain automatically. Here, the chosen example takes place in France where a new administrative division of the country into region areas has been officially adopted in January 2015 [16].

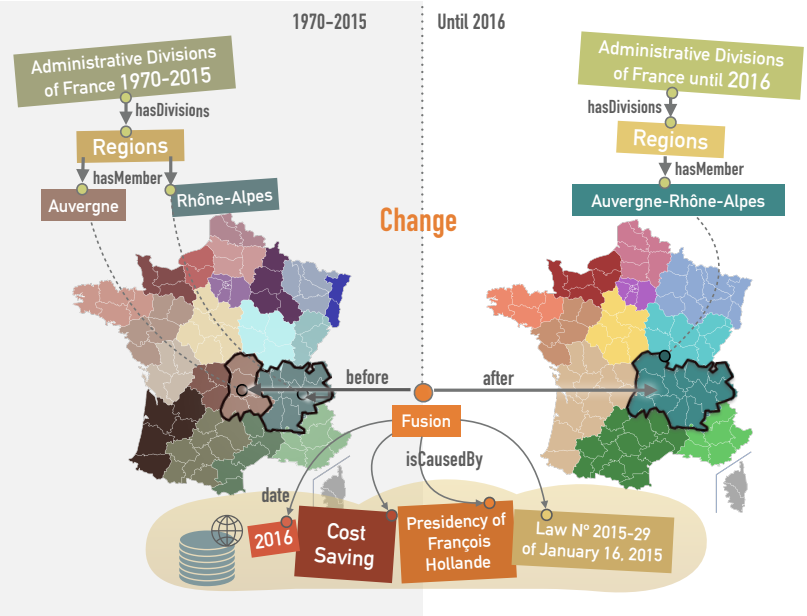


Figure 1.3 – Simplified illustration of the RDF Graph representing territorial changes we want to achieve.

As a first contribution, we present in this thesis report the **Theseus Framework** we have designed in order to supervise the whole life cycle of evolving TSNs on the

16. Law No 2015-29 of January 16th, 2015 https://www.legifrance.gouv.fr/eli/loi/2015/1/16/INTX1412841L/jo/texte

LOD Web: from the modeling of geographic areas to the exploitation of TSNs (and their successive versions) on the LOD Web. This framework is intended for Statistical Agencies, statisticians, or researchers who wish to publish on the LOD Web successive version of their TSN, as well as change and similarity descriptions between the versions *i.e.*, filiation links between the features throughout the versions. Its main objective is to automate the detection and semantic description of a set of well identified processes (merge, split etc.) that characterize the evolution of TSNs and of all their features (levels, TUs), adopting a multi-level approach for the description of the territorial structure, as well as for the description of the changes that impact the embedded features of TSNs.

As a second contribution, this Theseus framework encapsulates two ontologies we have designed, called **TSN-Ontology** and **TSN-Change Ontology**. Their goal is twofold: unambiguous identification of the statistical areas in time and space, and the description of their filiation links (comprising similarity and change descriptions) over time on the Linked Open Data (LOD) Web.

Theseus also embeds an implementation of an extended version of the *Algorithm for Automatic Matching of Two TSN Versions* of [Plumejeaud, 2011]. This adaptation, called the **TSN Semantic Matching Algorithm**, is our third contribution. It has been designed to automate both the detection and the description on the LOD Web of territorial similarities and changes among various TSNs. Together, all the software modules of the Theseus framework contribute to the publication on the LOD Web of **TSNs semantic history graphs**. Those latter constitute catalogs of evolving statistical areas that enhance the understanding of dynamics of the territories, providing statisticians, researchers, citizens with descriptions to comprehend the motivations and the impact of changes.

Theseus is then a step towards the generation of knowledge graphs of evolving geo-coded statistics that link several ontologies and data sets: RDF Data Cube data sets, geospatial TSN data sets, (historical) event data sets, law data sets... from which one could build intelligent tools for the analysis of the territorial dynamics (*e.g.*, analysis the territorial evolution together with demographic, economic, environmental indicators, information on the history of the territories...) . These tools could be capable of inferring new data (such as estimating statistical values in a new TSN version).

1.4 Thesis Outline

In Chapter II, we present in more details the topic of this thesis. We describe the current states of Territorial Statistical Information (TSI), an information made of two peaces: the statistical information (geo-coded statistics), and the geographic areas, organized into Territorial Statistical Nomenclature, used for the collection and dissemination of official territorial statistics. We present the problems caused by the evolution of TSN over time.
In Chapter III, we introduce the main spatiotemporal concepts involved in the modeling of evolving spatial entities, the fundamentals for the modeling of these entities,

and the existing approaches in the semantic Web.
In Chapter IV, we present the fundamentals for the spatiotemporal modeling of evolving TUs and TSNs, that are specific spatial entities, with a focus on the approaches implemented in the semantic Web.
In Chapter V, we evaluate the existing methods for automatically generating, then exploiting in the semantic Web, descriptions of evolving TUs (descriptions of their life, filiation and changes over time).
In Chapter VI, we make a synthesis of the state of the art, and we list the requirements that a system managing the whole life cycle of evolving TSNs in the Semantic Web should meet.
In Chapter VII, we introduce *Theseus*, a configurable framework designed for managing evolving TSNs in the semantic Web, according to a management process that consists of the following activities: Specify, Model, Generate, Publish, and Exploit.
In Chapter VIII, we present the main component of *Theseus*, a ontological model, made of two ontologies, called *TSN Ontology* and *TSN-Change Ontology*, designed in order to describe in the semantic Web any TSN hierarchical structure and its changes over time.
In Chapter IX, we describe the two workflows created in order to populate our TSN ontological model. We present, in particular, the *TSN Semantic Matching Algorithm* designed in order to automatically detect and describe, on the LOD Web, similarities and changes of TSN structures throughout their versions.
In Chapter X, we present experiments performed on three very different TSNs in order to evaluate our Theseus Framework and the performances of the *TSN Semantic Matching Algorithm*.
In Chapter XI, we show how to explore and exploit data created by Theseus, in particular how to explore territorial change descriptions and how to exploit them with other data on the LOD Web, such as Linked Open statistical Data.
In the concluding Chapter, we list the main contributions of the thesis. We discuss the limitations of our work that focuses on a specif type of geographic divisions. We present how one could extend our approach, and some perspectives for future work in this respect.

Part A

State of the Art

Chapter 2

Territorial Statistical Information

In this Chapter, we first describe in more details the current states of Territorial Statistical Information (TSI), an information made of two peaces: statistical (geo-coded statistics, statistics that are spatially referenced), and geographic information (geographic areas used for the collection and dissemination of geo-coded statistics).

Second, we define in this Chapter the focus of our research: namely these geographic areas, organized into Territorial Statistical Nomenclature. We focus on the problem caused by their evolution over time.

Please note that most of the concepts presented in this chapter are defined in the *TSN Ontology* (one of our proposition), available online at `http://purl.org/net/tsn#`. This ontology is an online glossary associated to this manuscript.

2.1 Current states of Territorial Statistical Information

In this manuscript, we focus on geo-coded statistical data that are aggregated data (not point-based data). Indeed, as noticed by [Cheng and Adepeju, 2014]: "*Observations of discrete geographic data are usually made at point locations, but are often aggregated into areal units for various reasons, such as confidentiality of individual records, data summary or to fit into an existing zoning system (e.g., districts, service areas, police beats etc.).*" And, as Openshaw [1984] says, for many purposes the zones in a zoning system constitute the basic units for the observation and measurement of spatial phenomena.

2.1.1 Not fully interconnected data

In Plumejeaud [2011], the author highlights the structure of geo-coded statistics that usually comes in the form of data sets. Statistical data sets can be understood in three levels of information:
(1) the data set level,
(2) the level of statistical indicator(s) (*e.g.*, Eurostat's indicators: *People living in jobless households, Early leavers from education and training,...*) composing the data set, and
(3) the level of data, which describes the indicator values for each TU, subdivision of a territory.

Although statistics produced by (N)SAs, and geographic information produced by National Mapping Agencies have a strong connection because geo-coded statistics measure some observed phenomenon that lives on a territory, most of the time, on the Open Data Web they are available in separated data sets. They are generally distributed in different formats, using different Web Services (SDMX REST Web services for the statistics [1], and OGC Web services (WFS [2], WMS [3], CSW [4]...) for the geographic information). And, even when they are distributed in the same format (CSV or XML for instance), most of the time the two types of information are described using different data models that are not fully interconnected.

We focus now on the European context because the European territory is an interesting case of study: it is a vast and multi-level territory where directives such as the *Open Data Directive* and the *INSPIRE Directive* [5] accelerate the sharing of standardized information across Europe. Some standards enable the description of geographic information (*e.g.* ISO-19115), while others enable the description of statistical information (*e.g.* ISO-17369/SDMX). Sometime, specifications address both statistical and territorial aspects, but every time one of the aspects is much more accurately described then the other. In Europe, standards data models address both statistical and territorial aspects, for example: the INSPIRE metadata models and, the Euro-SDMX Metadata Structure (ESMS) model.

The Eurostat organization has built the Euro-SDMX Metadata Structure (ESMS) model on a subset of the SDMX one [Gotzfried and Pellegrino, 2008]. The SDMX initiative fosters standards for the exchange of statistical information [6]. The underlying purpose is to improve the interoperability of systems that manage statistical data and metadata [SDMX, 2012]. SDMX defines a data and metadata structure adapted to the statistical datasets implementation. The ESMS model is available as a spreadsheet file [7] that lists the 21 SDMX key concepts chosen by Eurostat in order to describe their statistical data (see an extract of this model in Figure 2.1).

1. For instance, the REST SDMX API of Eurostat gives access to the Eurostat data (`https://ec.europa.eu/eurostat/web/sdmx-web-services/rest-sdmx-2.1`
2. `https://www.opengeospatial.org/standards/wfs`
3. `https://www.opengeospatial.org/standards/wms`
4. `https://www.opengeospatial.org/standards/cat`
5. The INSPIRE Directive requires producers of digital geographic data to apply rules for their harmonization, standardization and dissemination. More broadly, this directive aims at improving the access to spatial information across Europe by fostering cooperation between institutions, so that they share common tools and build together a network of SDI at European level. `https://inspire.ec.europa.eu/inspire-directive/2`
6. `http://sdmx.org/`
7. `http://ec.europa.eu/eurostat/data/metadata/metadata-structure`

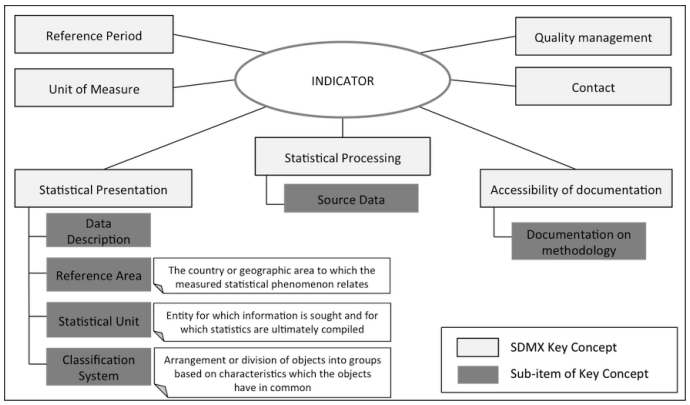


Figure 2.1 – Extract of the ESMS model (based on the Eurostat Metadata Structure, source: `http://ec.europa.eu/eurostat/data/metadata/metadata-structure`

The model supports key statistical metadata such as: *Unit of Measure*, *Source Data* (sub-item of the *Statistical processing* concept), *Documentation on methodology* (sub-item of the *Accessibility of documentation* concept) and data quality (concept entitled *Quality management*). Regarding the temporal dimension, the *Reference period* concept indicates the period of time or the instant to which the measured observation refers. The field *Reference Area*, sub-items of the *Statistical Presentation* concept, enables data producers to describe the studied geographic areas. However, no specific field in the metadata is dedicated to the description of the TSN these areas belong. Thus, this information is sometimes not available or sometimes mentioned in the fields *Classification System* or *Data Description*. Also, the modeling of TSN (their nested territorial levels and territorial units) is not addressed through the ESMS and SDMX model. Though ESMS and SDMX "*allow the publication of statistical information, they do not cover spatial representations of the statistical units*" [European Joint Research Centre, 2013b] ("*statistical units*" means "*Territorial Units* in our vocabulary). Then in SDMX, the spatial dimension of statistics is only succinctly described.

Conversely, the INSPIRE Directive focuses more on the spatial dimension of statistics through the themes *Statistical Units* (SUs). The SU theme deals with the geographic areas used in statistics [European Joint Research Centre, 2013b], which we previously called Territorial Units (TUs). Another INSPIRE theme, the *Population Distribution and Demography* (PDD) is dedicated to the description of statistical data associated with TUs [European Joint Research Centre, 2013a]. The PDD theme points out elements that enable the description of phenomena concerning the population living in some SUs [European Joint Research Centre, 2013a]. This theme has no direct spatial features. It needs to be linked to the SU theme to address the geographic nature of the statistical data [European Joint Research Centre, 2013b], as shown in Figure 2.2.

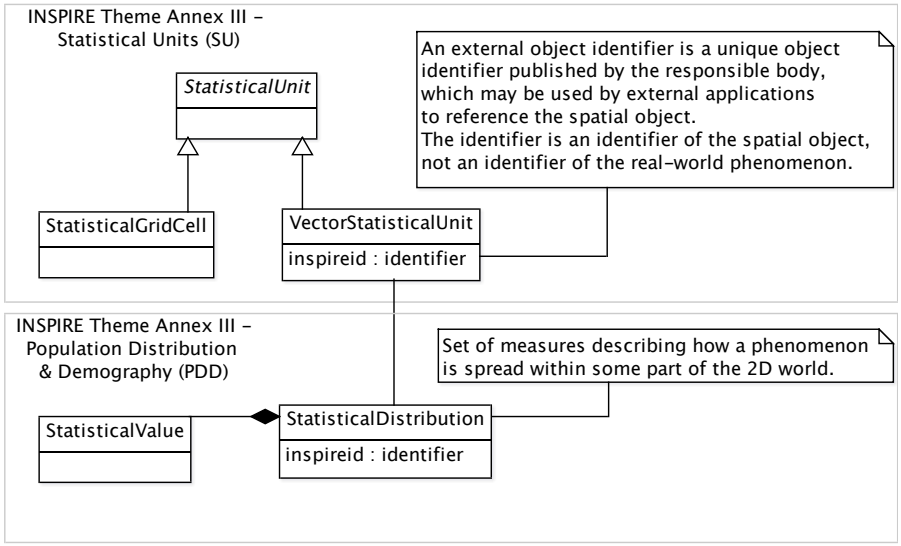


Figure 2.2 – INSPIRE data model extract (Based on: INSPIRE Consolidated UML Model `http://inspire.ec.europa.eu/data-model/approved/r4618-ir/html/`)

This Figure 2.2 shows the link between the SU and PDD themes established through a unique identifier called the *inspireid* of TUs: a set of measures (*StatisticalDistribution* element) is directly linked to a territorial unit (*VectorStatisticalUnit* element). Although INSPIRE tries to address the representation of statistical data with the PDD theme, not all the necessary metadata are represented. For instance, the PDD technical guidelines do not address the quality of statistical data, whereas these metadata are key elements for the user to estimate statistical data reliability. The ground rules are laid down and suggestions are made, such as the use of the SDMX model to reach the description of statistical data sets [European Joint Research Centre, 2013a].

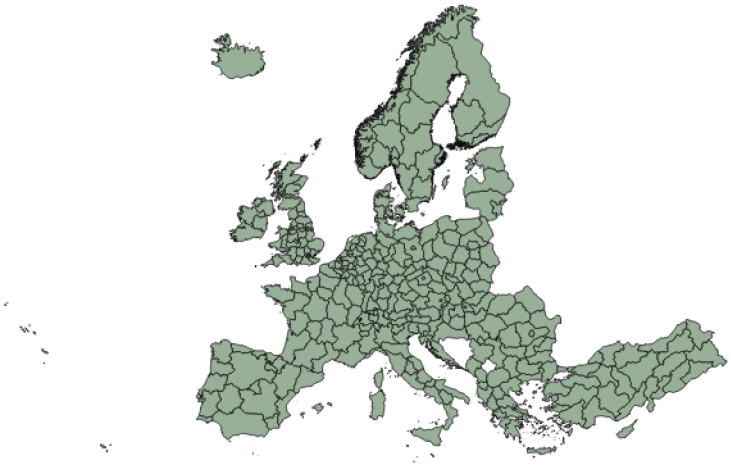
Then, most of the time, the only link that is made between these siloed statistical and geographic data sets relies on the code (such as the *inspireid* identifier of TUs) of the geographic area that the statistical observation refers to.

As a result, siloed data makes it more difficult for analysts to perform multi-criteria, multi-indicator, multi-source analyzes of a territory. In order to understand correlations or causalities on a studied territory, analysts have to download multiple isolated data:
(1) several statistical data sets (containing one or several indicator(s)), each of them resolving only a part of their problem (in tabular format most of the time);
(2) several metadata sets explaining the methodology used to computes the statistical indicators;
(3) several geospatial data sets, each of them containing the TUs of one TSN version (sometimes the file contains TUs at one territorial level only), while statistical data sets can refer to TUs that do not belong to the same TSN version (in shapefile ESRI

format most of the time);
(4) several metadata files describing the TSN versions, their TUs used for calculation of the statistical values and their changes over time (in tabular format most of the time).

	A	B	C	D	E	F	G
1	unit,age,sex,geo\time	2018	2017	2016	2015	2014	2013
2	PC,Y15-24,F,ES6	45.3	47.6	58.5	58.8	60.8	64.9
3	PC,Y15-24,F,ES61	46.5	49.3	59.9	58.9	61.0	66.8
4	PC,Y15-24,F,ES62	36.3	37.3	50.2	55.2	57.4	53.5
5	PC,Y15-24,F,ES63	65.3 u	: u	: u	81.4 u	79.5 u	69.7 u
6	PC,Y15-24,F,ES64	: u	: u	78.0 u	90.7 u	: u	59.5 u
7	PC,Y15-24,F,ES7	35.5	41.6	50.1	51.0	57.2	62.4
8	PC,Y15-24,F,ES70	35.5	41.6	50.1	51.0	57.2	62.4

Figure 2.3 – Eurostat Unemployment rates by sex, age and NUTS regions data set extract (source `https://ec.europa.eu/eurostat/web/products-datasets/-/LFST_R_LFU3RT`)



NUTS_RG_01M_2016_4326_LEVL_2 :: Features total: 332, filtered: 332, selected: 0

	LEVL_CODE	NUTS_ID	CNTR_CODE	NUTS_NAME
129	2	ES61	ES	Andalucía
130	2	ES62	ES	Región de Murcia
131	2	ES63	ES	Ciudad Autónoma de Ceuta
132	2	ES64	ES	Ciudad Autónoma de Melilla
102	2	ES70	ES	Canarias

Figure 2.4 – Eurostat NUTS 2016 Level 2 ESRI shapefile. Data was rendered using the QuantumGIS software (source `https://ec.europa.eu/eurostat/cache/GISCO/distribution/v2/nuts/download/ref-nuts-2016-01m.shp.zip`)

For instance, if one wants to analyze, compare and visualize on a map the unemployment rates in European regions (we restrict the analyze to a single indicator), one has to download on the Eurostat Website:
(1) the indicator *Unemployment rates by sex, age and NUTS 2 regions* `https://ec.europa.eu/eurostat/web/products-datasets/-/LFST_R_LFU3RT`. Figure 2.3 presents an excerpt of this indicator. The codes of the TUs that the statistical values describe are highlighted in yellow. These codes do not contain information on the reference period for the areas;
(2) metadata explaining the methodology used for the calculation of the indicator, available from the Web page `https://ec.europa.eu/eurostat/cache/metadata/en/reg_lmk_esms.htm`;
(3) metadata explaining which TSN and reference areas were considered for the calculation of the indicator values. These metadata are available at `https://ec.europa.eu/eurostat/web/nuts/background` (this link is referenced in the statistical metadata page (2)) . It is only in these metadata descriptions that the analyst will know which version of the NUTS was used for the calculation of the indicator values.
(4) the shapefile of the NUTS version used for the calculation of the indicator values, available from `https://ec.europa.eu/eurostat/cache/GISCO/distribution/v2/nuts/download/ref-nuts-2016-01m.shp.zip`. Figure 2.4 renders the shapefile of the NUTS version 2016, using the QuantumGIS software. The table under the map displays information on each of the TUs contains in the shapefile:
– column 1) the NUTS level the TUs belong to (NUTS 2016 level 2 in the present case),
– column 2) the TUs' code in the NUTS (*e.g.*, code ES63, highlighted in yellow),
– column 3) the code of the country the TUs belong to,
– column 4) the name of the TUs (*e.g.*, "Ciudad Autónoma de Ceuta").

As explained in the Chapter 1, the main drawback of using such codes for areas lies in their lack of accuracy in time and space, as they may designate a region that has changed over time in its boundaries or/and name. For instance, the code ES63, highlighted in yellow in both Figures 2.3 and 2.4, refers to a TU that was split in two areas, from the NUTS version 1999 to 2003, one of the areas keeping the identifier ES63 after the change (please see Figure 2.14). Then, the code ES63 in statistical data sets, without information on the reference period of the geographic area, may refer to several different geographic objects *i.e.*, **it no longer fulfills its role of unique identifier and it may lead to errors in analysis when it does not change while its boundaries change**. Even if sometimes the TUs are re-codified after a major change, the fact remains that in both cases (recoding or not), the statistical series referring to the codes are broken, as explained in Subsection 2.1.2.

2.1.2 Broken time-series

The boundaries of geographic areas defined by humans evolved over time, because of reforms, electoral concerns, alliances or conflicts between human groups.

This leads to broken statistical time-series since past data about a given statistical indicator (*e.g.*, unemployed number, income distribution, life expectancy) can no more be compared to more recent data if the geographic areas observed have changed. This problem, well known as the *Modifiable Areal Unit Problem* (MAUP) Openshaw and Taylor [1979]; Gotway Crawford and Young [2005]; Howenstine [1993] (defined below in Section 2.3), describes the phenomenon where data collected in different TSNs or versions of a TSN are not comparable due to potential differences between the geographic areas used as supports for the collected data.

In many cases, the single use of the latest States' borders leads to data inconsistency, if the data provider do not pay attention to boundaries changes. Let us take the example of an Open data set that contains the number of medals won per country, in the Olympic winter games over a period from 1924 to 2010 (available at `https://www.data.gouv.fr/fr/datasets/winter-olympics-medals/`). If one wants to represent this data set on a map divided by countries, the evolution of the borders of some countries and the disappearance of others (USSR, German Democratic Republic, Federal Republic of Germany, ...) is problematic if the base map represents only the last borders (Germany, Russia ...). Considering the case of Russia (created after the split of the USSR in several smaller countries), it is impossible to represent the evolution of the medals of Russia, from 1924 to 2010. Rather, one could represent the number of medals won by Russia since 1991 (the creation date of Russia) to 2010. Before 1991, it must be considered that Russia did not exist and not to award it medals that could just as well go back to other countries (such as Ukraine, Kazakhstan, etc.) from the former USSR, and therefore one should not count Russia as the former USSR. The time-series of this indicator, over the period from 1924 to 2010, is broken for geographic areas such as Russia. Beyond these examples, changes of geographic areas are, most of the time, more complex than a simple split of areas and very numerous the more we go down the territorial hierarchy (changes in the municipalities for instance).

2.1.3 Removal of territorial changes

In order to cope with the broken time-series problem, SAs mitigate traces of territorial changes, and provide users with (estimated) statistical data in the latest TSN version only. While, in general, such territorial changes are not meaningless because they are decided or/and voted by an authority pursuing some objectives. It is crucial to keep and enrich such information about territorial changes for several reasons:
1) first, it avoids errors in analysis when the areas keep their identifier while their boundaries have changed;
2) second, the nature of the change the areas undergone (*e.g.*, split, merge, redistribution of the areas) helps in estimating data in a new geographic division. Then, it helps in constructing long time-series and analyzing the evolution of a territory, using the latest boundaries;
3) conversely estimating data, such as electoral results, in former electoral areas helps in analyzing, for instance, electoral votes as if there was no redistricting and then observe the influence of the new areas on results;

4) and finally, knowing which area has changed, because of which event or law helps in understanding the reason for the change. For instance, the fusion of the French regions in 2016 was acted by a law that aims at reducing the number of administrative levels for cost saving purposes.

In this Section, we have identified three issues (not fully interconnected data, broken time-series, removal of territorial changes) related to the TSI. In the next Section, we describe in more details the geographic part of the TSI *i.e.*, the *Territorial Statistical Nomenclatures*.

2.2 Territorial Statistical Nomenclature Structures

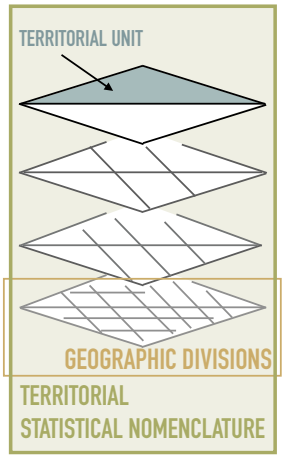


Figure 2.5 – Schema of a Territorial Statistical Nomenclature Structure.

Although *Geographic Divisions* vary in nature (*e.g.*, regular/irregular shapes, contiguous/non contiguous areas, etc.) or in the territories they cover (States, metropolises, etc.), they have in common to draw frontiers dividing the geographical space into *Territorial Units* (TUs).
A *Territorial Unit*, in this thesis manuscript, refers to a bounded geographic area (a portion of space on Earth *e.g.*, a Region in Europe) delimited by and under the control of a human group (in general a political or administrative authority) [Sack, 1983]. *Territorial Statistical Nomenclatures* (TSNs) are composed of one or several standard geographic divisions, built by (N)SAs to observe a territory at several geographic levels (*e.g.*, countries, regions, districts, municipalities, ...). They are sets of artifact (embedded) areas designed for statistical purpose, although they usually derived from a political or administrative structure. Together, the embedded geographic divisions of a TSN form a territorial hierarchy.

In the literature, the following terms may refer to geographic divisions: zoning, partition, mesh. In this manuscript, we will use these terms interchangeably.

In the literature, the following terms may refer to a TSN: Hierarchy of Census Geographic Entities [8] ; Hierarchy of standard geographic units/Hierarchy of geographic areas used for disseminating data [9] ; Statistical Geography Standard/Standard Geographical Classification [10] ; Nomenclature of territorial units for statistics [11]. In

8. Source: United States of America Census Bureau, the principal agency of the U.S. Federal Statistical Systemhttps://www.census.gov/geo/reference/hierarchy.html
9. Source: Statistics Canada, the national statistical office of Canada https://www12.statcan.gc.ca/census-recensement/2016/ref/98-304/chap12-eng.cfm
10. Source: Australian Bureau of Statistics https://www.abs.gov.au/websitedbs/d3310114.nsf/home/geography
11. Source: Eurostat, the European Statistical Office https://ec.europa.eu/eurostat/web/nuts/background

this manuscript, we will use the term TSN most of the time.

The terms *Hierarchy*, *Classification* or *Nomenclature* are sometimes used interchangeably while they do not mean the same thing [OECD, 2008]. In this manuscript, we make a difference between them. A hierarchy is "*a system in which people or things are arranged according to their importance*" [12]. Whereas the structure of a *Classification* can be either hierarchical or flat. For instance, the sex classifications are flat classifications, not hierarchical [OECD, 2008]. The OECD [2008] provides the following definition of classification, in the specific context of statistics: "*A set of discrete, exhaustive and mutually exclusive observations, which can be assigned to one or more variables to be measured in the collation and/or presentation of data.*". The definition of *Nomenclature* provided by OECD [2008] is the following one: a "*nomenclature is a systematic naming of things or a system of names or terms for things. In classification, nomenclature involves a systemic naming of categories or items.*". Nomenclature in this manuscript refers to a standard convention for naming things of a domain. The definition provided by Eurostat is closed to the definition we adopt. For Eurostat a nomenclature (or statistical classification) "*is an exhaustive and structured set of mutually exclusive and well-described categories, often presented in a hierarchy that is reflected by the numeric or alphabetical codes assigned to them, used to standardize concepts and compile statistical data*" [13].

As for TSN, in the literature there is a lot of terms that refer to the TUs of a TSN. We have observed the following denominations: areal units [14]; zonal objects [Openshaw, 1984]; census units; statistical units; geographic areas; census geographic entities; standard country or area for statistical use [15]. In this manuscript, we will use the term TU most of the time.

For the territory covered by a TSN, the term *Statistical Territory* is used and defined as follows "*The statistical territory of a country or area is "the territory with respect to which data are being collected"* [Nationen, 2004]. In this manuscript, we make a distinction between the study area of a TSN (*i.e.*, its geographic extent) and the **Territor-y/ies** it covers. The study area is implicit in TSNs: SAs do not provide the boundaries of this study area but they provide the boundaries of the territories, largest TUs of the TSN. Then, as for TUs, the territory concept describes an abstract representation of a portion of the geographic space that is claimed or occupied by a person, or a group of persons, or by an institution [16].

Hence, the study area is the geographic extent of the TSN (an implicit object) and a TSN may cover one or several *Territory/ies* inside this study area. Indeed, even if most of the time a TSN covers only one territory (*e.g.*, the European Union), sometimes it covers several ones in order to reflect on different political or economical

12. Source: Cambridge Dictionary https://dictionary.cambridge.org/dictionary/english/hierarchy
13. Source: Eurostat https://ec.europa.eu/eurostat/data/classifications
14. *e.g.*, http://www.restore.ac.uk/geo-refer/35236ceurs00y19880000.php
15. *e.g.*, https://unstats.un.org/unsd/publication/SeriesM/Series_M49_Rev3(1996)_en.pdf
16. Source: http://www.oxfordbibliographies.com/view/document/obo-9780199874002/obo-9780199874002-0076.xml

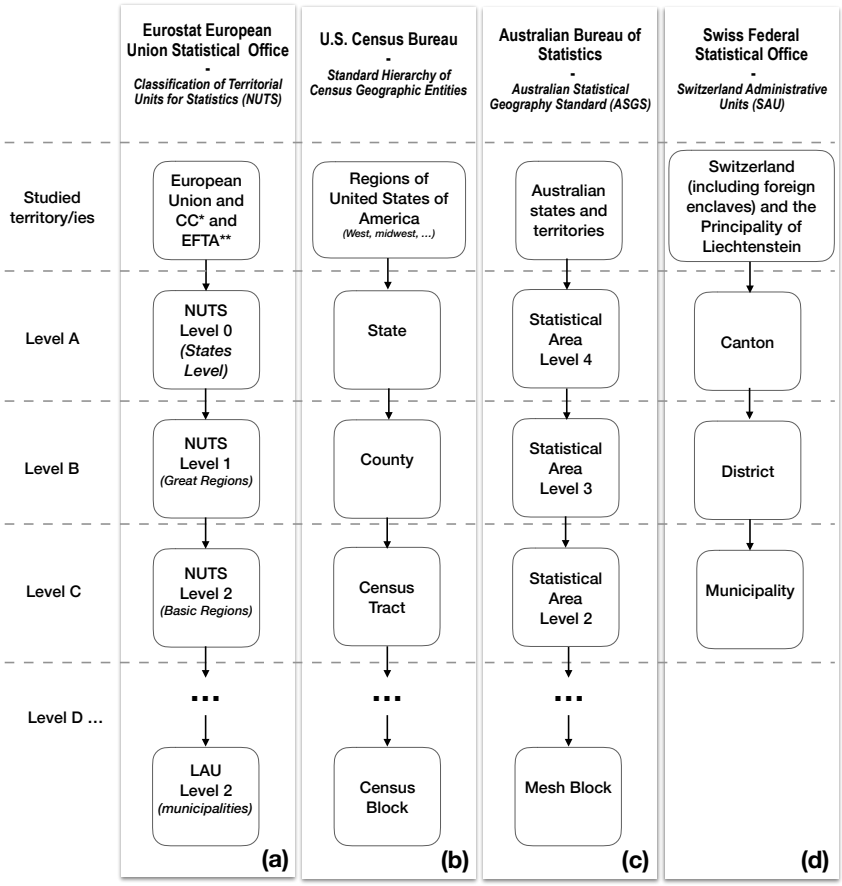
union of States, regions (*e.g.*, the European Union political union territory and the European Free Trade Association territory (a free trade area consisting of four European states: Iceland, Liechtenstein, Norway, and Switzerland)), or to observe the influence of a territory on another (*e.g.*, a TSN that observes two territories, the Grenoble Metropolis and the Gresivaudan *communauté de communes*, in order to simulate the influence one of this territory may have on the other). The geographic extent (or study area) of the TSN is then the union of the surface area of each territory.

In general, a TSN is created in order to harmonize regional statistics. For instance, on its website, Eurostat defines the purpose of the *Nomenclature of Territorial Units for Statistics* (NUTS) TSN as follows: "*The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the European Union for the purpose of collection, development and harmonization of European regional statistics.*" Indeed, all the members States of the European Union political organization, have their own NSA. Sometimes the statistics published by NSAs and the regions observed answer national interests. Those regional statistics, coming from several data providers, must be harmonized at European level to enable comparisons between countries and regions. Then, the European regional statistics, provided by Eurostat, enable policy-makers to compare their region or country to others in Europe, with regard to the political objectives defined by the European Union [Ysebaert et al., 2017].

On the basis of the definition of the TSN called the *Australian Statistical Geography Standard* (ASGS) from the Australian Bureau of Statistics [17], and on the basis of the definition of the term *Nomenclature* provided by Eurostat, we propose the following definition of a TSN:

> A TSN provides a catalog of TUs (also called statistical areas), often presented in a hierarchy that is reflected by the numeric or alphabetical codes assigned to them (see Figure 2.9). A TSN is used to compile statistical data on a or several studied statistical territory/ies, at one or several geographic division level(s). TSNs are used by (local or national) SAs or other organizations to compile and publish statistics that are comparable and spatially integrated.

17. http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS)



*CC: candidate countries for future membership of the European Union
**EFTA: European Free Trade Association, free trade area consisting of four European states: Iceland, Liechtenstein, Norway, and Switzerland

Figure 2.6 – Four TSNs Structures from four agencies producing, managing and delivering official statistics (Eurostat, the U.S. Census Bureau, the Australian Bureau of Statistics and the Swiss Federal Statistical Office).

Generally speaking, a TSN splits a territory into TUs, at one or several territorial level(s). Several nomenclatures may partition the same geographic space to answer various needs. For instance, a TSN may split a territory into administrative units or into medico-social or voting area units. Numerous TSNs exist throughout the world. The one from the European Union, called the *Nomenclature of Territorial Units for Statistics* (NUTS) [18], provides four nested divisions of the European Union (EU) territory for the collection of EU regional statistics and two other divisions for the collection of local statistics (Local Administrative Units (LAU), see Figure 2.6 (a)). Eurostat has created an extension to this NUTS nomenclature, and has also defined statistical regions for candidates countries awaiting accession to the European Union, and for countries of the European Free Trade Association (see Figure 2.6 (a), studied territories)*source: `https://ec.europa.eu/eurostat/we`*

18. `http://ec.europa.eu/eurostat/web/nuts/overview`

b/nuts/statistical-regions-outside-eu. In this manuscript, when we use the acronym "NUTS", we refer to this Eurostat extended NUTS. Sometimes, the statistical areas of TSNs are designed to be relatively homogeneous in terms of population characteristics, to avoid statistical bias. This is the case for the artifact areas called *Census Tracts*[19], built by the United States Census Bureau, on a basis of a population size between 1,200 and 8,000 people (see Figure 2.6 (b)). Another example is the *Australian Statistical Geography Standard* (ASGS)[20] (built by the Australian Bureau of Statistics) that provides seven nested divisions of the Australian territory (see Figure 2.6 (c)). For its part, the *Switzerland Administrative Units* (SAU)[21] is a Swiss Nomenclature updated every year and published by the Swiss Federal Statistical Office on the basis of the Swiss administrative boundaries. It allows provision of statistical data on the Swiss municipalities, districts and cantons.

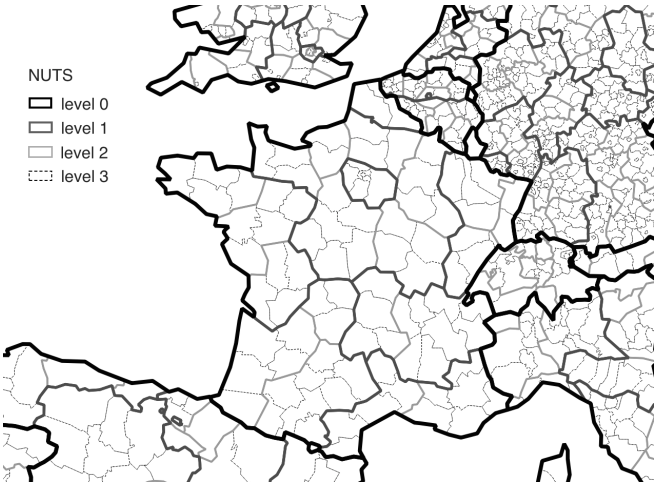


Figure 2.7 – The NUTS divisions of the French territory into four levels.

As explained before, the NUTS TSN divides the European territory into four hierarchical regional levels (forming nested geographical objects), as shown in Figure 2.7 for France:
Level 0 – State members;
Level 1 – major Regions;
Level 2 – basic Regions;
and Level 3 – small Regions.

19. https://www.census.gov/geo/reference/gtc/gtc_ct.html
20. http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS)
21. Please note that this name SAU is not the official TSN name. The official one is *Swiss official commune register* while under the term "commune" there are vector boundaries of municipalities, but also districts, cantons, large regions, etc. as explained by the Swiss Federal Statistical Office on its Web site https://www.bfs.admin.ch/bfs/fr/home/services/geostat/geodonnees-statistique-federale/limites-administratives/limites-communales-generalisees.html

The NUTS classification divides the European territory into contiguous TUs (*i.e.*, continuous mesh), irregularly shaped, covering the entire territory without overlapping of the TUs of a same level [Plumejeaud, 2011].

Other TSNs (see Figure 2.8) describe, through one territorial level, non-contiguous geographic objects (*i.e.*, discontinuous mesh) such as urban objects (*e.g.*, Urban Morphological Zones (UMZ), Morphological Urban Areas (MUA)). In this manuscript, we will address only TSNs with continuous TUs, and with multiple levels (non-flat TSNs).

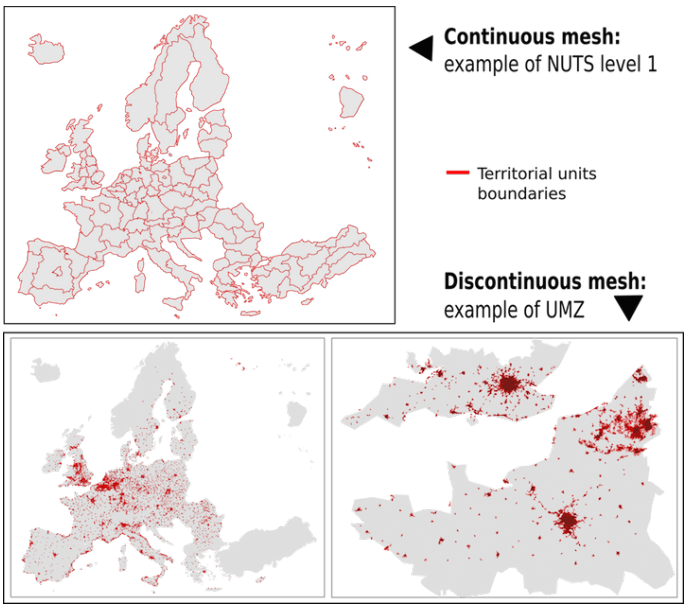


Figure 2.8 – Territorial nomenclatures heterogeneity: example of the NUTS and UMZ nomenclatures.

A more formal definition of the TSNs we address in this thesis is the following one: a TSN is a set of continuous TUs, with non-overlapping of the TUs of a same level, organized in several territorial division levels (at least one), whose spatial union of the TUs at each level forms the study area. Each level in this hierarchy consists of a given division of space (for instance, an administrative division). The levels are nested levels which means that they are spatially and hierarchically organized. For any couple L_i, L_{i+1} of two consecutive levels, the following relation and assertion stands: any TU of a level L_{i+1} is spatially included into one and only one TU of the level directly above L_i, as shown on Figure 2.9 (except for the TUs of the upper level). Hence, such a territorial nomenclature corresponds to a recursive spatial hierarchical division of the space. According to Plumejeaud et al. [2011], the lowest level of a TSN consists of the TUs which form the finest partition of the geographic space.

Then, tree data structure is well adapted to the representation of a TSN as a hierarchy of TUs: each TU being a node of the tree, the edges between the nodes

representing the spatial and hierarchical relations between the TUs directly above or below. Then, the TSNs we address are balanced tree, where the root node is the *study area* of the TSN, all the children being TUs (including *territories* that are the largest TUs, directly above the root node). Figure 2.9 presents, using the tree data structure, two versions of a fictive TSN. The code of the TUs are written inside the node, and reflect the hierarchical order between the TUs.

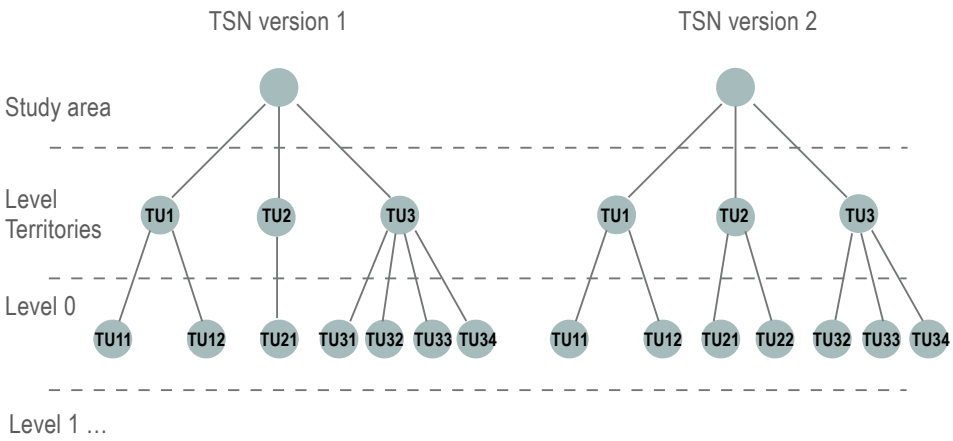


Figure 2.9 – Example of two versions of a TSN represented as tree data structures.

Furthermore, the vocabulary associated to *Online Analytical Processing* (OLAP) technologies suits also to the description of TSNs. In the OLAP technologies, data sets are multi-dimensional and the values of a dimension can sometimes be organized as a hierarchy (*e.g.*, a spatial hierarchy, a hierarchy of categories of products, etc.). There are different OLAP terms to refer to different kinds of hierarchy, for instance:

– the term *Covering* describes hierarchies where the parent TU of another one is at the territorial level immediately above. Conversely, in a *non-covering* hierarchy, some units may have superior unit(s) at a territorial level not immediately above [Plumejeaud et al., 2011]. Non-covering hierarchies occur when links between dimension values "skip" levels [Pedersen et al., 1999].

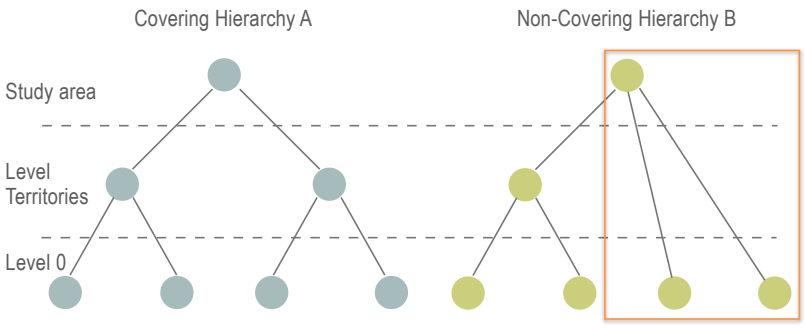


Figure 2.10 – Example of covering and non-covering hierarchies.

The Figure 2.10 shows an example of covering (on the right side of the figure) and non-covering (on the left side of the figure, see the rectangle in orange).

– the term *Strict* describes hierarchies where the units have exactly one superior unit, in contrast to *non-strict* hierarchy where certain units may have two superior units at the same upper level [Plumejeaud et al., 2011; Banerjee and Davis, 2009]. Non-strict hierarchies occur when one lower-level item has several parents [Pedersen et al., 1999]. For instance, Figure 2.11 shows an example of non-strict hierarchy, on the right side of the figure.

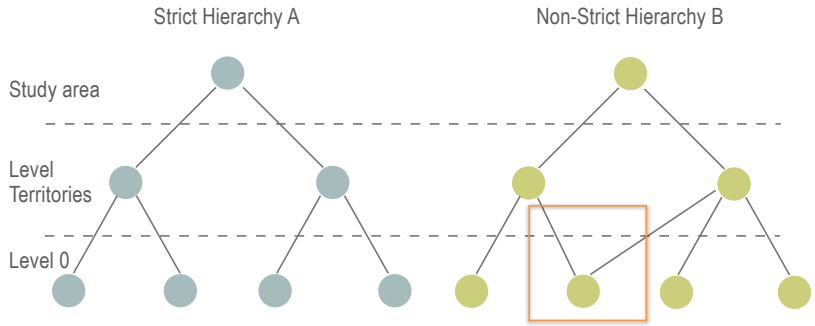


Figure 2.11 – Example of strict and non-strict hierarchies.

– the term *Onto* describes hierarchies where the path length from root to the leaves within the tree is balanced for all units. Non-onto hierarchies occur when the height of the hierarchy is varying [Pedersen et al., 1999].
The terms balanced, unbalanced are also used in the literature of tree data structures.

Figure 2.12 illustrates this hierarchy property.

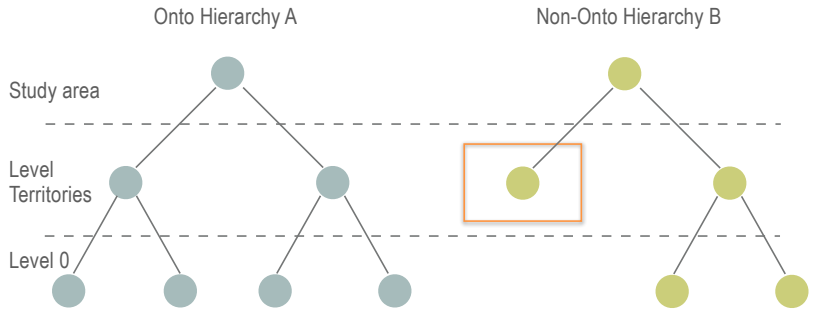


Figure 2.12 – Example of onto and non-onto hierarchies.

An example of a non-covering, non-strict, non-onto TSN is described by Plume-

jeaud et al. [2011]. It is the *Intercommunalities* TSN in France, illustrated in Figure 2.13. This TSN is composed of three territorial levels: Pays; EPCI ("*Etablissement Public de Cooopération Intercommunale*" in French – "Public Institution for Inter-Municipal Cooperation" in English) and Municipalities.

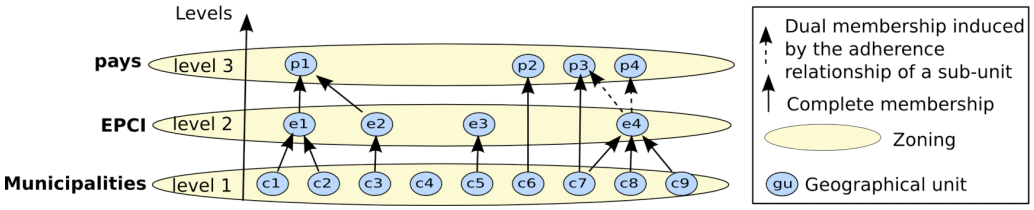


Figure 2.13 – Intercommunalities: a non-covering, non-strict, and non-onto nomenclature. Source: [Plumejeaud et al., 2011].

In this manuscript, we focus on TSNs with TUs that are irregular areal units. We do not address grid zoning. We focus also on TSNs with more than one territorial level, where each level *covers* the whole study area, and is composed of *non-overlapping* TUs, which corresponds to the structure of most of the administrative divisions in the world.
Regarding the TSN hierarchical structure, we focus on *covering*, *strict*, and *onto* hierarchies, using the OLAP vocabularies. In the tree data structure words, we study balanced trees, where the root is the study area (the geographic extent of the TSN) and the leafs constitute the finest subdivisions of the geographical space.

2.3 Territorial Statistical Nomenclature that change over time

The world is changing at an incredible speed: it jumps from one political, social, or technological order to another, at an ever-increasing rate [Plumejeaud et al., 2011]. Through times, the names, belonging, composition, and geometries of some regions all over the world, change. Think only how the map of Europe has changed during the last 100 years at the level of countries [Kauppinen and Hyvönen, 2007]. Maps showing the changes of a given territory offer an extremely useful measurement of these developments, and allow for an even better understanding of these dynamics [Plumejeaud et al., 2011].

Territorial changes are very frequent. As a result of territorial changes, TSNs also change over time: SAs usually update their geographic areas every 1, 3 or 10 years in order to reflect the real world evolution over time. Indeed, if a TSN wants to remain in adequacy with the territory it covers, it must also be revised, and account for any changes in the boundaries. This is, for instance, in order to meet the demands of elected representatives who want statistics on their territory within the current boundaries. As defined by the authors of [Plumejeaud et al., 2011],

by "change", we refer to the transformation of a TSN, which continues to exist as a whole, even though some of its internal parts may have had their boundaries modified, or their code or name changed. Changes lead to the creation of a new version of the TSN (see TSN version 1 and 2 on Figure 2.9). Very rarely, they lead to the creation of a new TSN because TSNs are standards, defined with the intention of being stable. The creation of a new TSN more takes place when new needs arise, in terms of observation of the territory, leading to the creation of new TUs, quite different in size or in their social or economic function (*e.g.*, a TSN to observe only the metropolitan areas).

In this thesis, we focus on changes that lead to the creation of a new TSN version (a revision). We do not processed the cases of major changes in the structure that lead to the creation of a new TSN.

All the SAs face this problem of evolution of the regions through time. As a matter of fact, on their websites, definition of their TSN is immediately followed by consideration on their evolutions over time. For instance, on the website of:

- the U.S. Census Bureau: "*Census tracts occasionally split due to population growth or merge when there is substantial population decline.*"[22]

- the Eurostat European Union Statistical Office: "*The regulation also specifies stability of the classification for at least three years. However, sometimes national interests require changing the regional breakdown of a country.*"[23]

- the Australian Bureau of Statistics: "*The regions that are defined in the ABS Structures are updated on a five yearly basis aligning with the Census of Population and Housing to provide a balance between stability and relevance to the changing underlying geography.*"[24]

SAs choose either to ignore this issue of evolution of TSN through time – then to publish only the latest version of their TSN – or to address it, by publishing several versions of the TSN, and in the best case, to describe also changes of the TUs, from one version to another.

22. https://www.census.gov/geo/reference/webatlas/tracts.html https://www.census.gov/geo/reference/boundary-changes.html
23. http://ec.europa.eu/eurostat/web/nuts/history
24. http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS)

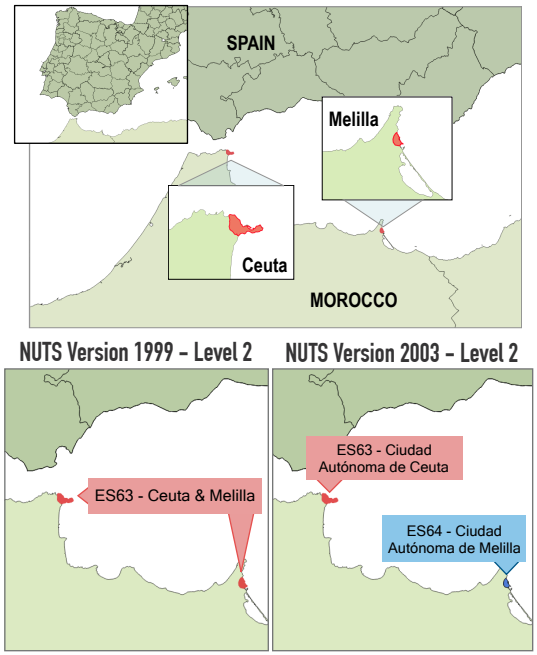


Figure 2.14 – Example of territorial change - Split of the territorial unit ES63 from NUTS Version 1999 to NUTS Version 2003, at Level 2.

Changes in boundaries are key information to understand the way territories evolve through time, and the history of these territories. For instance, within the NUTS Version 1999 at Level 2 (*i.e.*, basic regions in the Eurostat terminology), the two Spanish enclaves in North Africa, *Ceuta and Melilla*, were considered as one TU (with identifier ES63), part of a Spanish province. These enclaves are spatially distant though considered as one TU in NUTS Version 1999. The multi-polygon geometries of this TU are presented in red, on the left map, in Figure 2.14. In 1995, both Ceuta and Melilla became Autonomous cities of Spain, requiring a redistricting of the Spanish regions. It resulted in a split of the TU ES63 in two TUs, named *Ciudad Autónoma de Ceuta* (with the same identifier ES63) and *Ciudad Autónoma de Melilla* (with identifier ES64), in the NUTS Version 2003. The right side of Figure 2.14 shows the two created TUs ES63 and ES64, in the version 2003. Also, the sub-TUs of ES63 (ES631 and ES632) in the NUTS version 1999 both get new identifiers. Then, at the time of analysis, in order to fully understand this area and its specific features, statisticians must be provided with descriptions of its changes over time, and ideally with useful and meaningful information explaining the societal context of this event. Moreover, statisticians should be warned whether the statistical data sets, engaged to conduct analyses, refer interchangeably to the two versions of the TU ES63 (before or after the change event). This shall avoid misinterpretation of the values (and of their variations), here due to a decrease of the TU's area, not to a decrease in the number of inhabitants, for instance.

TSNs change over time, for political, or administrative reasons (*e.g.*, the names or geometries of districts may change) while most of the time boundaries of TU

are delineated with the intention of being maintained over a long time to ensure comparability of statistical observations. On the one hand, with regard to the problematic of evolution through time, if a TSN wants to remain in line with the territory it covers, then it has to change at the same time as the territorial partition evolves. On the other hand, statisticians need stability in the partition to make sure that data are comparable over the years (as they refer to the same TU). Stability allows them to construct long time-series (*i.e.*, series composed of several statistical observations of an indicator, indexed in time order, to observe the variations of the indicator values measured on a TU that does not change over the series) that are of huge importance for stakeholders to be aware of the impact of past policies, to understand territory at present time, and then, to better build the future. Thus, the alternative chosen in general by statisticians is to create a "snapshot" of the territorial divisions at one time and to amend this version at the end of a specified period of stability (*e.g.*, three years at least for the NUTS, 10 years for the *Census Tract*) to reflect changes in the regional breakdown of a territory.

By doing this, statisticians freeze territory boundaries for a period of time and make effective the changes undergone by the territory, in a new TSN version. Such a versioning approach solves data collection problems but, considering a long time period, leads to broken statistical series because data are collected for and stored in different versions of the same TSN. Then, territorial changes lead to new TSN versions causing ruptures in statistical time-series. Here, time-series refers to sets of observations of a statistical variable, measured at regular time intervals (*e.g.*, year, month, day) [25].

Let us assume that the version $V1$ of the NUTS N (*i.e.*, N_{V1}) is used, in 2014, to measure the indicator "Employment rate in Europe" and the version $V2$ of the NUTS N (*i.e.*, N_{V2}) is used to measure the indicator, in 2015. In this case, the two sets of measurement do not belong to the same time-series, since the territory may have changed between N_{V1} and N_{V2}. The time-series of the indicator is broken, in 2014, unless all data measured before 2015 are migrated from N_{V1} to N_{V2}, or vice-versa. Such a process implies that statisticians check, value-by-value, whether the described TUs are the same from one TSN version to another and if not, they compute estimates of values, considering the new TU boundaries. For instance, the estimation of values may be computed using the ratio established between the former and the new surface area of the TU, provided that there is a correlation between the surface area and the value. Thus, geographic information (among other data, such as the number of inhabitants) is crucial in territorial statistics [Gallego, 2010] as it enables statisticians to reconstruct broken time-series of indicators describing a territory.

The estimation of data in a new geographic support raises issues well known in the literature under the term *Change of Support Problem* (COSP) [Openshaw and Taylor, 1979; Howenstine, 1993; Gelfand, 2001; Gotway Crawford and Young, 2005]. In spatial statistics, this term relates to several types of change of data support, such as:

25. `https://www.insee.fr/en/metadonnees/definition/c2068`

– the change of data support from points to surfaces *i.e.*, aggregation/upscaling;
– and conversely, the change of data support, from surfaces to points *i.e.*, disaggregation/downscaling).
In [Gelfand, 2001] the *Change of Support Problem* is introduced as follows:"*In practice, spatial data are sometimes collected at points (i.e., point-referenced data) and at other times are associated with areal units (i.e., block data). The change of support problem is concerned with inference about the values of a variable at points or blocks different from those at which it has been observed.*" The data we are interested in

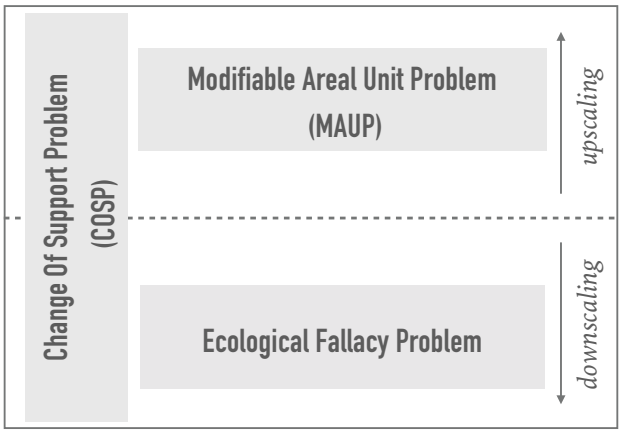


Figure 2.15 – Different Problems related to spatial sampling (Figure inspired from [Louvet et al., 2015])

are collected exclusively at blocks and inference is sought exclusively at new blocks versions [Gelfand, 2001]: this specific issue is known under the *Modifiable Areal Unit Problem* (MAUP). The COSP covers all the problems due to spatial aggregation or disaggregation whereas the MAUP is only an upscaling problem, since spatial information is aggregated by changing the geographic divisions or changing the scale of the zoning [Louvet et al., 2015]. Figure 2.15 illustrated the links between the three spatial data analysis problems. The ecological fallacy occurs when, at the opposite, it is inferred that results based on aggregate zonal (or grouped) data can be applied to the individuals who form the zones or groups being studied [Openshaw, 1984]. The MAUP describes the fact that spatially aggregated data are sources of statistical bias. It describes the influence of the zoning on data. Indeed, the aggregated point-based data vary according to how one draws the boundaries. The MAUP is composed of two separate but closely related problems, illustrated in Figure 2.16:
– the scale effect "*which is the variation in results that can often be obtained when data for one set of areal units are progressively aggregated into fewer and larger units for analysis. For example, when census enumeration districts are aggregated into wards, Districts, and Counties the results change with increasing scale.*" [Openshaw, 1984]
– the zoning effect (or aggregation problem) defined as "*the problem of alternative combinations of areal units at equal or similar scales. Any variation in results due to the use of alternative units of analysis when the number of units is held constant is termed the aggregation problem.*" [Openshaw, 1984]

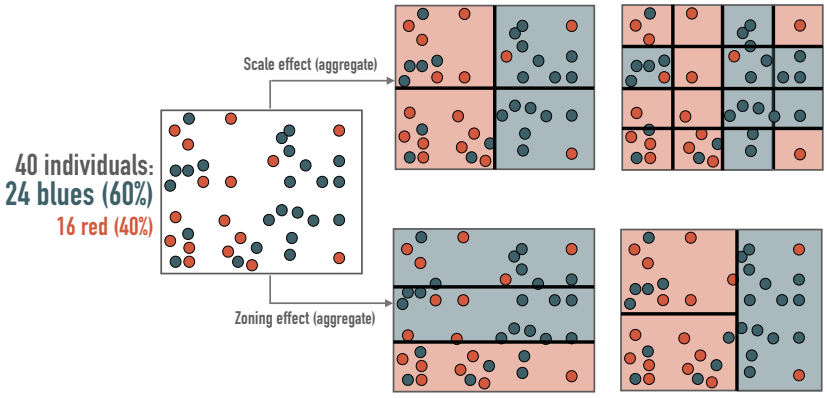


Figure 2.16 – Modifiable areal Unit Problem (MAUP) – scale effect and zoning effect (Figure inspired from [Loidl et al., 2016])

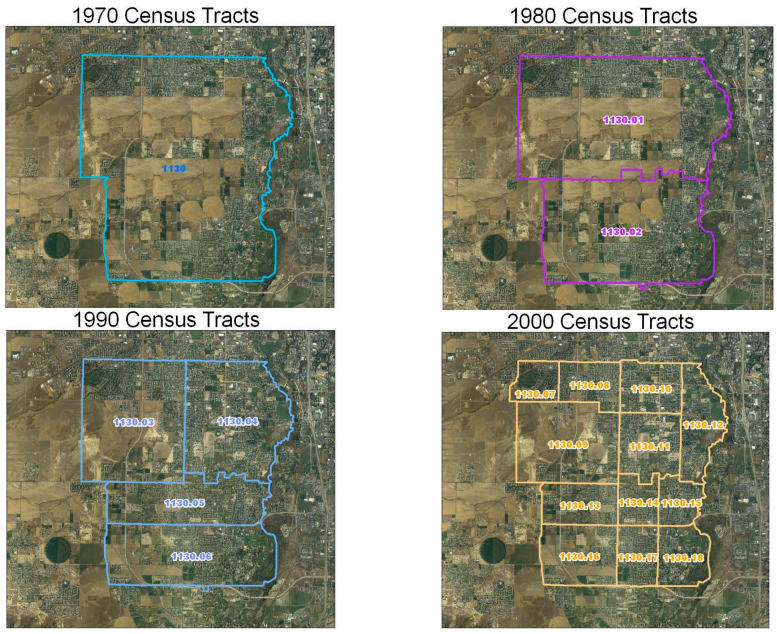


Figure 2.17 – State of Utah Census Tracts – Source [Census Tracts Geographic Products Branch, 2013]

Then, data collected using different zoning are not comparable due to potential differences between the geographic objects observed. Within the field of census data, this problem is known under the *Split tracts Problem* from Howenstine [1993] . The Figure 2.17 illustrates the changes over time of one U.S. census TU, of the U.S. Census Bureau. Census TUs are updated every 10 years, prior to each decennial census. When a tract is split, its TSN code (a number, in this TSN) changes. A two-digit extension is attached to the four-digit tract identifier [Howenstine, 1993]. The

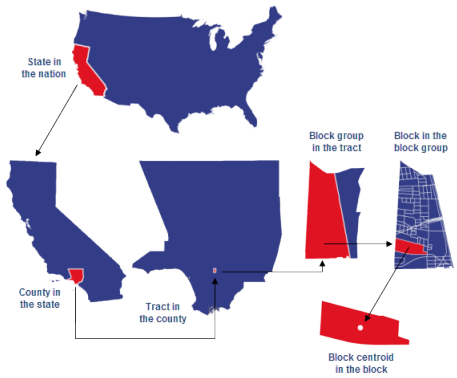


Figure 2.18 – United States census geographic units [ESRI, 2017]

U.S. census geographic entities are organized in a hierarchy of levels that starts with the level of States, then the level of Counties, Tracts, Block groups, and finally the level of Blocks. This hierarchy is illustrated in Figure 2.18. The tract TUs, called *Census tracts*, are rapidly growing areas with respect to the number of inhabitants, and are often split into two or more tracts between decennial censuses because of that [Howenstine, 1993].

The split tract problem focuses on territorial changes of type "split". To address this problem in a more comprehensive way (and cover other types of territorial changes), we adopt the term *MAUP in time* instead.

The problem we face is called **MAUP zoning effect over time** (illustrated in Figure 2.19). We define it as follows: When point data are aggregated into surface data according to the boundaries of areal units of a zoning A (version V') at time t, and a time $t+1$ alternative areal units are used to aggregate the data, after the change of the initial zoning A (resulting in a new version V'' of the zoning), the problem is that aggregated data observed with the zoning V' and V'' are not comparable, since even if point data are the same, the result of the aggregation could produce different surface values. This problem is referred as the *modifiable areal unit problem, zoning effect over time* or *evolving zoning problem.*

Here, a *zoning* refers to one territorial level of a TSN. Therefore, by extension to a hierarchy of zoning (*i.e.*, a TSN), the MAUP zoning effect over time describes the phenomenon where data collected **in different TSN versions** (*e.g.*, see Figure 2.19, zoning A version *V'* and zoning A version *V"*) are not comparable due to potential differences between the geographic objects observed, in analogy to the MAUP zoning effect at time t, that describes the phenomenon where data collected **in different TSNs** (*e.g.*, see Figure 2.19, zoning A and zoning B) are not comparable due to potential differences between the geographic objects observed.

The MAUP zoning effect over time is not the same as the *Modifiable Temporal Unit Problem* (MTUP) defined in analogy to the MAUP but for temporal units instead of spatial units. The MTUP focus on:

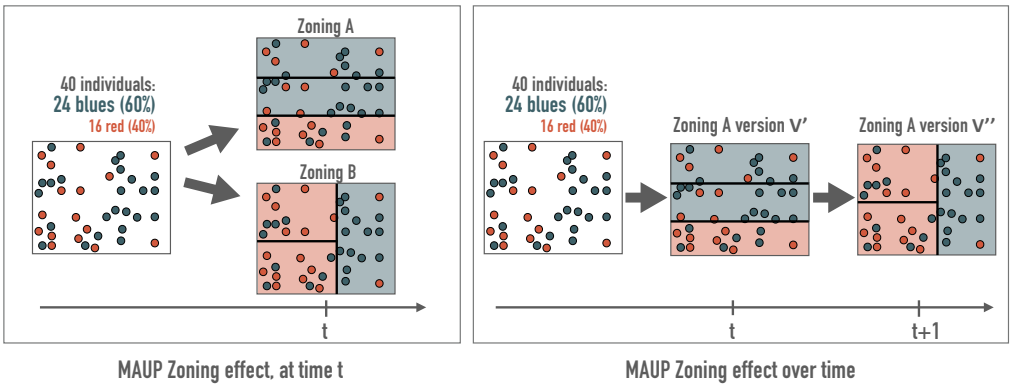


Figure 2.19 – Modifiable areal Unit Problem (MAUP) over time – zoning effect at time t and zoning effect over time (evolving zoning problem)

– the temporal aggregation of data, in analogy to the scale effect of the MAUP;
– and the temporal segmentation of data (*i.e.*, how the temporal dimension can be discretized into temporal units), in analogy to the zoning effect in the MAUP [Çöltekin et al., 2011; Cheng and Adepeju, 2014].

To cope with this problem, methods such as aggregation, disaggregation (*e.g.*, Simple Areal Weighted), and areal interpolation (*e.g.*, Inverse Distance Weighting, Kriging method), can be used to transfer data into another TSN or TSN version or to abolish the boundaries between TUs by using spatial smoothing [Flowerdew, 1991; Wang, 2014; Anderson et al., 2012; Plumejeaud et al., 2010]. However, these methods mitigate traces of changes, and thus do not help to understand the various logic of governance and planning behind these changes in partitions [Plumejeaud et al., 2011]. In general, such changes are not meaningless because they are decided or/and voted by an authority pursuing some objective (for instance, in case of an electoral zoning, the redistricting of a territory under some political purposes, is often used and known as *gerrymandering*).

Changes in boundaries are key information to understand the way territories evolve. Then, it is crucial not to lose such information, but, on the contrary, to enrich them with metadata and other resources on the Web that may explain these territorial changes (*e.g.*, societal reasons, historical events).

The following chapters of the State of the art focus on existing spatiotemporal approaches for the description of territorial changes and on the models (especially ontological models) that allow mirroring the complexity of TSN hierarchical structures that evolve over time. We present the existing tools in computer science that manage these TSNs and their changes over time, from their acquisition until their restitution as LOD on the Web. Indeed, more and more institutions need nowadays such tools for the online sharing of their TSNs (and their changes), and of their statistical data relying on these evolving TSNs.

Preliminary Remarks – Data management process in the Semantic Web

In order to ease the development and delivery of open data as Linked Open Data, the W3C sets out a series of best practices in an online document [Hyland et al., 2014]. This document includes recommendations such as, how to review existing ontologies and evaluate their usefulness with regard to the processed data.

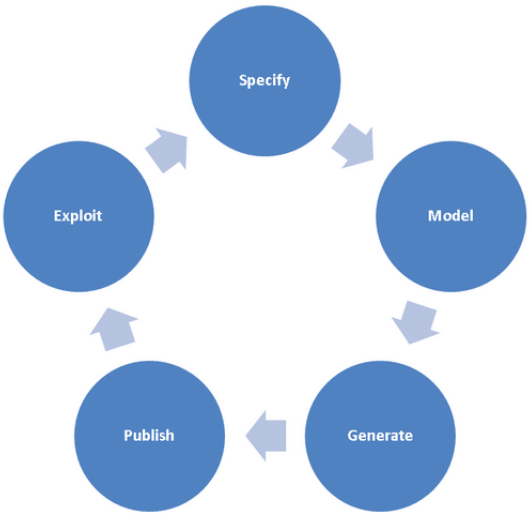


Figure 2.20 – Linked Government Data information management process, proposed in [Villazón-Terrazas et al., 2011].

This document also includes three LOD information management processes to help agencies in the delivery of data as Linked Open Data. The three management processes all share main steps that are: specify, model and publish data in standard open Web formats. The process established by [Villazón-Terrazas et al., 2011] addresses specifically Linked Government Data. It consists of the following activities: Specify, Model, Generate, Publish, and Exploit (illustrated in Figure 2.20). As TSN are standard hierarchical structures, most of the time establish by NSA, they fall under the methodological guidelines of [Villazón-Terrazas et al., 2011] for the publication and use of high quality Linked Government Data. The following state of the art chapters 3, 4, and 5 present the existing methods, models and/or tools, at each step of the [Villazón-Terrazas et al., 2011] life cycle, for the publication of TSN data on the LOD Web. The existing is approached from the perspective of our ultimate objective: providing Governmental Institutions with the necessary tools to automatically or semi-automatically publish on the LOD, descriptions of evolving areas under their authority.

Chapter 3

Specifying and Modeling spatiotemporal entities

In this Chapter, we focus on the two first steps of the [Villazón-Terrazas et al., 2011] life cycle: Specifying and Modeling. We first present the main spatiotemporal concepts involved in the modeling of evolving spatial entities. Second, we present the fundamentals for the modeling of these evolving spatial entities, and the existing approaches in the Semantic Web.

3.1 Specifying

The first step of the [Villazón-Terrazas et al., 2011] life cycle consist in *Specifying* the data *i.e.*, which data one wishes to publish on the LOD Web (see Subsection 3.1.1), which modeling problem one faced (see Subsection 3.1.2), from which discipline we can learn from (see Subsection 3.1.3).

3.1.1 Spatiotemporal entities

As explain in the main introduction, Chapter 1, the descriptions of territorial changes over time are critical for the interpretation of statistical data produced over several versions of a TSN. Description of the way territories evolve should help in understanding the events behind changes and the motivations for partitioning a territory. In [Beller, 1991; Claramunt and Thériault, 1995], it is stated that reproducing the dynamics of spatiotemporal processes consists in describing changes and relationships between entities, and in considering the events behind changes and the facts which enable observation of these changes. On these fundamentals, [Del Mondo et al., 2010] notice that modeling these dynamics of the world requires modeling the entities themselves but must of all, **the *spatial*, *spatiotemporal* and *filiation relations* between them (*e.g.*, topological relation, topological relation considered in time, ancestor/descendant relation).** The entities we address particularly are TUs spatiotemporal entities (composing a TSN), and the relations between them, over time and space. [Del Mondo et al., 2013] provide the following definition for *Spatiotemporal entities*:

> **Spatiotemporal entity** *A spatiotemporal entity (also denoted entity) represents an abstraction of the real world. An entity has a fixed identity and a type that describes its semantics; for example, types of*

> *entities are counties, cities, lakes, and so on. An entity can have also time-dependent thematic and spatial properties. For example, the geometry of an entity is a property that can vary in different time instants. Relations connect entities at the same or different time instants.*

In this citation, the expressions *fixed identity* and *a property that can vary* bring to attention the fact that even if an object change, its identity may endure over time. Although, in the context of spatial entities, one may wonder how far geometry, name, population or land-cover can change before a (spatial) entity loses its identity?

3.1.2 Identity concept

The term *identity* requires further definition, since in the context of evolving object, one may wonder, as noticed in [Harbelot et al., 2013], *"How far can an entity vary before losing its identity?"*.

In philosophy, this issue is often illustrated by *The Ship of Theseus* Greek legend: The Ship of Theseus was rebuilt entirely over centuries, since each broken plank was replaced one by one. Then, a philosophical question arises: Is the ship still the Ship of Theseus after all planks being replaced?

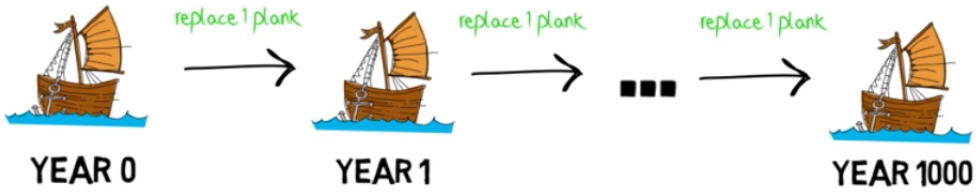


Figure 3.1 – Ship of Theseus illustration from [Vazirani, 2014].

In Fearon [1999], a philosophical sense of identity is provided: "*the identity of a thing (not just a person) consists of those properties or qualities in virtue of which it is that thing. That is, if you changed these properties or qualities, it would cease to be that thing and be something different.*"

This issue of entities that change and remain the same entities or not is a problem that the database community in computer science faced many years ago. Then, we should get some inspiration from results of this community. In the Subsection below, we investigate some fields of computer science to build upon since they deal with change of resources over time.

3.1.3 Versioning in computer science

As notice in [Noy and Klein, 2004], evolution or versioning of schema are a problem that the database community faced many years ago, and it was the subject

of schema-evolution research mainly. Schema-evolution research tries to preserve the integrity of data through the schema versions and to answer the following questions, for each processed database: "*How does the new schema affect the view of the old data? Will queries based on the old schema work with the new data? Can old data be viewed using the new schema?*" [Noy and Klein, 2004]. Database schema-evolution is "*a centralized process: Developers of the original schema usually make the changes and maintain the schema.*" While ontology development is a more decentralized process than database schema development since a change in one ontology may affect the other ontologies that use it. In [Noy and Klein, 2004], it is stated further that in the context of ontologies, one has to distinguish between "changes in the domain or changes in the real world" and "changes in conceptualization". Changes in subdivisions of a territory fall under the first type, as noticed by [Kauppinen and Hyvönen, 2007], with respect to geospatial changes over time. Then, clearly this kind of change should not be confused with ontology versioning, database schema-evolution or ontology evolution that deal with changes in the conceptualization.

Attention should be paid to the way semantic Web technologies deal with change over time. The RDF Specification [Cyganiak et al., 2014] states that the RDF data model is atemporal since RDF graphs are static snapshots of information. However, RDF graphs can express information about events and temporal aspects, given appropriate vocabulary terms (*i.e.*, ontological concepts). Whenever possible, it is recommended to use existing vocabularies (such as OWL-Time [1]) to achieve semantic interoperability. Another concern addressed by the W3C [Cyganiak et al., 2014] is the possibility to express different states of a resource over time through different RDF graphs enclosed in one resource. However, [Stefanidis et al., 2014] warns about possible duplications of data from one version to another. Then, this approach can rapidly increase the space memory requirements. An alternative solution is to store only one version (*e.g.*, the first one) and *deltas* (the differences between two consecutive versions).

Another field we have to explore is the field of *Software configuration management (SCM)* for managing the evolution of large software systems [Tichy, 1988]. Within this field, a *version* is defined as a state of an evolving item and different types of version are identified [Conradi and Westfechtel, 1998] such as: *revisions* that are versions intended to supersede their predecessor; *variants* that are versions intended to coexist. Versions of TSN are of this second type, they are *variants* that must coexist in order to observe the influence of the territorial divisions on statistical data. Another core concept of SCM is the *version model*. A *version model* defines the items to be versioned and the *delta*, which is the differences between two consecutive versions [Conradi and Westfechtel, 1998].

1. `https://www.w3.org/TR/owl-time/`

In this Section, we have identified several points a model that reproduces the dynamics of territories should address: (1) the modeling of spatiotemporal entities, (2) a definition of what makes the identity of these entities over time, (3) the representation of their relations over time and space and in filiation, (4) the description of their differences (the *delta*) and changes over time, (5) the consideration of events behind changes.

Thus, in the Section 3.2 below, we investigate how existing models address the question of the *identity* of spatiotemporal entities that change over time.

3.2 Modeling

The second step of the [Villazón-Terrazas et al., 2011] life cycle consists in *Modeling* data, and in our case, modeling the dynamics of territories over time. In this Section, we focus on Semantic Web models and standards to achieve the description of such dynamics. We first present the standard ontologies to model spatial and temporal entities, and spatiotemporal relations between entities. Then, we present ontologies which model evolving entities, filiation relations between entities, and sometime the changes of these entities over time.

3.2.1 Standard space and time ontologies

Nowadays, spatiotemporal modeling of objects on the LOD Web, necessarily implies two standards and fundamentals ontologies that are the *GeoSPARQL* and *OWL-Time* ontologies.

GeoSPARQL is an OGC standard. It supports two main actions: representing and querying geospatial data on the Semantic Web. It defines an ontology for representing spatial data in RDF, and an extension to the SPARQL query language for querying spatial data [2]. Two different namespaces exist, available from the URIs:
– `http://www.opengis.net/ont/geosparql#` to access the GeoSPARQL ontology concepts (the short form of this URI is the prefix `geo`);
– `http://www.opengis.net/def/function/geosparql/` to access the GeoSPARQL functions and query geospatial LOD (the short form of this URI is the prefix `geof`).

The GeoSPARQL ontology is made of three main classes (please see Figure 3.2) [Perry and Herring, 2012]:
– The `geo:SpatialObject` defined as the super class of every feature or geometry that can have a spatial representation;
– The `geo:Feature` defined as the super class of every feature;
– The `geo:Geometry` defined as the super class of every geometry.

The definitions of the concepts "Feature" and "Geometry" provided by OGC do not say much about these two concepts and the existing links between them. The

2. `http://www.opengeospatial.org/standards/geosparql`

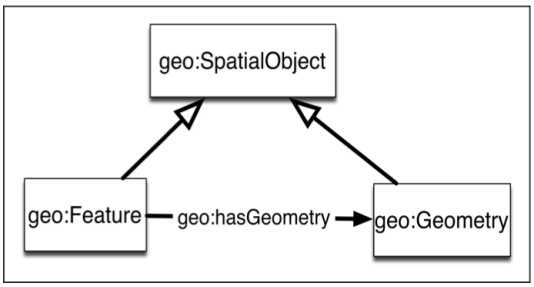


Figure 3.2 – GeoSPARQL main classes [Perry and Herring, 2012].

ArcGIS documentation provides the following definition of these two concepts[3]: "*Features consist of a geometry (point, line or polygon) and additional attributes (like a name) and a symbol that represents how the feature is rendered on the map.*"

The OGC GeoSPARQL ontology defines topological relations between the `geo:SpatialObject`. These relations are, for instance:
– `geo:sfTouches` defined as follows "Exists if the subject SpatialObject spatially touches the object SpatialObject";
– `geo:sfContains` defined as follows "Exists if the subject SpatialObject spatially contains the object SpatialObject".
Figure 3.3 presents these topological relations.

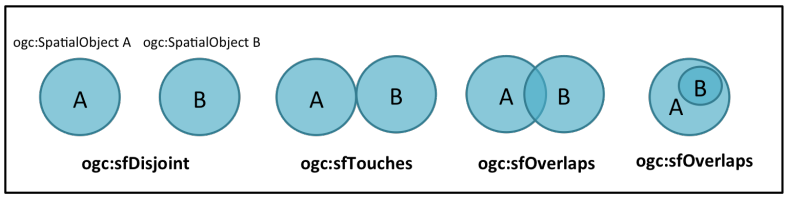


Figure 3.3 – GeoSPARQL main topological relations between geo:SpatialObject A and B [Perry and Herring, 2012].

Therefore, using the OGC GeoSPARQL concepts one may describe the topological relations between entities: one prerequisite to the modeling of the dynamics of the territories. Using the GeoSPARQL vocabulary, one may describe the geometries of TUs, and the topological relations between the TUs spatial objects composing a TSN. There exists also functions to query the spatial objects with regards to their topological relations (*e.g.*, the function `geof:sfContains` finds the spatial objects contained in another one).

The *Time Ontology in OWL* (also called *OWL-Time*) is a W3C Recommendation [Cox and Little, 2017]. The namespace of this ontology is `http://www.w3.org/2006/time#` and its prefix is usually `time`. It is introduced by the W3C as an "*OWL-2 DL ontology of temporal concepts, for describing the temporal properties of resources in the world or described in Web pages.*" The OWL-Time ontology is

3. `https://developers.arcgis.com/documentation/core-concepts/features-and-geometries/`

made of three main classes (please see Figure 3.4):
– `TemporalEntity` defined as "a temporal interval or instant with properties `time:hasBeginning` and `time:hasEnd` that link to the temporal instants that define its limits, and `time:hasTemporalDuration` to describe its extent.";
– `Instant` (subclass of `TemporalEntity`) defined as "a temporal entity with zero extent or duration";
– `Interval` (subclass of `TemporalEntity`) defined as "a temporal entity with an extent or duration".

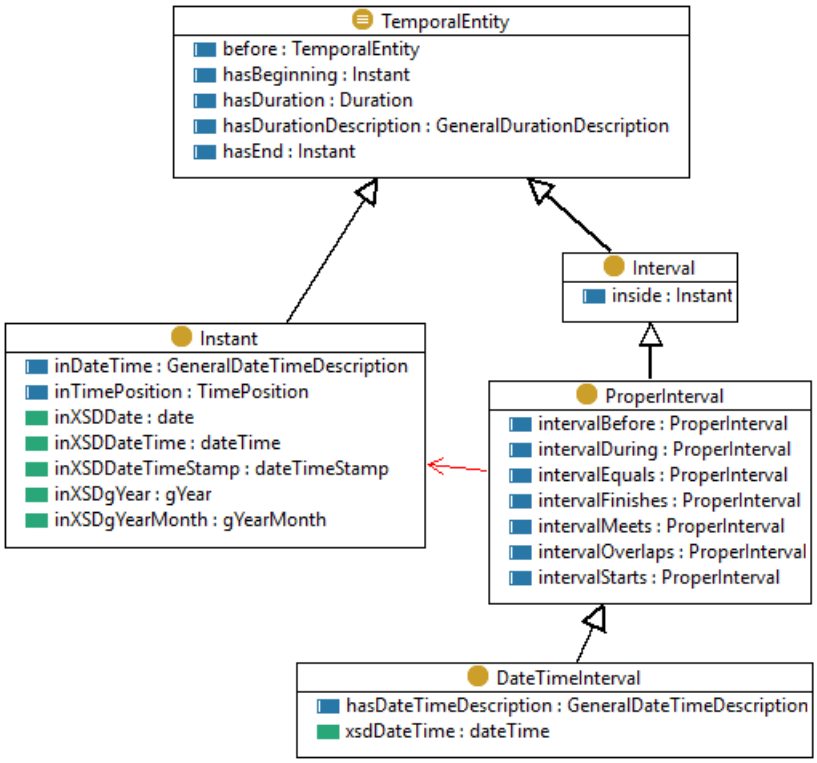


Figure 3.4 – The GeoSPARQL main classes [Cox and Little, 2017].

Also, the ontology "*provides a vocabulary for expressing facts about topological (ordering) relations among instants and intervals, together with information about durations, and about temporal position including date-time information.*" It supports the set of interval relations defined by Allen [1983]. Figure 3.5 shows the thirteen elementary relations between intervals implemented in the Time Ontology.

Therefore, using the W3C OWL-Time ontology one may represent the temporal relation between entities (another prerequisite to the modeling of the dynamics of territories) that belong to two different versions of a TSN. **Using both the OWL-Time and GeoSPARQL Ontologies, one may describe the geometries of TUs at a specific time interval and the topological relations as well as**

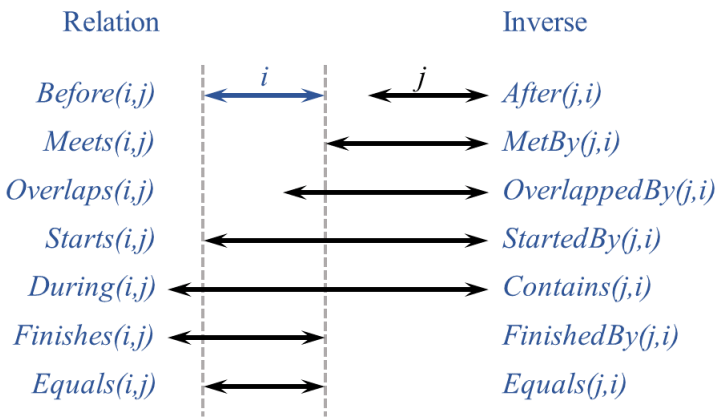


Figure 3.5 – Thirteen elementary possible relations between time periods [Cox and Little, 2017; Allen and Ferguson, 1997].

the topological relation considered in time between TUs spatial objects, belonging to two different versions of a TSN: the modeling of topological relation considered in time is one prerequisite to the modeling of territorial evolution.

3.2.2 Fundamentals for the modeling of spatiotemporal entities

We have seen, in the previous Subsection, that using the OWL-Time and GeoSPARQL standards, one achieves the modeling the spatial and temporal entities as well as the spatial and spatiotemporal relations between these entities. We focus, in this Subsection, on the modeling of the filiation relations between entities (*e.g.*, ancestor/descendant relation), and on the question of the identity which is preserved or not throughout the filiation.

3.2.2.1 The Identity of spatiotemporal entities over time

In [Khoshafian and Copeland, 1986], the authors define the *Identity* concept, in the context of database systems, as a "*property of an object which distinguishes each object from all others.* Based on this definition, [Hornsby and Egenhofer, 2000] address the issue of the identity of spatial entities that change over time. They explain that in the scenarios of change, the identity is a key factor in proving the existence or non-existence of an object as well as in being able to track similarities or differences in objects. Then, they examine change from the perspective of describing *identity changes* for objects. They first define a basic set of identity states of an object "*that can be combined to yield meaningful change operations*": non-existence without history ("*situation in which no object with identity is existing or has existed previously*"), existence, and non-existence with history ("*an object with identity previously existed, but has been eliminated and no longer exists*"). Then, they identify 9 transitions (or change operations) between these identity states: (a)

continue non-existence without history ("*transition between two identity states that are non-existing without history*"), (b) create ("*transition from a non-existing object without history to an existing object*"), (c) recall, (d) destroy, (e) continue existence, (f) eliminate, (g) forget, (h) reincarnate, and (i) continue non-existence with history.

In [Renolen, 1996] and [Renolen, 1997], the author is also interested in the description of the different states of spatiotemporal objects. He focuses on spatiotemporal modeling to understand how spatiotemporal objects "behave" in reality. He describes the different states of spatiotemporal objects (see Figure 3.6) that "*are either in a static state, in a changing state, or in a ceased state*". He introduce the state of change as a distinct state, in order to obtain a generic meta-model for spatiotemporal objects. In this model, the creation of an object may lead "*to a changing or to a static state depending on the type of object, hence the Conditional (C) transition. The model also allows an object that is ceased, to be reincarnated some time later*".

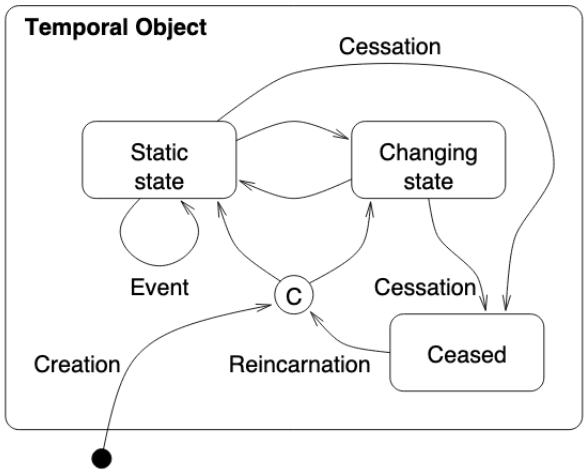


Figure 3.6 – Generic behaviour of spatiotemporal objects using the Statechart notation, from [Renolen, 1997].

Among these proposals from [Hornsby and Egenhofer, 2000] and [Renolen, 1997], we highlight in particular three **states of identity** generally employed: **creation, continuation and elimination** [Harbelot et al., 2013].

As one can observed in Figure 3.6, in this model objects may be described as a series of consecutive *static states* and transitions (*i.e.*, *changing states*). In the following Subsection, we investigate the different transition types defined in the literature.

3.2.2.2 Changes of spatiotemporal entities over time

In [Renolen, 1997], *events* are defined as transitions with zero duration (*i.e.*, when objects change suddenly *e.g.*, a cadastre that changes), whereas other objects may change continuously (*e.g.*, process of erosion of the coasts). And, it is stated that "*the change state is non-existent in a cadastre since all changes are events*", contrary to a process of erosion of the coasts for instance.

In [Claramunt and Thériault, 1995], events and processes are defined as follows: "*Events are things that happen; they are conditions, processes, or objects that exist and can be observed. We can thus model events as a set of processes that transform entities.*" On this basis, the authors present a typology of spatiotemporal processes (see Figure 3.7). This typology dissociates between the evolution a single entity may be subject to, the replacement and diffusion processes between entities, and the restructuring processes. With regard to TSN and their TUs, distinguishing between basic changes and transformations (see Figure 3.7, Ia) and Ib)), replacement processes (see Figure 3.7, IIa)), and restructuring processes (see Figure 3.7, III) is relevant. The *Movements* and *Diffusion* cases are less relevant since the TUs of a TSN are regions on the Earth whose boundaries may vary, which is unlikely the case with their position. However, this statement must be qualified, since the identity of the TUs may be based on an administrative status: for instance, within a TSN of administrative capital cities (composed of two levels *provinces* and *capital cities*), the change of the capital city in a province (from one city to another), may be characterized as a *Displacement*.

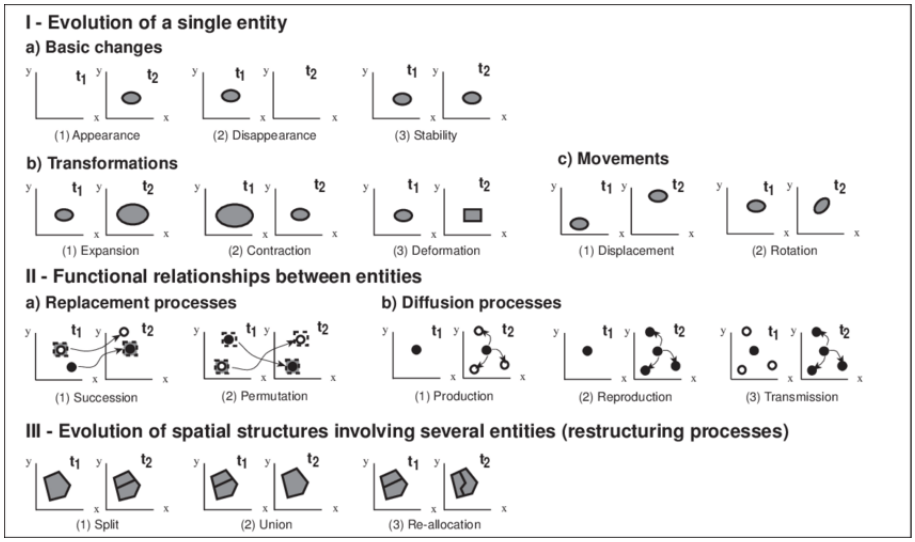


Figure 3.7 – Typology of spatiotemporal processes from [Claramunt and Thériault, 1995].

We have identified several terms (states and changes typologies) of the literature to characterize the evolution of spatiotemporal entities over time. In the next Sub-

section, we describe how the transitions from static states to changing states may be modeled to represent the history of entities over time.

3.2.2.3 Sequence of states and changes

In [Renolen, 1996], the author presents a notation called the *history graph* notation, that consists in creating a series of consecutive versions (*i.e.*, static states), and transitions (*i.e.* changing states) of an entity in order to represent its history over time. This sequence creates a *history graph* (see an example in Figure 3.8), that captures objects changes over time.

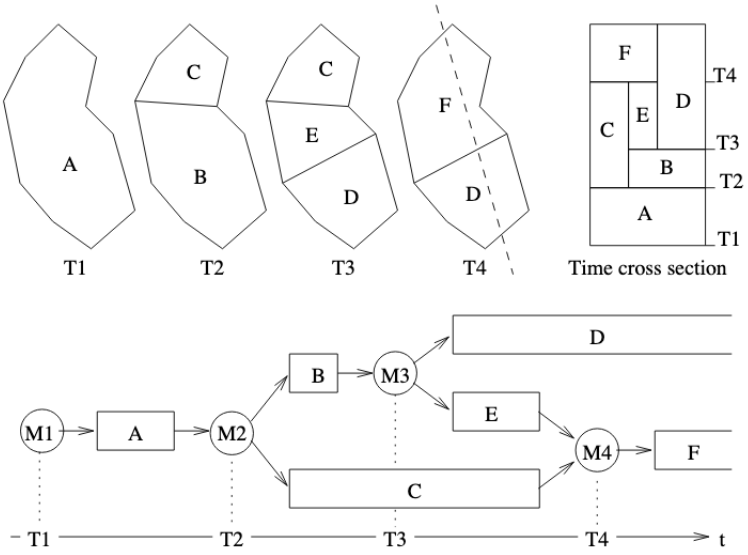


Figure 3.8 – The story of a land area (above) shown in the history graph notation (below), from [Renolen, 1997].

The graph approach is of particular interest in this context of representing entities (*i.e.*, vertices), and all their relations (*i.e.*, edges) with over entities (descendants, neighbors, or relations between all the temporal parts of the object that change over time). The graph model for spatiotemporal evolution of [Del Mondo et al., 2010] argues also in this direction. In [Del Mondo et al., 2013], the authors define the *spatiotemporal graph* as follows:

> **Spatiotemporal graph** *At an abstract level, entities (i.e., vertices) are related by spatial and filiation binary relations (i.e., edges).*

Then, spatiotemporal graphs focus on the links/relations between entities over time. These relations may be of two types: spatial or filiation relations. The filiation relations are links over time between the entities.

Also, in [Del Mondo et al., 2013], two filiation relation cases are identified, called

Continuation and *Derivation* relations defined as follows:

> **Continuation relation** *the first entity is the same as the second entity (e.g., one person at two times).*
> **Derivation relation** *the first entity creates (possibly with others) the second entity (e.g., a parent of a child)*

In the case of the *Continuation relation*, entities maintained their identity after change whereas, on the contrary, in case of the *Derivation relation*, the identity of the first element is not maintained after change, the two entities have not the same identity.

Let us consider once again the example of the ship of Theseus:

– Does the ship of Theseus at time 0 has a relation of type *Continuation* with the ship of Theseus at time 1?
– What if we consider the relation between the ship at time 0 and time 1000, all the planks having being replaced? (see Figure 3.1.2) In other word, it it still the same ship or a different one?

In the literature, there is no unique answer to this question, the response varies from one chosen definition of what makes the identity of the ship to another, and, more broadly, it depends on modeling choices. In the Subsections 3.2.2.4, 3.2.2.5, 3.2.2.6 different models and ontologies are presented. They all address the issue of the identity of evolving entities, but in different ways.

3.2.2.4 Versioning approach

The terms *Versioning* and *Version Control* are most of the time used in the context of computer science and more broadly for the management of modifications to text document. Rouse [2007] provides the following definition of these two terms:

> *Versioning is the creation and management of multiple releases of a product, all of which have the same general function but are improved, upgraded or customized. The term applies specially to operating systems (OSs), software and Web services.*
> *Version control is the practice of ensuring collaborative data sharing and editing among users of systems that employ different versions of a product. The terms "versioning" and "version control" are sometimes used interchangeably even though their technical meanings are different.*

As previously explained in Subsection 3.1.3, in the field of *Software configuration management (SCM)* (for managing the evolution of large software systems [Tichy, 1988]), a *version* is defined as a state of an evolving item. Another core concept of SCM is the *version model*. A *version model* defines the items to be versioned and the *delta*, which is the differences between two consecutive versions [Conradi and Westfechtel, 1998].

Version-Difference Spatiotemporal model Adopting the approach of *Software configuration management (SCM)*, Huibing et al. [2005] introduce the *Version-Difference Spatiotemporal model* developed with the requirement of using historical information in order to analyze changes of spatial objects over time.

Information systems based on this model store "the current state of an object (called the *default version*) and all historic changes (called *version difference*)." Then, using reconstruction operator and change descriptions, all the states of an entity over time can be obtained.

PAV and DC-terms Ontologies The *Provenance, Authoring and Versioning Ontology* (PAV) (URI: `http://purl.org/pav/`), and the Dublin Core Metadata Initiative Terms (DC-terms) (URI: `http://purl.org/dc/terms/`) also adopt the term *Version*. They focus on entities that are published on the Web: DC-terms provides terms to describe resources with metadata, whereas the PAV ontology is more devoted to the agents contributing in Web resources: contributors, authors, curators and digital artifact creators. Also, PAV provides terms for tracking the provenance of digital entities that are published on the Web [Ciccarese et al., 2013]. Both PAV and DC-terms use the predicate `hasVersion` to point towards a resource that is a version, edition, or adaptation of the described resource. The PAV ontology provides the following definition of the property *hasVersion*: "*This property is intended for relating a non-versioned or abstract resource to several versioned resources, e.g. snapshots.*

The term *Snapshot* is found in several models:
– in [Renolen, 1997], the following fundamental definition is given: *Objects that change suddenly would then be described by transitions with zero duration (i.e. events), while objects that change continuously would be described by version with zero duration (i.e. snapshots) describing intermediate states.*
– in the *Basic Formal Ontology* [Grenon and Smith, 2004], it is proposed to make a distinction between static *SNAP* (Snapshots) entities that change over time and dynamic *SPAN* entities. This model goes a step further than the models for versioning of entities, since it proposes to represent static *Snapshots* that change over time (*e.g.*, a person) but also the process behind these changes (*e.g.*, the aging process), *i.e.* the SPAN entities. This approach is presented below.

3.2.2.5 The SNAP and SPAN approach

The *Basic Formal Ontology* (BFO) is an upper ontology based on the approach of Grenon and Smith [2004] that accounts for both the static SNAP and the dynamic SPAN entities, while most of the existing models account only for one type at a time. Ontologies for **continuants** are called **SNAP**. **The terms 'continuant' and 'endurant' are used interchangeably** for those *entities that have continuous existence and a capacity to endure (persist self-identically) through time even while undergoing different sorts of changes (e.g., a person, the planet Earth)* Grenon and

Smith [2004].

Ontologies for *occurrents* are called *SPAN*. **Two different terms are used to refer to occurrent entities** (*e.g.*, a smile, the passage of a rainstorm over a forest) Grenon and Smith [2004]:
– *processes* that are occurrent entities which persist (perdure) in time *i.e.*, there are not instantaneous. The term *perdurant* is used for these occurrents;
– *events* that are occurrent entities which exhaust themselves in single instants of time.

The BFO framework addresses both the continuant and occurrent entities: it is a SNAP-SPAN framework. Indeed, while these two alternative views have traditionally been considered as incompatible, the authors argue that, as reality is essentially dynamic, an ontology must be capable of accounting for spatial reality both synchronically (entities that exist at a time) and diachronically (how the things unfold through time).

They introduce the notion of Trans-Ontological relation. A trans-Ontological relation is a "*relation between entities that are constituents of distinct ontologies.*"

The SNAP-SNAP Trans-Ontologies are ontologies that depict the world over time as a succession of temporally separated snapshots. Changes from one SNAP to another are described and may belong to one of the three main types that are:
– qualitative change: for instance, the color of a table becomes tarnished over time, yet there is still something which remains the same;
– locational change: for instance, "in one SNAP ontology the cup is on the table, in a later SNAP ontology it is on the floor. The cup underwent location change";
– substantial change: for instance, it is when a substance is divided up so as to produce a plurality of substances.
The SNAP-SPAN Trans-Ontologies are ontologies that depict the *life* or *history* of entities over time. The life line of an entity is a SPAN object, and the entity itself is a SNAP object. "*Histories might best be understood as four-dimensional settings.*"

The two approach SNAP-SNAP and SNAP-SPAN may also be combined in order to depict both the changes of a SNAP entity between two periods of time and its evolution process over time (*i.e.*, its life).

Then, [Grenon and Smith, 2004] address in particular the case of Geographical objects, and dissociate between the:

– *SNAP Geographical Object Ontology* from which the subcategory "*Boundaries and Geographical Regions*" is one of the 5 major subcategories to geographical SNAP entities. The authors explain that administrative boundaries are SNAP *fiat* objects, that is to say, they are constructed objects that exist according to an administrative, social or political convention (in opposition to *bona fide* object that correspond to "natural" objects, such as natural boundaries like mountains) [Smith and Varzi, 2000; Mathian and Sanders, 2014].

– *SPAN Geographical Process Ontology* that may be of two types depending on the processes described: physical processes or social processes. For instance, a change of the administrative boundaries of regions is a social process.

The BFO Ontology tree of concepts is presented in Figure 3.9 (this Figure is obtained from the OWL representation of the BFO Ontology, available at `https://raw.githubusercontent.com/BFO-ontology/BFO/master/releases/2.0/bfo.owl`). This tree synthesized all the above explanation. It is divided first in two sub-categories: continuant entities (SNAP) and occurrent entities (SPAN). The former has a sub-category that is called "immaterial entity". The spatial regions appear in this sub-category of fiat entities. There are artifact regions that have continuous existence in time. The category of occurrents has a sub-category that is called "spatiotemporal region", *i.e.* a region at a specific instant of time or time-period.

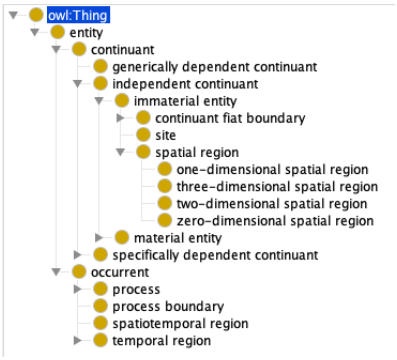


Figure 3.9 – The main concepts of the BFO ontology.

To conclude, as remarked by [Gantner, 2011] "since most geographical ontologies still take the view of a static world, Grenon and Smith developed the Basic Formal Ontology (BFO), an upper ontology that accounts for both the static SNAP and the dynamic SPAN entities."

In the following Subsection 3.2.2.6, we present another four-dimensional approach that focuses on the representation of a relation (between entities) that changes over time, as most of the time *relationships are diachronic, i.e. they vary with time*. This approach is called *Ontology for fluents*. In [Welty et al., 2006], the authors present an example that explains the 4D-Fluents approach they adopt: in this approach, statements such as "*Joe walked into the room*" (*i.e.*, the relationship between Joe and the room at one time), are represented as "*a temporal part of Joe walked into a temporal part of the room*".

3.2.2.6 The Ontologies for Fluents approach

Ontologies for fluents are based on the *perdurantist* approach (from D.K. Lewis philosopher) and represent, in OWL, relationships between entities that change with time [Welty et al., 2006]. In the *fluent* approach, entities are four dimensional with temporal parts. The 4D view maintains that all entities are perdurants (*i.e.*, an individual has distinct temporal parts throughout its existence). Thus, all entities have temporal parts and can be thought of intuitively as four dimensional *spacetime worms* whose temporal parts are slices of the *worm* [Sider, 2001; Welty et al., 2006].

Concepts in time are represented as 4-dimensional objects with the 4th dimension being the time [Batsakis et al., 2017]. Figure 3.10 shows on the left a SNAP approach, where representation of evolutions in ontology can only be described by a series of snapshot ontologies each superimposing itself on the previous version of the described reality. For its part, the 4D-Fluents (perdurantist) ontology (on the right) allows the concepts of time and change to become integral parts of the ontology [Batsakis et al., 2017].

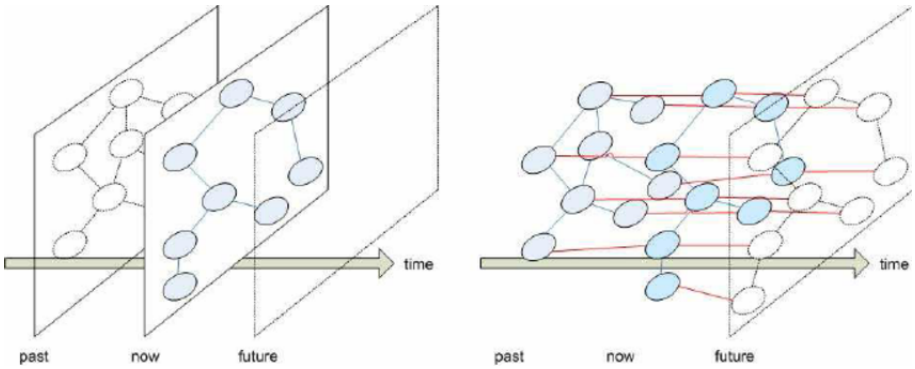


Figure 3.10 – Schematic representation of the concept of time-determined ontology[Baratis et al., 2009]. On the left: a series of snapshot ontologies. On the right: a 4D-Fluents (perdurantist) ontology.

Perdurants have two types of attributes: identity attributes and other attributes, valid for a period of time [Welty et al., 2006; Harbelot et al., 2015]. If the attributes that hold the identity change, a new entity is created. While, if the other attributes change, a new sub-object, called *TimeSlice*, linked to the main one is created: it holds all the attributes that may change over time [Welty et al., 2006]. The main issue addressed here is the representation of relations between entities that change over time. The authors call *Fluents* relations that hold within a certain time interval and not in others and the time interval of a time slice is defined to be the duration of the fluent holding. They take the example of the following sentence that hold information they want to register in a database system:

Sam Palmisano was named chief executive officer of the IBM Corporation effective March 1, 2002.

It is a common example in the literature to explain relationships that change over time, the example of the relation between an employee and a company, a relationship that is true for a period of time, since the person was not always an employee of this company.

There is different ways to address this problem. We present two of them here:

– the *reification* solution that reifies the relationships into an object, as shown in Figure 3.11. "*Reification is a general purpose technique for representing n-ary relations using a language such as OWL that permits only binary relations.*"[Batsakis and Petrakis, 2011]. Although this approach requires only one additional object for every temporal relation, it suffers from redundancy of the properties (*e.g.*, *employs, worksFor*) that participate in the reification of the relationships.

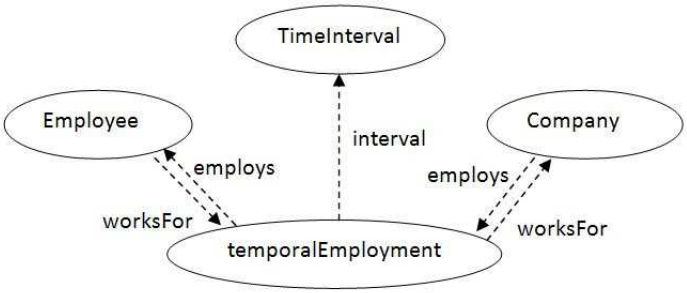


Figure 3.11 – Example of N-ary Relations from [Batsakis et al., 2017].

– the 4D-Fluents (perdurantist) approach where objects in time are represented by *TimeSlices*, and their temporary relationships are described between these timeslices. The main advantage of this approach is the possibility to describe changes of the entities on the timeslices sub-objects. However, this approach suffers from a proliferation of objects, as remarks by [Batsakis et al., 2017] since it introduces two additional objects (*e.g.*, *EmployeeTimeSlice* and *CompanyTimeSlice*) for each temporal relation (instead of one in the case of N-ary relations (*e.g.*, *TemporalEmployment*)).

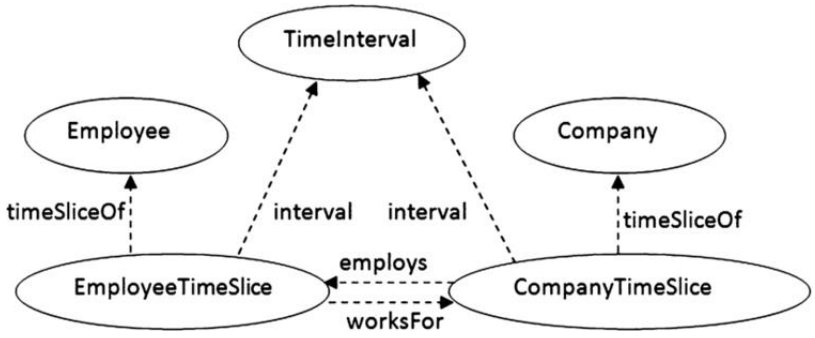


Figure 3.12 – Example of 4D-Fluents from [Batsakis et al., 2017].

The 4D-Fluents (perdurantist) approach offers a new way to answer the *identity* question over time, exposed previously with the Ship of Theseus story. In this approach, the Ship of Theseus life is represented through one *space-time worm* ship which has temporal parts (timeslices): a part at year zero, a part at year 1, ..., a part at year 1000 Sider [2001].

This 4D-Fluents approach, combined to the *OWL-Time* ontology (to assign time points or interval to a *TimeSlice*), is often used in the LOD Web to represent the evolution of resources [Batsakis and Petrakis, 2011], including geospatial ones [Harbelot et al., 2015; Tran et al., 2015]. For instance, the *Continuum Spatiotemporal Model* represents time evolving parcels, adopting the 4D-Fluents approach of [Welty et al., 2006]. This ontology is presented in the next Chapter 4.

Chapter 4

Modeling evolving geographic divisions and TSNs

In the previous Chapter 3, we have presented fundamentals for the modeling of evolving geospatial entities. In this Chapter, we focus on the modeling of evolving TUs and TSNs, in the literature.

4.1 Modeling evolving geographic divisions

In the literature, there are several models to represent the evolution of geographic units over time. We focus on this state of the art on the approaches that represent the evolution of geographic data that have irregular polygonal geometries (vector data). It is possible to draw inspiration from each of these approaches, although they address particular although each of them deals with data of a particular domain: land-cover parcels, historical regions, Jurisdictional Domains, administrative regions. All these geographic divisions have domain-specific peculiarities. They may be used as a support (collect and restitution) to statistical data. However, that does not mean that they match with TSNs, because they are sometimes less standardized and they are not organized in versions (*e.g.*, the changes are described as time goes by, while in the TSNs a period of artificial stability is created by the SAs, because of the time period needed for data collection).

4.1.1 Land-cover context - The Continuum Model

In [Harbelot et al., 2015], the spatiotemporal ontology-based model called Land Cover Change Continuum (LC3) Model is introduced. It is an extension of the Continuum Model introduced in [Harbelot et al., 2013] as follows:

The Continuum Model extends GeoSPARQL allowing it to represent spatiotemporal dynamics objects. This extension is achieved by combining the GeoSPARQL Ontology with an ontology of fluents.

The Continuum model have been applied to the Land Use Land Cover case study, in order to observe how humans modify the Environment over time: "*The land cover of a region varies along time due to a variety of factors. Being able to understand this process of evolution and identify trends and patterns would be helpful for scientists and policy makers to manage land territory*" [Harbelot et al., 2015].

In order to test the LC3 model, the authors present an implementation on the CORINE (COoRdinate INformation on the Environment) Land Cover (CLC) data set (from the European Environmental Agency (EEA)), that covers several countries, at three different years 1990, 2000, and 2006.

From a process of photo-interpretation of satellite images, polygonal geometries (vector data) are obtained associated with a class. This class describes the land cover or land use of the polygon. Indeed, within the CLC data set, a hierarchy of 44 classes (a nomenclature) of land-parcels is defined (see Figure 12.2). These classes allow to characterize the land use/cover as a residential, industrial, or natural parcel of forest for instance.

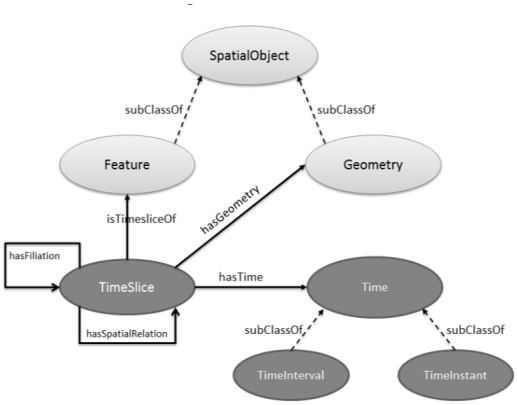


Figure 4.1 – Continuum Main concepts [Harbelot et al., 2013].

Figure 4.1 presents the main concepts mobilized in the Continuum model. The concept named `TimeSlice` inherits from the `Feature` concept of GeoSPARQL (see the GeoSPARQL main concepts in Figure 3.2). In this approach, geospatial objects may have two different kind of attributes: attributes that hold the identity of the feature (*identity component*); attributes that vary in time (hold by the *timeSlices* of the feature) that are: semantic components, spatial and temporal components. This means that, for instance, the localization or the boundaries of a feature may change while the feature maintained its identity after change (only if the identity components not change). As for the Ontologies for fluents (see Fig-

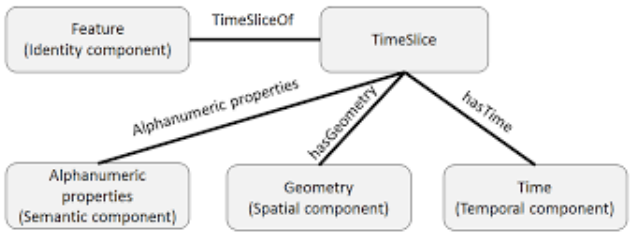


Figure 4.2 – Timeslice in the LC3 Model [Harbelot et al., 2015].

ure 3.12), the temporary relationships between `Features` (topological relations or filiation relations) are on the time-slices of the features.

In Figure 4.1, two main predicates are shown on the `TimeSlice` concept, `hasSpatialRelation` and `hasFiliation`. The relationship `hasSpatialRelation` expresses a topological link between two time-slices of different objects valid on the same interval or instant of time. The Continuum Model focuses on the second type of relation *i.e.*, the filiation relationship, defined as follows: "*This relation is established between two time-slices, it allows to track and define the evolution of an object in the case of a continuation, and to track and define the transformation of one object to another in the case of a derivation.*" [Harbelot et al., 2013]. The model proposes a hierarchy of filiation relations, presented in Figure 4.3 (*i.e.*, they are predicates in the RDF triples representation). These predicates are based on the filiation relations defined in [Del Mondo et al., 2013].

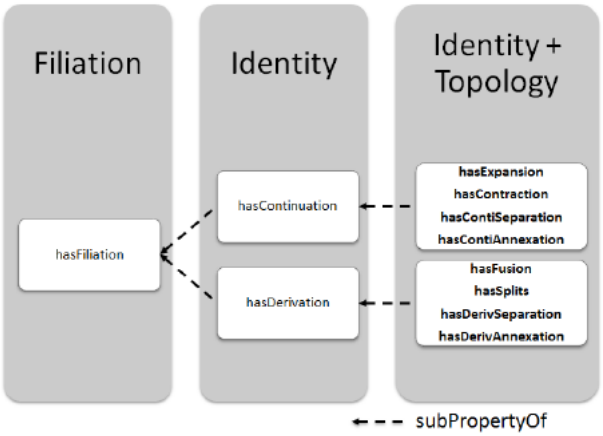


Figure 4.3 – The Continuum Model with different levels of filiation relationship [Harbelot et al., 2013].

The Continuum approach exploits the capabilities of both the Semantic Web and the Graph Model for Spatiotemporal evolution [Del Mondo et al., 2013] to **infer** filiation relationships (or better qualify them) on the basis of semantic constraints (*i.e.*, integrity constraints and inference-queries based on the constraints). "*Traditionally, the semantic Web approach is not intended for the management of integrity constraints*" [Harbelot et al., 2013] because the semantic Web approach follows the open world assumption, whereas the integrity constraints follow a close world assumption.

Then, a hybrid solution is presented, allowing to manage both reasoning capabilities under the open world assumption and data validation under closed world assumption.

In [Harbelot et al., 2015], an implementation of the LC3 model is presented, and called the *LC3 system*. The L3C system establishes filiation links between the three CLC data sets (1990, 2000, and 2006) searching for filiation links between

the different time-slices of the land-parcels. The CLC class associated with the parcel and the spatial properties of the parcel are taken into account to establish the filiation. The detection of the spatial relations between the time-slices are computed with a JAVA/Geotools application, developed to perform all the required spatial analysis (*e.g.*, parcels' polygons intersection).

The filiation links that qualify an evolution in the Continuum model are constructed according to a method presented in the Section 5.1.

At the end, the method produces graphs such as the one shown on Figure 4.4. The graph on the left is obtained after step 3, and the graph on the right is obtained after analysis of an expert after step 4. It links the land-parcels through time using the predicates of the Continuum Model.

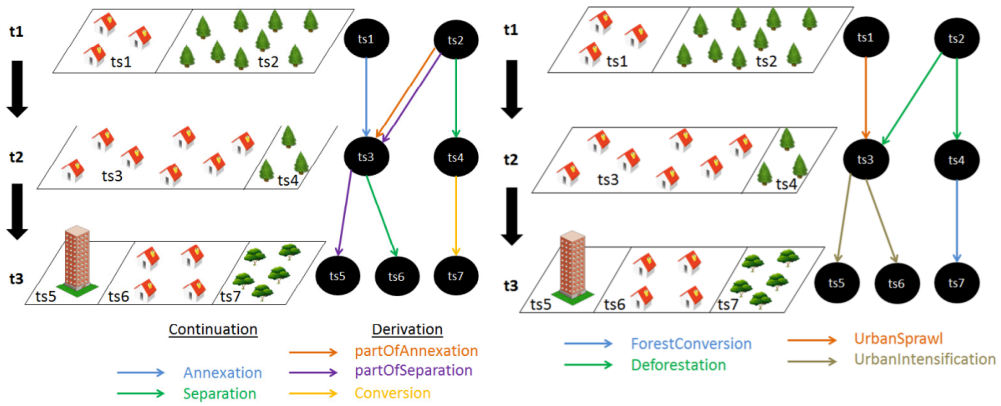


Figure 4.4 – Filiation Graphs of evolving land-parcels example [Harbelot et al., 2015].

This approach constructs a lineage of the land-parcels through time. Two attributes of the land-parcels are taken into account to determine if the filiation link throughout the versions of the CLC data set are of type *continuation* or *derivation*: the polygon geometry of the parcels and the nature of the land cover (*e.g.*, agricultural area, forest area, construction site, etc.) that comes from the *Corine Land Cover Nomenclature*. Thus, this approach suits well to CLC data sets, but it is not appropriated for administrative divisions, for instance. It does not take into account thematic attributes that may constitute the identity of the areas (*e.g.*, identifier, name of the entities). The identity through time is first based on the spatial component, since the method consists in determining the overlapping of entities then to qualify the filiation relationships in case of overlapping. This approach lacks flexibility when it comes to the definition of identity of the time-slices, if it was to describe changes in geographic divisions different from CLC.

The only implementation of the proposed model is the CLC implementation where the identity of the parcels is determined solely by two attributes: the geometry of the parcel and the nature of the land cover (one of the classes of the CLC hierarchy,

see Figure 12.2). As a consequence, the typology of patterns detected in the LC3 Model is a typology that suits only CLC polygons. It has to be extended to address administrative units that change over time, for instance, in order to describe name changes, capital changes, ...

Another problem of this approach is that the semantic of changes is only hold by the predicates that link two time-slices. If one wants to extend this approach in order to describe other changes such as identifier change, name change, inhabitant number change... then, whenever an attribute changes, a link has to be drawn between the two time-slices. Also, the direct links created between two time-slices in case of *Derivation* situations are questionable. Indeed, in many cases, direct links might not be relevant in the context of statistics. For instance, let us consider *Redistribution* events, where the identifiers and the boundaries of several TUs are modified simultaneously in a way that it is difficult to define the nature of the change (*e.g.*, in terms of *merge* or *split*). In this case, it might be more relevant to determine the set of impacted TUs and the set of created ones, then to link them all through a node that makes the switch from one version to another. In these cases, the *Change Bridge* approach of [Kauppinen and Hyvönen, 2007] has to be considered. This approach is presented in the next Subsection. It consists in indirectly linking *input* and *output* elements of a change event, through a *Change node* that describes the nature of the territorial change.

To conclude, the main drawback of the Continuum approach is that it makes no semantic distinction between the filiation relations and change descriptions. While both occur during change of the object from one state to another, they are two distinct information one should dissociate to better understand the dynamics. Thus, we propose in this manuscript to dissociate the filiation links (established between two time-slices of a same spatial entity) from the descriptions of changes to the entity.

4.1.2 Historical Context - The Finnish Spatiotemporal Ontology

As explained in the previous Subsection, there are two approaches to link entities that change over time: direct links, or indirect links using intermediate nodes that describe the changes of the entities (*e.g.*, merge of two entities). In this Subsection, we present the approach of [Kauppinen and Hyvönen, 2007] that have chosen the second approach in order to link historical regions of Finland over time, and to describe their changes.

In [Kauppinen et al., 2008], the authors introduce the notion of *Change Bridge* to chain former territories (*e.g.*, *East Germany* and *West Germany*) to their successors (*e.g.*, *Germany*) (see Figure 8.2, approach (2)), using the *Change Vocabulary* for the description of changes.

The *Change Vocabulary*[1] of [Kauppinen et al., 2008; Kauppinen and Hyvönen, 2007] is a lightweight spatiotemporal vocabulary made of two properties (*before*,

1. `http://linkedearth.org/change/ns/`

after) and five classes for the description of changes (*Establishment*, *Merge*, *Split*, *Namechange*, *Changepartof*).

In order to illustrate their approach, they take the example of the German reunification in 1991. Figure 4.5 shows how they represent this reunification, using the Change Bridge approach that consist of a change node between the regions, before and after the change event. In an online specification of the Change vocabulary, the following RDF representation (see Listing 4.1) of Figure 4.5 is provided.

In their model, the East and West Germany are two concepts of a Period Ontology O_1, and the new concept of Germany is introduced in a new Period Ontology O_2. A node (*merged42*) of type `merged` (at the middle of Figure 4.5) makes the bridge between the two old notions of East and West Germany, and the new concept *Germany*. The `merged` node holds also some other information such as the year of the reunification (1991).

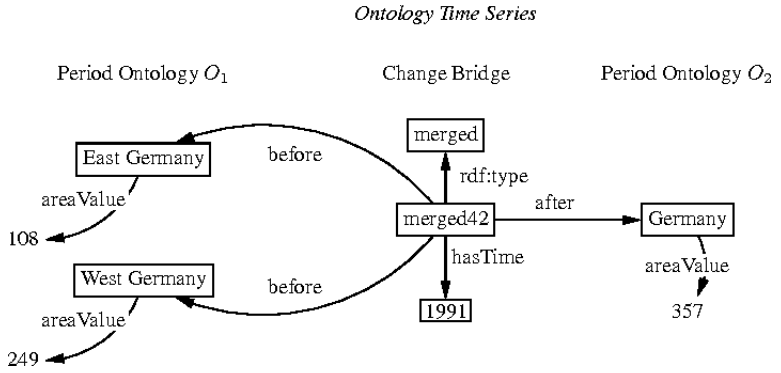


Figure 4.5 – A Change bridge example from [Kauppinen and Hyvönen, 2007].

```
@prefix dc: <http://purl.org/dc/terms/#>  .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix change: <http://linkedearth.org/change/ns#> .
@prefix tisc: <http://observedchange.com/tisc/ns#>  .

change:merge-of-germanies rdf:type change:Merge ;
    rdfs:label  "East Germany and West Germany merged to form Germany" ;
    change:before dbpedia:East_Germany ;
    change:before dbpedia:West_Germany ;
    change:after dbpedia:Germany ;
    dc:description "East Germany was dissolved upon joining the
                    institutions of West Germany in the German
                    reunification, on 3 October 1990" ;
    dc:date "1990-10-03" .
```

Listing Code 4.1 – Example of the use of the Change Vocabulary to model the merge of East Germany and West Germany in 1990.

Then, in the approach of [Kauppinen et al., 2008], each geographic area is a

concept that belongs to an ontology time-series, made of several Period Ontologies (*e.g.*, O_1, O_2). In [Kauppinen and Hyvönen, 2007], an ontology time series is defined as follows: "*Ontologies evolve when the underlying domain world changes at different points of time. The result then is a series of ontologies whose concepts are related with each other not only within one ontology valid at a moment but through the time.*" [Kauppinen et al., 2008] focus on the modeling of relations between geographic area concepts that evolved over long periods of time. The domain of application of their method is the historical geospatial reasoning. Thus, their *Change Bridge* approach has been implemented in order to create the *Finnish Spatiotemporal Ontology*, an **ontology time series** of the Finnish municipalities over the time interval 1865–2007 [Kauppinen et al., 2008].

In [Kauppinen and Hyvönen, 2007], the authors explain that they adopt a three-dimensional approach. They represent geospatial entities that change over time using a "*combination of a three-dimensional SNAP-ontology*". They do not have the four-dimensional notion of the same concept, for instance "Finland" or "Lappeenranta" (a city in Finland) that changes over time. Instead, they define different areas, each one having one temporal part (*e.g.*, `location:Lappeenranta(1967-1989)`, `location:Lappeenranta(1989-)`). Each time a change of boundaries or name occur, new area(s) (with a new temporal part) is/are created. This representation of changes over time suits well to the domain (historical data), and to the aim of their works, *i.e.*, enabling semantic search and browsing of historical regions (using the historical names of the regions for instance).

In order to create the ontology time series that represents the evolution over time of a country such as Finland, they propose a methodology in 3 steps [Kauppinen et al., 2008]:
– (1) modeling geospatial changes. They analyze the change types and construct a Metadata Schema of *Current Places*, one for *Historical Places* and another one for *Changes*. The Metadata Schema of Changes is composed of the following main fields: *identifier* for the change, *date* of the change, *change* type (either *establishment, merge, split, namechange, or changepart*).
– (2) populating the metadata schema. The metadata schemas were implemented as a spreadsheet table, edited by hand. This means that the authors do not automate the process of detecting and describing changes.
– (3) creating the ontology time series from the data spreadsheet table created after step (2). They create a JAVA program in order to transform the spreadsheet input into an RDF graph (output).

The created RDF graphs are available online from the Finnish Ontology Service of Historical Places and Maps [2] at `http://dev.hipla.fi/` and the HTML pages of the resources are accessible online *e.g.*, `http://www.yso.fi/onto/sapo/Heinavesi(1920-)`.

The modeling approach presented in [Kauppinen and Hyvönen, 2007; Kauppinen et al., 2008] is criticized by [Lacasta et al., 2014] as it may lead to a proliferation

2. description of the service at `https://seco.cs.aalto.fi/projects/histoplaces/en/`

of instances. Indeed, the complete status of a TU is described before and after the associated change event. We present in the next Subsection the approach from [Lacasta et al., 2014], deriving from [López-Pellicer et al., 2008].

4.1.3 Political context - The Jurisdictional Domain Ontology

[López-Pellicer et al., 2008] propose description of changes in political divisions, called Jurisdictional Domains, that change quite frequently. They approach is to minimize the description and to avoid duplication of data.

"*Jurisdictional Domains (JDs) are generally accepted* ***political divisions*** *of the earth surface that cover specific territorial and functional scopes over time. They are units of administration for local, regional, national, or international governance with specific roles separated by administrative boundaries.*" In [López-Pellicer et al., 2008; Lopez-Pellicer et al., 2012], the authors address two issues: the heterogeneity of political divisions and their evolving nature. They propose an ontology schema that combines, in a single model, the political structure, the spatial components, and the temporal evolution of units. Focusing on existing ISO standards, the authors conclude that no model has an appropriate semantic representation of the different types of administrative units, and of their spatial and temporal relations. They focus on JDs for local, regional, national or international governance.

They propose to connect to a higher ontology (the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)), and to allow others to create their own application ontology, by country, for the representation of their own administrative units.

> " *This scheme has three layers* as shown in Figure 4.6:
> *(1) A high-level ontology that defines data types and general relations which are independent of context.*
> *(2) A domain ontology which defines concepts and relations that can be reused in the context of the administrative models of different countries.*
> *(3) And an application ontology per country, which represents the specific types of administrative units of each country, along with specific instances of existing units.*"

In the Jurisdictional domain ontology, the *jurisdictional-domain* concept is specialized in *states* and *administrative-divisions* among others. "*States are territories under effective and civil government; administrative-divisions represent any division in a jurisdictional-domain (administrative-divisions hold a relation with their parent jurisdiction). The spatial information is modeled by the jurisdictional-geographic-object concept, which is defined as a kind of DOLCE politic-geographic-object.*"

In [Lopez-Pellicer et al., 2012], the authors identify different types of changes that a JD can undergone: creation, dissolution, change in its properties, and change in its relations. The latter has sub-types presented in Figure 4.7

The model supports the representation of JDs hierarchies using the predicate

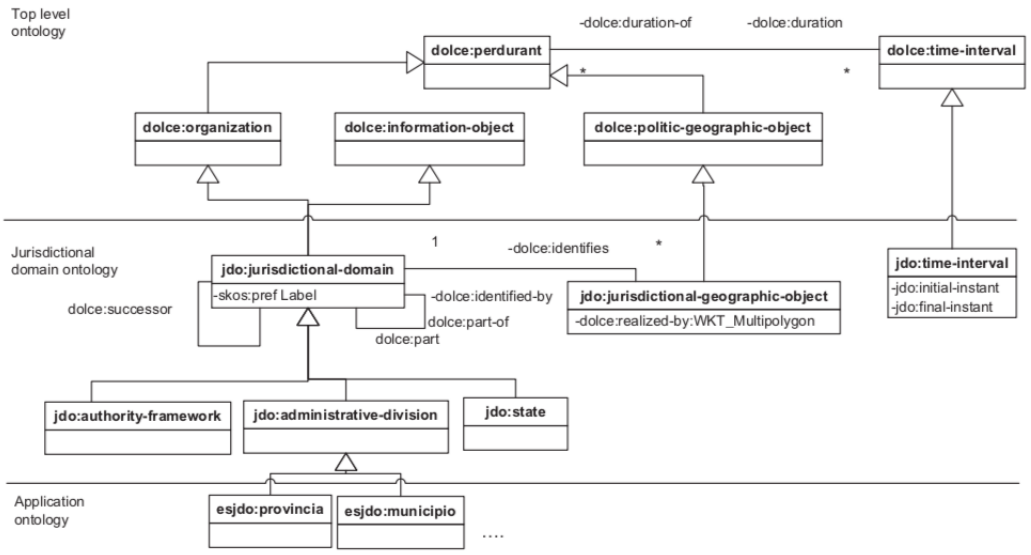


Figure 4.6 – The Jurisdictional domain ontology (upper, domain and application ontologies) from [Lopez-Pellicer et al., 2012].

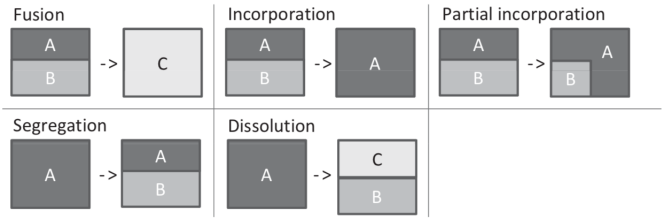


Figure 4.7 – Meaning of the different succession relations in the model of [Lopez-Pellicer et al., 2012].

`part-of`. As in [Harbelot et al., 2015], the authors choose to create links between entities over time, using predicates that hold both the semantic of filiation (typology of *successor* predicates), and the change descriptions. They criticize the approaches of [Kauppinen et al., 2008] and [Gantner, 2011] (presented in the following Subsection 4.1.4) as they may lead to a proliferation of instances. Indeed, the complete status of the geospatial entities are described before and after the change event. Then, the authors choose to describe changes in JDs at the property/attribute level (each property/attribute of a JD can have its own time span). This prevents from the creation of a new instance after each change.

In [Lacasta et al., 2014], the authors present a semi-automatic process to avoid redundancy and to merge several input RDF fragments (RDF fragments describing JDs and their changes (hand written) in an extended ISO 19112 metadata schema) into the JDs knowledge base. If two input JDs descriptions (in RDF) have the same id, name and upper JD (strict String equality), they are merged into a single entity *i.e.*, this step removes redundancy. The next step consists in constructing the JDs'

evolution descriptions. Types of change considered here are: *denomination change, fusion, incorporation, partial incorporation, segregation, dissolution* (presented in Figure 4.7).

The change of a property value of a JD that does not imply the creation of a new JD (administratively speaking) (*e.g.*, *Denomination change*) implies constructing one JD object that holds several values for this property, at different time span. On the contrary, the change of the administrative status of a JD implies the creation of a new JD object. Then, in Figure 4.7, the letters "A, B, C" are used to denote on an administrative status. If the JD's status changes, a new letter is used to signify this change of identity.

The objective of the authors here is to enhance the Information Retrieval (IR) system capabilities. Indeed, in [Lacasta et al., 2014], the authors explain that the continuous evolution of JDs complicates the search in collections containing historical data. They take the example of a project called the *Old Maps Online project* [Přidal and Žabička, 2008] where the access to historical maps, using place names, is performed using the current JDs names, not the historic ones.

> " *For example, a search about the old kingdom of Prussia (1525–1947) requires a query that includes current Germany, Poland, Russia, Lithuania, Denmark, Belgium, Czech Republic and Switzerland.* "

Then, in order to facilitate the navigation through entities that have existed in a place over time, they explicitly state in the model the *succession between entities*. They explain that this can be achieved automatically (using the boundaries of the areas, through geometrical intersection operations), however they do not automate this process since they do not have the exact geometry of all JDs.

For instance, the jurisdictional ontology of Spain has been created using as input the changes database of the Spanish registry of local entities, and other (older) changes have been manually added. The created RDF graphs were used within the Spanish Official State Gazette (the Spanish official publication collection for legislation and mandatory acts) with the objective of improving the search results of the IR system designed for the historical collection. The initial text-based search application returns all the documents containing the text of a query. If the text corresponds to the name of a JD, it returns documents containing the place name. The improved search component, presented in [Lopez-Pellicer et al., 2012], returns the documents containing the place name, as well as all the documents containing the alternative names used to identify the JD over time (information contains in the RDF graphs).

The results show an improvement of the IR possibilities. Let us take the example of an incorporation (please see Figure 4.7, case *Incorporation* of a JD B in a JD A), using information on JDs' changes over time, after querying data on a region B at time t, the system returns documents addressing also a bigger region A at time $t+1$, as the region B between time t and $t+1$ has merged with another JD, resulting in its incorporation in the JD A. Thus, the system provides users with information

about the evolution of the JDs (fusions, segregation, etc.). The main issue identified by the authors is the lack of available spatial historical information.

Then, the *Jurisdictional Domain Ontology* (JDO) [Lopez-Pellicer et al., 2012] describes changes in Jurisdictional Domains (JDs) at the property level (each property can have its own time span). This prevents from the creation of new instance after each change. However, the modeling of JDs changes over time is still a different issue than the modeling of TSN changes over time. Indeed, the objective is to report on JDs changes as soon as they occurred (quite frequently) whereas, in statistics, several versions of the whole TSN (geographic divisions) are created to report on TSNs changes all at the same time, that is to say not at the time where they occur but after a period of stability (see Chapter 2, Section 2.3). Our model, described in the Chapter 8, is trans-temporal (the time of the statistical data is dissociated from the time of the territorial partition) and adopts a versioning approach. In [Lacasta et al., 2014], the authors propose a semi-automatic process for creating and populating the evolution model of any country, with an approach that is also quite generic since the characteristics of the heterogeneous sources can be set dynamically. However, it does not produce any description of the impacts of changes on hierarchies of TUs. Furthermore, this approach requires the use of a **written dictionary of changes** to populate the ontology. This dictionary lists all of the individual changes as textual data described in the ISO-19112 standard, extended with elements to represent changes. This requirement may be seen as a drawback for current statistical information systems because the manual listing of changes takes times. Automating the change detection and annotation allows for processing many geographic divisions in the world (using the same semantics for the description of changes), whether or not there exists a catalog of changes. However, it should be noted that in the case of crowd-sourcing platform, this approach is a very interesting one, as each participant may edit the geometries of the JDs they know, the changes on the JDs' attributes, using the appropriate administrative terms of the JDs described.

Moreover, the main drawback of this approach is the lack of definition of what could make the identity of the JDs (over time). Thus, we assume that what is taken into account (when describing changes manually) for the distinction between a change of type continuation or derivation (*i.e.*, a change of the identity of the JDs or not, *e.g.*, Fusion *versus* Incorporation, see Figure 4.7) is a new act of law (or not), that indicates a change of the political identity of the JD.

The next Subsection presents an ontological approach for the representation of the evolutions of the Switzerland administrative regions. The model also deals with administrative divisions, with a much less important place to the political dimension. Conversely, in this approach it was decided to define clearly what makes the identity of the administrative regions over time.

4.1.4 Administrative Context - The SONADUS Ontology

In [Gantner, 2011], an ontology called *Spatiotemporal Ontology for the Administrative Units of Switzerland* (SONADUS) [3] is proposed in order to observe the evolution of the Swiss administrative units, between 1960 and 2010. The Ontology is written in German: "*Since potential users are mostly Swiss, the concepts in the ontology were given German names.*" It focuses on the specificity of the Swiss territory without adopting abstract terms so that other geographic divisions in the world could be described using the ontology. The author points out that the administrative units (AUs) of Switzerland evolve rapidly and can become unrecognizable over time because neither the *Historicized Municipality Register* nor the geometric data sets define what makes the identity of these AUs. This was the main motivation for creating the SONADUS ontology and the data sets based on it.

The SONADUS ontology and data set were created according to the following steps:

(1) The ontology was first created by transferring in OWL the conceptual data model of the *Historicized Municipality Register* (HMR) [4]. At this step, both concepts describing the Swiss administrative structure (cantons, districts, ...) and concepts describing changes were created based on the change types defined by the Swiss Federal Statistical Office in HMR publication [statistique, 2017]. The SONADUS typology of changes is presented in the Figure 4.8 (please, see the second column that is an English translation of the concepts of the SONADUS ontology.

Regarding these change types: "*SONADUS differentiates between 24 types of change, an increase of 11 types in comparison to the HMR+. This also contrasts sharply with the classification of [Kauppinen et al., 2008] that includes seven change types of the AUs of Finland. SONADUS, more precisely, makes a distinction between the hierarchical levels, expands the set of change types to the level of cantons, opens and ends the change history, corrects errors in the source data, and finally contains purely administrative change types such as formal renumbering.*" [Gantner, 2011].

The main issue regarding this typology is that in one concept several different kind of change, and several different semantic information are expressed. For instance, the concept "*Change of membership of a canton on the municipality level*" expresses to which Swiss administrative level the AUs that change belong to, as well as the Swiss administrative level of their super AUs, and finally the nature of the change (that in a more abstract way may be described as a change of super TU). **Also, all the concepts depend on the Swiss administrative context, and are written in German which may hinder their re-use.** In this respect, the author wonders: "*To what extent is SAPO [Kauppinen et al., 2008] applicable to the situation in Switzerland? In what way does the evolution of administrative units differ between Finland and Switzerland?*". Even if the

3. URI of the SONADUS ontology: `https://www.wsl.ch/sonadus#`

4. `https://www.bfs.admin.ch/bfs/fr/home/bases-statistiques/repertoire-officiel-communes-suisse/liste-historisee-communes.html` ; `https://opendata.swiss/en/dataset/historicized-municipalities-register`

Change types in SONADUS	English translation	Former change type in the HMR+	Difference
NeueBezirksgliederung	Restructuring of districts	Restructuring of districts	none
Ausgemeindung	Secession	Secession	none
Eingemeindung	Annexation	Annexation	none
Gebietsabtausch	Land exchange	Land exchange	none
Gemeindefusion	Merger	Merger	none
Gemeindetrennung	Separation	Separation	none
AenderungBezirkszugehoerigkeit	Change of membership of a district	Change of membership of a district	none
AenderungGemeindenamen	Change of municipality name	Change of municipality name	none
FormaleNeunummerierungGemeinde	Formal renumbering of a municipality	Formal renumbering of a municipality	none
ErsterfassungAufStufeBezirk	Initial entry on the district level	Initial entry	Breakdown according to the hierarchical level
ErsterfassungAufStufeGemeinde	Initial entry on the municipality level	Initial entry	Breakdown according to the hierarchical level
AenderungBezirksnamenAufStufeBezirk	Change of district name on the district level	Change of district name	Breakdown according to the hierarchical level
AenderungKantonszugehoerigkeitAufStufeBezirk	Change of membership of a canton on the district level	Change of membership of a canton	Breakdown according to the hierarchical level
FormaleNeunummerierungBezirkAufStufeBezirk	Formal renumbering of a district on the district level	Formal renumbering of a district	Breakdown according to the hierarchical level
AenderungBezirksnamenAufStufeGemeinde	Change of district name on the municipality level	Change of district name	Breakdown according to the hierarchical level
AenderungKantonszugehoerigkeitAufStufeGemeinde	Change of membership of a canton on the municipality level	Change of membership of a canton	Breakdown according to the hierarchical level
FormaleNeunummerierungBezirkAufStufeGemeinde	Formal renumbering of a district on the municipality level	Formal renumbering of a district	Breakdown according to the hierarchical level
NeugruendungKanton	Creation of a new canton	none	new
ErsterfassungKanton	Initial entry of a canton	none	new
LetzterfassungAufStufeBezirk	Final entry on the district level	none	new
LetzterfassungAufStufeGemeinde	Final entry on the municipality level	none	new
LetzterfassungKanton	Final entry of a canton	none	new
MutationsprozessUpdateGeometrie	Update of geometries	none	new
Eingemeindung und Gebietsabtausch	Annexation and land exchange	Annexation	new - error correction

Figure 4.8 – Change types of SONADUS [Gantner, 2011].

SONADUS typology distinguishes between more changes than SAPO, for the author, these differences in typologies are not an obstacle to the application of SAPO to the Swiss context. A common typology of changes may be created. The main difference is that SONADUS uses upper ontologies then the author argues that: "*the universality of upper ontologies facilitates the integration of different data sets, an issue that SAPO does not address.*"

(2) In a second time, all the OWL classes (created after step (1)) were attached to a concept of the upper ontology BFO (see Section 3.2.2.5) in order to account for the dynamic processes of evolution of the AUs using both SNAP and SPAN view on the AUs of Switzerland. "*AUs as being modeled in the HMR+ both have characteristics of continuant and occurrent entities.*" [Gantner, 2011]. In the SONADUS ontology, the AUs are then represented both as *whole objects* (WOs) with an **identity** and as *partial objects* that are different versions of a *whole object*. Thus, WOs are AUs that may undergo incremental changes while preserving their identity. Regarding changes, SONADUS differentiates between two types: *fundamental* changes (the AU ceases to exist); and *incremental* changes (the AU WO still exists after change).

(3) The third step consists of the creation of the RDF data set based on the SONADUS ontology. It consists of the transformation of data (AUs descriptions and their changes descriptions) into RDF/XML. In order to construct an RDF graph, SQL queries were run on the relational database containing the HMR data set. The HMR data set lists all the Swiss municipalities and describes their changes over time. At this step: "*AUs were endowed with a unique identity. Whereas an*

AU maintains its identity in the case of incremental change, fundamental change causes either the creation or the loss of an identity."

To conclude on the SONADUS approach, we want to emphasize here the quality of the design of this ontology, which takes into account both the SNAP and SPAN representations of the AUs, connects concepts dependent from the Swiss context to more abstract concepts of the BFO framework, and transfers a relational database to a representation in triplets, using a data-model that suits with the triple representation.

Since the work from [Gantner, 2011], the *Governmental Ontology Switzerland* (curated by Zazuko.com) has been created, and is available at `https://gont.ch/` [5]. This ontology is written in English, and aims at describing the HMR data set, as the SONADUS ontology. The data sets created using this ontology are available from `http://classifications.data.admin.ch/datasets/`. The different versions of the municipalities, cantons, and districts of Switzerland are represented. For instance, the predicate `districtVersion` is used to say that a `DistrictEntity` has a district version `DistrictEntityVersion` (*e.g.*, representation of a district and of its version under the number 10107 `http://classifications.data.admin.ch/district/1001`).

The concept `ChangeEvent` is defined as a subclass of the `Event` class defined in the *The Simple Event Model Ontology* [6]. Then, several `ChangeEvent` sub-concepts are defined: *InitialCompilation, NameChange, NewDistrictOrCanton, TerritorialChange, Renumbering, Dissolution, AbrogationOfMutation.*

One of the issues behind these graphs is that numbers are used in the graphs to refer to the type of change that has occurred on a canton, district, or municipality. This makes it difficult for humans to read graphs, because we have to refer to a correspondence table (number/change type). This suggests that the design of this ontology is not really based on a triplet approach but on a direct translation of a relational model into a triplet one. Another problem is that the concepts are, as in the case of SONADUS, very dependent from the Swiss administrative context, and the changes are not represented in a multilevel approach (*e.g.*, the impact of the changes of a canton on its sub-districts is not represented).

4.2 Modeling evolving TSNs and their changes

Various initiatives to disseminate Linked Statistical Data have emerged throughout the world, for instance: the *Aragón Statistical Office Open Data Portal* provides LOD statistics on municipalities of the Aragon region of Spain [7]; the Italian *Istat Linked Open Data Portal* [8] and the *e-Stat Japanese Portal* [9] disseminate statistical

5. The turtle of the ontology is on github `https://github.com/gontch/gont.ch/blob/master/gont.ttl`
6. URI of the ontology: `http://semanticweb.cs.vu.nl/2009/11/sem/`
7. `http://opendata.aragon.es/`
8. `http://datiopen.istat.it/index.php?language=eng`
9. `http://data.e-stat.go.jp/lodw/en`

LOD for the National SAs of Italia and Japon; the *European Union (EU) Open Data Portal*[10] gives access to (L)OD published by EU institutions such as Eurostat that provides official statistics on the European Union. The W3C *RDF Data Cube* ontology (QB)[11] is widely used to describe these LOD statistics in a way that is compatible with the *Statistical Data and Metadata eXchange* (SDMX), an ISO standard for exchanging and sharing statistical data and metadata among organization [Cyganiak and Reynolds, 2014]. Using the QB vocabulary, one can publish statistical observations and a set of dimensions that define what the observation applies to: time, gender and geographic areas, for instance. The *RDF Data Cube extensions for spatiotemporal components* (QB4ST)[12] [Atkinson, 2017] provides canonical terms to defined in a QB observation the space (`qb4st:SpatialDimension`) and time dimensions of the data. Also, the two concepts `qb4st:RefArea` (subclass of `qb4st:SpatialDimension`) and `qb4st:subdivides` provides a means to express nested features (for example countries containing administrative units) [Atkinson, 2017] (see Listing 4.2)

```
eg:country a qb4st:RefArea ;
  qb:codeList eg:Countries .
eg:admin1 a qb4st:RefArea ;
  qb:dimension qb4st:refArea ;
  qb:codeList eg:Admin1 ;
  qb4st:subdivides eg:country .
```

Listing Code 4.2 – Nested Spatial Reference Features using QB4ST [Atkinson, 2017]

The SKOS extension for representing statistical classifications (XKOS)[13] [Cotton et al., 2013] allows also for the representation of hierarchies of geographic levels (an ordered list of `xkos:ClassificationLevel`) but more broadly, this extension aims at representing any statistical classifications (*e.g.*, the Statistical classification of economic activities in the European Community (NACE)). In the online documentation of the ontology, it is stated that the precise definition of what constitutes a version of a classification is out of scope for this model, but validity and version information can be represented in simple ways: "*The succession in time of classifications and classification schemes is expressed by the xkos:follows property [...]*". The XKOS model focuses on the correspondences between the items of two classifications which is a particular case of correspondence. In their case, correspondences means more broadly, correspondences between two different classifications of activities in North America or Europe, for instance. The QB4OLAP vocabulary[14], for its part, allows for the representation of multi-dimensional data, including (for the spatial dimension) the description of hierarchy of geographic levels (using the concepts `qb4olap:Hierarchy`, `qb4olap:hasLevel`, `qb4olap:LevelProperty`), with OLAP functions to aggregate data [Etcheverry and Vaisman, 2012]. However, regarding the time dimension of data, it is a discrete approach that does not represent processes of evolution of elements over time, and the underlying changes and events.

10. `http://data.europa.eu/euodp/home`
11. `http://purl.org/linked-data/cube#`
12. `http://www.w3.org/ns/qb4st/`
13. `http://rdf-vocabulary.ddialliance.org/xkos#`
14. `http://purl.org/qb4olap/cubes`

With regard to the evolving nature of the geographic areas the observations cover and the TSN these evolving areas belong to, none of the ontologies described above provide the vocabulary required to achieve the description of the way these TSNs, their levels, and TUs evolved over time. The (N)SAs or National Mapping Agencies, for their part, often create their own ontology for the description of their areas (*e.g.*, the *Territorio Ontology*[15] of the *Italian National Institute for Statistics*, the *Geography Ontology*[16] of the Scottish Government, the Ordnance Survey Ireland's Ontologies[17] [Debruyne et al., 2017]), which results in a counterproductive proliferation of non-aligned vocabularies. Among these initiatives, we highlight here some of them. Then, we present the generic approach of [Plumejeaud et al., 2011; Plumejeaud, 2011] that addresses the TSNs specificities, and describes their evolution and changes over time.

4.2.1 Ontologies for TSN representation

4.2.1.1 INSEE - INSEE Ontologie geographique

The INSEE French National Institute of Statistics and Economic Studies publishes the *INSEE Geographic Ontology*[18]. However, this ontology is written in French and the concepts are strongly dependent from the French context.

4.2.1.2 Eurostat - Reference And Management Of Nomenclatures Ontology

The Eurostat Directorate-General of the European Commission publishes the *Reference And Management Of Nomenclatures Ontology (RAMON)*[19]. This ontology allows for the description of Local Administrative Units[20] and NUTS TSN. However, if one wants to describe TU coming from another TSN than NUTS and LAU, the ontology needs to be modified by adding new sub-concepts, subsumed by the "Geographical Region" concept.

15. http://datiopen.istat.it/odi/ontologia/territorio/
16. http://statistics.gov.scot/vocabularies/
17. Ireland's Administrative boundaries ontology:http://ontologies.geohive.ie/osi#; Ontology to describe the evolution of Ireland's boundaries: http://ontologies.geohive.ie/osiprov#
18. http://rdf.insee.fr/def/geo/insee-geo-onto.ttl
19. http://ec.europa.eu/eurostat/ramon/ontologies/geographic.rdf
20. http://ec.europa.eu/eurostat/web/nuts/local-administrative-units

4.2.1.3 Ordnance Survey - Administrative geography and civil voting area Ontology for UK

The Ordnance Survey national mapping agency in Great Britain publishes the *Administrative geography and civil voting area ontology for UK*[21]. As stated in [Correndo et al., 2010], the hierarchical nature of NUTS can be described with the Ordnance Survey ontology, but the temporal extent of a given geographical subdivision cannot. Furthermore, as for INSEE the concepts are strongly dependent from the UK context.

4.2.1.4 ONS - ONS Boundary Change Ontology

The UK Office for National Statistics proposes vocabularies[22] to represent the evolution of geographic areas in the context of statistical data publication. Even if we can notice a certain level of abstraction on the terms used to describe TUs[23], which makes it possible to describe other TUs than UK's ones, none of these vocabularies allows to describe any TSN structure and its levels. New concepts need to be added to the ontology *Geographical hierarchy ontology*[24] in order to describe new territories and new territorial meshes than those listed. Furthermore, none of the concepts of the *ONS Boundary Change Ontology*[25] provides naming for changes that impact several TUs at the same time. Finally, the concepts describing change events are few and limited to *Recoding change, Boundary change*.

4.2.1.5 Statistics Bureau of Japan - SAC Ontology

The Statistics Bureau of Japan proposes vocabularies[26] to represent geographic areas and their evolution, including the description of change events in the context of statistical data publication. In [Yamamoto et al., 2017], the authors propose a solution to describe changes, with the reason for the change. Nevertheless, they limit to description of changes in municipalities, using the following change event concepts: *absorption, abolishment, separation, establishment, division, name change, boundary change*, and *shiftToAnotherKindOfCity*. Also, the description of the reason for the change is limited to one literal.

21. http://data.ordnancesurvey.co.uk/ontology/admingeo/
22. http://statistics.data.gov.uk/vocabularies
23. http://statistics.data.gov.uk/def/statistical-entity
24. http://statistics.data.gov.uk/def/hierarchy/best-fit
25. http://statistics.data.gov.uk/def/boundary-change
26. e.g., http://data.e-stat.go.jp/lod/sac/, http://data.e-stat.go.jp/lod/terms/sacs#, http://data.e-stat.go.jp/lod/sace/

4.2.1.6 Food and Agriculture Organization of the United Nations - FAO Geopolitical Ontology

The Food and Agriculture Organization of the United Nations (FAO) publishes the *FAO Geopolitical Ontology*[27]. FAO produces also a data set composed of a list of countries in the world. This data set is semantically annotated thanks to geopolitical concepts of the ontology (*e.g.*, *self-governing*, *non-self-governing*, *disputed*). The *FAO Geopolitical Ontology* identifies successive versions of a TU (*i.e.*, predicates *predecessor*, *successor*, *valid since*, *valid until*). However, the changes undergone by TUs are not described (*e.g.*, spatial area changes).

4.2.2 TSN data sets as Linked Data

When publishing a TSN on the Web of data, we have to distinguish between the ontology used to describe the structure of a TSN and the data sets described via this ontology. With regard to the TSN data sets published on the LOD Web, we focus in this manuscript on the NUTS TSN data sets:

4.2.2.1 NUTS-RDF Geovocab

NUTS-RDF Geovocab[28] are published as part of the *Planet Data EU Network of Excellence*[29]. Nevertheless, only one version of the NUTS is published online.

4.2.2.2 Eionet - NUTS Data set

Eionet publishes NUTS versions in RDF format[30] and uses the Eionet RAMON Ontology for the description of NUTS. There are two types of links (*owl:sameAs*, *rdfs:seeAlso*) to connect TU from the NUTS version 2006 to TU from the 2003 version. Nevertheless, changes undergone by TU from one version to another are not described.

4.2.2.3 EnAKTing - Linked NUTS

The EnAKTing project[31] (funded by the Engineering and Physical Sciences Research Council) publishes NUTS versions in RDF format, under the *Linked NUTS*

27. http://www.fao.org/countryprofiles/geoinfo/en/
28. http://nuts.geovocab.org/
29. http://www.planet-data.eu/
30. http://rdfdata.eionet.europa.eu/ramon/nuts.rdfhttp://rdfdata.eionet.europa.eu/ramon/nuts2008.rdf http://rdfdata.eionet.europa.eu/ramon/nuts2003.rdf
31. http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/G008493/1

name [Correndo and Shadbolt, 2013]. The *Linked NUTS* represents the modification of NUTS TU over time. However, the description of the changes undergone by TU from one NUTS version to another is brief and restricted to code change, region splitting, and regions merging, while, as will be seen later, other changes that affect the TSN components should be taken into account to avoid misinterpretation of statistical data.

The Subsections 4.2.1 and 4.2.2 have shown that, despite the multiple initiatives to represent TSNs on the LOD Web, none of them is abstract enough to enable the description of any TSN (with a structure such as the one we address in this thesis, defined in Chapter 2, Section 2.2) and of its changes over time, on the LOD Web. The next Subsection presents an approach that is generic. Our work is a continuation of these works from [Plumejeaud, 2011], achieved in our STeamer research group. We propose an immersion of the models and algorithm in the semantic Web since they were originally designed specifically for relational databases, and not for triplets' representation and graphs.

4.2.3 A spatiotemporal model for TSN

The generic approach of [Plumejeaud et al., 2011; Plumejeaud, 2011] addresses TSNs specificities, and describes their evolutions and changes over time. In previous work of our research group [Plumejeaud et al., 2011], the model is presented as follows: *"A model based on the identity of geographic units which enables one to study the genealogy of units, and to observe a unit inside different kinds of territorial organizations, possibly changing over time."*

The model is introduced in this manuscript as in the paper [Plumejeaud et al., 2011], in two steps: first we present how the identity concept of TUs inside territorial hierarchies is modeled; second, we present how is modeled the evolution of the TUs over time, from the concept of identity to the concept of genealogy.

4.2.3.1 Identity and Hierarchy

The first part of the model, presented in Figure 4.9 enables the description of a TSN hierarchical organization: the concept `Nomenclature` represents the TSN used by statistical agencies, composed of `Geographic Units`.

The `Zoning` object refers to a set of Geographic Units which have the same scale, that is to say the same level, inside a given `Nomenclature`. Each Geographic Unit is linked to a nomenclature by an association class `IdentityAttributes` that holds the list of attributes that usually describe a Geographic Unit in a nomenclature:
– *Designation: refers to the official name of the unit in a given language.*
– *Code: specifies the code of the unit inside the Nomenclature.*
– *Center: represents the position of the center of the unit (a point or another GeographicUnit included inside the considered unit). The center can also have an official*

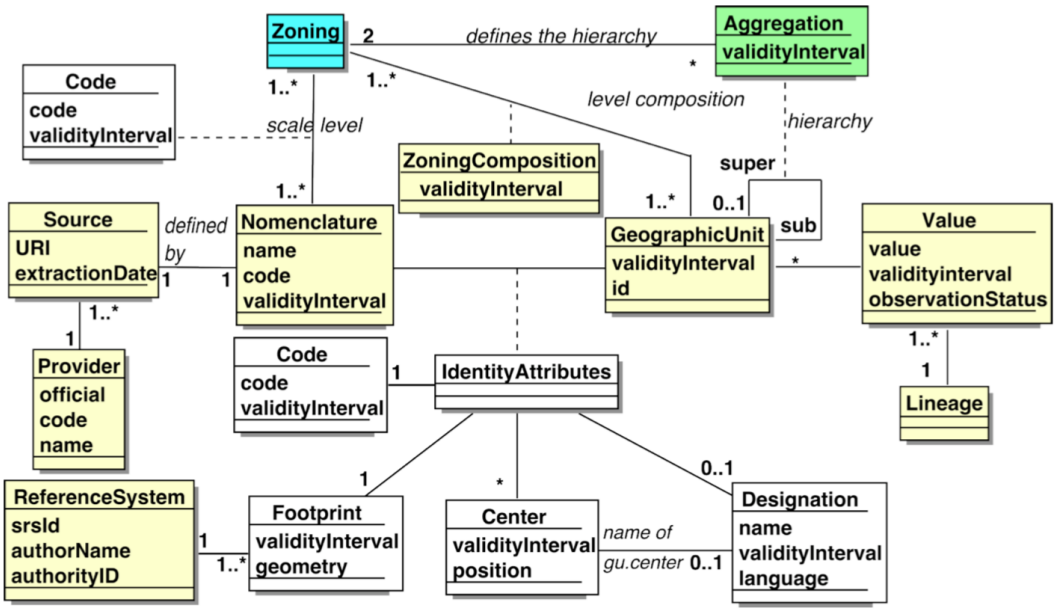


Figure 4.9 – A model for identifying geographic units into various nomenclatures, from [Plumejeaud et al., 2011].

Designation in various languages. Inside a Nomenclature, a unit can also have no Center or many Centers.
– Footprint: corresponds to the spatial representation of the unit. It is typically stored as a polygonal geometry, and its associated reference system, ReferenceSystem, is the one which is specified for the whole Nomenclature version.

It has to be noticed that each of the main components of the model has a `valididyInterval` attribute (each of the Identity attributes too) that corresponds to the life span of the element. Thus, each component of the TSN and each Identity attribute of a TU can change independently from each other.

The next paragraph presents how the changes of an element of the TSN, or of an attribute of the TU (that may lead to the change of identity of the TU) are modeled in the approach of [Plumejeaud et al., 2011].

4.2.3.2 From identity to genealogy

Based on the typology of [Claramunt and Thériault, 1995], Plumejeaud et al. [2011] construct a typology of *territorial events* (see an extract of this typology in Figure 4.10) that names the different types of change process and events the TUs of a TSN may undergone.

This typology has different levels of details in the event descriptions (please see Figure 4.10, *Level 1, Level 2, Life events*):

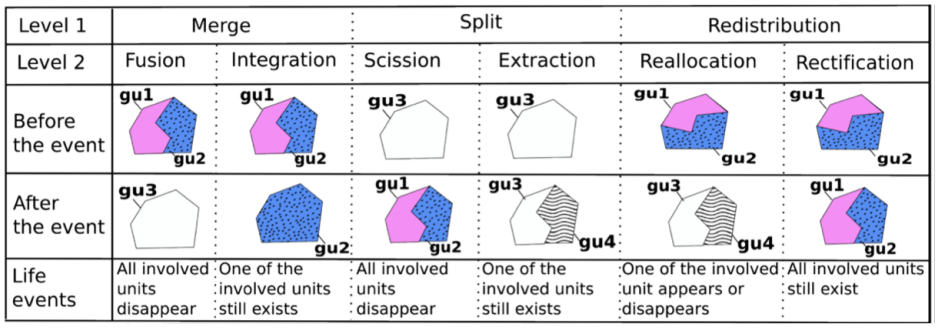


Figure 4.10 – Typology of territorial events in TSN, from [Plumejeaud et al., 2011].

(1) at Level 1, it distinguishes between three Territorial Events that are: *Merge, Split*, and *Redistribution*. This typology is similar to the typology III (*Evolution of spatial structures involving several entities*) of [Claramunt and Thériault, 1995], in Figure 3.7. For any of these three event types, *the combination of unit footprints that existed prior the event is equal to the combination of unit footprints that are still in existence after the event* [Plumejeaud et al., 2011];

(2) at Level 2, there is a refinement of the three *Territorial Event* tags (Merge, Split, and Redistribution) in order to describe whether the involved TUs change their identity after the event or not;

(3) the TUs may undergone also what is called *Life Event* (*i.e.*, *Appearance*, *Transformation*, *Disappearance*), caused by a *Territorial Event* (as shown on the Figure 4.11). This typology is similar to the typology I (*Evolution of a single entity*) of [Claramunt and Thériault, 1995], see Figure 3.7.

Thus, in an approach similar to [Beller, 1991; Claramunt and Thériault, 1995], Plumejeaud et al. [2011] do not limit territorial event to a single level of processes. With the difference that [Plumejeaud et al., 2011] chose to specify the semantic links between the processes of type I and the processes of type III, and use a semantics of causality (cause/consequence).

The Figure 4.11 below shows in UML the whole typology of events of [Plumejeaud et al., 2011].

In this model, the consequences of each territorial event, on each geographic unit that is to say, all the life events that affect a geographic unit over time are stored. And, as explained above, the *cause/consequence* semantic is used to link a life event to a territorial event.

For the modeling of sequences of events, Plumejeaud et al. [2011] introduce the concept `GenealogyEvent` composed of several territorial events, and which has a non-zero duration. In [Plumejeaud, 2011], the author takes the example of the restructuring of the hospital districts that can proceed in several places and times.

This model is adapted to the representation of TUs' genealogy and provides

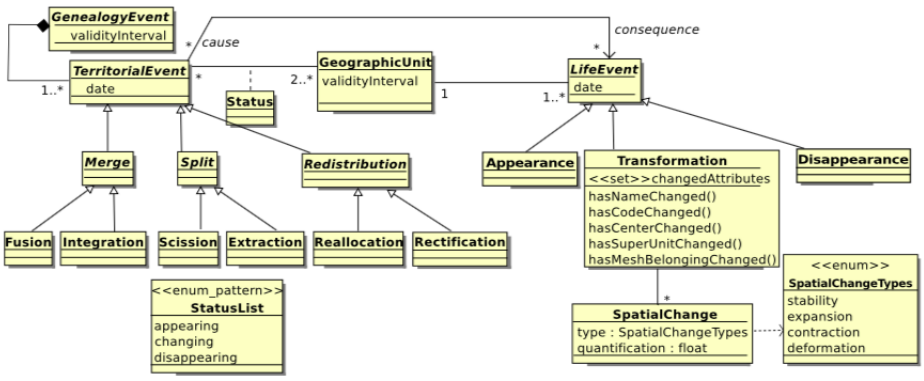


Figure 4.11 – The geographical units of a TSN represented in relation to the territorial events they are involved in and that lead to their change over time (LifeEvent), from [Plumejeaud et al., 2011].

users with a view on the whole chain of descents (ancestors and descendants). It makes a distinction between territorial events that transform the TUs and events that lead to the disappearance of the TUs. This is determined by the definition of the identity of the TUs, that depends upon the TSN specificities. Then, this identity definition varies from one TSN to another and has to be set at the implementation time, when generating (detecting and describing) the changes in the versions of one TSN. We come back to this issue when we present the algorithm of [Plumejeaud, 2011], in the following Chapter 5, and in the Contribution Part, Chapter 9.

Chapter 5

Generating and Exploiting descriptions of evolving TSN

In this Chapter, we evaluate the existing methods for automatically generating then exploiting, in the semantic Web, descriptions of evolving geographic areas (descriptions of their life, filiation and changes over time).

5.1 Generating

We focus in this Section on the process of generating information on the dynamics of the world (mainly generating similarities and changes description of geospatial objects that evolved over time). Generating LOD descriptions of geospatial objects at one time period is a trivial task if one has, for instance, a shapefile containing the geospatial features. Indeed, tools such as *GeoTriples*[1] exist, taking as input a shapefile and generating the RDF representation of the features inside the input shapefile. In the Chapter 9, we explain how to do this, using the GeoTriples program and a mapping file that associate the shapefile columns with ontological concepts.

We have seen in Chapter 3, Section 3.2 that in order to populate their ontological model, [Lopez-Pellicer et al., 2012], [Kauppinen et al., 2008], and [Gantner, 2011] use a written dictionary of changes, others such as [Plumejeaud, 2011], and [Harbelot et al., 2015] develop computer programs in order to detect the changes. In this Section, we focus on the second approach that automate the change descriptions in geospatial data. In the literature, the approaches consist mainly in computing similarity scores between each pair of geospatial objects (*e.g.*, a pair composed of a TU that belongs to a TSN version V' and of a TU that belongs to the next version V'' of the same TSN). However, the approaches differ in the similarity threshold calculation: sometimes it is based only on the spatial footprint (by calculating the intersection of geometries in the case of polygons), or, sometimes, it takes into account other information (such as the name of the TU, its capital, etc.). A further difference is that some approaches describe changes that affect only isolated features, while others try to detect and put a semantic label on changes that simultaneously affect several features (*e.g.*, merge of two TUs). In the following Subsection, we present conflation approaches that calculate similarities between data in order to combine information from two or more related data sets.

1. https://github.com/LinkedEOData/GeoTriples/wiki

5.1.1 Conflation Algorithms

As noted in [Li and Goodchild, 2011], many GIS require data from more than one source. However, in this case the establishment of a coherent GIS can be difficult since data may differ greatly: geographic data may initially be created by various actors, for a variety of purposes, in different formats, or at different scales...

Depending on the final objective of mobilizing various data sources, the processes to be apply on data that enter the GIS are different. Let us assume that the different data sources describe the same geographic objects:
– if there is a need to preserve all the information from these different data sources: then, there is a need for tools that detect and describe the differences and similarities between the data sources;
– if there is a need to create, using different data sources, one data set that merge data from different sources (because one data set may contain information the others don't have and conversely): then, in this case, there is a need for tools that resolve conflict between duplicated data for instance, and remove the duplicates.

Yet, both cases face first the same problem of **feature matching as in conflation approaches**: in the second case, only one of the features that match will be kept; in the first case, both of them will be kept, and similarities measure will be translated into semantic information for instance.

Based on the Conflation definition of [Longley et al., 2005], Shekhar and Xiong [2007] provide the following definition of the term *Conflation of Features*:

"In GIS, conflation is defined as the process of combining geographic information from overlapping sources so as to retain accurate data, minimize redundancy, and reconcile data conflicts".

One of the conflation steps consists in *feature matching*: "*Feature matching involves the identification of features in multiple data sets that represent the same entity in reality*" [Li and Goodchild, 2011]. The main problem of feature matching is "*due to different data quality, different scales, different schemata and different purposes, the same entity may be represented differently in terms of position, shape and level of detail.*" [Li and Goodchild, 2011].

"*There are generally two major steps in feature matching: First, we choose a similarity measurement to be used as a criterion for matching; second, we identify all matched pairs of features using this selected similarity criterion [...] similarity measures commonly used in feature matching can be classified into three types according to whether they are based on similarities of geometry, attribute or topology, or combinations of these.*" [Li and Goodchild, 2011]. Conflation algorithms, like those proposed by [Hastings, 2008] and [McKenzie et al., 2014], but also [Ruiz et al., 2011] take in consideration the spatial dimension. [Hastings, 2008] takes a multi-attribute approach to conflating digital gazetteers. [McKenzie et al., 2014] use also the multi-attribute approach from [Hastings, 2008] to perform Points of Interest conflation, namely identifying whether two information entities refer to the same place in the

physical world. They demonstrated a weighted multi-attribute matching strategy that can successfully match 97% of randomly selected POI from two different data sets. Metric results are then used for conflict resolution for gazetteers or POI (*e.g.*, choose the canonical form of a place name between multiple possibilities).

[Plumejeaud, 2011] also uses similarity measures on a list of TU's attributes, identified as relevant for comparison between two versions of a TSN. She designs an Algorithm presented in the following Subsection.

5.1.2 Algorithm for the automatic matching of two TSN versions

Please note that this algorithm and its main functions are more detailed in the Chapter 9 as our work enriches this algorithm to output semantic RDF descriptions of the territorial changes in TSNs. In this Section, we provide the reader with the main ideas of how the algorithm works.

The algorithm looks for TUs that are equal between two versions of a TSN. Knowing that two units are considered equal if a combination of criteria (relating to the attributes of the TUs) are satisfied. These criteria consist of similarity measures on a list of TU's attributes, identified as relevant (by an expert of the TSN) for comparison between two TSN's versions. Some similarity tests are more complex than others.

For instance, the distance test on the geometry of the TUs, implies first that the geometries are in the same coordinate system or, if not, it implies converting geometries into the same representation system. Second, the geometries may not have the same level of detail, *i.e.*, they are not exactly the same although the unit has not changed between the two versions. There are several solutions to compute a similarity score between two geometries, depending on the nature of the geometries *i.e.*, point, line or polygon. The score corresponds most of the time to a computed distance, usually determined by a Euclidean distance for the point geometries, or Hausdorff distance for the poly-lines geometries. In addition to Euclidean or Hausdorff distances, other distances are used in order to compute similarity scores. [Devogele, 2002] uses the maximal distance between two lines, called the discrete Frechet distance, that considers the locations and the ordering of the constituent points of poly-lines when calculating proximity. [Bel Adj Ali and Vauglin, 1999] propose an areal distance test for polygonal geometries. It computes for two geometries that intersect, the ratio of the shared area by the union of their areas. In [Plumejeaud, 2011], the proposed solution is to use the areal distance of [Bel Adj Ali and Vauglin, 1999] for the similarity test between two TUs' geometries, in two different versions of a TSN. If the ratio obtained after the [Bel Adj Ali and Vauglin, 1999] test, multiplied by 100 is smaller than a small value, denoted epsilon ϵ, it means that the spatial footprint has changed very little. It is therefore considered unchanged. This surface distance test has the advantage of being customizable, by varying the ϵ value. All the results of the areal distance test between each pair of TUs are registered into a SpatialMatch matrix.

In a second step, in case of a geometry change identified by the areal distance test,

the algorithm tries to determine precisely if only one TU's geometry has changed or more neighboring TUs with it. For instance, if a TU $u'1$ is split in two TUs ($u''2$ and $u''3$), the SpatialMatch matrix will have registered two intersections for the TU $u'1$ and will conclude to a territorial change of type *Split* (see the [Plumejeaud et al., 2011] typology of changes in Figure 4.11). An equality test is performed on the external geometries of the two sets of TUs before and after change (*e.g.*, equality test between the external geometry of $u'1$ and external geometry of the union of $u''2$ and $u''3$) in order to validate the aggregation of several TUs changes under one territorial change tag (*e.g.*, split, merge, redistribution), with a tolerance threshold.

For each pair of TUs that intersects, other distance tests, on the others TUs attributes are performed and summed at the end, in order to determine if the TUs of the two versions match or not. Please note that all these tests are described in more details in the Chapter 9.

The algorithm always proceeds from top to bottom, it first matches higher-level units before attempting to match the units at the next lower level. This algorithm has been implemented and tested on the NUTS TSN versions. The program retrieves all the territorial and life event changes in the NUTS TSN, in half an hour or less. The author concludes that the program is very beneficial to agencies such as Eurostat, as the manual listing of changes is time consuming for them.

The areal distance test proposed in [Plumejeaud, 2011] needs to test first the intersection between two geometries then to compute a percentage of intersection. This percentage is computed using the spatial functions of the PostGIS DBMS. In [Harbelot et al., 2015], the authors adopt the same methodology. They first compute a similarity score between the parcels of the CLC data set (in different versions), then depending on the score values, they attach semantic tags that describe filiation links between the parcels, or changes.

5.1.3 Methodology for constructing filiation links in CLC data sets

As noticed in [Harbelot et al., 2015], the Corine Land Cover data set does not provide any knowledge about filiation relationships and requires a new methodology in order to identify the relationships between the different parcels versions.

Then, the authors propose the following solution to semi-automate such description of the filiation links in the CLC data sets, using the Continuum models:
– (1) find filiation and link entities using `hasFiliation` Continuum predicates. It consists in finding all pairs of entities having an overlap (*i.e.*, intersection test between two geometries). At this step, an overlapping score is computed between each pair of features and stored. However, some of these intersections represent a filiation relationship while others should be considered as noise due to inaccuracies or negligible changes. A threshold is defined in order to filter spatial filiation and accept a correct filiation relationships only when the parent and child overlap exceeds a certain threshold.
– (2) find continuation links that may be of three different kinds referring to growth, reduction or equality of the geometries. In order to distinguish each case, the pro-

posed method is to compare the child and parent overlapping value (using the overlapping score obtained after step (1)).
– (3) find complex patterns of changes that may correspond to a division or a merge. This is done in two steps: First, one focuses only on the spatial component of time-slices by searching division or merging pattern. Second, the identity component is used in order to distinguish a merge from an annexation or a splitting from a separation. The list of these change patterns (annexation, split, ...) detected by the system are shown by Figure 5.1. – (4) find how deep changes are and discover

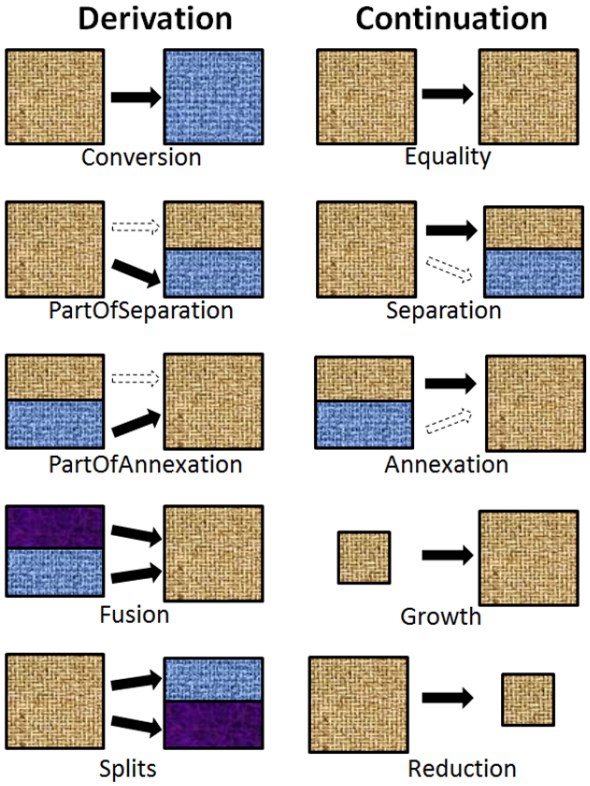


Figure 5.1 – List of patterns detected in the LC3 Model [Harbelot et al., 2015].

knowledge about changes (the causes) using specific patterns that depend on the data. The CLC data set is based on a hierarchy of classes of land-parcels. Using the depth of the land cover hierarchy, the depth of change is analyzed. For instance, the land-parcel that changes from Forest to artificial surface has undergone a strong derivation which may be semantically tag as a `deforestation`.

As explain in the Subsection 4.1.1, this approach suits well to CLC data sets, but it is not appropriated to administrative divisions, for instance, because it only takes into account the geometry of the parcel and the nature of the land cover when detecting changes. It does not take into account other attributes that may constitute the identity of the geospatial entities, such as the identifier or name of the entity. This approach lacks flexibility when it comes to the definition of identity of the time-slices, if it was to describe changes in geographic divisions different from CLC.
The approach of [Plumejeaud, 2011] is more flexible (the definition of what makes the identity of the TUs (a list of TUs' attributes) and in which proportion those attributes can vary are configurable). However, it is not immersed in the Semantic Web, contrary to the L3C method from Harbelot et al. [2015].

Here, we have reviewed the solution for conflation of geospatial data sets and the different measures of similarities between the geometries of geospatial data sets. Version Control Systems (VCS) propose another approach that intends to solve the problem of data conflict that occurs when several persons simultaneously edit the same source code text, for instance. As in the conflation approach, the VCS tools need first to identify the changes between two texts, then to resolve the conflict between different versions.

5.1.4 Version Control System

We focus in this Subsection on the existing Version Control System Tools for the versioning of Geospatial Data. With regard to software tools specialized in version control, semantic descriptions of changes in ontologies emerge [Redmond et al., 2008], [Völkel and Groza, 2012] in order to ease collaborative ontology editing and resolve conflicts. These ontology versioning systems describe the set of edit operations made on ontology versions. For instance, the SemVersion, plug-in to Protégé software "*provides structural and semantic diffs visualization of the added and removed statements*" [Völkel and Groza, 2012]. And, they plan for "*future developments, to create an intuitive format, for example, by displaying the two ontologies in parallel and create graphic connections to indicate the added and removed statements.*"

Conflict resolution in VCS, is most of the time a problem of collaborative construction of data set or software source code. With regards to Geospatial Data, for instance the software Ethermap [Fechner et al., 2015] is a real-time synchronized collaborative map editor, with multiple concurrent users. It allows users to review changes made to the map, using an "History" tool that tracks the edit (*e.g.*, descriptions such as "edited properties", "edited geometry", "created feature") or remove actions, and tracks which user is responsible for those actions on the map.

Other VCS tools, such as *GeoGIG*[2] or the *GitHub Inc. project*[3], focus on geospatial data with the very purpose of conflict detection, conflict resolution and merge

2. http://geogig.org/
3. https://blog.github.com/2014-02-05-diffable-more-customizable-maps/

in data when multiple contributors edit the data. Hereafter is the description of the GeoGIG tool on its official Website: "*Users are able to import raw geospatial data (currently from Shapefiles, PostGIS or SpatiaLite) into a repository where every change to the data is tracked. These changes can be viewed in a history, reverted to older versions, branched in to sandboxed areas, merged back in, and pushed to remote repositories.*" GeoGIG or Google limit their comparison to attributes independently.

Although tools such as Ethermap, *GeoGIG*[4] or the *GitHub Inc. project*[5] [Negretti, 2015; LocationTech, 2018; Boundless, 2014a; **?**,b] handle versioning of geospatial data and provide a way to visualize differences in geometries, feature by feature, they are unable to group changes that affect several geographic objects at the same time (*e.g.*, a merge of two TUs), or to attach semantics describing the context and nature of these territorial changes. Indeed, the very purpose of these tools is conflict detection and resolution, as explained before. Such software do not focus on the semantic of changes that help in understanding the reason for the change.

Several works try also to reconcile data conflicts and focus on the measurement of similarities between GeoSpatial Data [Angela, 2008], in order to improve the quality of volunteered geographic data for instance [Vandecasteele and Devillers, 2013], by avoiding duplication of data, among others [Zhang et al., 2007]. In [Vandecasteele and Devillers, 2013], the authors explain that, during the editing process, by automatically notifying contributors when two attributes are too similar or too dissimilar using similarity measures, their system reduces semantic heterogeneity of the edited data. They address the issue of semantic similarities between concepts such as 'road', 'highway' or 'river': "*Semantic similarity quantifies how similar, or dissimilar, concepts are, based on their meanings.*" Several methods for measuring semantic similarities exists such as the Network model where concepts are in a semantic networks. "*In such networks, concepts are represented as nodes and relations between concepts are represented by edges. The similarity between two concepts is measured using a distance function. This distance function is based on graph theory algorithms, such as the shortest path model.*" [Vandecasteele and Devillers, 2013]. These approaches based on similarity measures are another way of resolving conflicts in order to avoid unnecessary changes (to avoid the addition of similar data).

Other works propose solution (using relational database system) to keep track of multi-scale data evolution. Multi-scale means multi-representation of a geospatial object, at different levels of detail: "*in different scales the objects are usually represented in different ways, because each scale can have a convention of representation*". However, we haven't find software that keep track of multi-levels (*e.g.*, States, regions, districts) data evolution.

As a conclusion, none of the works presented in this section addresses the semantic description of changes (changes in feature attributes and more complex changes such as merge of two polygons) in multi-levels geospatial data. Nevertheless, they all investigate interesting options to deal with geospatial evolution over time, in real-time context different from the one of official TSNs data published every 3 (NUTS)

4. `http://geogig.org/`
5. `https://blog.github.com/2014-02-05-diffable-more-customizable-maps/`

or ten years (U.S. Census tract) for instance.

We have investigated in this section the algorithms and tools for conflation and description of changes in geospatial datatasets. In the next Section, we investigate the existing tools that exploit description of changes in geospatial data.

5.2 Exploiting

After generating and publishing LOD data on a triple store (using software tools such as Strabon, Stardog, GraphDB), the SPARQL language allows to query and explore data. Also, for the Geospatial data, the GeoSPARQL extension to SPARQL may be used to query data and discover their spatial relations according to the RCC-8 typology, for instance. Some visualization tools to explore the LOD on the Web offer view on nodes and edges of an RDF graph. For instance, the GraphDB triple store has a visualization module that draws the RDF graph, result of a `Construct` SPARQL query. This tool helps checking the content returned after a SPARQL query. Another more specific tool called Sextant `http://sextant.di.uoa.gr/` allows visualizing time-evolving linked geospatial data that are described with the stSPARQL ontological concepts (*i.e.*, not with the standard space and time ontologies *OWL Time* and *GeoSPARQL*, which limits its usability). The research project GeoPeuple `http://geopeuple.ign.fr/` provides a demo available at `http://www.rotefabrik.free.fr/geopeuple/en/onglets-33038.html` to visualize the evolution of administrative regions in France over time.

Exploration of isolated data restricts users to a closed world. Thus, more than providing tools to visualized data, one of the most important steps to benefit from the distributed Web database is to integrate first the published LOD with other data on the LOD Web. This adds more values for the end-users, because combining data may reveal relationships between data and provide context for their interpretation [Fahd and Yousuf, 2017].

The final goal of our work is to construct and publish on the LOD Web, RDF graphs that describe the evolution and changes of TUs over time. In order to find automatically the (historical) reason for the changes over time, we have to identify and link TSN data with other Link Open data sets. This information may provide context to territorial change descriptions, and help the users to understand the events at the origin of a territorial reform. Several tools help in interlinking data, such as SILK, LIMES, KnoFuss, RDF-AI, SERIMI, OKKAM [Fahd and Yousuf, 2017], but how to make links is not really the topic of the present study. Our issue is more: what to connect to, which are the existing source of information (encyclopedia, media, etc.) on the LOD Web, where and how the historical or societal information explaining the cause of change are available on the LOD Web (if they are)? Also, in the context of geographic divisions for statistics, a source of information to which we seek to relate is of course geo-coded statistics (Linked Open statistical data sets), which make it possible to characterize a territory, and its TUs in terms of number of inhabitants, for instance.

We describe below several sources of information to link with TSN descriptions as LOD: the first Subsection 5.2.1 presents statistical standards to publish geo-coded statistics as LOD. This Subsection presents also the limits of these standards with regard to the representation of the spatial dimension of statistical data (*i.e.*, TSNs); the second Subsection 5.2.2 presents some LOD sets one may connect to in order to provide users with the context where a territorial change happened, that is to say the cause(s) of the change(s).

5.2.1 Linked Open statistical data sets

The RDF Data Cube (QB) is a W3C recommendation. It enables the description and publication of multi-dimensional data, such as statistics on the Web [6]. It is built upon three models: (1) the core of the SDMX model; (2) the Dublin Core Terms for metadata; and (3) FOAF for the description of agents such as the data producers. Using the RDF Data Cube (QB) ontology, one can describe in RDF statistical data sets as a set of observations that consist of dimensions, measures and attributes.

RDF Data Cube "*is based on the popular SDMX standard and designed to represent multidimensional statistical data using RDF*" [Salas et al., 2012]. "*Multidimensionality means that a measured fact is described based on a number of dimensions, e.g. unemployment rate on different countries, years, and age groups. This type of data is compared to a cube [...]*" [Kalampokis et al., 2015].

	2004-2006		2007-2009		2010-2012	
	Male	Female	Male	Female	Male	Female
FR02	76	81	75	80	77	80
FR03	76.5	80	77	81	77.5	80
FR04	75	80	76.5	80	76	81

Figure 5.2 – Excerpt of a statistical data set measuring the Life Expectancy (in years), in France (the codes FR02, FR03, FR04 are the codes for french areas in the NUTS TSN).

Figure 5.2 presents an excerpt of the values of a statistical indicator measuring the Life Expectancy in France [7]. We explain and show below (in Listing 5.1) how to describe and encode this indicator in RDF, using the QB concepts. From lines 2 to 12, the structure of the data set is described, using the main concept `qb:DataStructureDefinition` (at line 2). This DataStructureDefinition is composed of the following elements: – the dimensions of the data set (lines 4 to 6) are declared using the predicate `qb:dimension`. Here, the dimensions are: time, region and sex;
– The `qb:measure` concept defines what the indicator measures: here, the Life Expectancy and Age Average (lines 8 to 9). Other indicators can be declared, if their dimensions are the same (for instance at line 9, a second indicator "Age Average"

6. `http://www.w3.org/TR/vocab-data-cube/`
7. This example is inspired by the one presented within the W3C Web page dedicated to the QB ontology (`http://www.w3.org/TR/vocab-data-cube/`)

is declared;
– The qb:attribute concept is used to add metadata descriptions to the statistical values, *e.g.*, unit of measure, data source, etc. (lines 11 to 12).

After the description of the structure of the data set, the data set itself is described with the concept qb:DataSet (lines 15 to 19). Then, the observations it contains are declared, using the qb:Observation concept (lines 21 to 26). Here for instance, one observation measures the life expectancy of males in FR02 for the period 2004-2006, and the value is 76 years.

```
1 #Define the data set structure
2 eg:dsd-dataset1 a qb:DataStructureDefinition;
3  #The dimensions
4  qb:component [qb:dimension eg:refArea];
5  qb:component [qb:dimension eg:refPeriod];
6  qb:component [qb:dimension sdmx-dimension:sex];
7  #The measures
8  qb:component [qb:measure eg:lifeExpectancy];
9  qb:component [qb:measure eg:ageAverage];
10 #The attributes
11 qb:component [qb:attribute sdmx-attribute:UnitMeasure];
12 qb:component [qb:attribute sdmx-attribute:DataSource];
13
14 #Define the data set and its observations
15 eg:dataset1 a qb:DataSet;
16  rdfs:label "France population"@en;
17  rdfs:comment "Description of the French population (Life Expectancy and
       age average)"@en;
18  qb:structure eg:dsd-dataset1;
19  sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year>.
20
21 eg:observation1 a qb:Observation;
22  qb:dataset eg:dataset1;
23  eg:refArea FR02;
24  eg:refPeriod <http://reference.data.gov.uk/id/year/2004>;
25  sdmx-dimension:sex sdmx-code:sex-M;
26  eg:lifeExpectancy 76;
27 ...
```

Listing Code 5.1 – RDF Data Cube Observation example in turtle [Atkinson, 2017]

In addition, the concept qb:Slice enables to define slices on data because, as noticed by the W3C, it is useful to fix all but one of the dimensions and be able to refer to all observations with these dimension values as a single entity [8]. Then, within the QB ontology, one can fix all the dimensions except the temporal dimension, in order to group values under one time-series. For example, by fixing the dimensions, region and sex, on "male" and "FR02", we extract from a cube of data a time series that shows the evolution of the life expectancy of males for the region with code FR02, from 2004 to 2012. This time series (*i.e.*, qb:Slice) becomes

8. http://www.w3.org/TR/vocab-data-cube/

addressable through an URI. To conclude, the QB ontology is well adapted to the description of statistical data sets and time-series. Several organizations use the QB Ontology (*e.g.*, Eurostat, European Environment Agency, OECD, INSEE, etc.).

The QB4ST extension [9] [Atkinson, 2017] provides canonical terms to defined in a QB observation the space (`qb4st:SpatialDimension`) and time dimensions of the data (as explained in 4.2). However, as explained in 4.2, with regard to the evolving nature of the geographic areas the observations cover, none of the QB, QB4ST, XKOS, and QB4OLAP ontologies provide the vocabulary required to achieve the description of the way the TSNs, their levels, and TUs evolved over time.

To conclude, with regard to the spatial dimension of the statistical QB observations, other ontologies than QB must be used in order to describe the geospatial areas the observations apply to (*i.e.*, the spatial dimension of statistical data). The issue with the existing vocabularies is that they do not represent hierarchies of TUs and/or they do not represent the evolution of the TUs over time. Then, most of the time, QB observations are declared for regions, with boundaries for which it is not said at which period of time they correspond, and if they were the boundaries of the TUs at the time of the data collection, and to which geographic division the TUs belong to. As a result, there is a limited trust in the statistical data published on the LOD because it is difficult to know which geographic object the observed values refer to. A model is still lacking that allows SAs to mention precisely for each statistical observation: what is the observed TU (time and space) and to which geographic division (hierarchy) it belongs. Is also lacking a model that can help to know how this area has evolved over time, and therefore to see how the statistical values measured on this evolving area have evolved too.

5.2.2 Contextual information about territorial changes

Various models have been proposed for representing events on the Semantic Web, such as the Simple Event Model [Hage et al., 2011], the Event Ontology [Raimond et al., 2007], the Linking Open Descriptions of Events Ontology (LODE) [Shaw et al., 2009], the F-Model (F) [Scherp et al., 2009], the Event Ontology used in CultureSampo [Ruotsalo and Hyvönen, 2007] and the CIDOC Conceptual Reference Model Ontology (`www.cidoc-crm.org`), an event centric model, designed for the representation of historic events [Doerr, 2003]. The CIDOC ontology mediates between different sources of cultural heritage information (from museum for instance). For a comparison of these models the reader should refer to [Hage et al., 2011].

The DBPedia Ontology define a class named *Event* [10] but, as noticed by [Hienert and Luciano, 2015], existing events in the DBPedia data set are distributed over only a few categories: military battles and sport events...

9. `http://www.w3.org/ns/qb4st/`
10. `http://dbpedia.org/ontology/Event`

Because the historical events in Wikipedia are listed in the article body, they are not yet included in DBPedia and cannot be queried in a structured way. Then, to automate the contextualization of territorial changes, using the LOD Web, one may encounter some problems using the DBPedia service. The work by [Hienert and Luciano, 2015] offers a solution to this issue. They automate the extraction of historical events from articles in Wikipedia, using language-dependent patterns to identify the text section with events. For instance, to extract events, their system extracts sentences containing a date.

In Europe, the European Legislation Identifier (ELI) [11] provides Web identifiers (URIs) for legal information. For instance, in France the Law No 2015-29 of January 16th, 2015, redefining the regional administrative areas for France, has an ELI URI that is `https://www.legifrance.gouv.fr/eli/loi/2015/1/16/INTX1412841L/jo/texte`. The implementation of ELI is still a work in progress in Spain, as the Spanish legal system is complex and diverse, comprising rules corresponding to different territorial levels (national, Autonomous Community and local) [12].

We have reviewed the existing works for the management of TSNs on the LOD: from the existing modeling approaches and ontological models to the exploitation of the TSNs as LOD, interlinked with other information that give sense to the territorial redistribution. In the following Chapter 6, we make a synthesis of the state of the art and try to answer how this existing solves or not the issues identified in the main introduction, Chapter 1.

11. `https://eur-lex.europa.eu/eli-register/about.html`
12. https://eur-lex.europa.eu/eli-register/spain.html

Chapter 6

Synthesis

The state of the art has presented the existing spatiotemporal approaches, models, tools and algorithms, at each step of the [Villazón-Terrazas et al., 2011] life cycle, in view of our fundamental objective: the publication of evolving TSNs on the LOD Web, with semantic representations of their changes over time.

As explained in the Chapter 1, such an objective raise four research challenges. Below, we summarize the main elements of the existing approaches to meet these challenges, we list shortcomings and requirements for a system managing the whole life cycle of evolving TSNs in the Semantic Web. We briefly describe our proposals to meet these challenges.

– Challenge (1): To overcome the lack of interoperability between systems for TSNs exchange in the semantic Web.

Many Statistical Institutes (SAs) through the world now publish their statistics as LOD (*e.g.*, Italian statistical data [1], Scottish statistical data [2], UK Department for Communities and Local Government statistical data [3], Aragonese statistical data [4] or Japanese statistical data [5]). However, for the description of the geographic regions they administrate, they often create their own ontology (*e.g.*, the *Territorio Ontology* [6] of the *Italian National Institute for Statistics*, the *Geography Ontology* [7] of the Scottish Government), which results in a counterproductive proliferation of non-aligned vocabularies. The most thoroughgoing initiative is the one from UK Office for National Statistics (ONS) that proposes vocabularies [8] to represent geographic areas in the context of statistical data publication. Even if we can notice a certain level of abstraction on the terms used to describe TUs [9], which makes it possible to describe other TUs than UK's ones, none of these vocabularies allows one to describe any TSN structure and its levels. One needs to add new concepts to the ontology *Geographical hierarchy ontology* [10] if one wants to describe new territories and new territorial hierarchies than those listed. Similarly, in the Subsection 4.2.1,

1. http://datiopen.istat.it/
2. http://statistics.gov.scot/
3. http://opendatacommunities.org/data
4. http://opendata.aragon.es/sparql/
5. http://data.e-stat.go.jp/lodw/
6. http://datiopen.istat.it/odi/ontologia/territorio/
7. http://statistics.gov.scot/vocabularies/
8. http://statistics.data.gov.uk/vocabularies
9. http://statistics.data.gov.uk/def/statistical-entity
10. http://statistics.data.gov.uk/def/hierarchy/best-fit

we have shown that none of the existing ontologies of SAs or Mapping Agencies, has a sufficient level of abstraction to semantically describe any TSN hierarchical structure. The model created by [Plumejeaud et al., 2011] (a relational database model), previous work of our research group, has, for its part, a sufficient level of abstraction to semantically describe any TSN (with a structure such as the one we address in this thesis, defined in Chapter 2, Section 2.2).

▶ Hence, one first requirement to meet the first challenge is to immerse the model of [Plumejeaud et al., 2011] in the semantic Web.

With regard to TSN data sets published on the LOD Web, the RDF NUTS data sets (see Subsection 4.2.2) provide only one representation (*i.e.*, one resource with one URI) for TU such as ES63 Version 1999 and Version 2003 where two distinct resources are preferable, to refer to them in a statistical data set by their respective URI. Indeed, statisticians need to easily dissociate observations measured on these two distinct geographic objects since the surface is reduced by almost half from one version to another. More broadly, they need to easily distinguish each geographic area, in time and space, especially distinguish between different versions of a TU that changes while it maintains its name, or code.

▶ The second requirement identified is to create unique identifiers (URI on the LOD web) for each element of a TSN version.

We propose and present in a next Chapter 8, the *TSN Ontology* that enables the representation of any TSN hierarchical structure (structure defined in Section 2.2) on the LOD Web. This ontology immerses the generic model of [Plumejeaud et al., 2011] in the semantic Web.
We make the following design choices according to the [Grenon and Smith, 2004] approach, we decide to create a SNAP-SPAN Trans-Ontologies that depict the life (or history) of entities (TUs, Territories, Levels, TSNs) over time. Our model distinguishes between SNAP (continuant) entities (TUs, Territories, Levels, TSNs), and SPAN (occurrent) entities (TU versions, Territory versions, Level versions, TSN versions). The SNAP entities represent the TUs, Territories, Levels and TSNs as one would have represented a *person* that "*endures, or continues to exist through time while maintaining its identity*". The SPAN entities, along the versions, draw the life of these elements (a process): they are 4D representation of them (space time worm), that depict their history over time (using the 4D-fluents approach).
We demonstrate, in the Implementation Chapter 10 that our vocabulary is abstract and generic enough to cover any TSN hierarchical structure cases. It allows (N)SAs to describe their own territorial hierarchical structure, as well as their TUs, for each version of the TSN. We propose patterns for the creation of unique URIs for each element of a TSN version. Using our vocabulary, SAs may create LOD dictionaries of spatial units for statistics, all TSNs being described using the same terms, for the sake of interoperability.

– **Challenge (2): To provide a description of TSNs evolution over time in a way that it helps understanding the reason for the changes,**

and assists statisticians in the operation of transferring statistical data from one TSN version to another.

Most of the agencies contributions limit their description to isolated attribute change and do not propose any vocabulary to describe and group together changes that are involved in the same territorial event. Among these initiatives, we have highlighted several of them (please see Subsection 4.2.1). Here we come back to two of them that address the problem of evolution through time:
– the UK (ONS) proposes a vocabulary [11] to represent evolution of geographic areas in the context of statistical data publication. Nevertheless, none of the concepts of the *ONS Boundary Change Ontology* makes it possible to name changes that impact several TUs at the same time. The concepts describing change events are few and limited to *Recoding change* and *Boundary change*.
– vocabularies of the Statistics Bureau of Japan [12] represent geographic areas and their evolution, including description of area changes, in the context of statistical data publication. Their proposal is limited to description of changes in municipalities. Also, the description of the reason for the change is limited to a literal while we propose to provide URI links to discover new things in other data sets on the LOD, such as DBpedia.

With regard to software tools specialized in version control, although tools such as GeoGIG or the GitHub Inc. project handle versioning of geospatial data and provide a way to visualize differences in geometries, feature by feature [Negretti, 2015; LocationTech, 2018; Boundless, 2014a; **?**,b], they are unable to group changes that affect several geographic objects at the same time (*e.g.*, a merge of two TUs), or to attach semantics describing the context and nature of this territorial change.

Hence, changes occurring in TSN structures are rarely described and whenever they are, the descriptions are made TU by TU, with no link between changes. This makes it difficult to identify each of the TSN's features (territories, levels and TUs) that changes because of a same event. The model created by [Plumejeaud et al., 2011] includes a *Typology of Territorial Changes* in order to characterize changes in TSNs over time, whether they are isolated changes (*e.g.*, a name change) or changes that impact multiple TUs at the same time (*e.g.*, merge of two TUs). Regarding the description of changes, in [Plumejeaud et al., 2011], the authors dissociate *Derivation* cases (*e.g.*, Figure 8.7, (a)) from *Continuation* cases (*e.g.*, Figure 8.7, (b)), depending on whether the identity of the object is impacted during the change event or not. This approach is also the one adopted by *ontologies for fluents* (based on the *perdurantist* approach) which represent, in OWL, relationships between entities that change with time [Welty et al., 2006]. The *Typology of Territorial Changes* of [Plumejeaud et al., 2011] is not published on the LOD Web, contrary to the *Change Vocabulary* [13] of [Kauppinen et al., 2008; Kauppinen and Hyvönen, 2007], a lightweight spatiotemporal vocabulary made of two properties (*before*, *after*) and five classes for the description of changes (*Establishment*, *Merge*, *Split*, *Namechange*,

11. `http://statistics.data.gov.uk/def/boundary-change`
12. e.g., `http://data.e-stat.go.jp/lod/sac/`, `http://data.e-stat.go.jp/lod/terms/sacs#`, `http://data.e-stat.go.jp/lod/sace/`
13. `http://linkedearth.org/change/ns/`

Changepartof). In [Kauppinen et al., 2008], the authors introduce the notion of *Change Bridge* to chain old territories (*e.g.*, *East Germany* and *West Germany*) to the following ones (*e.g.*, *Germany*), using the *Change Vocabulary* for the description of changes.

▶ The third requirement identified is to immerse in the semantic Web the *Typology of Territorial Changes* of [Plumejeaud et al., 2011], adopting the *Change Bridge* approach of [Kauppinen et al., 2008] to chain former territories to the following ones, and the 4D-fluent perdurantist approach in order to describe the life process of spatiotemporal entities.

> Using typologies of spatiotemporal processes from [Claramunt and Thériault, 1995; Del Mondo et al., 2010; Plumejeaud et al., 2011], we propose the *TSN-Change Ontology* that defines a set of tags to describe territorial changes whether they modify the identity of the geographic objects or not. Contrary to [Harbelot et al., 2013], our model dissociates the filiation links (established between two time-slices of a same spatial entity) from the descriptions of changes to the entity, and it adopts the *Change Bridge* approach of [Kauppinen et al., 2008]. The change descriptions are SPAN objects that may be link with other resources on the LOD Web (LOD sets describing historical events for instance), in order to contextualize the territorial changes. We believe that our model, by offering a more detailed description of territorial changes in TSN, contributes to a better understanding of territorial evolution over time.

– Challenge (3): To take into account the vertical/hierarchical dimension of TSNs in their evolution.

Works from [Kauppinen et al., 2008] and [Lacasta et al., 2014] adopt an ontological approach for the description of changes of geographic features over time, but these approaches lack a description of changes that propagates level-to-level (*e.g.*, change of the boundaries of a TU may impact also the boundaries of its nested TUs) considering the nested nature of the geographic elements of a TSN.

▶ The fourth requirement identified is to construct chains of changes that is to say, create links between change descriptions when the changes impact embedded features of the TSN.

> We propose to describe changes that affect several levels of the TSNs. In order to automate the detection and semantic description of territorial changes, we decide to adapt and immerse the algorithm of [Plumejeaud, 2011] in the semantic Web. Our algorithm, called the *TSN Semantic Matching Algorithm*, detects TUs changes and chains together changes through the TSN levels to offer a view on embedded changes, probably caused by the same initial event (societal, historical or political event). Then, our automatically generated descriptions of territorial changes are by far more accurate and complete than the one presented in the cultural heritage portal of [Kauppinen et al., 2008].

– Challenge (4): To populate automatically such an ontological model, by preserving its genericity, since the characteristics which make up the identity of geographic areas may vary form one country to another.

In [Lacasta et al., 2014], the authors propose a semi-automatic process for creating and populating the evolution model of any country, with an approach that is also quite generic since the characteristics of the heterogeneous sources can be set dynamically. However, this process does not produce any description of the impacts of changes on the hierarchy of TUs. Furthermore, this approach requires the use of a written dictionary of changes as input, listing all of the individual changes. This requirement may be seen as a drawback for current statistical information systems because the manual listing of changes is time consuming. Then, we decide to automate the detection and description of such a dictionary.

▶ The fifth requirement identified is to create a configurable system the users can set up according to the characteristics of the TUs identity.

To meet this challenge, we get inspiration from conflation approaches [Hastings, 2008; McKenzie et al., 2014; Ruiz et al., 2011]. We create a configurable framework, called the *Theseus Framework*, that computes similarity measures on TUs attributes. The Theseus framework uses the concept of *Identity* attached to each TU, in an approach similar to the one of Huibing et al. [2005]. A weighting function is created in order to determine if the identity of a TU is preserved in the following version of the TSN. While metrics results are used for conflict resolution for gazetteers or POI (*e.g.*, choose the canonical form of a place name between multiple possibilities), we chose to transform metrics in semantics. Each result of the distance tests on each attribute is transposed into semantic descriptions using the TSN-Change Ontology. Then, the strength of our approach stands not only in the computation of similarities, but also, and mostly, in the underlying semantic model for changes representation and spatial objects lineage descriptions.

As a conclusion, we have identified in this state of the art that no ontological model enables the representation of any TSN hierarchical structure and of its evolution over time on the LOD Web. However, we have identified modeling approaches (SPAN, perdurantist and 4D-Fluent approaches) and upper ontologies to build upon (PAV, BFO, GeoSPARQL, OWL Time), as well as algorithm [Plumejeaud, 2011] that proves the feasibility of detecting and describing automatically complex changes in TSNs.

We present in the following Chapter 7 our contributions for the publication and exploitation on the LOD Web of TSNs in such a way that it gives access to the life of the TSNs over time, as well as the life of each of the features that compose it, considering their belonging to a territorial hierarchy.

Part B

Contributions

Chapter 7

The Theseus Framework

7.1 Introduction

In this manuscript, we present *Theseus*, a framework designed for managing TSNs in the semantic Web, according to a management process that consists of the following activities: Specify, Model, Generate, Publish, and Exploit. It is called *Theseus* with reference to the identity philosophic issue raised by the *Ship of Theseus* that changes over time and was rebuilt entirely over the years, every plank being broken one by one. This framework automates the publication on the LOD Web of any TSN (structure defined in Section 2.2), the detection of their changes over time, and the publication of such changes on the LOD Web. It encapsulates two ontologies we have designed in order to describe TSNs and their changes over time: the ontologies are called *TSN-Ontology* and *TSN-Change Ontology*. It also encapsulates an implementation of the algorithm of [Plumejeaud, 2011] that detects and describes similarities and changes between two consecutive versions of a TSN. However, the [Plumejeaud, 2011] algorithm has been modified to output RDF descriptions of a TSN and of its changes over time, based on the concepts defined in the TSN-Change ontology. This adaptation of the algorithm of [Plumejeaud, 2011] is called the *TSN Semantic Matching Algorithm*. The major challenge we face when automating the detection and semantic description of TSN changes is the implementation of our concepts and algorithm with regard to the heterogeneity of the input data. Indeed, the definition of the identity of TUs varies from one TSN to another TSN. Does a TU, its name or geometries or both are changed, remain the same TU, or no longer exist? There is no unique answer to this question, again the response varies from one TSN to another. Thus, the Theseus Framework is made of several configurable software modules that implement the *TSN Semantic Matching Algorithm*, allowing to define the list of attributes that compose the identity of a TU in a processed TSN, and in which proportion each of these attributes can vary before a TU loses its identity and is therefore considered as a new one. This semantic framework is, as far as we know, the first one that handle TSNs' structure versioning. But, far more than just linking the TSN elements throughout the versions, this framework provides users with semantic descriptions of TUs filiation links over time (including similarities and changes description).

7.2 Motivations and requirements

Below, we present the issues behind the four challenges we have identified in Chapter 1, and the requirements [1] our system must meet to address these issues:

Issue 1: Lack of semantic interoperability between systems for TSNs exchange.
Requirement 1.1: Adopting Semantic Web technologies.
Requirement 1.2: A pivot semantic model to describe any TSN hierarchical structure.

Issue 2: Lack of semantic interoperability between systems for TSNs change descriptions exchange in a way that it assists statisticians in the operation of transferring statistical data from one TSN version to another.
Requirement 2.1: A semantic model to describe any TSN change over time.
Requirement 2.2: A "Gateway" modeling approach that may bridge data from different versions.

Issue 3: Lack of semantic interoperability between systems for TSNs change descriptions exchange in a way that it helps understanding the impact of a territorial change on all the levels of the territorial structure.
Requirement 3: A modeling approach to construct chains of changes, that is to say, create links between change descriptions when the changes impact embedded features of the TSN.

Issue 4: Lack of semantic interoperability between systems for TSNs change descriptions exchange in a way that it helps understanding the reason for the changes.
Requirement 4: A modeling approach that connects territorial changes to the events that have caused these changes.

Issue 5: Lack of tools to populate automatically such a descriptive model of TSNs and of their evolution over time, by preserving its genericity, since the identity characteristics may vary depending on countries and on the quality of the input TSN data sets.
Requirement 5: A configurable system the users can set up according to the characteristics of TUs.

To meet these requirements, we adopt the semantic Web technologies for the description of the TSN's areas and of their changes. These technologies guaranty the syntactic interoperability and, most of all, the semantic interoperability between systems exchanging TSN information. Thus, we make sure that the semantics of changes are shared and understood in the same way, by humans and software agents, that share the same concepts and definitions. Another benefit is the LOD Web, which allows us to connect and contextualize data from our system to other data on the LOD Web. As a matter of fact, the OGC and W3C recommend for spatial data to be on the Web, to be connected or linked, to other resources. They suggest linking spatial data with URIs (using a Linked Data approach) from popular repositories

1. The requirements presented here are more precises then the ones presented in the previous Chapter 6

to improve discoverability of data [Tandy et al., 2017].

We have shown in the state of the art Chapter 4, Subsection 4.2.1, that there is no reference ontology, on the LOD, abstract enough to describe any TSN any TSN hierarchical structure that exists in the world (*i.e.*, **Issue 1**). In this respect, we have first created the *TSN Ontology* which allows for the description of any TSN hierarchical structure and of its successive versions, rather than to provide users only with the latest one, as usually done. It ensures unambiguous identification, in time and space, of the TUs. Then, we have created the *TSN-Change Ontology* which contains a set of tags that describe the nature of the changes undergone by TUs and assist analysts in understanding the context of such changes (*i.e.*, **Issue 2, 3, 4**). In order to automatically populate the *TSN-Change Ontology*, we have created a configurable *TSN Semantic Matching Algorithm* based on the Algorithm of [Plumejeaud, 2011]. This algorithm automates the detection and description of similarities and changes between two TSN versions. It characterizes the changes by creating filiation links and *bridges* in order to navigate from one version to its subsequent and conversely (*i.e.*, **Issue 2, 5**).

7.3 Use Cases

The Theseus Framework carries out a number of functions that we present in this Section as use cases. This framework is intended first for the SAs, statisticians,

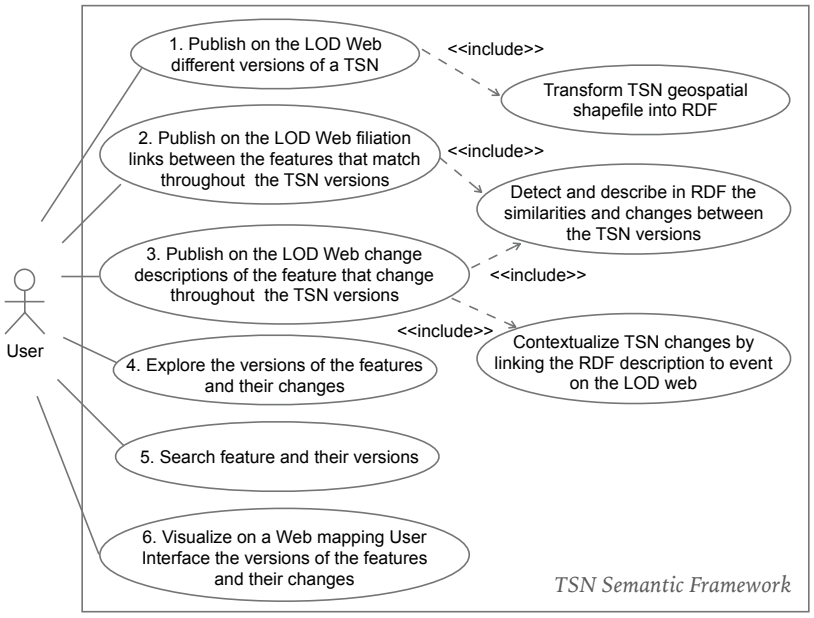


Figure 7.1 – The Theseus Semantic Framework Use Cases.

and researchers users who wish to publish on the LOD Web successive version of their TSN(s), change and similarity descriptions between the versions *i.e.*, filiation links between the features throughout the versions (see use cases 1, 2, 3 on Figure 7.1). The RDF graphs created by the framework are intended for a wider audience

of statisticians, researchers, or citizens. Indeed, the created RDF graphs provide statisticians, researchers, or citizens with unambiguous description of areas they can, besides, explore, search, and visualize on a Web mapping Graphical User Interface (GUI) (see use cases 4, 5, 6 on Figure 7.1).

The use case 1 (see Figure 7.1) describes the primary need for statistical agencies due to open data directives, namely, to publish on the LOD Web different versions of their nomenclature (from shapefiles to RDF graphs). This use case 1 includes a second use case that consists in transforming the TSN geospatial shapefiles into RDF, before publishing them on the LOD Web. The use cases 2 and 3 describe the situation where, in addition to TUs of each version, the user needs to access the filiation links (continuation or derivation) between these TUs over time, as well as the change descriptions if the TUs have changed. The publication on the LOD Web of such bridges between the TSN versions implies, first for the Theseus Framework to automatically create such descriptions (without the use of a written dictionary of changes) (use case "Detect and describe in RDF the similarities and changes between the TSN versions"), and, then, to contextualize these changes by interlinking the change descriptions with data sets on the LOD Web that describe historical events or law acts (use case "Contextualize TSN changes by linking the RDF description to event on the LOD Web"). The three use cases 4, 5, 6 are achievable only once the RDF data are published on the LOD Web, because they consist of the exploration (by following the links on a Web browser for example), of their search (using SPARQL requests for instance), and the visualization (through Web mapping GUI for instance) of data.

7.4 Prerequisites

The Theseus Framework takes as input several geospatial files (ESRI shapefiles), one for each version of the TSN, and transform them in RDF graphs.

Regarding the description of changes: one prerequisite for optimal results is that all the geometries of the TUs being at the same generalization level, in the same spatial reference system. Also, in the framework, we make the assumption that what constitutes the identity of a TU is a list of geographic and nomenclature attributes (*e.g.*, area geometry, area surface, or location of the capital, spatial structure information, toponymic information). The Theseus Framework de-correlates the geographic information from all other thematic information associated with a TU, such as the number of inhabitants, the elected representatives, etc.

This means that the framework, as it stands, aims at describing geographic changes. It is not yet generic enough to describe over kind of changes such as the number of inhabitants in an area. However, in order to deal with such changes to thematic information in the future, we make the framework configurable (it takes as input a list of parameters that could be extended in the future in order to be more generic).

The framework detects geographic changes but, as we will see later, the geome-

tries of the TUs shall not be considered when calculating the changes: the change descriptions could only apply to the name of the TUs for instance. This will depend on the identity definition of the TUs set by an expert that configures the framework (the list of parameters).

The framework, as it stands, could proceed TSNs that divides a territory into contiguous territorial units (*i.e.*, continuous mesh), irregularly shaped, covering the entire territory without overlapping of the TUs of a same level. Regarding the structure, it processes *covering*, *strict*, and *onto* hierarchies (as explained in Section 2.2).

7.5 Overall Architecture

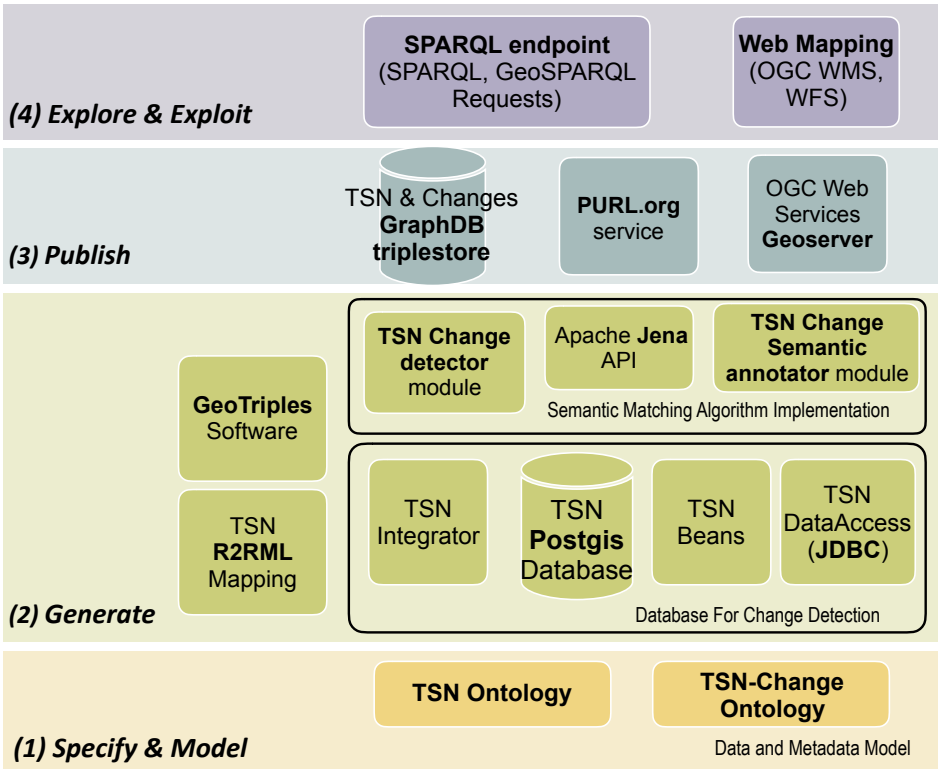


Figure 7.2 – The Theseus Framework Modules.

The Theseus Framework is composed of several modules (organized into four functional levels, see Figure 7.2) to handle the whole TSN data life cycle on the LOD Web: from the modeling of data to the exploitation of data on the LOD Web (see Appendix A). To achieve this, we follow the W3C Best Practices recommendations for Publishing Linked Data [Hyland et al., 2014]. We follow in particular, the life cycle established by Villazón-Terrazas et al. [2011], for the process of publishing Government Linked Data (see Preliminary remarks, in 2.3). We choose to present the Theseus Framework architecture as a stack, starting from the first activities recommended by Villazón-Terrazas et al. [2011] (*i.e.*, specify and model), and ending

by the Exploit activity. The Figure 7.2 presents, the modules of the framework, at each steps of the life cycle process:

– *(1) Specify and Model*: at the very heart of the Theseus Framework lies its data model (composed of the *TSN and TSN-Change ontologies*). We have specified this model according to the methodology for designing ontologies proposed in [Bachimont et al., 2002]: a corpus of TSNs and existing ontologies (based on the corpus presented in Section 4.2) helped us to determine the vocabulary to be used to be close to the designations used by the SAs, and thus to facilitate the use of the model. The TSN ontological model provides the modules of the framework with a formal representation of TSNs and how they change over time.

– *(2) Generate*: during the activity of generating RDF data that apply to the *TSN and TSN-Change ontologies*, several software modules are used, such as: the *Geotriples Software*[2] that aims at transforming the TSN Geospatial files (also called *shapefile*) into RDF triples, using the TSN ontology concepts for the descriptions of the TSN elements; the *TSN Change Detector* and the *TSN Change annotator* modules, we have developed in order to detect changes in TSN's Geospatial shapes, and describe these changes using the RDF model and the TSN-Change ontology concepts.

– *(3) Publish*: at this step, three existing software are used: the purl.org service (to provide persistent URI to data); the GraphDB triplestore (to publish the RDF data created); the GeoServer software (to publish the TUs' geometries linked to their LOD representation). We will come back to the usefulness of these three software in the Section 9.2 dedicated to the description of the workflows implemented within the Theseus Framework.

– *(4) Explore and Exploit*: in order to provide users with tools to explore and exploit the published data, we set up various search engines, using standard Web mechanisms: a SPARQL enpoint for performing RDF data manipulations through SPARQL and GeoSPARQL requests; a Web Graphical User Interface (GUI) to visualize and query the geospatial data, using standard OGC requests to the Web Map Service (WMS) and Web Feature Service (WFS) published by our GeoServer that connects to the TSN PostGIS database of our framework.

7.6 Conclusion

In this Chapter, we have introduced the Theseus Framework, the requirements it answers, its main use cases, and its overall architecture. The main component of the framework is its model, composed of two ontologies, in order to dissociate between the TSN elements in each version and the filiation links between these elements over time. The *TSN Ontology* aims at describing the TSN elements and TSN's versions elements. The *TSN-Change Ontology* aims at describing the filiation links between the TSN versions elements over time.

2. `http://geotriples.di.uoa.gr/`

In order to populate this model, two different workflows have been set up within the framework: the first workflow creates the RDF catalogs of TUs' versions (please see Workflow A, Figure 9.1); the second workflow detects and describes the filiation links between the TSN elements throughout the versions (please see Workflow B, Figure 9.1). These two workflows of the Theseus Framework are presented in the Chapter 9.

Prior to this, the following Chapter 8, present the two ontologies core components of the Theseus Framework.

Chapter 8

The TSN ontological model

8.1 Introduction

We present, in this chapter, the TSN ontological model we have created. It is made of two ontologies: the *TSN Ontology* and the *TSN-Change Ontology*. Both the *TSN* and *TSN-Change* ontologies are geographic ontologies that define geographic concepts (*e.g.*, a town, TU) and not geometric concepts [Klien and Probst, 2005], contrary to the *NeoGeo Geometry Ontology*[1], for instance.

In order to be generic, the ontologies described the main attributes (and their changes) of the areas for statistics observed in TSNs in the world, that is to say: the TU's identifier in the TSN (also called its *code*), name, geometry, level of belonging (*e.g.*, district level), and a free text description. The two ontologies are domain ontologies, that are built on upper ontologies.

We first present in Section Specify 8.2 our design choices, based on the review of the existing works (please see Chapters 3, 4). Second, in the Sections 8.3 and 8.4, we present the two ontologies that together constitute a SNAP-SPAN trans-ontology, according to the approach of [Grenon and Smith, 2004].

8.2 Specifications

We have specified the TSN ontological model according to the methodology for designing ontologies proposed in [Bachimont et al., 2002]: a corpus of TSNs and existing ontologies (based on the corpus presented in Section 3.2) helped us to evaluate the most relevant vocabulary to be close to the designations used by the SAs, and thus to facilitate the adoption of the model. The result of this *Specify* step corresponds to the state of the art part of this manuscript (Chapters 3, 4).

We get inspired by Version Control Systems and the way they manage changes throughout the versions (see Subsection 5.1.4). We also get inspired by the BFO framework, and have decided to represent the life of TUs over time, that is to say four-dimensional entities also called *occurrent* entities (SPAN is the ontological theory of those entities in the approach of [Grenon and Smith, 2004]) (see Subsection 3.2.2.5). More precisely, we adopt the *Ontology for fluents* [Welty et al., 2006] approach that is a perdurantist approach. In the BFO framework, perdurants are processes, occurrent entities which persist through time. Our choice is to detect and

1. `http://geovocab.org/geometry`

describe processes of evolution of geographic units over time, in territorial statistical nomenclatures. According to the approach of [Grenon and Smith, 2004], we have decided to create our ontological model as a SNAP-SPAN Trans-Ontology that depicts the life (or history) of entities over time. The ontological model we have created handles both SNAP and SPAN views on the geographic divisions. The TSN and its components are *continuant* entities (in analogy to a *person*) (please, see Figure 8.1 and the TSN Ontology concepts *Unit, Nomenclature, Level* that inherit from the BFO concept *Continuant*). All the versioned components of a TSN are occurrent entities (that depend on the continuant objects) as well as the change nodes (see Figure 8.1). The concepts *Change* and *Version* of the TSN and TSN-Change Ontologies inherit from the BFO concept *Occurrent*. Together, the versioned components and the change nodes constitute 4D objects that depict the life of the TUs, over time (in analogy to the life of a *person*).

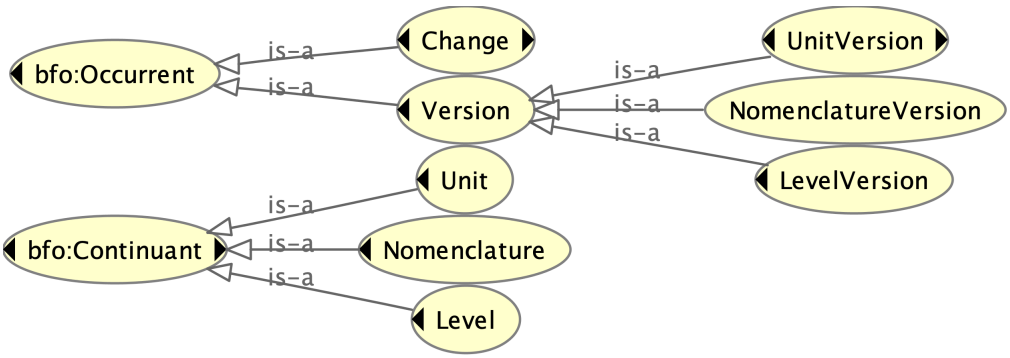


Figure 8.1 – TSN Ontological Model main concepts inherited from the BFO ontology (Occurrent and Continuant concepts from [Grenon and Smith, 2004]).

As explained above, we adopt the *perdurantist* approach of ontologies for *fluents*, for the description of the TSN elements that vary in time (Figure 8.2, Approach (1)). However, where ontologies for *fluents* use the term *time-slice*, we would rather use the term *Version*, to be as close as possible to the statisticians targeted users (Figure 8.2, Approach (3)). Also, we adopt the *Change Bridges* approach for managing the union of successive *Versions* (see Figure 8.2, approach (2)). Within these *Change Bridges* (that we call *XChange-Bridges*, X for eXtended), we describe the differences between two versions and characterize the nature of the territorial changes, using a typology of changes (presented in Figure 8.5) based on the typologies of [Claramunt and Thériault, 1995] and [Plumejeaud et al., 2011]. This typology extends the *Typology of Territorial Changes* of [Plumejeaud et al., 2011], in order to describe changes in the boundaries of territories (*e.g.*, EU enlargement). Like [Plumejeaud et al., 2011], we consider that a change is rarely isolated and independent from the other changes that occur simultaneously within the other units inside a given area.

Based on this observation, we create *Multi-level Territorial Change Graphs* (also called *Change Graphs*) which describe and link together concomitant changes that impact on different levels of the territory. Then, we provide analysts with a detailed

representation of a change event, and enable them to follow links to discover consequences of this change (on each level of the TSN) and causes of this change by linking to resources on the LOD (historical events for instance).

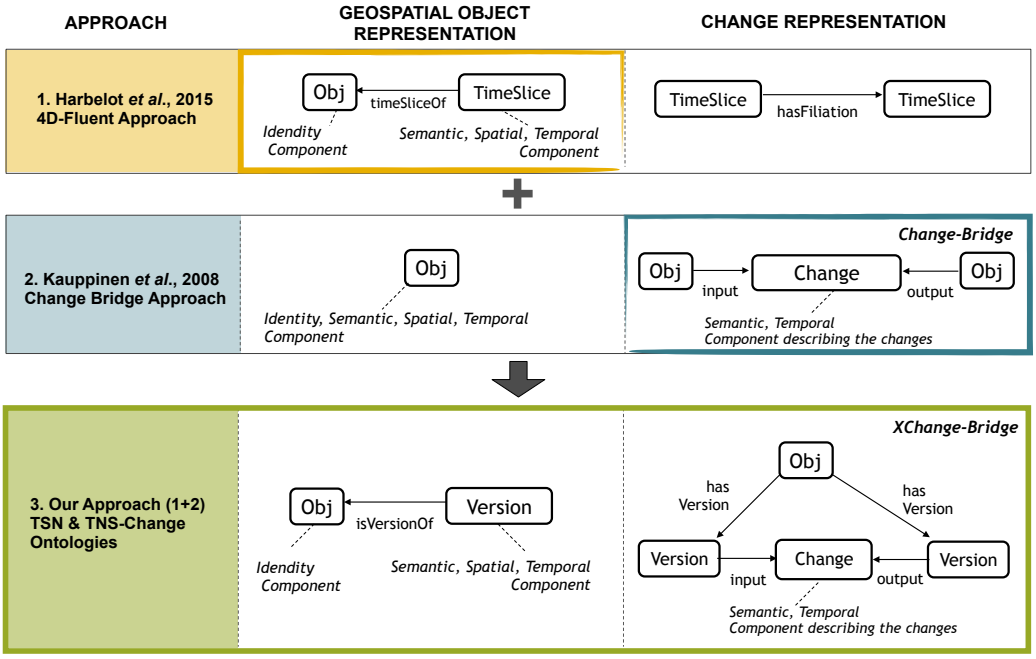


Figure 8.2 – Combining approaches for the TSN/TSN-Change Ontological Model.

The Change graphs (nodes and predicates) we automatically create are SPAN entities too, which depict the life of the spatial regions over time. Then, our approach goes a step further compared to the models for versioning such as PAV, since we propose to represent a geographic unit that changes over time (`Unit` and its versions `UnitVersion` in our vocabulary, *i.e.* the successive versions of the geographic unit) but also the processes of evolution and changes (4D view on the evolution of the geographic units over time that is drawn through a RDF graph).

The Theseus framework is developed to answer the problematic presented in Chapter 1: how to enhance the understanding of the dynamics of territories, providing users with tools to comprehend the evolution of spatial entities. As noticed by Del Mondo et al. [2010] (see Subsection 3.1.1), modeling the dynamics of territories requires modeling the entities and the spatial, spatiotemporal and filiation relations between them. We develop this framework (and the model behind it) in this respect. The *TSN Ontology* is based on two standard ontologies for space and time: the GeoSPARQL Ontology [Perry and Herring, 2012] designed by the Open Geospatial Consortium (OGC) and the Time Ontology in OWL [Cox and Little, 2017], designed by the W3C. The *Region Connection Calculus* (RCC8) has been implemented in GeoSPARQL. Similarly, the Time Ontology supports the set of interval relations defined in Allen [1983]. Therefore, the RDF graphs created by our

framework, based on the TSN ontological model, based itself on the GeoSPARQL and Time Ontologies, provide a representation of the spatial and temporal relations between the geospatial entities of a TSN. To model the filiation relations, we have designed the *TSN-Change Ontology*, as no existing ontology was able to represent these relations within TSN elements. Thus, using the SPARQL endpoint module of our framework, at `http://steamerlod.imag.fr/sparql`, users may query these three types of relations (spatial, temporal and filiation relations) and discover the relations between two TUs of a TSN. Then, our framework handles the whole life cycle of data describing evolving TSNs on the LOD Web, from representation to exploration, in order to enhance the understanding of territorial evolution.

The next Section introduces in more details the TSN Ontology: its main concepts and predicates, and how it enables the description of any TSN hierarchical structure in the world.

8.3 The TSN Ontology

The *TSN Ontology* allows for the description of any TSN hierarchical structure (structure defined in Section 2.2), while existing ontologies apply to TSN of a particular State or region of the world (*e.g.*, *Administrative geography and civil voting area Ontology for UK*[2]). It may be used as a *shared vocabulary* since it contains basic terms of the domain, such as *Unit* (*i.e.*, TU *e.g.*, *ES63* Unit within the NUTS), *Level* (*e.g.*, *NUTS Level 1* composed of major socio-economic regions), etc. Thanks to its immersion into the LOD Web, humans but also machines can share the same definition of a TSN structure. Therefore, the *TSN Ontology* enhances interoperability between systems in the domain of TSN. For instance, the following Figure 8.3 shows how to declare the NUTS, ASGS, U.S. Census Tract TSN using the TSN `http://purl.org/net/tsn#Nomenclature` main concept. Then, the text 8.1 provide the RDF/turtle version of this figure.

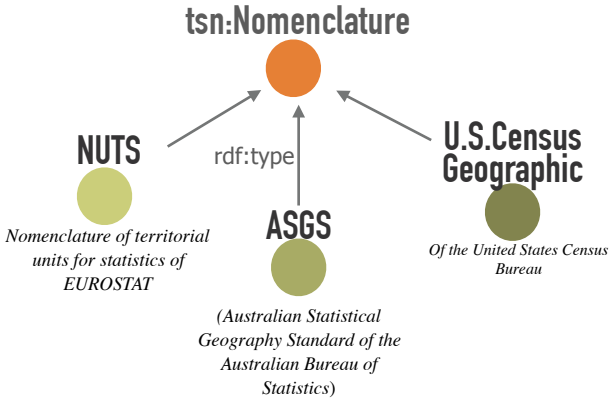


Figure 8.3 – TSN Ontology - Example of three TSN declared using the TSN *Nomenclature* main concept.

2. `http://data.ordnancesurvey.co.uk/ontology/admingeo/`

```
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:       <http://www.w3.org/2000/01/rdf-schema#> .
@prefix tsn:        <http://purl.org/net/tsn#> .

<http://purl.org/steamer/nuts>
  a  tsn:Nomenclature ;
  tsn:hasIdentifier "NUTS"^^xsd:string ;
  tsn:hasName
      "Nomenclature of Territorial Units for Statistics"^^xsd:string ;
  tsn:hasAcronym "NUTS"^^xsd:string ;
  tsn:hasDescription
      "The NUTS classification
      (Nomenclature of territorial units for statistics)
      is a hierarchical system for dividing up
      the economic territory of the EU for
      the purpose of: collection, development and harmonisation
      of European regional statistics ; Socio-economic analyses of
      the regions ;  Framing of EU regional policies
      (http://ec.europa.eu/eurostat/web/nuts).
      This nomenclature specification contains
      territorial units definition for
      the EU, EFTA (Norway, Switzerland, Liechtenstein
      and Iceland) and Candidate Countries."^^xsd:string ;

   tsn:hasVersion <http://purl.org/steamer/nuts/V1999> ;
   tsn:hasVersion <http://purl.org/steamer/nuts/V2003> ;
   tsn:hasVersion <http://purl.org/steamer/nuts/V2006> ;
   tsn:hasVersion <http://purl.org/steamer/nuts/V2010> ;
   tsn:hasVersion <http://purl.org/steamer/nuts/V2013> .

 <http://purl.org/steamer/asgs>
    a  tsn:Nomenclature ;
    tsn:hasIdentifier "ASGS"^^xsd:string ;
    tsn:hasName
        "Australian Statistical Geography Standard"^^xsd:string ;
    tsn:hasAcronym "ASGS"^^xsd:string ;
    tsn:hasDescription
        "The Australian Statistical Geography Standard (ASGS)
        provides a framework of statistical areas used by
        the Australian Bureau of Statistics (ABS) and
        other organisations to enable the publication of
        statistics that are comparable and spatially
        integrated."^^xsd:string ;
    tsn:hasVersion <http://purl.org/steamer/asgs/2011> ;
    tsn:hasVersion <http://purl.org/steamer/asgs/2016> .

 <http://purl.org/steamer/usCensusTract> TODO.
```

Listing Code 8.1 – Description of the NUTS and ASGS nomenclatures using the TSN Ontology vocabulary (RDF-Turtle syntax)

Figures 8.4 and 8.5 are simplified RDF graphs showing the main classes and

predicates of the two ontologies TSN Ontology and TSN-Change Ontology [3]. The *Predicates* label the arrows. An arrow links the *domain* to the *co-domain or range*.

We have designed the *TSN Ontology* generic enough so that it allows for the description of TSNs hierarchical structure, while other existing ontologies only apply to TSN of a particular State or region of the world (*e.g.*, *Administrative geography and civil voting area Ontology for UK* [4]) (as explained in Chapter 4).

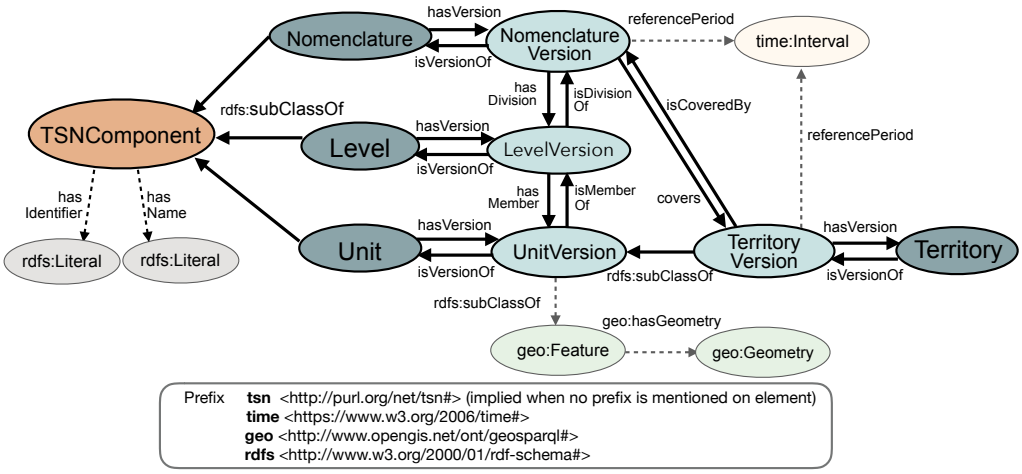


Figure 8.4 – TSN Ontology main concepts.

The *TSNComponent* concept is the super class of all the concepts of the TSN Ontology (see Figure 8.4). On this concept are defined properties that all elements of a TSN can hold (`hasIdentifier`, `hasName`). Figure 8.4 depicts *Continuant* elements composing a TSN, from the *Nomenclature* (*e.g.*, *NUTS*) broad concept to its *Level*s (*e.g.*, *NUTS Level 0*) and *Unit*s (*i.e.*, TUs *e.g.*, *NUTS Level 0 ES63*). There is no relation (no predicate) between these three continuant concepts because their relations are *fluents* that hold within a certain time interval and not in others (see Subsection 3.2.2.6). For instance, a TU may exist in two versions of a TSN, then disappear in the next TSN version after a redistricting. Hence, we design a second hierarchy of occurent concepts that depicts the hierarchy of elements composing a *NomenclatureVersion*, that hold within a certain time interval, from the *NomenclatureVersion* (*e.g.*, *NUTS Version 1999*) broad concept to the smallest elements that may compose it, *i.e.*, *UnitVersion* (*e.g.*, *NUTS Level 0 Version 1999 ES63*). The *OWL-Time* vocabulary is used to assign a *reference period* to a *NomenclatureVersion* and another reference period to a *TerritoryVersion*. We choose to assign this period only to the *NomenclatureVersion* parent element in the hierarchy of component of a TSN version since it applies, by propagation, to child elements (*LevelVersion* and *UnitVersion*). The reference period of a TerritoryVersion is dissociated from the reference period of a NomenclatureVersion, because a TerritoryVersion has its own existence in reality, that does not depend on the NomenclatureVersion (whereas

3. Definitions of all concepts are available from the URI `http://purl.org/net/tsn#` and `http://purl.org/net/tsnchange#`.

4. `http://data.ordnancesurvey.co.uk/ontology/admingeo/`

the other elements UnitVersion, LevelVersion only exist because of the TSN: they are units designed for statistical purposes). The ontology introduces the *Territory* concept (*e.g.*, *European Union politic organization* (EU) Territory) and makes a difference between this upper TU and all the other TUs it encompasses. Whereas this concept seems to be implicit and does not exist in other models, we decide to make it explicit. Since, once a temporal approach is adopted, the territory covered by the TSN, is no longer an implicit object that do not change (*e.g.*, EU of 15 (1995-2004), EU of 25 (2004-2006) or EU of 27 (2007-2013) member states)). Then, its evolution must be described, as for other elements.

The concepts *TerritoryVersion* and *UnitVersion* are subclasses of the *geo:Feature* concept defined by the Open Geospatial Consortium (OGC) within the *Geosparql* ontology [5].

The main predicates the *TSN Ontology* encompasses are:

1. `hasVersion(x, y)` which stands for "the TSNComponent x has a Version y".
2. `isCoveredbBy(x, y)` which stands for "the TerritoryVersion x is covered by the NomenclatureVersion y".
3. `hasDivision(x, y)` which stands for "the NomenclatureVersion x has a geographic division LevelVersion y".
4. `hasMember(x, y)` which stands for "the LevelVersion x has a UnitVersion y member".

For each predicate, an inverse property is defined, in order to go up or down in the hierarchy of the TSN (*e.g.*, `isDivisionOf` vs `hasDivision`).

The predicates `hasSubFeature` and `hasSuperFeature` express a hierarchical order between two elements that belong to the same TSN class: between two *LevelVersion* (*e.g.*, in Switzerland (SAU) the Level of Cantons (version 2017) `hasSubFeature` the Level of Districts (version 2017)) ; between two *UnitVersion* (*e.g.*, in SAU the canton of Neuchâtel (version 2017) `hasSubFeature` the District of Boudry (version 2017)).

8.4 The TSN-Change Ontology

Our second ontology, the *TSN-Change Ontology*, allows for the description of territorial changes over time, through semantic graphs. For the representation of TSN elements evolution over time, we adopt a hybrid approach inspired from the *Ontologies for fluents* [Welty et al., 2006] and from the *Change Ontology* [Kauppinen and Hyvönen, 2007].

Adopting the widely-used *typology of spatiotemporal processes* from [Claramunt and Thériault, 1995], the *territorial change typology* from [Plumejeaud et al., 2011], and the concept of identity like in [Huibing et al., 2005], we propose the *TSN-Change Ontology* (available at `http://purl.org/net/tsnchange#`) that defines a set

5. `http://www.opengis.net/ont/geosparql`

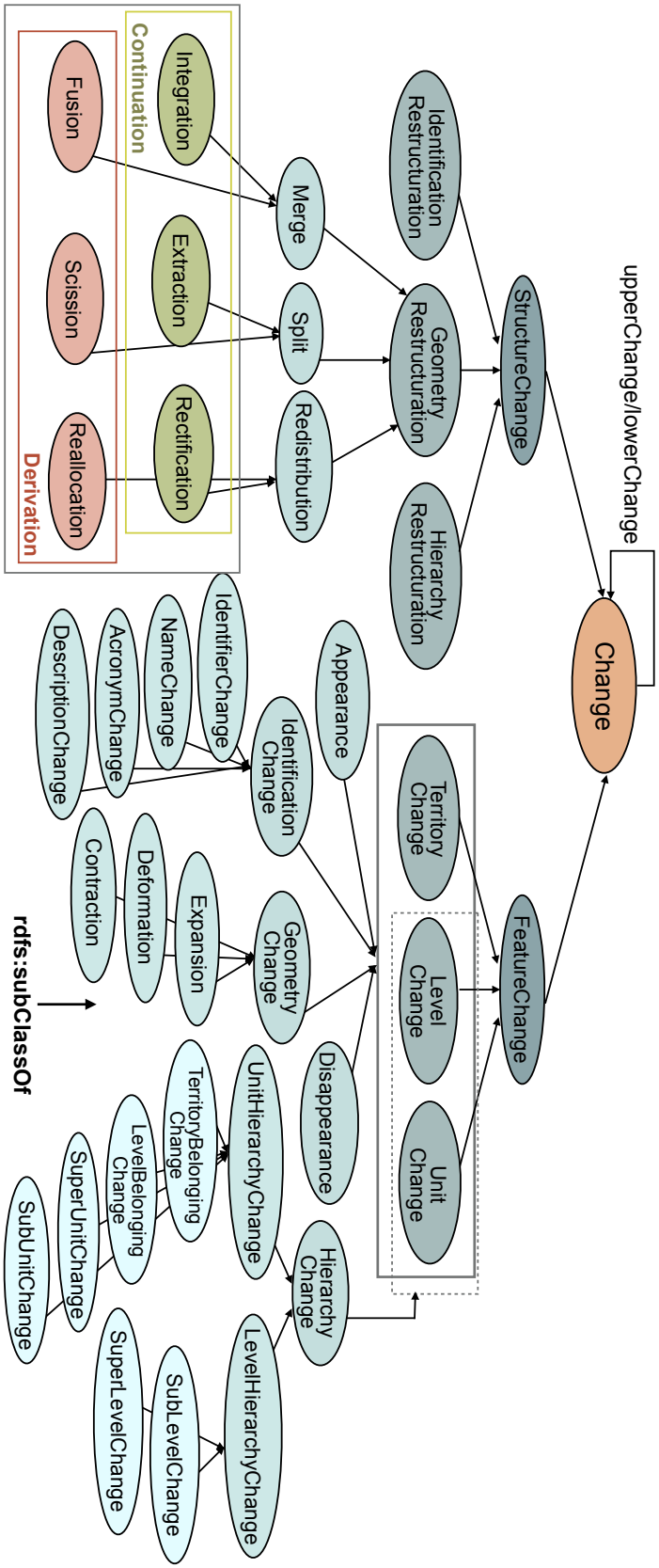


Figure 8.5 – TSN-Change Ontology main elements.

of tags to describe changes whether they modify the identity of the geographic objects or not. We extend the *Change Bridges* approach in order to manage the union of successive *versions* and describe additional useful types of territorial changes. Indeed, by combining several change nodes and dissociating the Territory main TUs from the over TUs, we can describe complex changes such as the *Expansion* of the areal surface of a territory (*e.g.*, the European Union), and the impact of this expansion on each TU that composes it. The *TSN-Change Ontology* divides the types of change into two broad categories:

1. *StructureChange*: sub-concepts which belong to this category describe changes that affect simultaneously several elements of the TSN (*e.g.*, a *Merge* event affects at least two elements that merge).

2. *FeatureChange*: sub-concepts which belong to this category describe changes undergone by one element of the TSN. For instance, the *Merge* of two or more TUs (that is a *StructureChange* event) causes several *FeatureChange* events: the *Disappearance* of one or more TU(s) and the *Expansion* of another one, if it keeps its identity after the merge event, otherwise the *Appearance* of a new one as the result of the merge.

The strength of the created typology of changes (available at `http://purl/org/net/tsnchange#`) lies in simple concepts (describing only one TU attribute that changes for instance) one has to combine in order to describe complex changes. A change that takes place in one level of the TSN is linked to others, at lower or higher levels, if they all occur on nested components. For instance, a change of a district boundaries will be linked to the changes of its sub-districts if any (using predicates `lowerChange`, `upperChange`). *Change Bridges* of [Kauppinen and Hyvönen, 2007] are called *XChange-Bridges* in our approach (X for eXtended). The *XChange-Bridges* structure is presented in Figure 8.6.

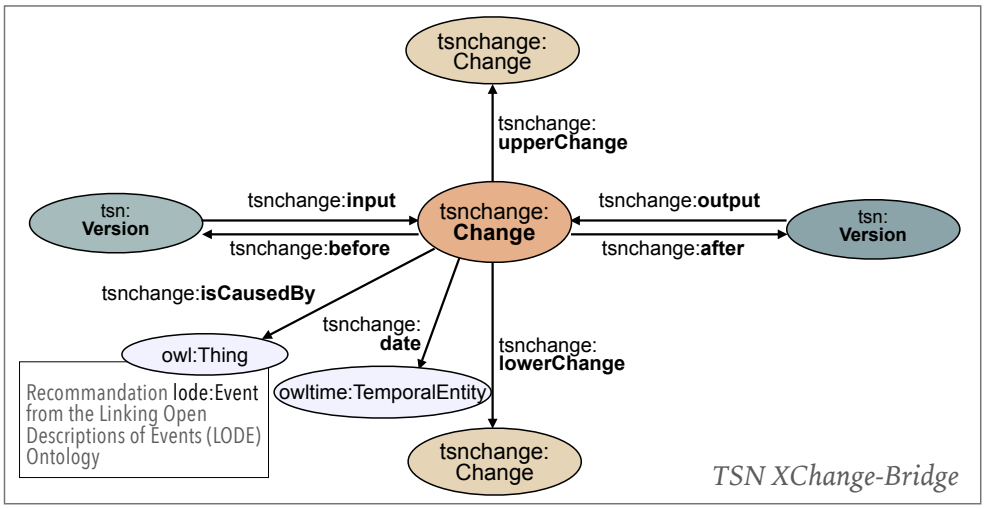


Figure 8.6 – TSN-Change Ontology X-ChangeBridge Model.

The predicate `tsnchange:input` (inverse property `tsnchange:before`) points to a *TSNComponent* (LevelVersion and/or UnitVersion) that changes; `tsnchange:output` points to a *TSNComponent* created or modified after the change event; the predicate `tsnchange:lowerChange` (inverse property `tsnchange:upperChange`) indicates another change event on a sub-element of the current described element that changes; and *isCausedBy* indicates the contextual reasons for the change (provided that the information is available on the LOD, on DBpedia for instance). We recommend the use of the *Linking Open Descriptions of Events Ontology* (LODE) [6] [Shaw et al., 2009] for the representation of (historical) events behind areas changes, because it is a domain-independent and a light weighted structure as remarked by [Hienert and Luciano, 2015].

A `XChange-Bridge` is defined by the tuple:

$$< t, input, output, lowerChange, upperChange, isCausedBy > \tag{8.1}$$

where t is the time instant when the change happens; *input* is the set of *TSNComponents* (Territories, Level and/or TUs) that are subject to change; *output* is the set of *TSNComponents* created or modified after the change event, *lowerChange* is a set of *Change* caused by the current described *Change*; *upperChange* indicates a set of upper *Change* that causes the current described *Change*; and *isCausedBy* indicates the contextual reasons for the change (provided that the information is available on the LOD). Using this tuple, any Territorial Change can be described. We show here how it allows us to dissociate the two `Split` subtypes: `Scission` (`Split` of a TU that causes its disapearance, i.e., *Derivation* case) and `Extraction` changes (`Split` of a TU that still exists after the change, i.e., *Continuation* case). The Figure 8.7 presents schematically the difference between these two change tags.

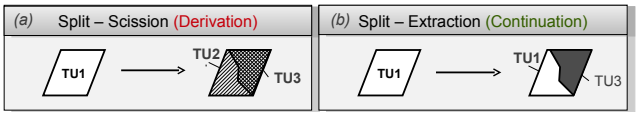


Figure 8.7 – Split of a TU - Scission and Extraction.

The *Scission* of the TU with identifier $TU1$, on the left side (a) of the Figure 8.7, is defined by the tuple:

$$\begin{aligned} Scission(TU1) \equiv \quad & < t, \{TU1\}, \{TU2, TU3\}, \{Disappearance(TU1), \\ & Appearance(TU2), Appearance(TU3)\}, *, ** > \end{aligned} \tag{8.2}$$

where $*$ represents the set of *upperChange* (not described here) and $**$ represents a resource (not described here) that explain why the current

6. http://linkedevents.org/ontology/

Scission change occurred, and:

$$\begin{aligned} &Disappearance(TU1) \equiv < t, \{TU1\}, \emptyset, \emptyset, Scission(TU1), ** > \\ &Appearance(TU2) \equiv < t, \emptyset, \{TU2\}, \emptyset, Scission(TU1), ** > \\ &Appearance(TU3) \equiv < t, \emptyset, \{TU3\}, \emptyset, Scission(TU1), ** > \end{aligned} \tag{8.3}$$

The *Extraction* of the TU with identifier $TU1$, on the right side (b) of the Figure 8.7, is defined by the tuple:

$$\begin{aligned} Extraction(TU1) \equiv \quad &< t, \{TU1\}, \{TU1, TU3\}, \\ &\{Contraction(TU1), Appearance(TU3)\}, *, ** > \end{aligned} \tag{8.4}$$

where :

$$\begin{aligned} &Contraction(TU1) \equiv < t, \{TU1\}, \{TU1\}, \emptyset, Extraction(TU1), ** > \\ &Appearance(TU3) \equiv < t, \emptyset, \{TU3\}, \emptyset, Extraction(TU1), ** > \end{aligned} \tag{8.5}$$

Two predicates of the *TSN-Change Ontology* express a chronological order in the succession of versions (of TU, Territory or Nomenclature) to go directly from one to another:

1. `hasNextVersion(x,y)` stands for "x has a next version y".
2. `hasPreviousVersion(x,y)` stands for "x has a previous version y".

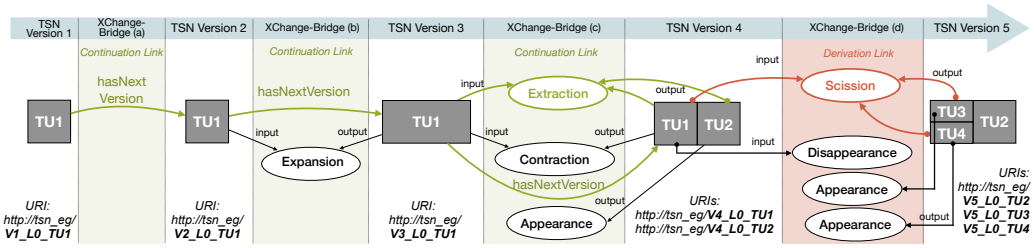


Figure 8.8 – TSN-Change Ontology TU Life line.

Such a filiation link is observed between two versions of a TSN's element only if the identity of the element is preserved after the change event. Considering that the identity is supported through a set of attributes values, this filiation link exists when the attributes that define the identity of the element have slightly or not changed between the two versions. [Del Mondo et al., 2010] uses the term *Continuation* link in this case. *Derivation* links also exist where, on the contrary, the identity of the TSN element is not maintained after the split event. In Figure 8.8, we suppose that the identity [7] of the TU is hold solely by the TU's identifier attribute (*e.g.*, label *TU1* in version 1 Figure 8.8). In case of a Derivation, in our model (contrary to [Harbelot et al., 2013]), no direct link `hasNextVersion` is created between the entities. For instance, there is no direct link between the *TU1* in version 4 and *TU3*, *TU4* in version 5 (*XChange-Bridge (d)*), because the identity of the input *TU1* is not maintained after the change (it changes its identifier). Conversely, the *XChange-*

7. We discuss identity issues page 124

Bridge (c) describes a split event as a structure change of type *Continuation* where the identity of the input *TU1* is maintained after the change, although the TU splits in two.

This figure shows, within the graph of versions of a TSN, nodes and links that draw the life line of one TU, and provides a view of the three main states (*creation*, *continuation*, *elimination*) as expected in such a model that addresses identity evolution issues [Harbelot et al., 2013].

8.5 Conclusion

We have presented, in this Chapter, the TSN Ontology for the description and publication on the LOD Web of any TSN hierarchical structure. We have underlined the importance of adopting a spatiotemporal approach when analyzing territories and have proposed to publish all the versions of a TSN rather than to provide users only with the latest one. Then, we have introduced the TSN-Change Ontology for the description of (boundaries) changes from one TSN version to it subsequent. Together, the TSN and TSN-Change ontologies constitute an innovative approach (combining the Change Bridge approach with the fluents one) for the modeling of territorial changes on the LOD Web through graphs we call *Multi-level Territorial Change Graphs*.

Chapter 9

Populating the TSN ontological model

9.1 Introduction

In this Chapter, we present our approach to populate our two ontologies, TSN and TSN-Change. We present the workflows and our methodology to populate automatically the TSN-Change ontology by taking into account the heterogeneity of the existing TSNs.

9.2 Workflows

There exists two different workflows within the Theseus Framework (see Figure 9.1): Workflow A creates the RDF catalog of all the TUs of the proceeded TSN versions; Workflow B creates RDF descriptions of the filiation links between the TSN versions. We present in this Chapter, the *TSN Semantic Matching Algorithm*, an adaptation of the Algorithm of [Plumejeaud, 2011] created to detect and semantically describe these filiation links, using the TSN-Change Ontology. At the end of the two workflows, while there are two distinct RDF outputs (TUs, levels descriptions and changes, filiation descriptions), all the created data are available from a single triplestore and are linked together because of the HTTP URIs assigned to each TSN's elements, URIs which are also referenced when describing the changes of these elements.

Both Workflows A and B take input files (Figure 9.1, *Input* box) that are:
– a tabular *metadata* file containing (*e.g.*, CSV, Excel©): metadata information on the TSN versions to publish as RDF *i.e.*, the names, acronym and description of the TSN and of its versions, the periods covered by the versions, the name and period of reference of the territories covered by the TSN versions; the list of TUs' attributes for which the algorithm will perform a comparison of the values from one version to another, as well as the weights assigned to each of these attributes; the value assigned to the three thresholds, run parameters of the algorithm. This tabular file is filled up by the expert of the TSN. (In the future, we have planned to design a GUI where the expert will be able to enter this information).
– a *data* shapefile for each version of the TSN containing the list of the TUs with their identifier in the version, their name, the language used for this name, their super-TU, and the level they belong to. For both the metadata and data input files, information comes from the TSN responsible authority. The shapefiles of the

different TSN versions have to be available online (most of the time, it is the case because of Open Data directives). In the future, we will extend the framework so that it can take as input other data formats than a tabular format (such as RDF data, more and more adopted by the SAs to publish their data). As it stands, the framework does not depend on a lot on the shapefile format since data from the shapefiles are immediately transferred in a PostGIS database and transform into RDF data.
– the OWL representation of the ontological model *i.e.*, the TSN and TSN-Change Ontologies. In Telechev and Le Rubrus [2013], further details are provided on the

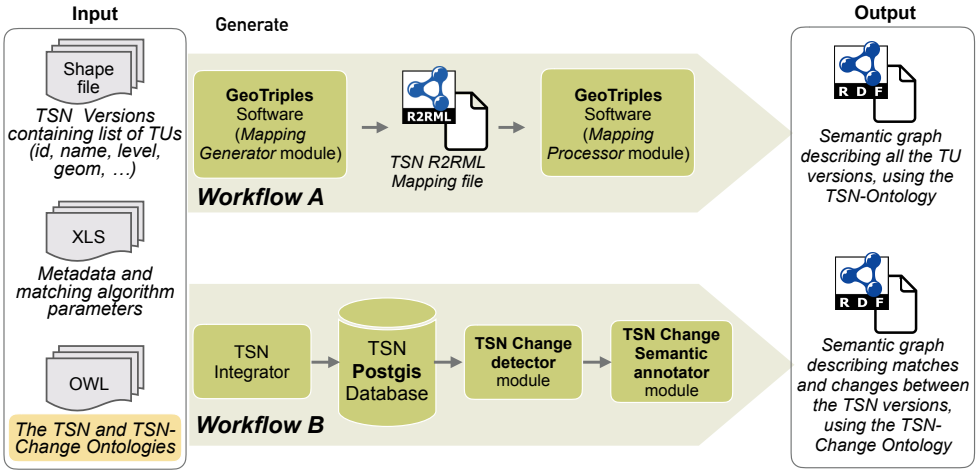


Figure 9.1 – The workflows of the Theseus Framework to populate semi-automatically the TSN ontological model.

requirements for the data and metadata files expected by our system. To transform these inputs (metadata and shapefiles) into RDF graphs (Figure 9.1 *Output* box) using the *TSN Ontology* and *TSN-Change* ontologies, our system runs two independent processes on data, described in the Sections 9.3 and 9.4 below.

9.3 Populating the TSN Ontology

In order to populate the TSN Ontology, we implement in the Theseus Framework a workflow that transforms each element composing the TSN versions (each version being described in a shapefile) into RDF triples (see Figure 9.1, Workflow (A)). It consists first of the alignment of the columns of the input shapefiles with the *TSN Ontology* concepts using a "mapping" file written in R2RML [1] which matches each column of the input shapefile with one concept of the TSN Ontology. This mapping file is generated using the *Mapping Generator* module of the GeoTriples Software. An example of one R2RML mapping file, for the NUTS TSN is available in the

1. "*R2RML is a language for expressing customized mappings from relational databases to RDF datasets*", source: `https://www.w3.org/TR/r2rml/`. The GeoTriples software "*supports the mapping languages R2RML and RML and extends them for modeling the transformation of geospatial data into RDF graphs.*", source `http://geotriples.di.uoa.gr/`

Appendix B.

Then, we automatically create the RDF graphs of all the versions of the processed TSN, using the *GeoTriples*[2] tool (*Mapping Processor* module) that reads the R2RMLmapping file and transforms the shapefile into RDF, using the information on the mapping file. We make these graphs available on-line from the SPARQL endpoint at `http://steamerlod.imag.fr//repositories/tsn`.

At this step, we have achieved the *Versioning* of all the elements composing the TSN versions. We follow the W3C Data on the Web Best Practices recommendation for the publication of data as LOD on the Web and especially, recommendations on Data versioning [Farias Lóscio et al., 2017]. Then, we provide, as required, a unique identifier for each element of the TSN, such as TUs, in order to differentiate between TUs of different versions (in particular, TUs that have undergone changes from one version to another while keeping the same identifier in the nomenclature). For instance, using the following patterns in Listing 9.1, each TSN feature has an URI that points to an immutable snapshot of the resource.

```
1  {base} TSN URI base (e.g., http://purl.org/steamer/nuts/ or @prefix nuts:)
2  {base}V{Accronym TSN Version} TSNVersion URI (e.g., nuts:V1999)
3  {base}V{Accronym TSN Version}_L{Accronym Level} LevelVersion URI
       (e.g., nuts:V1999_L2)
4  {base}V{Accronym TSN Version}_L{Accronym Level}_{Code TU} UnitVersion
       URI (e.g., nuts:V1999_L2_ES63)
```

Listing Code 9.1 – URIs' Patterns for the elements of a TSN.

Prior to the presentation of the second workflow that populate the TSN-Change Ontology, we present in more details in the next Section, the central element for the successful completion of this workflow: the *TSN Semantic Matching Algorithm* we have created in order to detect changes and filiation links between the element of a TSN throughout its versions.

2. `https://github.com/LinkedEOData/GeoTriples/wiki`

9.4 Populating the TSN-Change Ontology

9.4.1 Methodology

The major challenge we face when automating the detection and semantic description of TSN changes is the implementation of our concepts with regard to the heterogeneity of the input data. Indeed, the definition of the identity of TUs varies from TSN to TSN. In the context of spatial entities delimited by and under the control of a human group, one can wonder how far these entities, their geometries, names, population or land-cover can change before they are considered as new entities with a new identity? Is an area whose name is changed still the same area? Is an area that has merged with another area, still the same or has it disappeared although its name is preserved after the merge event? There are ontological questions that require the intervention of an expert who is expected to define in our framework, for each TSN, which attributes hold the identity of the TUs in the TSN, and in which proportion they can vary before TUs lose their identity.

The Theseus Framework uses the concept of *Identity* attached to each TU u, in an approach similar to the one of Huibing et al. [2005]. A weighting Function F, in Equation 9.1, is used to determine whether the identity of a TU u' in V' is preserved in the version V'' *i.e.*, a unit u' in V' matches a unit u'' in V''. Two TUs match if the function F returns a match score p greater than a global threshold β *i.e.*, there exists a filiation link of type *Continuation* between the two TUs, the identity of the TU u' is preserved in the version V''.

This weighting test allows for considering relevant C_a attributes, and their weights α_a that quantify how important each attribute is in the definition of the identity of a unit. These lists of C_a attributes and weights α_a are defined by an expert of the TSN, during a configuration step of the framework. S/He configures the framework according to h-er/is knowledge on the criteria for the evaluation of TU identity continuation and distributes the weights so that the sum of the r weights α_a equals 1 (see Equation 9.1).

$$\boxed{\begin{array}{l} F((\alpha_1, C_1), (\alpha_2, C_2), ..., (\alpha_r, C_r)) where \sum_r \alpha_k = 1 \\ F((\alpha_1, C_1), (\alpha_2, C_2), ..., (\alpha_r, C_r)) : \\ (V' \times V'') \rightarrow \mathbb{R}, (u', u'') \mapsto p, p \in [0, 1],\ \forall u' \in V' \text{and } \forall u'' \in V'' \end{array}} \tag{9.1}$$

$$
\begin{aligned}
&F(\alpha_1, identifier), (\alpha_2, geom), (\alpha_3, name), (\alpha_4, super) \\
&\qquad where \qquad \alpha_1 = 1/3, \alpha_2 = 1/3, \alpha_3 = 1/6, \alpha_4 = 1/6 \\
&F(\alpha_1, C_1), (\alpha_2, C_2), ..., (\alpha_r, C_r) : \\
&(V' \times V'') \rightarrow \mathbb{R}, \\
&(u', u'') \mapsto 1/3 * (\left|u'.identifier - u''.identifier\right|) \\
&+1/3 * (\left|u'.geom - u''.geom\right|) + 1/6 * (\left|u'.name - u''.name\right|) \\
&+1/6 * (\left|u'.super - u''.super\right|
\end{aligned}
\tag{9.2}
$$

Equation 9.2 shows an example of implementation of the weighting Function F based on the following TU's attributes: identifier (*identifier*), geometry (*geom*), designation (*name*), and super-TU (*super*) (please, see Chapter 10, Section 10.3.3 for a discussion of the values assigned to the weights α_a).

Not far from conflation approaches, this weighting test requires to measure first, for each of the TUs' attribute, how far they have changed from V' to V'' using a distance test. The framework contains a library of distance tests that could be enriched to processed new attributes or apply a new method (*e.g.*, there exists several method to measure the distance between two strings). We have written several distance tests, such as:

(1) a distance test on the *name* attribute (formalized in Equation 9.2 by the notation $|u'.name - u''.name|$), performed using the Levenshtein Distance [Levenshtein, 1966];

(2) a distance test between two geometries in V' and V'' ($|u'.geom - u''.geom|$ in Equation 9.2), calculated using the areal distance of Bel Adj Ali and Vauglin [1999]. It computes, for two geometries that intersect, the ratio of the shared area by the union of their areas. If the result of this test exceeds a spatial threshold $\varepsilon_{spatialFeature}$ (set by an expert of the TSN), the algorithm concludes that the geometries intersect but are not equal. We will see in the following subsection 9.4.2 that there exist several other solutions to compute a similarity score between two geometries.

Because the distance tests are input parameters of our algorithm, other distance tests, on other attributes than the ones presented in Equation 9.2 can be defined in order to address other kinds of TSN and feature. For instance, we are currently experiencing a new configuration of the framework in order to describe changes in the Corine Land Cover data sets (available online from the Copernicus web site `https://land.copernicus.eu/pan-european/corine-land-cover`). We come back to this perspective in Section 12.2.

To summarize, in order to adapt to any TSN hierarchical structure (structure defined in Section 2.2), we make our framework configurable. This is a step towards a generic framework, as we propose in the perspective to these works 12.2. There are several sets of parameters to adjust for the matching of the TSN versions, depending on characteristics of the geospatial data:
(1) a list of TUs' attributes considered for matching;

(2) several distance tests on each attributes;
(3) the weights carried by each attribute of a TU (*e.g.*, $\alpha_{geom}, \alpha_{name}$);
(4) spatial and global thresholds used during the test of matching.

9.4.2 The TSN Semantic Matching Algorithm

The *TSN Semantic Matching Algorithm* aims at automating the population of the *TSN-Change Ontology*. This algorithm compares the two successive versions V' and V'' of a TSN in order to determine underlying similarities and changes. It takes as input two lists of TUs each corresponding to version V' and V'' of the TSN. For instance, Tables 9.1 and 9.2 show two lists for the NUTS TSN. For each TU is given, at least, the following list of attributes: identifier of the TU in the TSN, name, level in the TU hierarchy, super-TU and geometries (as Multi-polygon geometries). This list of attributes can be expanded (as we will see later). From these two lists of TUs, the algorithm builds the RDF graph which describes TSN version-to-version filiation relations according to the *TSN-Change Ontology*.

TU id	***TU name***	***TU level***	***TU super***	***TU geom***
ES6	SUR	1	ES	multipolygon
ES63	Ceuta y Melilla	2	ES6	multipolygon

Table 9.1 – Extract of the list of TUs of the NUTS version 1999 (V') (Algorithm 1 input).

TU id	***TU name***	***TU level***	***TU super***	***TU geom***
ES6	SUR	1	ES	multipolygon
ES63	Ciudad Autonoma de Ceuta	2	ES6	multipolygon
ES64	Ciudad Autonoma de Melilla	2	ES6	multipolygon

Table 9.2 – Extract of the list of TUs of the NUTS version 2003 (V'') (Algorithm 1 input).

In order to explain how our algorithm works, a concrete example of change that took place in the NUTS between versions 1999 (V') and 2003 (V'') is presented. Figure 9.3 shows an abstract representation of the two hierarchies to which belong the TUs ES63 and ES64 in the version 1999 and/or 2003: at Level 1, in both hierarchies the super-TU is ES6; and at Level 3, their sub-TUs differ. In the version 1999, at Level 2 (*i.e.*, basic regions in Eurostat terminology), the two Spanish enclaves in North Africa, *Ceuta and Melilla*, were considered as one TU (with identifier ES63), part of a Spanish province. In 1995, both Ceuta and Melilla became Autonomous cities of Spain, requiring a redistricting of the Spanish regions. It resulted in a split of the TU ES63 in two TUs, named *Ciudad Autónoma de Ceuta* (with identifier ES63) and *Ciudad Autónoma de Melilla* (with identifier ES64), in the NUTS Version 2003. Also the sub-TUs of ES63 (ES631 and ES632) in the NUTS version 1999 both change their identifiers. The identifiers of TUs ES631 and ES632 became

ES630 and ES640 respectively. These enclaves are spatially distant though considered as one single TU in Version 1999. The multi-polygon geometry of this TU is presented in red, on the left map, Figure 9.2. The right side of Figure 9.2 shows the two TUs ES63 and ES64, in the version 2003.

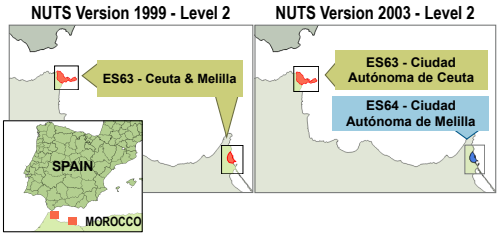


Figure 9.2 – Ceuta and Melilla TUs in the NUTS TSN.

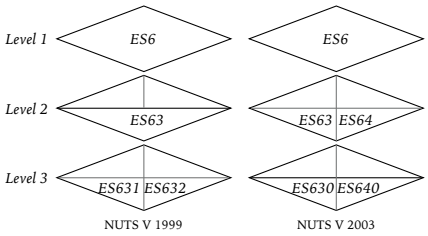


Figure 9.3 – NUTS Hierarchy of levels linked to the example of Figure 9.2.

Algorithm 1 presents the sequence of the algorithm steps, as an overview. We provide detailed explanations and descriptions of all the steps of this Algorithm 1, below, in this Section. The Table 9.3 lists the constants used in Algorithm 1.

Constants
$INTERSECTION$ = -1
$EMPTY = 0$
$MATCH = 1$
$MERGE = 2$
$SPLIT = 3$
$REDISTRIBUTION = 4$
$IENTIFICATION\_RESTRUCTURATION = 5$

Table 9.3 – TSN Semantic Matching Algorithm Constants.

Input: A list *V'* of m TUs *u'* and a list *V"* of n TUs *u"*, all TUs being at one Level l.
Output: RDF graph (*RDFout*) describing Similarities and Changes on TUs

Begin
  Step 1 - Identification of Feature changes
  for *i in m* **do**
    for *j in n* **do**
      SpatialMatch[*i*][*j*] ← `SpatialMatchTest` ($u'_i, u''_j, \varepsilon_{spatialFeature}$) ;
      foreach $u'_i.attribute$ **do**
        GlobalMatch[param.attribute][*i*][*j*] ← `DistanceTest` ($u'_i.attribute, u''_j.attribute$) ;
      end
      GlobalMatch[param.global][*i*][*j*] ← `WeightingTest` (u'_i, u''_j, C_a, β) ;
      (Also fills the SuperMatch$_l$ *map at current level l)*
    end
  end
  Step 2 - Identification of clusters of changes (Structure Changes)
  for *i in m* **do**
    for *j in n* **do**
      (Identifying each clusters)
      if SpatialMatch[*i*][*j*] == $INTERSECTION$ **then**
        StructureChangeK[*i*][*j*] ← `FindStructureChange` ($u'_i, u''_j, \varepsilon_{spatialStructure}$) ;
        (Finds StructureChange events and Assigns to each StructureChange an ID, registered in StructureChangeK[i][j]. Also refines SpatialMatch*[i][j] type (MERGE, SPLIT, ...))*
      end
    end
  end
  Step 3 - Association of a meaning with clusters of changes
  for *i in m* **do**
    for *j in n* **do**
      if SpatialMatch[*i*][*j*] > $MATCH$*;*
      (means that a StructureChange event is registered, see the Table of constants 9.3.)*;*
      then
        if GlobalMatch*[param.global]*[*i*][*j*] == $MATCH$ **then**
          SpatialMatch[i][j]← `RefineChangeType` (
          Continuation, SpatialMatch*[i][j]*) ;
          (Refines SpatialMatch*[i][j] type (INTEGRATION, EXTRACTION, ...))*
        end
        else
          SpatialMatch[i][j]← `RefineChangeType` (
          Derivation,SpatialMatch*[i][j]*) ;
          (Refines SpatialMatch*[i][j] type (FUSION, SCISSION, ...))*
        end
      end
    end
  end

```
41
42  Step 4 - Creation of the RDFOUT file
43  for i in m do
44      for j in n do
45          if GLOBALMATCH[param.global][i][j] == MATCH then
46              RDFOUT← CreateTUGenealogyLink(u'_i, u''_j);
47          Step 4.1 - Description of the StructureChange event
48          if STRUCTURECHANGEK[i][j] > 0 then
49              RDFSTRUCTURECHANGE←
                CreateChangeNode(TsnchangeOntology.StructureChange,
                SPATIALMATCH[i][j], u'_i, u''_j);
50              if u'_i.super != null then
51                  RDFSTRUCTURECHANGE←
                    ChainingChangeNode(RDFSTRUCTURECHANGE, u'_i.super);
52              else
53                  RDFSTRUCTURECHANGE←
                    ChainingChangeNode(RDFSTRUCTURECHANGE, u'_i.level);
54              RDFOUT← Write( RDFSTRUCTURECHANGE);
55          Step 4.2 - Description of the FeatureChange event
56          foreach u'_i.attribute do
57              if GLOBALMATCH[param.attribute][i][j] == EMPTY;
58              (means the attribute has changed.) then
59                  RDFFEATURECHANGE←
                    CreateChangeNode(TsnchangeOntology.FeatureChange,
                    param.attribute, u'_i, u''_j);
60              if RDFSTRUCTURECHANGE != null then
61                  RDFFEATURECHANGE←
                    ChainingChangeNode(RDFFEATURECHANGE,
                    RDFSTRUCTURECHANGE);
62                  (l − 1 means level up from l)
63              else
64                  if u'_i.super != null then
65                      RDFFEATURECHANGE←
                        ChainingChangeNode(RDFFEATURECHANGE, u'_i.super);
66                  else
67                      RDFFEATURECHANGE←
                        ChainingChangeNode(RDFFEATURECHANGE, u'_i.level);
68                  (l − 1 means level up from l)
69              RDFOUT← Write(RDFFEATURECHANGE);
```

Algorithm 1: The TSN Semantic Matching Algorithm (SMA).

The TSN Semantic Matching Algorithm (SMA) 1 detects similarities and changes between two successive versions of a TSN, achieving traversal level by level, starting by the highest one (*e.g.*, Territory Level), down to the lowest (*e.g.*, Level 3 in the NUTS). The two input versions of TSN are used to build some temporary matrices that store knowledge about the TUs similarities and changes, before building the output RDF graph.
The principle of the TSN SMA is to compare each attribute and measure their difference (*i.e.* their distance) (function *DistanceTest* called at line 7 in Algorithm 1 and detailed in Algorithm 2).

For instance, for the *name* attribute, we use one of the standard approaches [Santos et al., 2018] that consists of an *Edit Distance* measurement using the Levenshtein Distance [Levenshtein, 1966], the most common method for measuring the distance between two sequences of text [McKenzie et al., 2014].

Input: Parameters to the StringMatchTest function on the TUs' name attribute: $u'_i.name, u''_j.name$
1 **if** `StringMatchTest(`$u'_i.name, u''_j.name$`)` **then**
2 | GlobalMatch[param.name][i][j] $\leftarrow MATCH$
3 **end**
Algorithm 2: The TSN SMA *StringMatchTest* function (that encapsulates a *DistanceTest* between the string names) (called at line 7 in Algorithm 1) on TUs' name.

The function `StringMatchTest` (see Algorithm 2, and the JAVA code of this function in Listing 9.2) encapsulates an implementation of the Levenshtein distance Algorithm. This method computes a distance score and transforms the distance result into a boolean value indicating whether the compared string is very similar to the original one. Strings are considered here to be similar if no more that 1/3 of the characters in the original string must be replaced to obtain the compared one.

The Levenshtein Distance corresponds to the number of editions (*i.e.*, insertions, removals, replacements) required to transform one string into another. We convert the distance metric to a similarity score, that equals 1 when strings are equal, tends to 1 if they are very similar, 0 else. This similarity value is computed by dividing the Levenshtein distance by the greatest possible Levenshtein distance for the given strings (*i.e.*, the length of the longer string), and subtracting the resulting value from 1 [Recchia and Louwerse, 2013]. There are alternatives to the Levenshtein similarity as demonstrated in [Recchia and Louwerse, 2013; Santos et al., 2018] (*e.g.*, the Damerau-Levenshtein distance that counts the number of transpositions between two strings, the Hamming Distance, the Jaro Distance...). As data we compare version-to-version are in the same language and are in standardized data sets (official geographic hierarchy), we have observed that the Levenshtein Distance on the processed data set (NUTS versions and SAU versions) are highly consistent.

```
1public boolean StringMatchTest(String original, String compared) {
2    if (original == null || compared == null){
3        return false;
4    }
```

```
5     original = original.trim().toLowerCase();
6     compared = compared.trim().toLowerCase();
7     if (original.length() == 0 || compared.length() == 0){
8         return false;
9     }
10    int[] cost;
11    int[] back;
12    cost = new int[original.length() + 1];
13    back = new int[original.length() + 1];
14    for (int i = 0; i <= original.length(); i++){
15        cost[i] = i;
16    }
17    for (int j = 0; j < compared.length(); j++) {
18        final int[] t = cost;
19        cost = back;
20        back = t;
21        cost[0] = j + 1;
22        for (int i = 0; i < original.length(); i++) {
23            final int match = (original.charAt(i) == compared.charAt(j)) ? 0 : 1;
24            cost[i + 1] = Math.min(back[i] + match, Math.min(cost[i] + 1, back[i +
                  1] + 1));
25        }
26    }
27    return cost[original.length()] <= original.length() / (original.length() > 6 ?
           3 : 4);
28 }
```

Listing Code 9.2 – An implementation of the Levenshtein distance algorithm that allows to make an approximate comparison of the matching between two given strings (Author: Christine Plumejeaud).

Regarding the distance test between two geometries, we use the areal distance [Bel Adj Ali and Vauglin, 1999] (test implemented within the function *Spatial-MatchTest* called at line 5 in Algorithm 1 and detailed in Algorithm 3). This distance computes, for two TUs u' and u'' that intersect, the ratio of the shared area by the union of their areas (see the PostGIS implementation of this test in Listing 10.7). When geometries are equal or almost, this ratio is close to 1, else it tends to a 0 value. A threshold ε is used as a parameter to express how far geometries can be distant before being seen as different, which is equivalent to define the proportion of area that two geometries have to share to be considered as identical. We have to tolerate few small differences between the geometries of two versions, because of errors in the data set after a human modification, or sometimes because when a polygon in the network changes, the geometry of the neighbors polygons are impacted whereas they have not changed in reality. An equality test like the ST_equals of OGC standard is too strict to identify a non-exact matching between two geometries. Our solution allows for non strict equality between the geometries. And, because it computes a ratio between TUs that are at the same scale (in the same Coordinate Reference System), it is efficient at each scale, even on very small

units.

Input: Parameters to the SpatialMatchTest function: $u'_i, u''_j, \varepsilon_{spatialFeature}$
if $u'_i.geom \cap u''_j.geom \neq \varnothing$ **then**
  SpatialMatch[i][j] $\leftarrow INTERSECTION$;
  $diff \leftarrow u'_i.geom - u''_j.geom$ *(simplified representation of the Areal Distance test)*;
  if $diff \leq \varepsilon_{spatialFeature}$ **then**
    SpatialMatch[i][j] $\leftarrow MATCH$;
  end
end

Algorithm 3: The TSN SMA *SpatialMatchTest* function (called at line 5 in Algorithm 1).

As noticed in [Harbelot et al., 2013], a major issue arises: *"How far can an entity vary before losing its identity?"*. To address this question, we propose the weighting Function (Equation 9.1) explained previously in the methodology section 9.4.1. The matching algorithm applies a weighting between the result of each distance measurement (that computes a value between 1 (equality) and 0 (non equality)) of the r attributes and computes a result between 0 and 1, 1 when the TU u' in version V' is exactly the same as the TU u'', in version V'' (function *WeightingTest* called at line 8 in Algorithm 1 and detailed in Algorithm 4).

Input: Parameters to the WeightingTest function: u'_i, u''_j, C_a, β
$C_a[param.geom] = \alpha_{geom} = 0.4$;
$C_a[param.id] = \alpha_{id} = 0.4$;
$C_a[param.name] = \alpha_{name} = 0.1$;
$C_a[param.super] = \alpha_{super} = 0.1$;
valueGlobal = 0;
foreach $u'_i.attribute$ **do**
  if GlobalMatch$[param.attribute][i][j] > 0$ **then**
    $valueGlobal = valueGlobal + 1 * C_a[param.attribute]$;
  end
end
if $valueGlobal \geqslant \beta$ **then**
  GlobalMatch[param.global][i][j] $\leftarrow MATCH$;
  SuperMatch$_l$.put $(u'_i.DBid, u''_j.DBid)$;
  (Register in a map SuperMatch all the TUs that do not change their identity, in order to access this information when traversing the lower levels)
end

Algorithm 4: The TSN SMA *WeightingTest* function (called at line 8 in Algorithm 1).

The TSN SMA 1 is applied to fill up two matrices made of n rows listing units u' and m columns listing units u''. The first matrix named SpatialMatch contains the values of the *spatial match test* between geometries, or 0 if they do not intersect. The second matrix named GlobalMatch contains the result of the *weighting test* (see Equation 9.1) between all criteria for each couple of units that intersect, 0 if they do not intersect. In this first step, the algorithm can determine which units have

not changed by comparing the computed values with the global threshold β given in the configuration. Conversely, units that have changed can be identified, and are associated with the corresponding *FeatureChange* tag from the *TSN-Change* Ontology. When applied to our case study in Spain, at the end of this step, it has been detected that the TU $ES63'$ intersects the TU $ES63''$, but that their geometries are not equal. Also, the appearance of the TU $ES64''$ is detected. Still, the set of changes have to be grouped together under *StructureChange* tags (from the *TSN-Change* Ontology) such as *Split*.

The second traversal of SpatialMatch aims at building k sets of TUs involved in the same *StructureChange* (each one being denoted as *StructureChangeK*, *i.e.*, "K" stands for "Knowledge") (function *FindStructureChange*, Algorithm 1,more details about this function can be found in [Plumejeaud, 2011]). By reading the SpatialMatch matrix, the function *FindStructureChange* (Algorithm 1, line 14) tries to determine precisely if only one TU's geometry has changed or more neighboring TUs with it. For each *StructureChangeK* event, the algorithm builds through iterations two sets: the beforeK set of TUs preceding the event, and the afterK set of TUs succeeding the event. For instance, if a TU $u'1$ splits in two TUs ($u''2$ and $u''3$) from version V' to V'', the SpatialMatch matrix will have registered two intersections for the TU $u'1$ and will conclude that a *StructureChange* of type *Split*. If two TUs $u'1$ and $u'2$ merge into one TU $u''3$ from version V' to V'', the SpatialMatch matrix will have registered two intersections for the TU $u''3$ and will conclude that a *StructureChange* of type *Merge*. Figure 9.4 presents schematically an example of the iterations of the algorithm, in case of a *Split* event. An equality test is performed on the external geometries of the two sets of TUs before and after change (*e.g.*, equality test between the external geometry of $u'1$ and external geometry of the union of $u''2$ and $u''3$) in order to validate the aggregation of several TUs changes under one semantic tag (*e.g.*, split, merge, redistribution), with a tolerance threshold $\varepsilon_{spatialStructure}$.

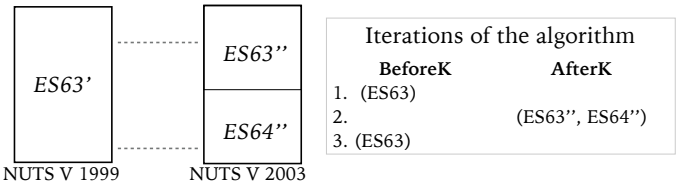


Figure 9.4 – Iterations of the SMA Algorithm in order to find a set of TUs involved in a same *StructureChange*.

The first iteration starts with the first TUs that have an areal distance different from 0 in SpatialMatch. For instance, on the TUs $ES63'$ and $ES63''$ the score in SpatialMatch is of 0.5. Thus, a set named beforeK is initialized with the TU $ES63'$ identified as a TU that intersects another one, here the TU $ES63''$. Starting from this seed, here $ES63'$, the second iteration adds to the set named afterK the whole set of TUs that intersects $ES63'$ in V'' ($ES63''$, $ES64''$ here) (an information stored in the SpatialMatch matrix). Conversely, starting from afterK, the third iteration checks if any TU of afterK intersects others in V'.

Here, $ES63''$ and $ES64''$ intersect $ES63'$ (information in SPATIALMATCH also). And so on, until the two sets, in BEFOREK *or* AFTERK, remain identical between two iterations. At the end of the iterations, an equality test on the external geometries of the two sets BEFOREK and AFTERK is performed, in order to validate the result (with a tolerance $\varepsilon_{spatialStructure}$).

The cluster of changes *StructureChangeK* can be identified by the type of event: if BEFOREK has only one element, the event is a division of a TU (`Split`); when AFTERK has only one element, the event is a `Merge`; otherwise, it is a `Redistribution`. Because in our example, BEFOREK contains one element ($ES63'$), the territorial event here is identified as a *Split* event. After the discovery of structure changes, the third step of the algorithm aims at distinguishing between *Continuation* and *Derivation* cases in the *StructureChange* model. For this purpose, the reading of the kind of *FeatureChange* (in the matrix GLOBALMATCH) that affects each TU being involved in territorial events helps to refine the label given to a structure change, depending on whether the involved TUs have change their identity or not (Function *RefineChangeType*, Algorithm 1, more details about this function can be found in [Plumejeaud, 2011]). For instance, on the basis of the *TSN-Change Ontology*, we distinguish between two `Split` subtypes: `Scission` (`Split` of a TU that causes its disappearance, *i.e.*, *Derivation* case) and `Extraction` (`Split` of a TU that still exists after the change, *i.e.*, *Continuation* case) (see Figure 8.7).

The fourth step of the algorithm transforms the output matrices into semantic RDF graphs, using concepts of the *TSN-Change Ontology*. This step consists of the following operations:

1. Linking TUs that do not change over time or change while maintaining their identity, using a triple representation in which the predicate `tsnchange:hasNextVersion` connects the two identifiers of the TU in V' and V'' (*e.g.*, `V1999_L2_ES63 hasNextVersion V2003_L2_ES63`) (function *CreateTUGenealogyLink* called at line 46 in Algorithm 1 and detailed in the Algorithm 5).

Input: Parameters to the CreateTUGenealogyLink function: u'_i, u''_j
1 RDFU'← `CreateResource(GetUri(`u'_i`))`;
2 RDFU"← `CreateResource(GetUri(`u''_j`))`;
3 RDFU'← `AddProperty(`*TsnchangeOntology.hasNextVersion,* RDFU");
4 RDFU"← `AddProperty(`*TsnchangeOntology.hasPreviousVersion,* RDFU');
5 RDFOUT← `Write(`RDFU');
6 RDFOUT← `Write(`RDFU");
Algorithm 5: The TSN SMA *CreateTUGenealogyLink* function (called at line 46 in Algorithm 1).

2. Building a *XChangeBridge*, each time a *StructureChange* or *FeatureChange* is found (function *CreateChangeNode* called at line 49 and 59 in Algorithm 1, detailed in Algorithm 6). Within these *Change Bridge*, we describe the V' features that come as *inputs* and the V'' features that come as *outputs* of the change node (using predicates `tsnchange:input`, `tsnchange:output`) (function *CreateChangeBridge* called at line 4 and 10 in Algorithm 6, detailed in Algorithm 7). We also charac-

terize the nature of the territorial changes (using the typology of changes in the *TSN-Change Ontology*, see Figure 8.5).

Input: Parameters to the CreateChangeNode function: current change's type (StructureChange/FeatureChange), SpatialMatch[i][j], current attribute Compared, u'_i, u''_j

```
1  if StructureChange then
2  |  RDFstructureChange← CreateChangeNodeURI(TsnchangeOntology.StructureChange, SpatialMatch[i][j], u'_i, u''_j);
3  |  RDFstructureChange← AddProperty(RDF.type, SpatialMatch[i][j]);
4  |  RDFstructureChange← CreateChangeBridge(RDFstructureChange, RDFu', RDFu")
5  end
6  else
7  |  (FeatureChange described)
8  |  RDFfeatureChange← CreateChangeNodeURI(TsnchangeOntology.FeatureChange, attributeCompared, u'_i, u''_j);
9  |  RDFfeatureChange← AddProperty(RDF.type, attributeCompared+"Change");
10 |  RDFfeatureChange← CreateChangeBridge(RDFfeatureChange, RDFu', RDFu")
11 end
```

Algorithm 6: The TSN SMA *CreateChangeNode* function (called at line 49 and 59 in Algorithm 1).

Input: CreateChangeBridge function parameters: RDFcurrentChange, $RDFu', RDFu''$

```
1  RDFu'← AddProperty(TsnchangeOntology.inputUnitVersion, RDFcurrentChange);
2  RDFu"← AddProperty(TsnchangeOntology.outputUnitVersion, RDFcurrentChange);
3  RDFcurrentChange← AddProperty(TsnchangeOntology.unitVersionBefore, RDFu');
4  RDFcurrentChange← AddProperty(TsnchangeOntology.unitVersionAfter, RDFu");
```

Algorithm 7: The TSN SMA *CreateChangeBridge* function (called at line 4 and 10 in Algorithm 6).

3. Chaining the changes that propagates through the partition levels (function *ChainingChangeNode* called at line 51, 53, 61, 65 and 67 in Algorithm 1 and detailed in Algorithm 8).

Input: ChainingChangeNode function parameters RDFcurrentChange. Optional parameters: RDFparentChange, $u'_i.super, u'_i.level$

```
1  if RDFparentChange != null then
2  |  RDFparentChange ← AddProperty(TsnchangeOntology.lowerChange,
   |   RDFcurrentChange);
3  |  RDFcurrentChange ← AddProperty(TsnchangeOntology.upperChange,
   |   RDFparentChange);
4  end
5  else
6  |  (Linking the RDFcurrentChange to an element directly above in the territorial
   |   hierarchy)
7  |  if u'_i.super != null then
8  |  |  (Inform the Super TU of u' that its Sub-TU u' has changed)
9  |  |  RDFsubUnitChange←
   |  |   CreateChangeNode(TsnchangeOntology.FeatureChange, SubUnitChange,
   |  |   u'_i.super, null);
10 |  |  RDFsubUnitChange← ChainingChangeNode(RDFsubUnitChange,
   |  |   RDFcurrentChange);
11 |  end
12 |  else
13 |  |  (Inform the Level of u' that the current TU u' has changed)
14 |  |  RDFlevelChange← AddProperty (RDF.type,
   |  |   TsnchangeOntology.StructureChange); RDFlevelChange←
   |  |   CreateChangeBridge (u'_i.level, u''_j.level, RDFchangeGraphEntryPoint,
   |  |   RDFlevelChange); RDFcurrentChange←
   |  |   ChainingChangeNode(RDFlevelChange, RDFcurrentChange);
15 |  end
16 end
```

Algorithm 8: The TSN SMA *ChainingChangeNode* function (called at line 51, 53, 61, 65 and 67 in Algorithm 1).

By always informing a super element of the changes undergone by its sub-elements, using information on the TSN hierarchy (as shown in Figure 8.4), our algorithm for hierarchy matching always constructs a chain of change nodes that starts from the TUs change node at the lowest level and reaches the parent elements that are *NomenclatureVersion* elements, in version V' and V''. Thus, the parent change node of a *Multi-level Territorial Change Graph* is always of type `StructureChange` and takes as *input* and *output*, resources that are instances of the class *NomenclatureVersion* (the highest versioned element in the NUTS hierarchical structure, see Figure 8.4). This way, a machine can recognize the entry points (*i.e.*, parent nodes) to the *Multi-level Territorial Change Graphs* (see Figure 9.5 top of the (b) graph).

For the chaining of changes, we use two predicates of the *TSN-Change Ontology* `upperChange` and `lowerChange` (bold arrows in Figure 9.5). Please note that for the readability of Figures, only one of the predicates is used at a time, whereas in each case, the inverse property is drawn using our algorithm.

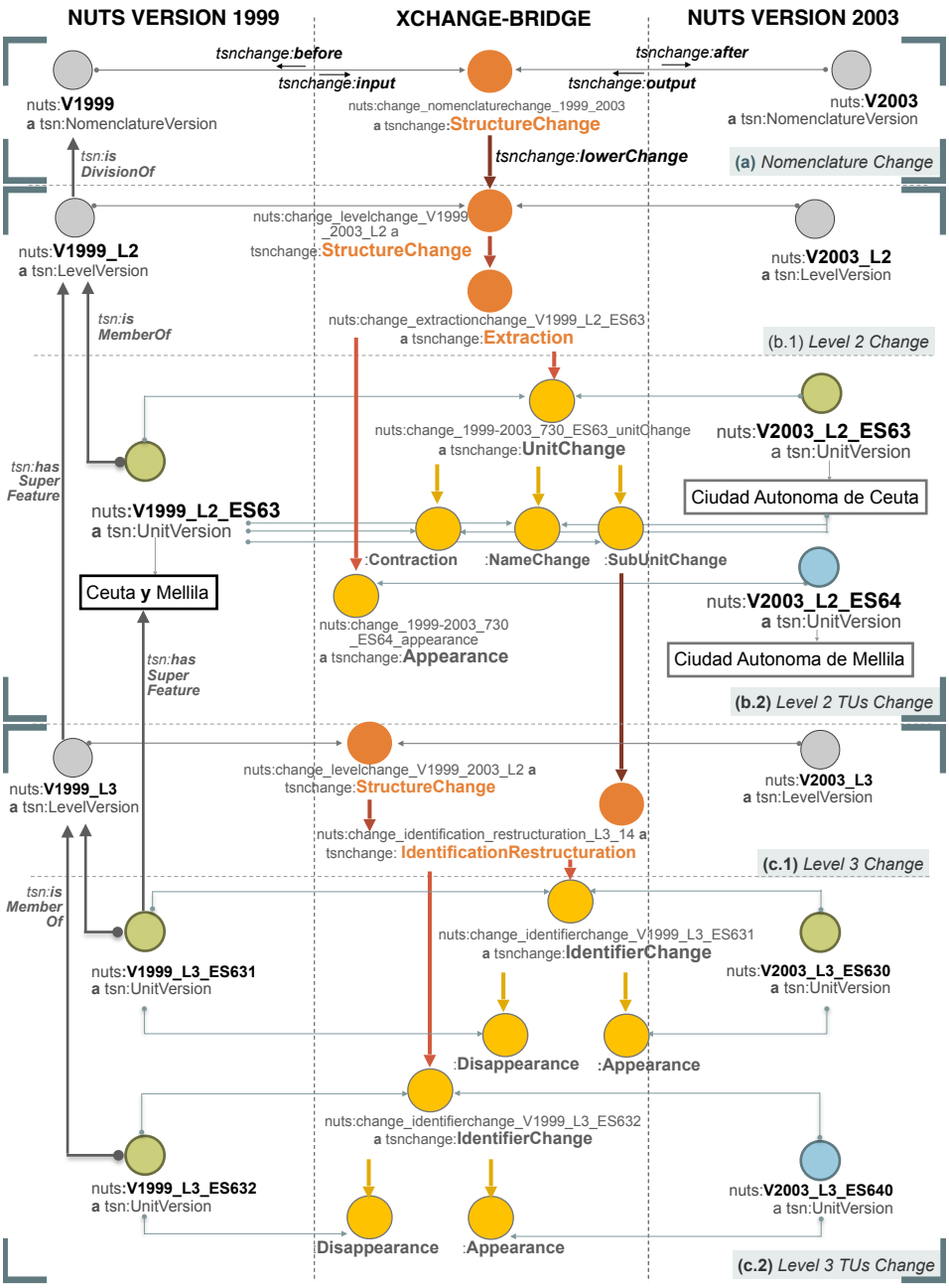


Figure 9.5 – Example of a multi-level change graph created with Theseus, linked to the example Figure 9.2.

Figure 9.5 shows the simplified *Multi-level Territorial Change Graphs* of the change of the TU ES63, created using our algorithm. The NUTS Version 1999 is represented on the left side of the Figure 9.5 (please, note that the TUs and levels' hierarchy is represented only for this version, to provide the reader with an overview of this part of the graph); the NUTS Version 2003 on the right side; and the territorial change is described in the middle of the figure. The *ES63 Change graph* is

designed as follows:
(a) at the top of the graph, the parent change node is of type *StructureChange* (because this *Split* event impacts several TSN features). The *NomenclatureVersion* resource *NUTS Version 1999* comes as an *input* resource to this change node, and the *NUTS Version 2003* comes as an *output* resource, as the TSN versions themselves change after this event on ES63.
(b.1) this parent change node has a lower change at *Level 2* (linked to it using the predicate *tsnchange:lowerChange*) labeled *Extraction* (*i.e.*, *Split* change where *Continuation* of the input TU is observed after the change, see Figure 8.7).
(b.2) each modification on each impacted feature, caused by this *Extraction* event is then represented, thanks to *FeatureChange* nodes (see the *FeatureChange* sub-types in Figure 8.5). Here, the split of the TU *ES63* results in: the *Appearance* of the TU *ES64* (*Ciudad Autónoma de Melilla*) ; several changes on the TU *ES63* (all these changes are grouped under one *UnitChange* node): its *Contraction*, *NameChange*, and also the change of its sub-units. This node *SubUnitChange* is essential: it chains the change nodes through the levels.
(c.1) then, by visiting the $ES63'$ node, one may directly know the changes its sub-TUs have undergone, without having to go through all the list of its sub-TUs to check if they have changed or not. By following this *SubUnitChange* node, one discovers that the sub-TUs of $ES63'$ change their identifier (*IdentificationRestructuration*) after the *Extraction* event.
(c.2) each modification on each impacted feature, caused by this *IdentificationRestructuration* event is then represented. Here, it consists of several *IdentifierChange* descriptions, for each of the sub-TUs of $ES63$ that changes its nomenclature identifier (also called *code*).

The following Subsection presents the workflow implemented in the Theseus Framework to populate the TSN-Change ontology, using the TSN SMA.

9.4.3 The Workflow

The second workflow implemented in the Theseus Framework consists in performing the matching of all the TSN elements, version-to-version, and describing their changes by means of RDF triples using the concepts of the *TSN-Change* Ontology (see Figure 9.1, Workflow (B)), and an implementation of the TSN SMA.

The *TSN SMA* has been implemented in two Java modules (see modules *TSN Change Detector* and *TSN Change Semantic Annotator*, Workflow (B), Figure 9.1).

The *TSN Change Detector* module takes as input a tabular metadata file, set up by an expert of the TSN with the list of the parameters of the TSN SMA Algorithm, as defined in the Methodology Section 9.4.1. This file contains: the list of TU's attributes to be considered during the matching tests on TUs, the weight assigned to each attribute, the value of the three thresholds. In order to compute the changes between two versions of a TSN, a PostGIS spatial database is built.

Using the *TSN Integrator* module we have developed, we insert in the *PostGIS*

spatial database of the framework, the shapefiles of the TSN, the metadata description of the TSN (contained in an tabular file) and the algorithm parameters values, set by the expert (in an tabular file also). Using the Jena Plugin *Schemagen*[3], we generate a Java representation of the TSN ontological model (all the concepts are translate into Java constants to be used in our Java programs).

Then, the *TSN change Detector* module accesses the PostGIS spatial database of the framework, and according to the values of the weights and thresholds set by the expert in the CSV file, it computes similarities measures (registered in temporary matrices).

The *TSN Change Semantic Annotator* reads the matrices and according to the values, assigns a TSN-Change tag to each change detected, and constructs the RDF change bridges between all the elements of the two TSN's versions processed, using the Apache-Jena RDF API[4] and the Java representation of the TSN and TSN-Change ontologies.

At the end of Workflow B, the TSN RDF history graph is created. This graph is available online from the SPARQL endpoint at `http://steamerlod.imag.fr//repositories/tsn` (please note that the *SPARQL Query* GUI provided by our GraphDB triplestore, at `http://steamerlod.imag.fr/sparql`, may be more convenient to human tests).

9.5 Conclusion

In [Wiemann and Bernard, 2016], the authors describe their approach for spatial data fusion and/or versioning management of network data in Open Street Map (OSM) data sets. In an approach very similar to ours, they sum several weighted similarity measurements (the Damerau–Levenshtein distance to calculate the distance between road names, the Hausdorff distance, the angle difference and the length difference between poly-line geometries of the network) and construct RDF graphs describing changes over time, in OSM data sets. They use the ChangeSet vocabulary[5] to describe the delta between two versions of a resource, in terms of *additions* and *removals* only. Consequently, the description of changes are very limited, as the aim is more to track, feature by feature, who has modified a feature and how (*e.g.*,the change of the type of a road from "footway" to "path" will be described as a removal of the "footway" type and addition of the new type "path"). This vocabulary does not fit our objective of precisely describe changes in TSN that may impact several features at the same time (*e.g.*, merge of two features in one). Our TSN-Change Ontology inherits from the Provenance, Authoring and Versioning Ontology[6], and contains more tags resulting in an accurate description of feature and spatial structure changes.

Because the distance tests are input parameters to our algorithm, other distance

3. `https://jena.apache.org/documentation/tools/schemagen-maven.html`
4. `https://jena.apache.org/`
5. `http://purl.org/vocab/changeset/schema#`
6. `http://purl.org/pav/`

tests, on other attributes than the ones presented in Equation 9.2 can be defined in order to address other kinds of TSN and feature. For instance, we are currently experiencing a new configuration of the framework in order to describe changes in the Corine Land Cover data sets (available online from the Copernicus web site `https://land.copernicus.eu/pan-european/corine-land-cover`). We come back to this perspective in the Chapter 12.

In order to adapt in the future to any TSN hierarchical structure in the world, we make our framework configurable and expandable. We propose, in Section 12.2, a solution to make our framework more generic.

So far, the *Theseus Framework* has been used on 3 different TSNs *i.e.*, 3 TSN RDF history graph have been created:
– the NUTS TSN (from Eurostat) versions 1999, 2003, 2006, 2010. The output graph (containing all the TUs and changes descriptions) is available online at `http://purl.org/steamer/nuts` (entry points of the change graphs are available at `http://purl.org/steamer/nuts/change_nomenclaturechange_1999_2003`; `http://purl.org/steamer/nuts/change_nomenclaturechange_2003_2006`; `http://purl.org/steamer/nuts/change_nomenclaturechange_2006_2010`);
– the SAU TSN (from the Swiss Federal Statistical Office) versions 2017 and 2018. The output graph (containing all the TUs and changes descriptions) is available online at `http://purl.org/steamer/sau` (the entry point of the change graph is available at `http://purl.org/steamer/sau/change_nomenclaturechange_2017_2018`);
– the ASGS TSN (from the Australian Bureau of Statistics) versions 2011 and 2016. The output graph (containing all the TUs and changes descriptions) is available online at `http://purl.org/steamer/asgs` (the entry point of the change graph is available at `http://purl.org/steamer/asgs/change_nomenclaturechange_2011_2016`).

In the next Chapter of this manuscript, we present first the two tests on the NUTS and SAU TSNs. Then, we present the ASGS case study as it covers the huge territory of Australia. It brings new challenges, such as big data processing by our software tools, and questions on the generalization of the geometries version to version. Finally, these three case studies allow us to assess the complexity of our algorithm.

Chapter 10

Case studies and discussion

10.1 Introduction

In this chapter, we present first three tests of our *Theseus Framework* developed to describe and publish on the LOD Web the elements of a TSN and its changes over time. We focus on the description of changes created by our program that implements our *TSN Semantic Matching Algorithm*. As described in the previous Chapter 9, the program consists of two Java modules (the *TSN Change Detector* and *TSN Change Semantic Annotator*) of the *Theseus Framework*. We ran three tests on three very different TSNs: the NUTS, ASGS, and SAU TSNs. Regarding the two data sets NUTS and SAU, although they both focus on Europe, they are quite different because of the size of the TUs they contain. The NUTS biggest TUs are the EU member states whereas the biggest TUs of the SAU are cantons, equivalent to the NUTS smallest regions (Level 3). The smallest units in the SAU are municipalities, much smaller units and more numerous than in the NUTS. Thus, these two data sets allow us to open up on interesting perspectives, such as the matching of two different TSNs, using the *TSN-Ontology* and *TSN-Change Ontology* to create bridges between the Level 3 of the NUTS and the Level 0 of the SAU. The third Australian data set ASGS was chose for experiment as it covers a very different region of the world, a vast territory whose TUs' geometries are composed of 4,000 vertices on average (compared to an average of 28 vertices in the NUTS). This was a challenge for our software program, allowing us to assess its ability to generate similarities and changes descriptions even on large data set.

In a second time, we assess the complexity of the *TSN Semantic Matching Algorithm*. We report about the problems we have encountered when testing our program on these three data sets, and we discuss the genericity of our approach, meaning its capability to handle various TSNs or not.

10.2 Case Studies

Please, note that the value of the input parameters (*i.e.*, weights on TUs attributes, tolerance thresholds for variation) of our program that implements the *TSN Semantic Matching Algorithm* are not discussed in this Section, but in the Section 10.3. These parameters and the definition of what makes the identity of the processed TUs have been fixed by experts of the TSNs.

10.2.1 Main characteristics of the three TSNs

The NUTS divides the European territory into 5 hierarchical levels: level 0 corresponds to the division of the European territory into States (European Union member states (EU), candidate countries for future membership of the EU and countries members of the European Free Trade Association (EFTA)), level 1 splits each of the level 0 areas into major Regions, level 2 corresponds to basic Regions, and level 3 to small Regions. The NUTS 1999, 2003, 2006, 2010 versions are available on the Web [1] as geospatial vector files containing the list of TUs of each version with their attributes (code, name, level, geometry, ...).

The ASGS main structure divides the Australian territory into 6 hierarchical levels: level State/Territory (level 5) corresponds to the State members, areas at level 4 are aggregations of the Statistical Areas at level 3 (*i.e.*, the abbreviation used by the ASGS to refer to this level is SA3), and have a population above 100,000 persons, areas at level 3 generally have a population between 30,000 and 130,000 persons, and so on until the Mesh Blocks level, the smallest geographical areas defined by the ABS. As for the NUTS, the ASGS (2011 and 2016) versions are available on the Web [2] as geospatial vector files.

The SAU nomenclature [3] created by the Swiss Federal Statistical Office on the basis of administrative boundaries, describes the cantons, districts and municipalities of Switzerland in 2017 and 2018. The SAU (2017 and 2018) versions are available on the Web as geospatial vector files [4].

Using the *TSN Ontology*, we have described with the same terms these 3 rather different nomenclatures. The Figure 10.1 shows how to describe these 3 (NUTS, ASGS, and SAU) spatial structures, using the `tsn:Nomenclature` main concept. The Table 10.1 shows the number of triples (feature versions and change descriptions) generated at the end of the process, for each of the 3 TSNs.

http://purl.org/steamer/nuts triples number	156,162
http://purl.org/steamer/sau triples number	76,504
http://purl.org/steamer/asgs triples number	89,974
Total number of triples	322,640

Table 10.1 – TSN triplestore numbers of triples

1. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts
2. Version 2016: https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument
3. Please note that SAU is not the official TSN name that is *Swiss official commune register*. We found this name misleading and reductive as under the term "commune" there are vector boundaries of municipalities, districts, cantons, large regions, etc.
4. https://shop.swisstopo.admin.ch/en/products/landscape/boundaries3D

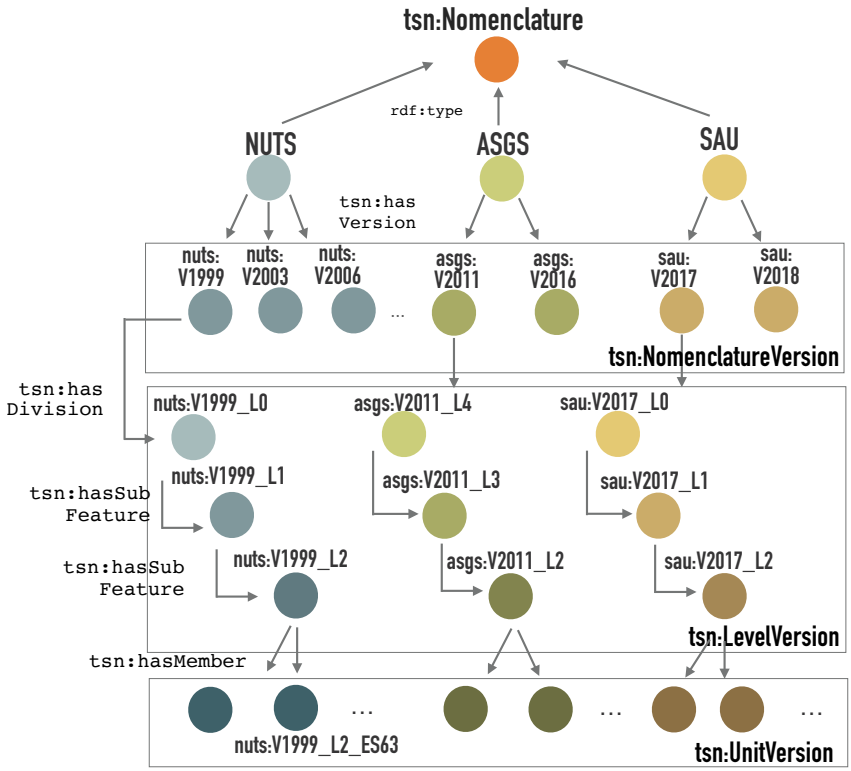


Figure 10.1 – Example of three TSNs' structures declared using the TSN Ontology.

10.2.2 The NUTS Eurostat Nomenclature

This Subsection presents the results we obtain using our framework on four versions of the NUTS (1999, 2003, 2006, 2010) [5]. The Table 10.2 lists the characteristics of the inputs, of the RDF outputs created after the execution of Workflows (A) and (B) described in the Chapter 9, and indicates the official dictionary of changes used to check the output change descriptions.

5. available from the Eurostat Web site at `http://ec.europa.eu/eurostat/fr/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts`

Input	Shapefiles	http://ec.europa.eu/eurostat/fr/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts
	Versions	NUTS versions 1999, 2003, 2006, 2010
	Levels	Level territory: European Union; Level 0: States; Level 1: Great regions; Level 2: Regions; Level 3: Small regions
	TUs total Number	V1999: 1861; V2003: 1902; V2006: 1933; V2010: 1923
	Weights	$\alpha_{id} = 0.4, \alpha_{geom} = 0.4, \alpha_{name} = 0.1, \alpha_{super} = 0.1$
	Threshold	$\beta = 0.5, \varepsilon_{spatialFeature} = 0.005, \varepsilon_{spatialStructure} = 0.005$
Output	RDF Output Graph	http://purl.org/steamer/nuts
	Official Changes Dictionary used to check the RDF Output graph	https://ec.europa.eu/eurostat/web/nuts/history

Table 10.2 – Configurations for test of the Theseus Framework on the NUTS TSN.

More than 122,000 resources have been generated and published online after the execution of the Workflow (A) (see Figure 9.1). These resources include the description of the NUTS structure (hierarchy of territorial levels) and of all the TUs, at each level and for each version of the NUTS. TUs' geometries are also available as multi-polygons (please, consult for instance `http://purl.org/steamer/nuts/Geometry_3512`). Listing 10.1 presents an example of the RDF description of a TU automatically computed by our system after Workflow A: here the TU ES63 in NUTS Version 1999 is described using the TSN ontological model and the GeoSPARQL ontology. First, the identity of the TU version is defined (by an identifier, a name, a super TU, a level of belonging) (line 7 to 10), then the lineage of the TU is described (line 11 and 12), and the geometry of the TU, for the version of the NUTS observed, is defined from line 15 to 20. The geometry is written in Well-known text (WKT), in EPSG 4326, for instance here.

```
1 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2 @prefix tsn: <http://purl.org/net/tsn#> .
3 @prefix tsnchange: <http://purl.org/net/tsn#> .
4 @prefix nuts: <http://purl.org/steamer/nuts/> .
5 @prefix geo: <http://www.opengis.net/ont/geosparql#>
6 nuts:V1999_L2_ES63 a  tsn:UnitVersion ;
7     tsn:hasIdentifier "ES63"^^xsd:string ;
8     tsn:hasName "Ceuta␣y␣Melilla" ;
9     tsn:hasSuperFeature nuts:V1999_L1_ES6 ;
10    tsn:isMemberOf nuts:V1999_L2 ;
11    tsn:isVersionOf nuts:L2_ES63 ;
12    tsnchange:hasNextVersion nuts:V2003_L2_ES63 ;
13    <http://dbpedia.org/ontology/languageCode> "es"^^xsd:string ;
14    geo:hasGeometry nuts:Geometry_3245 .
15 nuts:Geometry_3512 a  geo:Geometry ;
16    geo:asWKT "<http://www.opengis.net/def/crs/EPSG/0/4326>␣MULTIPOLYGON␣
          (((-2.84␣35.2937,␣-2.913␣35.2633,␣-2.9522␣35.3492,␣-2.8959␣
          35.3657,␣-2.893␣35.3619,␣-2.8593␣35.3185,␣-2.84␣35.2937))
```

```
        ,((-5.2888␣35.7974,␣-5.3748␣35.7767,␣-5.4141␣35.8383,␣-5.334␣
        35.8586,␣-5.288835.7974)))"^^geo:wktLiteral ;
17    geo:is3D "false"^^xsd:boolean ;
18    geo:isEmpty "false"^^xsd:boolean ;
19    geo:isSimple "true"^^xsd:boolean ;
20    geo:spatialDimension  2 .
```

Listing Code 10.1 – The NUTS 1999 ES63 TU description using concepts from the TSN and GeoSPARQL Ontologies.

Regarding now Workflow (B), using the *TSN Change Detector* and *TSN Change Semantic Annotator* of our framework, we perform the matching of the NUTS versions. The list of attributes considered for the matching of the versions of the NUTS TUs (see Equation 9.2) are: TU code in the TSN (id), geometry (geom), designation (name), and superior unit (super). The weights assigned to these criteria for comparison are: $\alpha_{id} = 0.4, \alpha_{geom} = 0.4, \alpha_{name} = 0.1, \alpha_{super} = 0.1$ (please, see Chapter 10, Section 10.3.3 for a discussion of the values assigned to these weights). The program creates the RDF graph between two versions of the NUTS with a running time[6] between 15 and 20 minutes depending on the input versions, given that each version is composed of a total number of TUs closed to 1,900. More than 34,000 resources describing matching and changes have been generated and published online after the execution of the Workflow (B) from our triplestore at `http://purl.org/steamer/nuts`. Regarding the nature of changes, between the NUTS version 2006 and version 2010 (see Table 10.3), 1,024 Feature Changes and 69 Structure Changes[7] have been detected, for a total of 1,093 change nodes. This is significantly greater than the number of changes described in Eurostat's changes lists (see Table 10.2, line "Official Changes Dictionary"), but our figures considerably depend on the representation of changes we use in our approach. For instance, we construct *SubUnitChange* and *SuperUnitChange* nodes which are not described by Eurostat. This addition of information provide genuine added value at the time of data retrieval, since it provides the users with descriptions of all the changes each element of the TSN undergone, including, from one element, the changes undergone by its sub(s) and super features.

Matching of the versions	NUTS 1999-2003	NUTS 2003-2006	NUTS 2006-2010
Number of Feature Change	2,017	1,118	1,024
Number of Structure Change	87	74	69
Total Number of Change	2,104	1,192	1,093

Table 10.3 – The NUTS TSN Change Graph – number of change nodes.

6. Our software runs on a iOS system with 16GB of memory and 2,5 GHz Intel Core i7 processor.

7. 8 reallocations, 14 fusions, 9 scissions, 33 identification restructurations, and 5 structure changes hub nodes that group all the changes occurring at one NUTS levels (territory, L0, L1, L2, L3).

Matching of the versions	NUTS 2006 - 2010
Number of GeometryChange	62
Number of NameChange	10
Number of IdentifierChange	165
Number of SubUnitChange	55
Number of SuperUnitChange	33
Number of Split	9
Number of Merge	14
Number of Redistribution	8
Number of IdentificationRestructuration	33

Table 10.4 – The NUTS TSN Change Graph – main change types distribution from version 2006 to 2010.

Then, for a more relevant evaluation of the results, we assess the quality of the changes descriptions considering their semantics. Looking at the automatically generated description of the change event (an excerpt is provided in Listing 10.2) we can highlight the capabilities of our algorithm:
– it retrieves *all* the changes manually listed by Eurostat (see Eurostat online catalog of changes in Table 10.2, line "Official Changes Dictionary");
– it refines the description of changes (*e.g.*, Listing 10.2, line 3 the split of the TU ES63 is described as an *Extraction* change, not as a *Split* event, as described by Eurostat). Then, semantically speaking, we have automatically enriched the description of changes provided by Eurostat, because *Extraction* means that the TU continues to exist after the change event (see Split Types, Figure 8.7). Our algorithm constructs the life line of the TU $ES63$. Statisticians can follow the links and discover this lineage and the changes of entities over time;
– it links inter-level changes: the change that affects the TU ES63 - at level 2 - is linked to sub-changes that affect its sub-TUs at Level 3 through a *SubUnitChange* node (Listing 10.2, line 14) that has a lower change of type *IdentificationRestructuration* (Listing 10.2, line 18). This node groups several identifier changes on TUs (here, both $ES631'$ and $ES632'$ have undergone a change of their TSN official identifier). This particular case allows us to highlight additional features of our algorithm: even on very small TUs, it is effective to detect changes, and even if the TU is a multi-polygon, the *Split* event is detected. Also, even if the TU ES63 is a two-part enclave, each part being spatially distant from each other, the changes each part undergone are linked to changes on the super-TU since the TSN hierarchy of elements is taken into account by our algorithm (and not only spatial information). Please, consult the following URI `http://purl.org/steamer/nuts/change_extractionchange_V1999_L2_ES63` to access a full description of the change undergone by the TU `http://purl.org/steamer/nuts/V1999_L2_ES63`.

```
1   @prefix tsnchange:  <http://purl.org/net/tsnchange#> .
2   @prefix nuts:  <http://purl.org/steamer/nuts/> .
3   nuts:change_extractionchange_V1999_L2_ES63 a tsnchange:Extraction;
4     tsnchange:lowerChange nuts:change_appearance_V2003_L2_ES64,
5     nuts:change_namechange_V1999_L2_ES63 , nuts:
          change_geometrychange_V1999_L2_ES63 ;
```

```
6       tsnchange:unitVersionAfter nuts:V2003_L2_ES64, nuts:V2003_L2_ES63 ;
7       tsnchange:unitVersionBefore nuts:V1999_L2_ES63 ;
8       tsnchange:upperChange nuts:change_levelchange_V1999_2003_L2 ,
9       nuts:change_subunitof_V1999_L1_ES6_change .
10    nuts:change_geometrychange_V1999_L2_ES63 a tsnchange:Contraction;
11      tsnchange:unitVersionAfter nuts:V2003_L2_ES63 ;
12      tsnchange:unitVersionBefore nuts:V1999_L2_ES63 ;
13      tsnchange:upperChange nuts:change_extractionchange_V1999_L2_ES63 .
14    nuts:change_subunitof_V1999_L2_ES63_change a tsnchange:SubUnitChange;
15      tsnchange:lowerChange nuts:
            change_identification_restructuration_L3_14 ;
16      tsnchange:unitVersionBefore nuts:V1999_L2_ES63 ;
17      tsnchange:upperChange nuts:change_levelchange_V1999_2003_L2 .
18    nuts:change_identification_restructuration_L3_14 a
          tsnchange:IdentificationRestructuration;
19      tsnchange:lowerChange nuts:change_identifierchange_V1999_L3_ES632 ,
20      nuts:change_identifierchange_V1999_L3_ES631 ;
21      tsnchange:unitVersionAfter  nuts:V2003_L3_ES640 , nuts:
            V2003_L3_ES630 ;
22      tsnchange:unitVersionBefore  nuts:V1999_L3_ES632 , nuts:
            V1999_L3_ES631 ;
23      tsnchange:upperChange  nuts:change_subunitof_V1999_L2_ES63_change .
24    nuts:change_identifierchange_V1999_L3_ES632 a
          tsnchange:IdentifierChange;
25      tsnchange:lowerChange nuts:change_appearance_V2003_L3_ES640 ,
26      nuts:change_disappearance_V1999_L3_ES632 ;
27      tsnchange:unitVersionAfter nuts:V2003_L3_ES640 ;
28      tsnchange:unitVersionBefore nuts:V1999_L3_ES632 ;
29      tsnchange:upperChange nuts:change_subunitof_V1999_L2_ES63_change ,
30      nuts:change_identification_restructuration_L3_14 .
```

Listing Code 10.2 – The RDF change description of TU ES63 from NUTS version 1999 to 2003 – please note that the Figure 9.2 is a map representation of this event.

10.2.3 The Swiss Administrative Units

This subsection presents our results using the Theseus Framework on two versions of the Swiss *Administrative Units* (SAU). This TSN is created by the Swiss Federal Statistical Office on the basis of administrative boundaries. The administrative boundaries do not cover the entire Swiss territory at the district level because, in Switzerland, each canton has its own constitution and legislature, resulting in a variety of structures and terminologies for the subnational entities between Canton and Municipality. Then, the Swiss Federal Statistical Office builds a version for statistics purpose by homogenizing the TUs of this territorial level under the term "districts". The list of attributes considered for the matching of the SAU TUs are: TU identifier in the TSN (id), geometry (geom), designation (name), and superior unit (super). The weights assigned to these criteria for comparison are: $\alpha_{id} = 0.4, \alpha_{geom} = 0.4, \alpha_{name} = 0.1, \alpha_{super} = 0.1$ (please, see Section 10.3.3 for a discussion of the values assigned to these weights). The Table 10.5 lists the characteristics of the inputs, of the RDF outputs created after the Workflow (A) and (B) described in the Chapter 9, and indicates the official dictionary of SAU changes used to check the output change descriptions.

Input	Shapefiles	https://www.bfs.admin.ch/bfs/fr/home/services/geostat/geodonnees-statistique-federale/limites-administratives/limites-communales-generalisees.html
	Versions	SAU versions 2017, 2018
	Levels	Level territory: Switzerland, Liechtenstein, German and Italian Enclaves Level 0: Cantonal boundaries Level 1: District boundaries Level 2: Municipal boundaries
	TUs total Number	V2017: 2457; V2018: 2419
	Weights	$\alpha_{id} = 0.4, \alpha_{geom} = 0.4, \alpha_{name} = 0.1, \alpha_{super} = 0.1$
	Threshold	$\beta = 0.5, \varepsilon_{spatialFeature} = 0.005, \varepsilon_{spatialStructure} = 0.005$
Output	RDF Output Graph	http://purl.org/steamer/sau
	Official Changes Dictionary used to check the RDF Output graph	https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/publications.assetdetail.4123244.html

Table 10.5 – Configurations for test of the Theseus Framework on the SAU TSN.

The program creates the RDF graphs between two versions of the SAU with a running time[8] of 21 hours, given that each version is composed of a total number of TUs closed to 2,400. This processing time is much more important than the average processing time for the matching between two NUTS versions. This can be explained first by the number of TUs processed (almost twice as in the NUTS's one),

8. Our software runs on a iOS system with 16GB of memory and 2,5 GHz Intel Core i7 processor.

secondly because of the geometries of the Swiss TUs that are much more precise than the geometries of the NUTS TUs (this issue is discussed in Section 10.3.1).

More than 76,000 resources (describing the SAU elements, their similarities and changes) are generated and published online from our triplestore at `http://purl.org/steamer/sau`. Regarding the nature of changes, between the SAU version 2017 and version 2018, 336 feature changes and 25 structure changes were detected, for a total of 361 change nodes.

Matching of the versions	SAU 2017 - 2018
Number of Feature Change	336
Number of GeometryChange	65
Number of NameChange	0
Number of IdentifierChange	30
Number of SubUnitChange	19
Number of SuperUnitChange	7
Number of Structure Change	25
Number of Split	0
Number of Merge	14
Number of Redistribution	1
Number of IdentificationRestructuration	6
Total Number of Change	361

Table 10.6 – The SAU TSN Change Graph – main change types distribution from version 2017 to 2018.

This is quite mush more than the number of changes described in the official catalog of changes of the Federal Statistical Office (see Table 10.5, line "Official Changes Dictionary"), but as for the NUTS TSN, our figures considerably depend upon the representation of changes we use in our approach. As for the NUTS TSN, we want to highlight the capabilities of our TSN Semantic Matching Algorithm regarding the richness of the semantic descriptions automatically generated. Let us take the example of the change undergone by the canton of Neuchâtel. In 2017, this canton was composed of 6 districts. In 2018, all these districts have disappeared and are replaced by a single new district with identifier 2400. This merge event caused the re-codification of all the sub-units at Municipality level, in order to align with the new super-unit identifier (2400). Then, the dissolution of the districts impacts the TSN at two levels: the District and Municipality levels (see Figure 10.2 where all the TUs composing the canton of Neuchâtel, at District and Municipality levels are highlighted in yellow, in version 2017 (on the left) and version 2018 (on the right)). In the Official catalog of changes, of the Swiss Federal Statistical Office, this event is described as follows: *"Change No. 3591 to 3620: In the canton of Neuchâtel, the territorial division into districts will be abolished on 1 January 2018. Following this, the distribution of the municipalities of the canton of Neuchâtel into districts will change in the official list of the municipalities of Switzerland. The previous distribution of the municipalities of 6 districts is replaced by the following distribution: No. 2400 canton of Neuchâtel".*

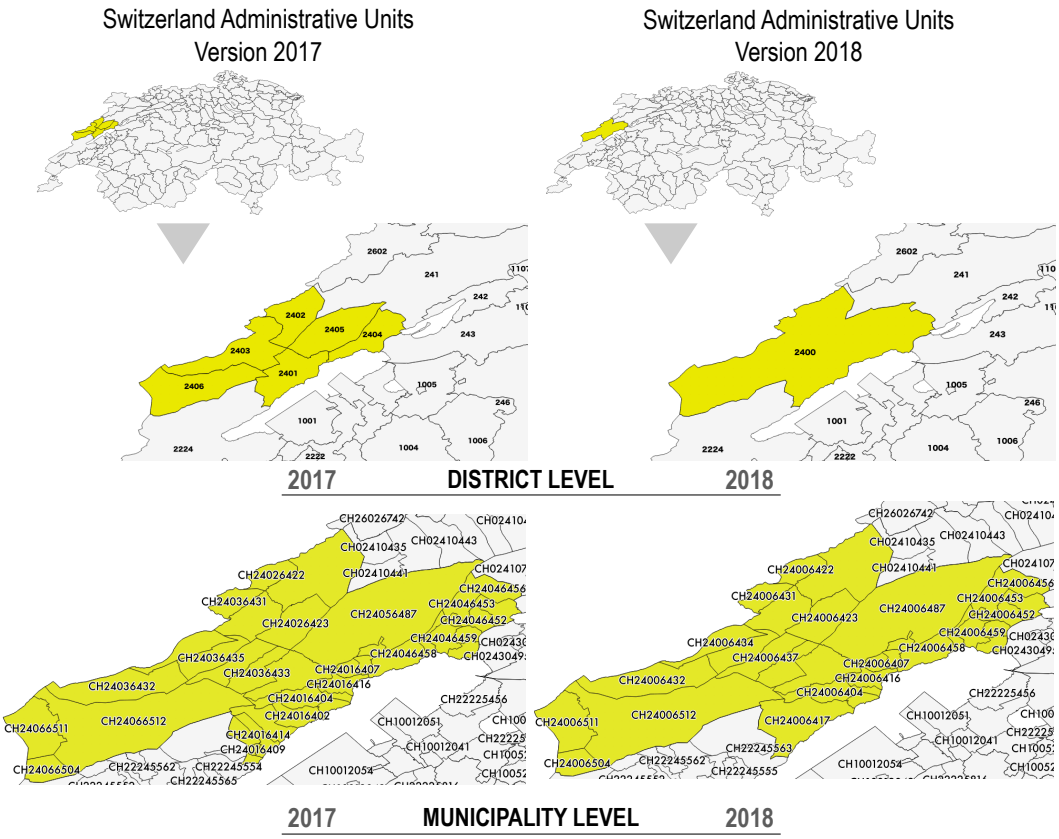


Figure 10.2 – SAU Districts Neuchâtel changes.

Considering our automatically generated description of the change event (see an excerpt in Listing 10.3), we can highlight the capabilities of our algorithm:
– it retrieves all the changes listed by the Federal Statistical Office (we have created an online file that compares for each change the description provided by the Federal Statistical Office and the description automatically generated by our TSN matching program [9]);
– it refines the description of changes as it links changes through the TSN levels. For instance, regarding the change of the canton of Neuchâtel, our program describes:

(1) first the change of the canton of Neuchâtel, at the Canton Level 0, that consists of the change of its sub-TUs (see Listing 10.3, lines 3 to 6) ;

(2) the change of the sub-TUs of the canton of Neuchâtel, at the District Level 1, that consists of the fusion of the 6 districts into one (see Listing 10.3, lines 7 to 20);

(3) visiting each of the 6 change nodes describing the changes of districts (an example on one district, with identifier 2401, is provided in Listing 10.3, lines 20 to

9. this catalog is available at `https://github.com/camillebernard/tsndata/blob/master/src/site/resources/SAU/sau_2017_2018_tsn_change_descriptions.pdf`

24), one discovers that each of them disappear (line 24), but also that their sub-units (municipalities) change (line 23));

(4) visiting this lower change node at `http://purl.org/steamer/sau/change_subunitof_V2017_L1_2401_change` (lines 25), one discovers that the 12 sub-units of the district 2401 in version 2017 have undergone two different changes:
– some of them change their nomenclature identifier (line 33) and align with the new super-TU identifier 2400 in version 2018 (lines 31 to 40). Please note that this listing of municipalities that change their identifier is not explicitly provided within the official catalog of changes of the Federal Statistical Office;
– some of them have merged (line 29 and description of the fusion lines 41 to 48). This change is described in the Official catalog of changes of the Swiss Federal Statistical Office. But this official catalog provides no link between this change and the change of the canton of Neuchâtel, unlike what we do here.

```
1   @prefix tsnchange:  <http://purl.org/net/tsnchange#> .
2   @prefix sau:  <http://purl.org/steamer/sau/> .
3   sau:change_unitchange_V2017_L0_24 a tsnchange:UnitChange;
4     tsnchange:unitVersionBefore  sau:V2017_L0_24 ;
5     tsnchange:upperChange  sau:change_levelchange_V2017_2018_L0 ;
6     tsnchange:lowerChange sau:change_subunitof_V2017_L0_24_change .
7   sau:change_subunitof_V2017_L0_24_change a tsnchange:SubUnitChange;
8     tsnchange:unitVersionBefore sau:V2017_L1_2401 , sau:V2017_L1_2406 ,
9     sau:V2017_L1_2402 , sau:V2017_L1_2403 , sau:V2017_L1_2404 , sau:
          V2017_L1_2405 ;
10    tsnchange:upperChange  sau:change_unitchange_V2017_L0_24 ;
11    tsnchange:lowerChange  sau:change_fusionchange_V2018_L1_2400 .
12  sau:change_fusionchange_V2018_L1_2400 a tsnchange:Fusion;
13    tsnchange:unitVersionAfter sau:V2018_L1_2400 ;
14    tsnchange:unitVersionBefore sau:V2017_L1_2406, sau:V2017_L1_2405,
15    sau:V2017_L1_2404, sau:V2017_L1_2403, sau:V2017_L1_2402, sau:
          V2017_L1_2401 ;
16    tsnchange:upperChange sau:change_subunitof_V2017_L0_24_change ;
17    tsnchange:lowerChange sau:change_appearance_V2018_L1_2400 ,
18    sau:change_unitchange_V2017_L1_2401, ...,
19    sau:change_unitchange_V2017_L1_2406 .
20  sau:change_unitchange_V2017_L1_2401 a tsnchange:UnitChange;
21    tsnchange:unitVersionBefore sau:V2017_L1_2401 ;
22    tsnchange:upperChange sau:change_fusionchange_V2018_L1_2400 ;
23    tsnchange:lowerChange sau:change_subunitof_V2017_L1_2401_change,
24    sau:change_disappearance_V2017_L1_2401 ;
25  sau:change_subunitof_V2017_L1_2401_change a tsnchange:SubUnitChange;
26    tsnchange:unitVersionBefore sau:V2017_L2_CH24016402 , ...,
27    sau:V2017_L2_CH24016416 ;
28    tsnchange:upperChange sau:change_unitchange_V2017_L1_2401 ;
29    tsnchange:lowerChange sau:change_fusionchange_V2018_L2_CH24006417,
30    sau:change_identification_restructuration_L2_14 .
31  sau:change_identification_restructuration_L2_14 a
        tsnchange:IdentificationRestructuration;
32    tsnchange:lowerChange sau:
          change_identifierchange_V2017_L2_CH24016413,
33    sau:change_identifierchange_V2017_L2_CH24016416, ... ;
34    tsnchange:unitVersionAfter sau:V2018_L2_CH24006408, sau:
```

```
          V2018_L2_CH24006412 ,
35      sau:V2018_L2_CH24006413 , sau:V2018_L2_CH24006404 ,
36      sau:V2018_L2_CH24006407 , sau:V2018_L2_CH24006416 ;
37      tsnchange:unitVersionBefore sau:V2017_L2_CH24016412 , sau:
          V2017_L2_CH24016416 ,
38      sau:V2017_L2_CH24016408 , sau:V2017_L2_CH24016413 ,
39      sau:V2017_L2_CH24016404 , sau:V2017_L2_CH24016407 ;
40      tsnchange:upperChange sau:change_superunit_V2017_L1_2401_change .
41    sau:change_fusionchange_V2018_L2_CH24006417 a tsnchange:Fusion;
42      tsnchange:lowerChange sau:change_appearance_V2018_L2_CH24006417 ,
43      sau:change_disappearance_V2017_L2_CH24016409 , ... ;
44      tsnchange:unitVersionAfter sau:V2018_L2_CH24006417 ;
45      tsnchange:unitVersionBefore sau:V2017_L2_CH24016414, sau:
          V2017_L2_CH24016402,
46      sau:V2017_L2_CH24016410, sau:V2017_L2_CH24016411, sau:
          V2017_L2_CH24016415,
47      sau:V2017_L2_CH24016409 ;
48      tsnchange:upperChange sau:change_subunitof_V2017_L1_2401_change .
```

Listing Code 10.3 – The RDF change description of the Canton of Neuchâtel from SAU version 2017 to 2018.

Here, we have shown that our algorithm automatically creates linked description of changes that help analysts in understanding the territorial changes and their impacts on each sub-element, starting from the main element that changes, the canton of Neuchâtel, and zooming in to visualize changes on each sub-elements (districts and municipalities). Please note that the links we have created between changes are not causal ones. The semantics they hold is a hierarchical information about the changes. Nevertheless, the *TSN-Change Ontology* defines a predicate to link a change node to some causal information (`tsnchange:isCausedBy`). The subsection 11.3 shows how we use the LOD Web to search for this causal information: historical, societal information that may explain the change event.

10.2.4 The Australian Statistical Geography Standard

This subsection presents our results using the Theseus Framework on the two existing versions (2011 and 2016) of the Australian Statistical Geography Standard (ASGS). This TSN is created by the Australian Bureau of Statistics. The ASGS nomenclature is complex and made of several branches, as shown in Figure 10.3. For this implementation, we focus only on the main structure (in blue) and have processed the first levels only: State/Territory level; Statistical Area Level 4 (SA4); SA3; SA2.

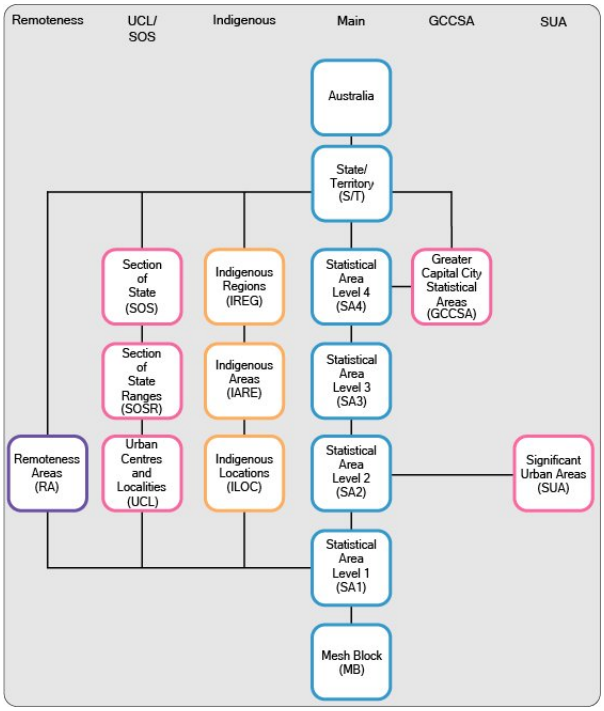


Figure 10.3 – The ASGS Structures (source: `https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS)`).

The list of attributes considered for the matching of the ASGS TUs are: TU identifier in the TSN (id), geometry (geom), designation (name), and superior unit (super). The weights assigned to these criteria for comparison are: $\alpha_{id} = 0.4, \alpha_{geom} = 0.4, \alpha_{name} = 0.1, \alpha_{super} = 0.1$. The Table 10.7 lists the characteristics of the inputs, of the RDF outputs created after the execution of Workflows (A) and (B) described in the Chapter 9, and indicates the official dictionary of ASGS changes used to check the output change descriptions.

Input	Shapefiles	https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument
	Versions	SAU versions 2011, 2016
	Levels	**Level territory**: New South Wales; Victoria; Queensland; South Australia; Western Australia; Tasmania; Northern Territory; Australian Capital Territory; Other Territories **Level 4**: Statistical Areas Level 4 (SA4s) have a population above 100,000 persons **Level 3**: SA3s generally have populations between 30,000 and 130,000 persons **Level 2**: SA2s generally have a population range of 3,000 to 25,000 persons
	TUs total Number	V2011: 2680; V2018: 2784
	Weights	$\alpha_{id} = 0.4, \alpha_{geom} = 0.4, \alpha_{name} = 0.1, \alpha_{super} = 0.1$
	Threshold	$\beta = 0.5, \varepsilon_{spatialFeature} = 0.1, \varepsilon_{spatialStructure} = 0.1$
Output	RDF Output Graph	http://purl.org/steamer/asgs
	Official Changes Dictionary used to check the RDF Output graph	http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument (one document per level *e.g.*, file Correspondence, 2011 Statistical Area Level 4 to 2016 Statistical Area Level 4

Table 10.7 – The TSN SMA Algorithm 1 implementation – configurations for tests on the SAU TSN.

The program creates the RDF graphs between two versions of the ASGS with an execution time[10] of 09 days 05 hours 25 minutes. Please, note that we discuss this processing time for the matching between two ASGS versions in the Section 10.3.1. More than 89,900 resources (describing the ASGS elements, their similarities and changes) have been generated and published online from our triplestore at `http://purl.org/steamer/asgs`.

10. Our software runs on a Debian system with 141GB of memory and Intel Xeon CPU E5-2640 2.40GHz processor.

Matching of the versions	ASGS 2011 - 2016
Number of Feature Change	632
Number of GeometryChange	93
Number of NameChange	14
Number of IdentifierChange	71
Number of SubUnitChange	69
Number of SuperUnitChange	13
Number of Structure Change	81
Number of Split	63
Number of Merge	3
Number of Redistribution	4
Number of IdentificationRestructuration	11
Total Number of Change	717

Table 10.8 – The ASGS TSN Change Graph – main change types distribution from version 2011 to 2016.

Let us take the example of a split event in the ASGS. Indeed, while in the SAU the merge of TUs is the predominant change type, in the ASGS we observe few merge events but a total of 63 splits, as explained on the ABS website [11].

"*Where possible the boundaries are split to accommodate growth.*". Also, on this web page, the change to ASGS TU codes is explained: "*ASGS users should be aware that the codes and names that are associated with statistical areas can also change when an area changes. In addition, where changes occur in the larger Statistical Area regions this can result in changes to the code for the smaller areas contained within them, even if one of these smaller areas has not changed itself. This occurs because the ABS Structures within the ASGS have a hierarchical coding system, meaning the smaller areas carry the codes associated with the larger areas.*"

We have chosen the following scenario to show how our program gives to see such split changes and changes in the hierarchical coding system. In order to observe the automatically generated descriptions of a Split event in the ASGS nomenclature, we first extract from our graphs all nodes of this type that impact the TUs at the highest level of the nomenclature. The SPARQL query 10.4 returns a list of changes of type "split of a TU", composed of only one change node, the node with URI `asgs:change_scissionchange_V2011_L4_508`, as at level 4 between the versions 2011 and 2016 only one event of type "split of a TU" took place.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX tsnchange: <http://purl.org/net/tsnchange#>
PREFIX tsn: <http://purl.org/net/tsn#>
PREFIX asgs: <http://purl.org/steamer/asgs/>
SELECT DISTINCT ?change
WHERE {
```

11. https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1270.0.55.001~July%202016~Main%20Features~Stability%20and%20Change%20in%20the%20ASGS~10018

```
7     ?change rdf:type tsnchange:Change .
8     ?change rdf:type tsnchange:Split .
9     ?inputTU tsnchange:input ?change .
10    ?inputTU tsn:isMemberOf ?level .
11    ?level  tsn:hasIdentifier  "ASGS_V2011_L4" .}
```

Listing Code 10.4 – A SPARQL query that returns all the change nodes of type "Split" in the ASGS change graph version 2011 to 2016 at the highest level SA4.

As shown on the maps of Figure 10.4 at level 4 (Statistical Areas 4) from the version 2011 to 2016, the TU with code 508 splits in two TUs, with code 510 and 511 respectively.

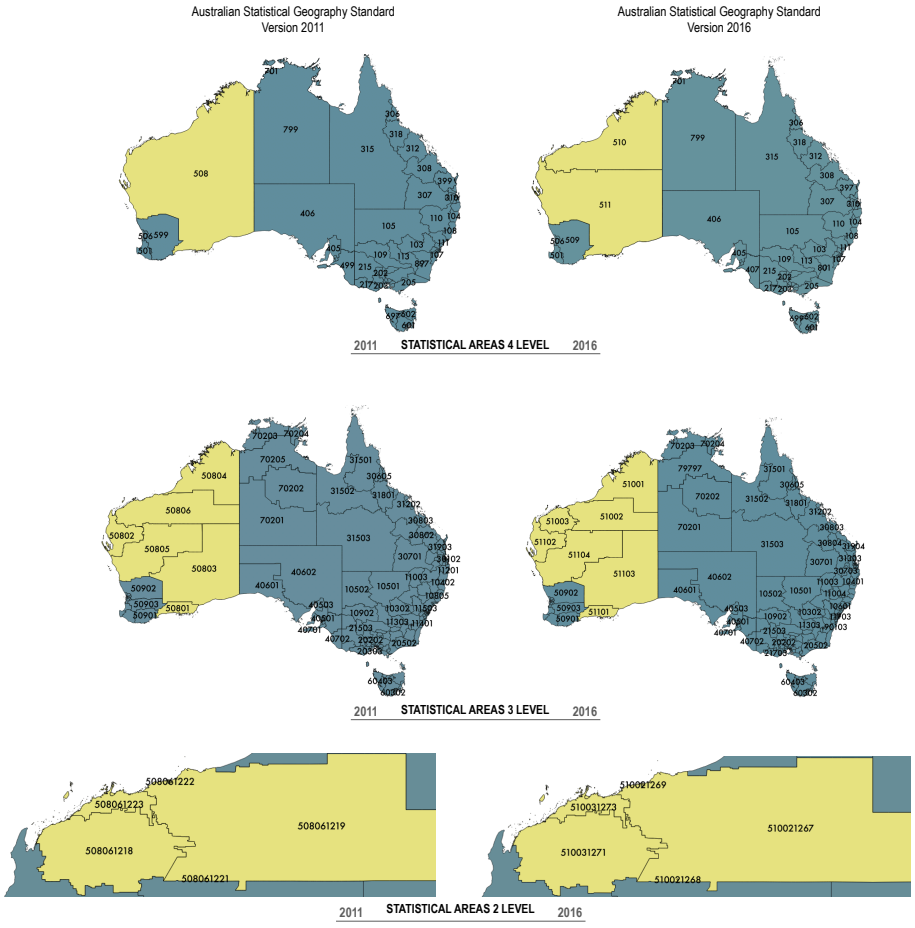


Figure 10.4 – The ASGS map representation of the Split of the TU 508 at level 4 and representation of the sub-changes on the sub-TUs of TU 508.

Using the query 10.5 that follows the links of type `tsnchange:lowerChange`, one can extract from the ASGS change graph the whole chain of changes (of type `tsnchange:StructureChange`), that occurred on the sub-TUs of the TU 508 that splits. This chain starts from the StructureChange node that impacts the TU with URI `asgs:V2011_L4_508` (see Listing 10.5, line 8) and ends by a StructureChange descriptions on the smallest sub-TUs of `asgs:V2011_L4_508`. Please note that in the next Chapter 11, we come back to this kind of requests that allows for an horizontal reading of the TSN change graphs.

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX tsnchange: <http://purl.org/net/tsnchange#>
3 PREFIX tsn: <http://purl.org/net/tsn#>
4 PREFIX : <http://purl.org/steamer/asgs/>
5 SELECT DISTINCT ?s
6 WHERE {
7   values ?p {tsnchange:inputUnitVersion tsnchange:lowerChange}
8   :V2011_L4_508 (tsnchange:inputUnitVersion|tsnchange:lowerChange)* ?s .
9   ?s ?p ?o .
10  ?s rdf:type tsnchange:StructureChange
11 }ORDERBY DESC(?s)
```

Listing Code 10.5 – A SPARQL query that returns the whole chain of changes that occurred on the sub-TUs to the TU with code 508 that split in ASGS version 2006.

The chain of changes, result of the query in Listing 10.5, is presented in the table 10.9. From this chain, we can identify two changes that affect the `asgs:V2011_L4_508` sub-TUs. At level SA3: there is a scission of the TU `asgs:V2011_L3_50806`, and a re-codification of all the other sub-TUs (described under the node `asgs:change_identification_restructuration_L3_7` that gathers under one node the re-codification of several TUs). At level SA2, the only changes the sub-TUs undergo are again changes of type re-codification (*e.g.*, `asgs:change_identification_restructuration_L2_73`).

	StructureChange nodes chain
SA4	asgs:change_scissionchange_V2011_L4_508
SA3	asgs:change_scissionchange_V2011_L3_50806
	asgs:change_identification_restructuration_L3_7
SA2	asgs:change_identification_restructuration_L2_73
	asgs:change_identification_restructuration_L2_72
	asgs:change_identification_restructuration_L2_71
	asgs:change_identification_restructuration_L2_70
	asgs:change_identification_restructuration_L2_69
	asgs:change_identification_restructuration_L2_68

Table 10.9 – The result of the query 10.5 that returns a chain of StructureChange nodes, starting from the change that affects the TU asgs:V2011_L4_508

Here, we show that our algorithm automatically accounts for a change rule in the ASGS described by the ABS on its website *i.e.*, "*where changes occur in the larger Statistical Area regions this can result in changes to the code for the smaller areas contained within them, even if one of these smaller areas has not changed itself*".

We have decided to group under one node of type *tsnchange:Identification-Restructuration* all the re-codifications of several TUs that all have the same super-TU. Another way of doing this would have been to search for all re-codifications, but by spatial proximity. We would then have grouped under the same node the 6 *tsnchange:IdentificationRestructuration* changes, at level SA2 (see Table 10.9). However, this approach may lead to encompass other neighboring TUs that change their code, while they do not share the same super-TU.

We have tested this search by spatial proximity in the NUTS, and we have obtained results that group under one change node, re-codifications taking place in different European countries. It seems to us that this choice is an unwise one, since it does not respect the national choices for changes in the codes of the European TUs. For instance, Listing 10.6 shows one RDF result that groups under one *tsnchange:IdentificationRestructuration* node, several changes of TUs' code in Poland, Germany, Estonia, Latvia, Lithuania because the TUs are neighbors. The map in Figure 10.5 is a geo-visualization of the TUs listed in 10.6 that change their code in the NUTS from version 1999 to 2003.

```
1 nuts:change_identification_restructuration_L2_8
2  a  tsnchange:IdentificationRestructuration ;
3 tsnchange:after
4     nuts:V2003_L2_LT00 , nuts:V2003_L2_EE00 , nuts:V2003_L2_LV00 ,
5     nuts:V2003_L2_PL34 , nuts:V2003_L2_PL33 , nuts:V2003_L2_PL43 ,
6     nuts:V2003_L2_PL41 , nuts:V2003_L2_PL61 , nuts:V2003_L2_DEF0 ,
7     nuts:V2003_L2_PL42 , nuts:V2003_L2_PL52 , nuts:V2003_L2_DE60 ,
8     nuts:V2003_L2_PL51 , nuts:V2003_L2_PL21 , nuts:V2003_L2_PL62 ,
9     nuts:V2003_L2_DE80 , nuts:V2003_L2_PL32 , nuts:V2003_L2_PL12 ,
10    nuts:V2003_L2_PL31 , nuts:V2003_L2_PL11 , nuts:V2003_L2_PL22 ,
11    nuts:V2003_L2_PL63 ;
12 tsnchange:before
13    nuts:V1999_L2_PL0C , nuts:V1999_L2_PL08 , nuts:V1999_L2_PL07 ,
14    nuts:V1999_L2_PL03 , nuts:V1999_L2_PL09 , nuts:V1999_L2_PL0D ,
15    nuts:V1999_L2_LV , nuts:V1999_L2_EE , nuts:V1999_L2_PL04 ,
16    nuts:V1999_L2_PL0F , nuts:V1999_L2_PL0E , nuts:V1999_L2_DE6 ,
17    nuts:V1999_L2_PL0B , nuts:V1999_L2_PL0A , nuts:V1999_L2_PL06 ,
18    nuts:V1999_L2_PL05 , nuts:V1999_L2_DE8 , nuts:V1999_L2_PL02 ,
19    nuts:V1999_L2_PL01 , nuts:V1999_L2_PL0G , nuts:V1999_L2_DEF ,
20    nuts:V1999_L2_LT ;
21 tsnchange:lowerChange
22    nuts:change_identifierchange_V1999_L2_PL05 ,
23    nuts:change_identifierchange_V1999_L2_LV ,
24    nuts:change_identifierchange_V1999_L2_DE8 ,
25    nuts:change_identifierchange_V1999_L2_DEF ,
26    nuts:change_identifierchange_V1999_L2_PL04 ,
27    nuts:change_identifierchange_V1999_L2_PL0B ,
28    nuts:change_identifierchange_V1999_L2_PL0C ,
29    nuts:change_identifierchange_V1999_L2_EE ,
30    nuts:change_identifierchange_V1999_L2_PL01 ,
```

```
31     nuts:change_identifierchange_V1999_L2_LT ,
32     nuts:change_identifierchange_V1999_L2_PL0F ,
33     nuts:change_identifierchange_V1999_L2_PL03 ,
34     nuts:change_identifierchange_V1999_L2_PL09 ,
35     nuts:change_identifierchange_V1999_L2_DE6 ,
36     nuts:change_identifierchange_V1999_L2_PL02 ,
37     nuts:change_identifierchange_V1999_L2_PL08 ,
38     nuts:change_identifierchange_V1999_L2_PL0G ,
39     nuts:change_identifierchange_V1999_L2_PL0E ,
40     nuts:change_identifierchange_V1999_L2_PL0D ,
41     nuts:change_identifierchange_V1999_L2_PL07 ,
42     nuts:change_identifierchange_V1999_L2_PL0A ,
43     nuts:change_identifierchange_V1999_L2_PL06  .
```

Listing Code 10.6 – One result of the implementation of IdentificationRetructuration search by spatial proximity (in the NUTS V1999 to 2003).

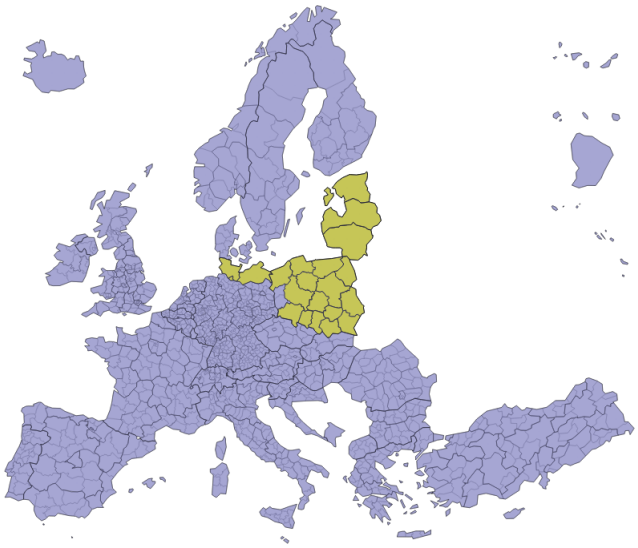


Figure 10.5 – Visualization of the TUs listed in Listing 10.6 that change their code in the NUTS from V1999 to V2003.

10.3 Discussion

As a reminder, the algorithm presented in this manuscript is developed for the purpose of improving the understanding of territorial evolution. Then, our challenge is to automatically find for any territorial change: *Where* and *When* the change occurs, *What* changes, *How* entities change but also, *Why* the boundaries change (because of a reform?) and *Who* (which institution, social, political, environmental, etc.) is responsible for this. In this Section, we discuss the scope and added value of our results. We first examine the capability of our implementation to answer the Where, When, What and How questions *i.e.*, we evaluate the effectiveness (algorithm complexity evaluation) and the generic behavior of our implementation that enables the management and versioning of any TSN hierarchical structure in the world. Then, in the Chapter 11 we study the feasibility of automating the contextualization of change events (*i.e.*, answer the Why and Who questions) relying on the current content of the LOD Web.

10.3.1 Algorithm Complexity

Regarding the analysis of the time complexity of the *TSN Semantic Matching Algorithm* 1, the running time of the first step of the algorithm (line 1 to 8) is the one that really interests us. We will focus on the evaluation of this first step that constructs the two main matrices `SpatialMatch` and `GlobalMatch`. Indeed, the remainder of the algorithm consists mainly in partially reading these matrices, for instance from indexes where are registered in the matrix the intersection of two TUs' polygons (*e.g.*, `FindStructureChange` function, line 14). We concentrate on finding the worst-case running time in order to provide users with an upper bound on the running time for any input. Within this first step of the Algorithm 1, we suppose that the number of TU's attributes (*e.g.*, TU's name, identifier, ...) is limited to a number that never exceeds a constant. Also, we suppose that the size of the TU's name is limited to a given number of characters. Then, we will ignore the real cost of the distance tests on each attribute and, in particular, the cost of the Levenshtein test on the name attribute, using the constant k to represent these costs.

The running time of our algorithm is the number of primitive operations executed on a particular input. In our case, there are two different sets of input parameters to consider. The first set of parameters is the set of TUs in each of the TSN versions V' and V'', of size m and n, respectively. The second set of parameters is the set of vertices composing the geometry of each of the TUs in V' and V'' (multi-polygon geometries). The notation v'_i is used to represent the number of vertices of one TU multi-polygon in the version V' of the TSN. The average number of vertices observed for TUs coming from one TSN version V', is noted $\bar{v'}$. The size of this set may be very large, for instance the Tasmania region of the ASGS TSN, version 2011 is made of 626,930 vertices (see line 2 Table 10.10). As notice by [Sun et al., 2003], processing these types of polygons is very expensive. The vertices of TUs are input parameters of the areal distance test [Bel Adj Ali and Vauglin, 1999] between two geometries (function *SpatialMatchTest* of our Algorithm 1, line 4).

	NUTS				SAU		ASGS	
	1999	2003	2006	2010	2017	2018	2011	2016
1. Average Number of vertices per TU ($\bar{v'}$ or $\bar{v''}$)	28.3	28.4	28.7	28.5	822.4	832.8	4,911.1	4,945.7
2. Max Number of vertices observed for a TU	472	473	473	483	44,960	44,842	626,930	625,334
3. TUs total number	1,861	1,902	1,933	1,923	2,457	2,419	2,784	2,680

Table 10.10 – Key Figures on several TSNs data sets, processed by the TSN SMA Algorithm 1.

Our implementation of this test results in the following SQL request (Listing 10.7). It computes a new geometry (if the geometries intersect) at the intersection of the two processed geometries (*ST_Intersection() function*[12] PostGIS function in our implementation).

```
1    SELECT CASE WHEN ((ST_Area(ST_Union(geom1,geom2))))= 0 THEN 0 ELSE
2    ((ST_Area(ST_Intersection(geom1,geom2))/(ST_Area(ST_Union(geom1,geom2
         ))))) END AS area_intersect
3    FROM unit_version WHERE ST_intersects(geom1,geom2) ");
```

Listing Code 10.7 – The Areal Distance Test implementation – The PostGIS DBMS SQL request executed during the *SpatialMatchTest* (see Algorithm 1 line 4) ("geom1" and "geom2" are input parameters – multi-polygon geometries of the two processed TUs in V' and V'').

Computing a new geometry at the intersection of others is a heavy and time-consuming task, regardless of the spatial data management system used. Indeed, as noticed in Furieri [2011], resolving the intersection for triangles or square is a simple and quick task whereas resolving highly complex polygons intersection (as for the Tasmania region of ASGS TSN version 2011) requires an impressive amount of floating point trigonometric functions to be calculated. In [Berg, 2008], the authors prove that, if one wants to compute the intersection of two polygons P_1 with v_1 vertices and P_2 with v_2 vertices (let $v = v_1 + v_2$), then $P_1 \cap P_2$ can be computed in $O(v \log v + q \log v)$ time, where q is the complexity of the output polygon at the intersection of P_1 and P_2 that depends on v. Thus, the algorithm is prohibitive for data sets that contain large polygons. Then, supposing the numbers of attributes and the number of characters in the name of TUs are limited (let us use the constant k to represent these costs and let $\bar{v} = \bar{v'_i} + \bar{v''_j}$, where $\bar{v}$ is an average number of vertices), the running time of the Algorithm 1 is:

12. https://postgis.net/docs/ST_Intersection.html

$$\sum_{i=1}^{m}\sum_{j=1}^{n} k + (\bar{v}\log\bar{v} + q\log\bar{v}) \tag{10.1}$$

Thus, Algorithm 1 has a worst-case running time of $\Theta(n\ m\ (\bar{v}\log\bar{v} + q\log\bar{v}))$.

Our implementation of the TSN Semantic Matching Algorithm creates the RDF graphs between the NUTS versions 1999 (with $\bar{v'} = 28.3$) and 2003 (with $\bar{v''} = 28.4$) with an execution time of 25 minutes, on an iOS system with 16GB of memory and 2,5 GHz Intel Core i7 processor. Since we do not have the constraint of generating descriptions of changes in real time, but rather a requirement for an accuracy of descriptions (calculated only once), result (running time) can be considered as satisfactory. However, in order to compute the graph between the ASGS version 2011 (with $\bar{v'} = 4,911.1$) and version 2016 (with $\bar{v''} = 4,945.7$), we have had to switch to a higher computing power, a Debian system with 141GB of memory and Intel Xeon CPU E5-2640 2.40GHz processor. Despite this higher computing power, the program creates the RDF graph between two versions of the ASGS with an execution time of 09 days 05 hours 25 minutes. It should be noted that in case of incremental changes (*e.g.*, if one constructs a new TSN version from a software and describes one change at a time), the problem is not the same. Indeed, with an incremental approach, one does not have to recalculate the full TSN change graph from the whole matrix of TUs in V' and V''. Since the system would receive gradually a small set of changed TUs, it could focus only on the local description of these changes. The complexity of our algorithm would then be quite different (and lower) and would be limited to the complexity of the number of vertices per polygon. Nevertheless, if one needs to calculate the entire change graph more rapidly, one should consider using other technologies such as spatial MapReduce algorithm solutions [Guo et al., 2015; Aji et al., 2013], or a solution that modifies the structure of the input data (instead of moving to a new technology).

For instance, the solution presented in Furieri [2011] (illustrated by Figure 10.6) splits the original complex TUs' polygon into many simpler polygons (such as squares), and then computes the intersections between square grids. This solution seems to be effective and implemented, but it still needs to be adapted to our version matching problem.

Another solution could be to reduce the number of vertices by simplifying the features (*i.e.*, generalize the geometries). But, as explained in the following subsection 10.3.2, simplifying the input geometries may reduce the precision of the change descriptions and introduces errors unless one follows the recommendations we make in the next subsection 10.3.2.

10.3.2 Geometries Generalization Problem

During the experiments and tests of our implementation of the TSN Semantic Matching Algorithm, we have faced two problems that are related to the generalization of geometries:

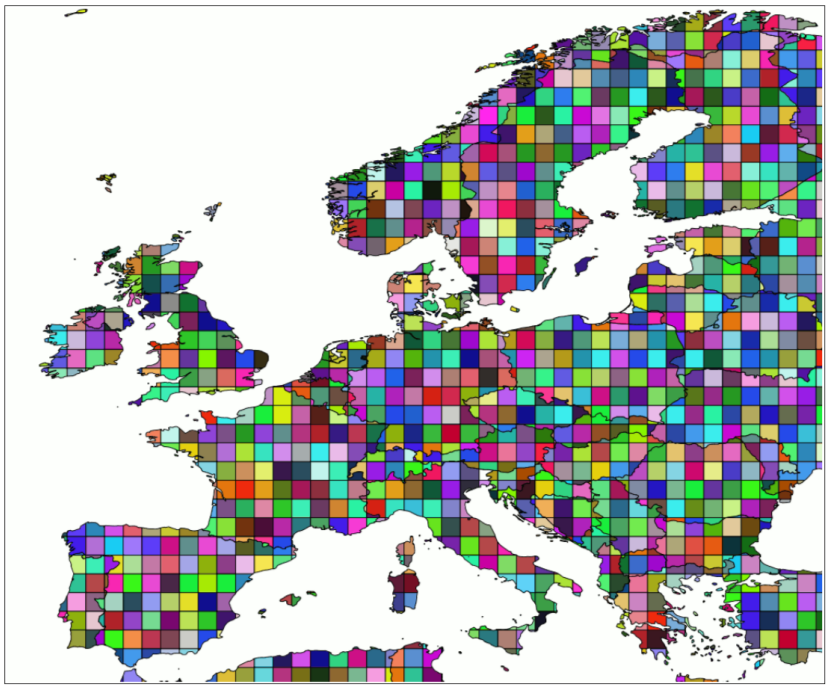


Figure 10.6 – A solution to reduce the calculation time of ST_Intersects() on complex polygons (more than 100 vertices) by reducing somewhat the number of vertices required to represent elementary polygons [Furieri, 2011].

– A. The first problem is about SA's online shapefile, available only as simplified/generalized geometries. Although simplified geometries are preferred to reduce the time needed to calculate the areal distance between two TUs that intersect (see subsection 10.3.1), the simplification of the features most of the time reduces the precision of data and introduces errors in the change descriptions computed by our program. For instance, the SAU shapefile provided by the Swiss Federal Statistical Office [13] are only distributed as generalized features. The problem is that, even if TUs have not changed between two successive versions, they are not generalized in the same way. Therefore, we have used other data available online and published by the SwissTopo organization. We have faced the same problem when looking for a solution to reduce the cost of calculating the intersection of two ASGS TUs containing a lot of vertices.

– B. The second issue is, at the opposite, that the geometries contain in the SAs' shapefile were too precise. Then, we have tried to reduce the number of vertices, facing the two following challenges:
– (1) preserve the topological consistency of boundaries between TUs;
– (2) do not introduce false TUs changes *i.e.*, generalizing in the same way an unchanged TU between two successive versions.

13. https://www.bfs.admin.ch/bfs/fr/home/services/geostat/geodonnees-statistique-federale/limites-administratives/limites-communales-generalisees.html

We have solved the issue (1) but not the (2) (we explain the consequences of that issue below).

The first challenge (1) has been solved using PostGIS functions. A SQL script was created and used on the ASGS data in order to reduce the number of vertices (please note that we may have used other tools such as the ARCGIS tool [14]).

```
1 -- Step 1.  Create a topology
2 SELECT CreateTopology('unit_version_topo', find_srid('public', '
    unit_version_simpl', 'geom_simpl'));
3 -- 2. Add a layer
4 SELECT AddTopoGeometryColumn('unit_version_topo', 'public', '
    unit_version_simpl', 'topogeom', 'MULTIPOLYGON');
5 -- 3. Populate the layer and the topology
6 UPDATE unit_version_simpl SET topogeom = toTopoGeom(geom_simpl, '
    unit_version_topo', 1);
7 -- 4.  Create Function SimplifyEdgeGeom
8 CREATE OR REPLACE FUNCTION SimplifyEdgeGeom(atopo varchar, anedge int,
    maxtolerance float8)
9 RETURNS float8 AS $$
10 DECLARE
11   tol float8;
12   sql varchar;
13 BEGIN
14   tol := maxtolerance;
15   LOOP
16     sql := 'SELECT␣topology.ST_ChangeEdgeGeom(' || quote_literal(atopo)
         || ',␣' || anedge
17       || ',␣ST_Simplify(geom,␣' || tol || '))␣FROM␣'
18       || quote_ident(atopo) || '.edge␣WHERE␣edge_id␣=␣' || anedge;
19     BEGIN
20       RAISE DEBUG 'Running␣%', sql;
21       EXECUTE sql;
22       RETURN tol;
23     EXCEPTION
24      WHEN OTHERS THEN
25       RAISE WARNING 'Simplification␣of␣edge␣%␣with␣tolerance␣%␣failed:␣%'
           , anedge, tol, SQLERRM;
26       tol := round( (tol/2.0) * 1e8 ) / 1e8; -- round to get to zero
           quicker
27       IF tol = 0 THEN RAISE EXCEPTION '%', SQLERRM; END IF;
28     END;
29   END LOOP;
30 END
31 $$ LANGUAGE 'plpgsql' STABLE STRICT;
32 -- 5. Simplify all edges up to 10000 units
33 SELECT SimplifyEdgeGeom('unit_version_topo', edge_id, 0.01) FROM
    unit_version_topo.edge;
34 -- 6. Convert the TopoGeometries to Geometries for visualization
35 UPDATE unit_version_simpl SET geom_simpl = topogeom::geometry;
36 UPDATE unit_version SET geom =  unit_version_simpl.geom_simpl  FROM
```

14. http://desktop.arcgis.com/en/arcmap/10.3/manage-data/editing-existing-features/simplifying-a-feature-generalize-.htm

```
unit_version_simpl WHERE unit_version.gid = unit_version_simpl.gid ;
```

Listing Code 10.8 – The SQL script that recalculates the topological network of a TSN and simplifies each TU geometry (polygon) while maintaining a network where simplified units are adjacent.

The second challenge (2) has not been solved. We came to the conclusion that since the boundaries of one TU change into the topological network, through neighborhood relation, the entire topological network is impacted and the generalization calculation is not done in the same way, even for some units that do not change between the two TSN' versions. This result is expected since the calculation of each point to be removed or kept during the generalization process is done considering all the points in the network. Since the network changes, the TU generalization from one version to another may change, even if the two TUs are exactly the same in both version.

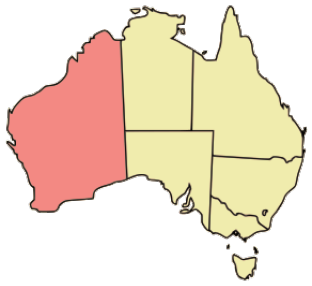


Figure 10.7 – A Map of Australia highlighting the Western Australia TU (source: https://commons.wikimedia.org/wiki/File:Western_Australia_locator-MJC.png.

On the Australian data set, the false changes introduced are of such a large size that even if the offset seems minimal, its order of magnitude is actually in hundreds of kilometers. For example, on the border of the TU called "Western Australia" whose perimeter (or frontier) is a straight line of over 1,874 Km [15] (see red TU on Figure 10.7), the offset created on the east straight line (see right side of the Figure 10.8 where the offset between the original (in green) and simplified (in red) geometries) is very important when one takes into consideration the entire frontier.

Even relaxing the spatial constraint when running our matching program on the simplified geometries of the ASGS, by modifying the spatial threshold parameter of the algorithm, does not lead to satisfactory results.

The only possible solution to both problems A and B is to start from a file (the first version of the TSN for instance), to simplify this file, and to report in a copy

15. Source: Australian Government Geo-science Australia - Border Lengths States and Territories http://www.ga.gov.au/scientific-topics/national-location-information/dimensions/border-lengths

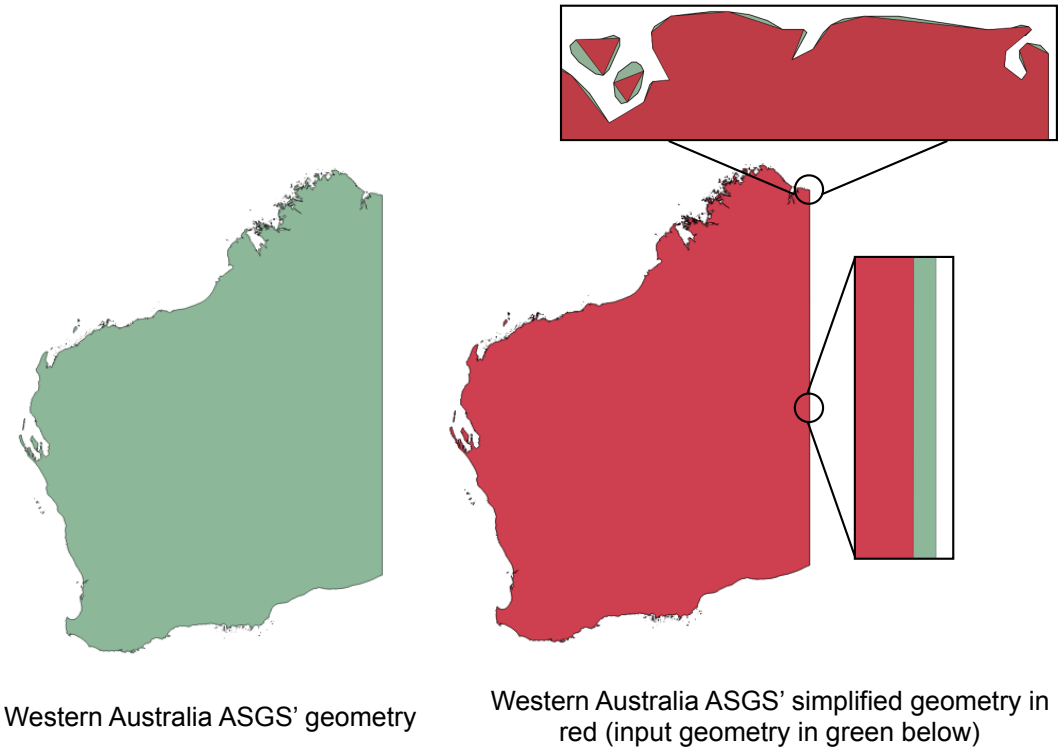


Figure 10.8 – Western Australia TU (ASGS version 2011) – The offset between the original geometry (in green) and the simplified one (in red).

of this generalized file, the changes of the new TSN version. In our opinion, this is the only way to ensure the quality of the TSN data between two successive versions, even if they are generalized.

10.3.3 Genericity of the approach

We discuss here the genericity and efficiency of our approach, with regard to other geographic divisions in the world. Conceptually, the approach tends to be as neutral and generic as possible. In fact, the *TSN Ontology* has a sufficient level of abstraction to semantically describe any hierarchical structure of geographic divisions. Using the *TSN Ontology*, our TSN Semantic Matching Algorithm may also process what is called *Socio-economic Units* in [Frank et al., 2003] or *Jurisdictional Domains* in [Lopez-Pellicer et al., 2012]. As far as some requirements are fulfilled, the approach implemented in this algorithm is generic, and makes it adaptable to the specificities of many TSN given as input. We free ourselves from the heterogeneity of shapefile sources by specifying the columns needed as inputs to our system [Telechev and Le Rubrus, 2013].

Indeed, in order to adapt to any TSN hierarchical structure in the world, we make our TSN Semantic Matching Algorithm configurable. There are two sets

of parameters to adjust for the matching of the TSN versions, depending on the characteristics (features) of data:

(1) the spatial thresholds ($\varepsilon_{spatialStructure}$ and $\varepsilon_{spatialFeature}$). They are used during the test of intersection and equality of two geometries. They represent a tolerated difference in the geometries of two TUs that intersect. If after calculating the areal distance between the two TUs, this distance is smaller than the threshold, the algorithm concludes that the geometries are equal. The $\varepsilon_{spatialFeature}$ threshold is used to identify, in the first step of the algorithm, whether the geometry of a TU has changed or not. The $\varepsilon_{spatialStructure}$ threshold is used in a second step to validate the cluster of changes found. The greater the spatial tolerance threshold is, the more units (that intersect), different in their geometry, will be considered unchanged in their boundaries. This tolerance is of huge importance when the two versions of the processed TSN have not the same level of detail (resolution) for their geometries. Thus, the use of the areal distance is a key to handle data quality issues linked to geometries.

(2) the weights carried by each attribute of a TU (*e.g.*, $\alpha_{id}, \alpha_{geom}, \alpha_{name}, \alpha_{super}$). A weight determines in which proportion this attribute is part of the identity of the TU. Once attributes are considered unchanged, their weights are summed to obtained a global score of matching of the two TUs u' and u''. If the global score is greater or equal to the β threshold: the algorithm concludes that the continuity of the identity of the TU u' with u''. By modifying the global β threshold, the nature of the identified territorial events may vary. For instance, we have used a β threshold fixed to 0.5 for the NUTS data sets, which leads us to restore all the changes described by Eurostat. But, we have experimented that by modifying its value to 0.7, the *Extraction* event of the TU ES63 is no longer recognized as an *Extraction* but as a *Scission* since the matching between the two TUs ES63 V1999 and V2003 is no longer established. Indeed, using the following weights on attributes: $\alpha_{id} = 0.4, \alpha_{geom} = 0.4, \alpha_{name} = 0.1, \alpha_{super} = 0.1$, a continuation link is described between $ES63'$ and $ES63''$, for a β threshold fixed to 0.5, as the two versions of the TUs match with a score of 0.5, obtained thanks to the *id* and *super* attributes only ($0.4(\alpha_{id}) + 0.1(\alpha_{super}) = 0.5$). If the β threshold is fixed to 0.7, the $ES63''$ TU is no longer considered as a continuation of the TU $ES63'$ (the score $0.5 < 0.7$), modifying also the nature assigned to the cluster of changes in which they are involved.

For the NUTS and SAU TSNs, the same weights and thresholds have been chosen (see Tables 10.2 and 10.5). This is because, first, the two TSNs share the same definition of what makes the identity of their TUs. Thus, the list of attributes considered for comparison of the TUs' identity over time, and the weights assigned to these attributes, are the same for these two TSNs. Secondly, the two TSNs shapefiles contain no error: their geometries are perfectly identical version to version, unless there has been a "real" change in the TSN (*i.e.,* "real change" means here a change listed in the official catalog of changes of the TSN, by opposition to "false" changes, that are errors in a shape file version most of the time due to offsets after generalization of the geometries). Thus, for these two TSNs, the spatial thresholds are set to a minimum, to ensure the detection of all spatial changes.

We have however processed a third data set that we have found interesting to consider here: the Australian ASGS that comes with two versions 2011 and 2016. In this TSN, there are false changes, in the form of minimal geometry changes, that are not listed in the official catalog of changes to this TSN (available online, see Table 10.7, line Official Dictionary). Figure 10.9 shows an example of these differences of geometry from one version to another: in green the TU Salisbury in the ASGS version 2016 and under, in yellow, the TU Salisbury in the ASGS version 2011. There is here an offset between the two versions of the geometry of the TU. This offset seems to be an error between the two data sets since this change in the geometry of the TU is not registered in the official catalog of changes of the ASGS.

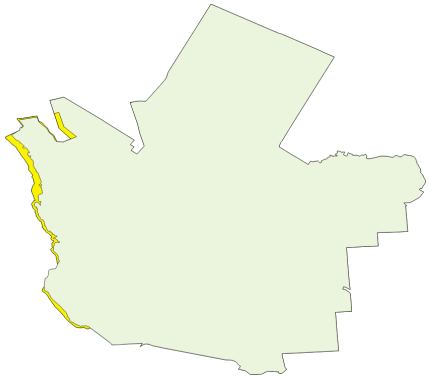


Figure 10.9 – The two TUs named Salisbury (id 40204) of the ASGS TSN in version 2011 and 2016 (the TU version 2016 is displayed in yellow, under the TU in the version 2011 to observe the boundaries differences).

We have run the algorithm a first time with the same thresholds and weights as for the TSNs NUTS and SAU: here, the RDF outputs describe those *false* changes of geometries for these TUs while it would be better if those minimal changes of geometries were not described. Therefore, we have decided to relax the spatial constraints: to allow a small difference between geometries, and consider them similar despite this small difference. Obviously, the more we relax the spatial constraint (or other), the less we find change. We have therefore modified the spatial thresholds, from 0.005 to 0.1. We have run the algorithm on a small data set composed of 3 TUs of the ASGS (named "Albion Park Rail", "Chermside West" and "Salisbury" in both version 2011 and 2016) that share the same characteristics: they have undergone no official change from version 2011 to 2016, although their geometries in the shapefile version 2016 suffer from small modifications. When running the algorithm with the relaxed spatial constraint (thresholds at 0.1), no *GeometryChange* node is created. Instead, only lineage links *hasPreviousVersion*, *hasNextVersion* are created. One then may perceive the usefulness of the flexibility of our algorithm in such cases.

	Run parameters	Run parameters (less binding)
Vbefore	ASGS 2011	ASGS 2011
Vafter	ASGS 2016	ASGS 2016
Threshold beta	0.5	0.5
Threshold spatialStructure	0.005	0.1
Threshold spatialFeature	0.005	0.1
Weight geometry	0.4	0.4
Weight identifier	0.4	0.4
Weight super	0.1	0.1
Weight name	0.1	0.1
Number of TUs	3 (2011) ; 3 (2016)	3 (2011) ; 3 (2016)
Number of geometry change detected	3	0

Table 10.11 – Variability of the TSN matching Algorithm outputs after modifying the Spatial Thresholds.

To conclude, as expected, the algorithm has a certain variability in its results which depends on the parameters. This allows to adapt to the specificity of each TSN and to generate different semantic graphs depending on the definition of the identity of the TUs chosen. It requires the intervention of experts to specify how to recognize the continuation of the identity of a TU. We have discussed here the genericity of our approach, whose counterpart can be the variability of the results obtained by our algorithm which is configurable to adapt to the specificity of each TSN. We have proved the usefulness of the parameters and of the thresholds in the case of small offsets in the geometry of TUs between two versions of the TSN. In such special cases, a small spatial tolerance to change makes it possible not to semantically annotate the changes that were made by mistake in the data set, and that are not real changes in the TSN.

10.4 Conclusion

Regarding first the impact of the resources presented in the chapter, we think that the created RDF graphs constitute a knowledge base of data not yet available on the LOD Web with such a level of semantics. Moreover, these graphs allow to extract patterns of change, simulate scenarios of evolution, and comparison of the TSNs. Also, our approach and the resources created from it, foster SAs' systems interoperability in the domain of change description of geospatial data.

We make our TSN Semantic Matching Algorithm configurable to adapt to the specificity of each TSN. In that respect, we have designed it in a highly configurable way. It takes as input a definition of what makes the identity of the geographic objects processed. This definition consists of a weighted list of attributes that constitute the identity of the TUs (*e.g.*, the name or the geometry attributes of TUs). There are two sets of parameters (the weights associated with the list of TUs' attributes and the spatial thresholds) that can be used for the matching of the TSN versions, depending on the characteristics of the data.

There is also a potential for extensibility to meet future requirements. For instance, currently we are working on several versions of the Corine Land Cover Data set and we are experimenting the expansion of the TSN-Change Ontology and TSN Semantic Matching Algorithm in order to take into account another attribute in the definition of the identity of the TUs: the nature of the land surface Harbelot et al. [2015].

Regarding the design quality, as explained previously, we follow the W3C *Data on the Web Best Practices* recommendations Farias Lóscio et al. [2017], the W3C series of best practices for publishing Linked Data Hyland et al. [2014], and the methodological guidelines of Villazón-Terrazas et al. [2011] for the publication and use of high quality Linked Government Data. For each concept defined, we use (when relevant) vocabularies such as *Dublin Core Metadata Initiative terms*[16] (for metadata such as the source of data) or *PAV - Provenance, Authoring and Versioning Ontology*[17]. For instance, we refine the property *pav:hasVersion*, by specifying the domain and range of this property using the *TSN Ontology* concepts. The concepts *TerritoryVersion* (*e.g.*, *EU of 15*) and *UnitVersion* are subclasses of the *geo:Feature* concept defined within the OGC *Geosparql* ontology. The *OWL-Time* vocabulary is used to assign a *reference period* to a *NomenclatureVersion* and *TerritoryVersion* resources.

In the following Chapter 11, we show how to query these graphs and exploit them with other data available on the LOD Web.

16. http://purl.org/dc/terms/
17. http://purl.org/pav/

Chapter 11

Exploring and Exploiting the TSN model and data

11.1 Introduction

The created RDF graphs from the NUTS, ASGS and SAU data sets (presented in the previous Chapter 10) constitute catalogs of TUs versions. Most of all, they draw the lineage of each TU over time (*horizontal reading* of the graphs) and provide a representation of the propagation of a change event through the divisions levels (*vertical reading*).

In order to highlight the potential of the created graphs, we come back here to the change example of Ceuta and Melilla (already presented in Chapter 9) that happened in the NUTS. We show here how, by querying our graphs, users are given access to rich descriptions of this change. In the NUTS version 1999, at Level 2, the two Spanish enclaves in North Africa, *Ceuta and Melilla*, were considered as one TU (with identifier ES63) part of a Spanish province (please see Figure 9.2). In 1995, both Ceuta and Melilla became Autonomous cities of Spain, requiring a redistricting of the Spanish regions. It resulted in a split of the TU ES63 in two TUs named *Ciudad Autònoma de Ceuta* (with identifier ES63) and *Ciudad Autònoma de Melilla* (with identifier ES64), in the NUTS Version 2003. Also, the sub-TUs of ES63 (ES631 and ES632) in the NUTS version 1999 both change identifiers. We try to find in our graph how this territorial change is described, first by a vertical reading of the graph, second by a horizontal reading.

Please note that an online data set at `http://purl.org//steamer/tsnoc/resources/tsn_sparql_requests.pdf` lists several SPARQL queries on the 3 TSN change graphs ASGS, SAU and NUTS.

11.2 Knowledge extraction from TSN history graphs

11.2.1 Vertical reading of the change graphs

The following query, in Listing 11.1, addressed to our SPARQL endpoint `http://steamerlod.imag.fr/sparql?&query=`, shows how a program may automatically extract from the `http://purl.org/steamer/nuts` graph all the TUs that have changed between two versions, at a specific level. Here, we ask for the list of all the NUTS 1999 major regions that change from 1999 to 2003. This request

URI of the TUs that change
nuts:V1999_L1_ES3
nuts:V1999_L1_ES6
nuts:V1999_L1_ES
nuts:V1999_L1_FI1
nuts:V1999_L1_FI2
nuts:V1999_L1_FR1
nuts:V1999_L1_FR3

Table 11.1 – Excerpt of the list of TUs that change at NUTS Level 1, from version 1999 to 2003, result of the query 11.1.

returns, among others, the URI of the TU ES6 (`http://purl.org/steamer/nuts/V1999_L1_ES6`), the super-TU of ES63.

```
1    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2    PREFIX tsnchange: <http://purl.org/net/tsnchange#>
3    PREFIX tsn: <http://purl.org/net/tsn#>
4    SELECT DISTINCT ?TU WHERE {
5       ?TU rdf:type tsn:UnitVersion .
6       ?TU tsnchange:input ?change .
7       ?change a tsnchange:Change .
8       ?TU tsn:isMemberOf ?level .
9       ?level  tsn:hasIdentifier  "NUTS_V1999_L1" . } ORDERBY ?TU
```

Listing Code 11.1 – A SPARQL query that returns all the TUs that change between two TSNs versions at a specified territorial level.

From the list of TUs (an excerpt is shown on Table 11.1) returned by the query of Listing 11.1, we can explore the sub-changes graph of each of the TUs in the list, and discover how their sub-units change or not. Listing 11.2 shows how to explore the sub-changes graph of the TU ES6. At line 10, we ask for the change that affect the TU `nuts:V1999_L1_ES6` (if there is one), and for all its sub-changes. At line 12, we ask to stop the chain of changes on the sub-TU `nuts:V1999_L3_ES631`. This is to focus on the Ceuta and Melilla case. Then, this query searches for change paths between two embedded TUs, at levels L1 and L3 of the NUTS, using SPARQL path expression.

```
1    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2    PREFIX tsnchange: <http://purl.org/net/tsnchange#>
3    PREFIX tsn: <http://purl.org/net/tsn#>
4    CONSTRUCT { ?s ?p ?o .
5      ?TU_before tsnchange:intputUnitVersion ?o .
6      ?TU_after tsnchange:outputUnitVersion ?o .}
7    FROM <http://purl.org/steamer/nuts> WHERE{
8    SELECT DISTINCT ?s ?p ?o ?TU_before ?TU_after where {
9     values ?p {tsnchange:inputUnitVersion tsnchange:lowerChange }
10    nuts:V1999_L1_ES6 (tsnchange:inputUnitVersion|tsnchange:lowerChange)
          * ?s .
```

```
    ?s ?p ?o .
    ?o (tsnchange:lowerChange)*  nuts:change_unitchange_V1999_L3_ES631 .
   Optional{ ?TU_before tsnchange:inputUnitVersion* ?o .
             ?TU_before a tsn:UnitVersion .
             ?TU_after tsnchange:outputUnitVersion* ?o .
             ?TU_after a tsn:UnitVersion .}}}
```

Listing Code 11.2 – A SPARQL query that returns the multi-levels change graph of a TU.

Figure 11.1 – The Multi-levels change graph of the NUTS 1999 TU ES6 (result of the query in Listing 11.2, also online at `http://purl.org/steamer/nuts_V1999_ES6_change_graph`).

Figure 11.1 provides an automatically generated graphic representation of the graph constructed after the query in Listing 11.2 thanks to the GraphDB visualization tool. The red nodes represent the TUs that change (please note that the graph on Figure 9.5 gives a more accurate view on this graph and on all the NUTS elements that change). At the top of the graph, the super-TU `nuts:V1999_L1_ES6`

is linked to a change node of type `tsnchange:SubUnitChange`. This node informs ES6 that its sub-TU changes and it is linked to a sub-change that describes the split of the sub-TU `nuts:V1999_L2_ES63`. The term used to describe this change is *tsnchange:Extraction* that is a sub-class of the *tsnchange:Split* concept, but more precise as it means that the initial TU that splits still exists after change. Finally, the graph describes the *Identification restructuration* of all the sub-TUs to ES63 (purple node).

For path search between two TUs (*e.g.*, a chain of change nodes between a super-TU and one of its sub-TUs), we ran into SAPRQL language limitations, including its weakness in terms of finding paths and especially shortest paths. The SPARQL query 11.2 uses property path expressions to search for a path between two resources. Property paths give a way to write parts of basic graph patterns [Harris and Seaborne, 2013]. And, as explained in [Sirin, 2017], these property paths "*can be used for queries that recursively traverse the RDF graph and find two nodes connected via a complex path of edges. But the result of a property path query is only the start and end nodes of the path and does not include the intermediate nodes. To find the intermediate nodes additional or more complex queries are needed.*" The query 11.2 illustrates such complex queries that return intermediate nodes of changes between two TUs (a super-TU and one of its sub-TU). Then, in an online documentation [Sirin, 2017], the author introduces an extension to the Stardog triplestore software, called *Pathfinder*, with the goal of providing an elegant, concise syntax for path queries in SPARQL. This extension allows, among others, to search for the shortest path between two resources, using the following concise syntax shown in Listing 11.3:

```
1PATH ?p FROM ([startNode AS] ?s) TO ([endNode AS] ?e) {
2   GRAPH PATTERN
3}
4[ORDER BY condition]
5[LIMIT int]
```

Listing Code 11.3 – Pathfinder syntax of path queries, source Sirin [2017].

Such extension would be particularly useful in the case of TSN change graphs and facilitate the search for chains of changes.

Please, note that one can also search for a sub-graph of changes without specifying an end node, and, conversely, by specifying the type of changes one wants to follow, as shown in the following query in Listing 11.5. This query returns the sub-graph of StructureChange (see filter on the type of changes at line 15) from the Australian super-TU `asgs:V2011_L4_508`.

```
1PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2PREFIX tsnchange: <http://purl.org/net/tsnchange#>
3PREFIX tsn: <http://purl.org/net/tsn#>
4PREFIX : <http://purl.org/steamer/asgs/>
5CONSTRUCT {
```

```
    ?s ?p ?o . ?TU_before tsnchange:intputUnitVersion ?o .
    ?TU_after tsnchange:outputUnitVersion ?o .}
FROM <http://purl.org/steamer/asgs>
WHERE{
    select distinct ?s ?p ?o ?TU_before ?TU_after
    where {
        values ?p {tsnchange:inputUnitVersion tsnchange:lowerChange}
        :V2011_L4_508 (tsnchange:inputUnitVersion|tsnchange:lowerChange)*
             ?s .
        ?s ?p ?o .
        ?o rdf:type tsnchange:StructureChange .
        Optional{
            ?TU_before tsnchange:inputUnitVersion* ?o .
            ?TU_before a tsn:UnitVersion .
            ?TU_after tsnchange:outputUnitVersion* ?o .
            ?TU_after a tsn:UnitVersion .        }
    }
}
```

Listing Code 11.4 – A SPARQL query that searches for the sub-graph of changes of a super-TU.

Also, the following query in Listing 11.5 shows how to traverse the graph but in the ascending direction this time:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX tsnchange: <http://purl.org/net/tsnchange#>
PREFIX tsn: <http://purl.org/net/tsn#>
PREFIX : <http://purl.org/steamer/asgs/>
CONSTRUCT {
 ?s ?p ?o .
 ?uBefore tsnchange:intputUnitVersion ?o .
 ?uAfter tsnchange:outputUnitVersion ?o .}
FROM <http://purl.org/steamer/asgs> WHERE{
select distinct ?s ?p ?o ?uBefore ?uAfter where {
  values ?p {tsnchange:inputUnitVersion tsnchange:upperChange }
  :V2011_L2_508011194 (tsnchange:inputUnitVersion|tsnchange:upperChange)*
       ?s .
  ?s ?p ?o .
  ?o (tsnchange:upperChange)*  :change_unitchange_V2011_L4_508.
  Optional{
  ?uBefore tsnchange:inputUnitVersion* ?o .
  ?uBefore a tsn:UnitVersion .
  ?uAfter tsnchange:outputUnitVersion* ?o .
  ?uAfter a tsn:UnitVersion .
        }
  }}
```

Listing Code 11.5 – A SPARQL query that, from a TU in the lowest level, goes up the chain of changes until a super-TU.

11.2.2 Horizontal reading of the change graphs

The following query shows how to extract from the `http://purl.org/steamer/nuts` graph (a graph composed of a succession of 4 versions of the NUTS TSN), the life line of one TU using a filter on its identifier "ES63" in the NUTS TSN.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX tsnchange: <http://purl.org/net/tsnchange#>
PREFIX tsn: <http://purl.org/net/tsn#>
CONSTRUCT { ?TU_input tsnchange:inputUnitVersion ?change ;
            tsnchange:hasNextVersion ?TU_output  .}
FROM <http://purl.org/steamer/nuts> WHERE{{
    ?TU_input tsnchange:inputUnitVersion ?change .
        ?change   tsnchange:unitVersionAfter ?TU_output .
        ?TU_input tsn:hasIdentifier "ES63".}
UNION {?TU_input tsnchange:hasNextVersion ?TU_output .
    ?TU_input tsn:hasIdentifier "ES63".}}
```

Listing Code 11.6 – A SPARQL query that returns the life line of a TU.

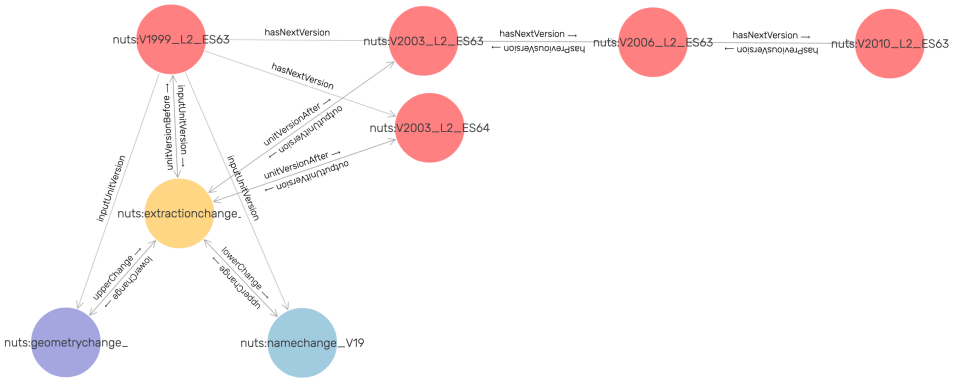


Figure 11.2 – The life line of the NUTS TU ES63 (result of the query in Listing 11.6, also online at `http://purl.org/steamer/nuts_ES63_lifeline`).

Figure 11.2 is an automatically generated graphic representation of the graph constructed by the query of Listing 11.6. The red nodes represent the successive versions of the TU with code ES63 in the NUTS. Also, the TU `nuts:V2003_L2_ES64` is shown as the result of the *Extraction* change, that occurred between the NUTS 1999 and 2003. As can be seen from this view, the *Extraction* change is the only change that has affected the TU ES63 during its life.

11.2.3 Towards a user interface to geo-visualize changes

We would like to highlight here the potential of the TSN ontological model for geo-visualization, and, in particular, for geo-visualization like in Version Control

Systems [Ruparelia, 2010; Negretti, 2015]. In Version Control Systems, there is a code of colors that consists in coloring in green the elements that do not change from one version to another, in orange the elements that change but that do not disappear, and in red the elements that disappear, as shown in the following GitHub Figure 11.3 [Balter, 2014; Storey et al., 2005]. This picture is an example of the views GitHub Inc [1] is able to create for geospatial data. Please, note that changes in geometries only are displayed and that there is no semantic of changes in GitHub.

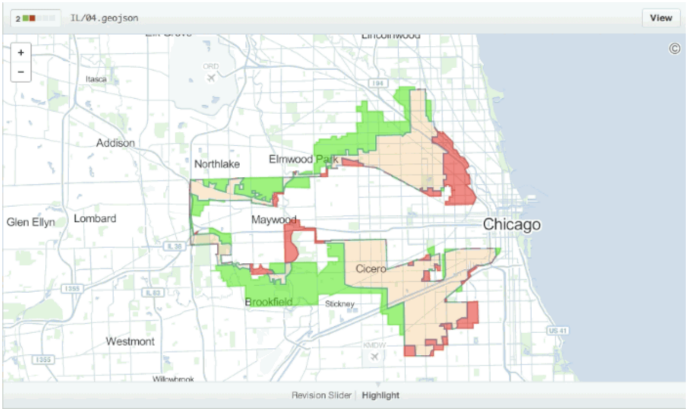


Figure 11.3 – Visualization of evolving geospatial data with GitHub [Balter, 2014].

We believe that combining geo-visualization capabilities of tools such as GitHub with our semantic of changes, in a geo-visualization software, may help users in easily understanding the dynamics of the territories over time. We list below the three queries that the software could run in order to color the TUs, in the origin version, into the appropriate green, orange, or red color.

Green The following SPARQL query (Listing 11.7) requests all the TUs at a specific level that do not change, from an origin version. The filter, line 9, excludes all TUs linked to a change node.

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX tsn: <http://purl.org/net/tsn#>PREFIX nuts: <http://purl.org/
    steamer/nuts/>
3 PREFIX tsnchange: <http://purl.org/net/tsnchange#>
4 SELECT DISTINCT ?TU
5 WHERE {
6     ?TU rdf:type tsn:UnitVersion .
7     ?TU tsn:isMemberOf ?level .
8     ?level  tsn:hasIdentifier  "NUTS_V1999_L1" .
9 FILTER (
10    !EXISTS {
11    ?TU tsnchange:input ?change .
```

1. https://github.com/

```
12})}
```

Listing Code 11.7 – Query the TUs that do not change from one version to another at a specific territorial level.

Orange The following SPARQL query (Listing 11.8) requests all the TUs at a specific level that change, from an origin version. Line 8, the query asks for all TUs linked to a change node, whatever its type. However, line 11, the "MINUS" SPARQL keyword excludes all TUs linked to a change node of type *Disappearance*. This is because, this change type, means that the TUs has disappeared after the change event.

At line 8, one can further specify the type of changes s/he wants to see, and color in orange, for example, TUs having changed their name (`?change a tsn-change:NameChange`).

```
1PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2PREFIX tsnchange: <http://purl.org/net/tsnchange#>
3PREFIX tsn: <http://purl.org/net/tsn#>
4PREFIX nuts: <http://purl.org/steamer/nuts/>
5SELECT DISTINCT ?TU WHERE {
6   ?TU rdf:type tsn:UnitVersion .
7   ?TU tsnchange:input ?change .
8   ?change a tsnchange:Change .
9   ?TU tsn:isMemberOf ?level .
10  ?level  tsn:hasIdentifier  "NUTS_V1999_L1" .
11  MINUS
12    {?change a tsnchange:Disappearance . }
13  }
```

Listing Code 11.8 – Query the TUs that change from one version to another at a specific territorial level.

Red The following SPARQL query (Listing 11.9) requests all the TUs at a specific level that disappear, from an origin version to a destination version:

```
1PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2PREFIX tsnchange: <http://purl.org/net/tsnchange#>
3PREFIX tsn: <http://purl.org/net/tsn#>
4PREFIX nuts: <http://purl.org/steamer/nuts/>
5SELECT DISTINCT ?TU
6WHERE {
7    ?TU rdf:type tsn:UnitVersion .
8    ?TU tsn:isMemberOf ?level .
9    ?level  tsn:hasIdentifier  "NUTS_V1999_L1" .
10   ?TU tsnchange:input ?change .
```

```
11    ?change rdf:type tsnchange:Disappearance .}
```

Listing Code 11.9 – Query the TUs that disappear from one version to another at a specific territorial level.

11.3 Automatic contextualization of territorial changes

By exploiting the distributed LOD Web, we can enrich our descriptions of TSN and of their changes, by looking for other resources and understand more deeply the evolution of the territory. Different data sets exist on the LOD cloud, but the ones we found the most relevant in our context are those from DBPedia [2] and WikiData [3]. Indeed, they are generalists and provide encyclopedia-style information, such as historical data. Then, it may be possible to find the historical cause of a territorial change in these data sets. A trivial example, addressing the DBPedia service within a SPARQL request allows to discover that the Ceuta TU of the NUTS TSN (using only the information on the TU identifier within the NUTS version 2003, *i.e.*, ES630) is an Autonomous city of Spain (`dbo:type dbr:Autonomous_cities_of_Spain`), since 1995 (`dbo:foundingDate 1995-03-14^^xsd:date`).
We also get, among other information, the total population of this TU (`dbo:populationTotal 823768sd:integer`), information that may help statisticians at the time of comparing this TU with others in Europe. In particular, the low number of inhabitants of this TU, compared to the average number of inhabitants of TUs at NUTS Level 3, should lead analysts to be careful in their comparative studies of the European regions. Indeed, as stated on the Eurostat's Website, "*despite the aim of ensuring that regions of comparable size all appear at the same NUTS level, each level still contains regions which differ greatly in terms of population*", such as Luxembourg that is a small country in Europe, which appears at NUTS level 0, despite its low number of inhabitants.

```
1     PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2     PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3     PREFIX geosparql: <http://www.opengis.net/ont/geosparql#>
4     PREFIX tsn: <http://purl.org/net/tsn#>
5     PREFIX dbo: <http://dbpedia.org/ontology/>
6     PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
7     SELECT *
8     WHERE {
9     ?tu a tsn:UnitVersion;
10    tsn:hasIdentifier "ES630";
11    tsn:hasName ?name;
12    tsn:isMemberOf ?level;
13    geosparql:hasGeometry [
14    geosparql:asWKT ?geom; ].
15    ?level tsn:hasIdentifier ?id_level ;
16           tsn:isDivisionOf ?tsn_version .
```

2. https://wiki.dbpedia.org/
3. https://www.wikidata.org/wiki/Wikidata:Main_Page

```
17    ?tsn_version tsn:hasIdentifier "NUTS_V2003"^^xsd:string .
18    ?tsn_version tsn:hasAcronym ?tsn_acronym .
19    OPTIONAL{
20        SERVICE <http://dbpedia.org/sparql> {
21        ?place rdf:type dbo:Place ;
22               rdfs:label ?label ;
23               dbo:abstract ?abstract ;
24               dbo:type ?placeType ;
25               dbo:foundingDate ?foudingDate ;
26               dbo:populationTotal ?population .
27     FILTER (lang(?label) = 'en') FILTER (lang(?abstract) = 'en')
28     FILTER (str(?label) = str(?name))}}}
```

Listing Code 11.10 – A SPARQL request to the DBPedia Service in order to find contextual information about a territorial change.

Table 11.2 below shows an excerpt of the response obtained from the request in Listing 11.10. Please note that we could have also used the WikiData service that provides similar information (at `https://www.wikidata.org/wiki/Q5823` for Ceuta).

TU URI	http://purl.org/steamer/nuts/V2003_L3_ES630
TU Name	Ceuta
TU's Level	http://purl.org/steamer/nuts/V2003_L3
TU's Version	http://purl.org/steamer/nuts/V2003
Dbo Place URI	http://dbpedia.org/resource/Ceuta
Dbo Place Type	http://dbpedia.org/resource/Autonomous_cities_of_Spain
Dbo Abstract	Ceuta, along with the Spanish exclave Melilla, is one of two permanently inhabited Spanish territories in mainland Africa. It was part of Cádiz province until 14 March 1995 when the city's Statute of Autonomy was passed. Ceuta, like Melilla, was a free port before Spain joined the European Union. As of 2011, it has a population of 82,376.
Dbo Founding Date	1995-03-14
Dbo Population	82376

Table 11.2 – Part of the result of the SPARQL query 11.10.

While the DBPedia predicate `dbo:foundingDate` (or the predicate `wd:-inception` from WikiData) points to a date in time when Ceuta was founded, it is in the free text parts, such as in the Dbo abstract (`dbo:abstract`), that a human learns more about Ceuta and understands the event that has caused its change in the NUTS. In particular, in the abstract free text of DBPedia, we learn about the link between Ceuta and Melilla. Yet, in both DBPedia and WikiData, we did not find any explicit RDF triple description of the (administrative) split of Ceuta and Melilla (administrative reform), after obtaining their status of autonomous city of Spain. Regarding the disappearance of the 6 districts of the canton of Neuchâtel,

in the SAU, neither DBPedia nor WikiData have information about this territorial change. We presume this is because the event occurs quite recently on 1 January 2018.

Various models have been proposed for representing events, as explained in the State of the Art part, Subsection 5.2.2. In order to automate the contextualization of territorial changes, using the DBPedia service we encounter problems because listed events in DBPedia are few and limited to military battles and sport events. They cannot be queried in a structured way. In [Hienert and Luciano, 2015], the authors present a solution that automates the extraction of historical events from articles in Wikipedia, using language-dependent patterns to identify sentences containing a date. For the representation of these events, they have chosen the LODE ontology from [Shaw et al., 2009] because it is a domain-independent and a light weighted structure. Nevertheless, it seems that the created SPARQL endpoint does no longer exist [4]. They have also created a GUI and a Web API to visualize and query the historical events extracted from Wikipedia articles. For instance, the following query `http://www.vizgr.org/historical-events/search.php?query=Ceuta` returns in XML a list of historical events that take place in Ceuta. Yet, the autonomy of Ceuta in 1995 is not in this list.

The autonomous status of Ceuta and Melilla was introduced by the Law 1/1995 and 2/1995, March 13, of the Statutes of Autonomy of Ceuta and Melilla [5]. In Europe, the European Legislation Identifier (ELI) [6] provides Web identifiers (URIs) for legal information. For instance, in France the Law No 2015-29 of January 2015, the 16th, redefining the regional administrative areas for France, has an ELI URI that is `https://www.legifrance.gouv.fr/eli/loi/2015/1/16/INTX1412841L/jo/texte`. The implementation of ELI is still a work in progress in Spain, since the Spanish legal system is complex and diverse, comprising rules corresponding to different territorial levels (national, Autonomous Community and local) [7]. Thus, the URI of the Laws 1/1995 and 2/1995, March 13, may be soon available. Everyone will then be able to automatically link to these URIs, using the ELI patterns for URIs construction, and therefore provide users with the legal context for the territorial change. In Wikidata, though there is the predicate *significant event* [8] to express a significant or notable event associated with the subject, we notice that it is not used for Ceuta for instance, and no description is returned on the element *history of Ceuta* [9]. Yet, in both DBPedia and WikiData, we have not found any explicit triple description of the administrative division of Ceuta and Melilla, after obtaining their status of autonomous city of Spain. Then, our proposal for future work is to:

(1) reuse the automatic extraction method of [Hienert and Luciano, 2015] to

4. The URL provided in the paper `http://lod.gesis.org/historicalevents/sparql` return a 404 error code.

5. `http://noticias.juridicas.com/base_datos/Admin/lo1-1995.html` `http://noticias.juridicas.com/base_datos/Admin/lo2-1995.html`

6. `https://eur-lex.europa.eu/eli-register/about.html`

7. https://eur-lex.europa.eu/eli-register/spain.html

8. https://www.wikidata.org/wiki/Property:P793

9. `https://www.wikidata.org/wiki/Q19632088`

extract historical event from Wikipedia;

(2) create a triple representation of these historical events using the LODE Ontology [Shaw et al., 2009] for their description or the CIDOC CRM Ontology, designed for the representation of historic events [Doerr, 2003];

(3) link TSN change nodes with event nodes created after step (2), using the predicate `tsnchange:isCausedBy` in order to express in RDF that a territorial change is caused by this historical event.

Regarding the disappearance of the districts of the canton of Neuchâtel, as noticed previously, no information was available in January 2019 neither in WikiData nor in DBPedia (e.g., `http://dbpedia.org/resource/Canton_of_Neuchatel`). However, this event is described in Wikipedia as follows: "*Until 2018 the Canton was divided into 6 districts. On 1 January 2018, the districts were dissolved and all municipalities were placed directly under the canton*". Since this description contains a date, we think that the method from [Hienert and Luciano, 2015] is relevant in this case to extract automatically a description of this event. Thus, by linking this event to SAU TUs representations before and after the event, described using the *TSN Ontology* and *TSN-Change Ontology*, we should obtain accurate (in time and space) and semantic representation of the territory and of its evolution over time.

11.4 Exploitation of TSN history graphs with Linked Open statistical Data

The W3C QB vocabulary is widely used to describe LOD statistics in a way that is compatible with the *Statistical Data and Metadata eXchange* (SDMX), an ISO standard for exchanging and sharing statistical data and metadata among organization Cyganiak and Reynolds [2014]. Using the QB vocabulary, one can publish statistical observations and a set of dimensions that define what the observation applies to: time, gender and geographic areas, for instance. The QB4ST extension [10] Atkinson [2017] provides canonical terms to defined in a QB observation the space (`qb4st:SpatialDimension`) and time dimensions of the data.

As explained in the Subsection 5.2.1, regarding the spatial dimension of data, the QB and QB4ST approaches do not deal with the representation of the evolution of the features over time, the underlying changes and events, as well.

Thus, we suggest, for the statistical observations described with the *RDF Data Cube* (QB) [11] vocabulary, to declare the spatial dimension of data using both the *RDF Data Cube extensions for spatiotemporal components* (QB4ST) [12] (for a canonical declaration of the spatial dimension of data), and our TSN ontology to describe the TUs spatial dimension of data, *i.e.*, the TU versions on which observations are

10. `http://www.w3.org/ns/qb4st/`
11. `http://purl.org/linked-data/cube#`
12. `http://www.w3.org/ns/qb4st/`

made (*e.g.*, `qb4st:SpatialDimension nuts:V2003_L2_ES63`). In the following Listing 11.11, we show how the two versions of the TUs ES63 in NUTS 1999 and 2003, are described using the TSN and the GeoSPARQL ontologies.

```
@prefix rdf:        <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:       <http://www.w3.org/2000/01/rdf-schema#> .
@prefix tsn:        <http://purl.org/net/tsn#> .
@prefix tsnchange:  <http://purl.org/net/tsnchange#> .
@prefix nuts:       <http://purl.org/steamer/nuts/> .
@prefix geosparql:  <http://www.opengis.net/ont/geosparql#>
nuts:V1999_L2_ES63 a tsn:UnitVersion  ;
    tsn:isMemberOf nuts:V1999_L2 ;
    tsn:hasIdentifier "ES63" ;
    tsn:hasName "Ceuta y Melilla" ;
    tsn:hasSuperFeature nuts:V1999_L1_ES6 ;
    tsnchange:hasNextVersion nuts:V2003_L2_ES63 ;
    tsn:isVersionOf nuts:L2_ES63 ;
    tnschange:inputUnitVersion
        nuts:change_extractionchange_V1999_L2_ES63 ;
    geosparql:hasGeometry nuts:Geometry_3245 ;
    <http://dbpedia.org/ontology/languageCode>
        "es"^^xsd:string .
nuts:V2003_L2_ES63 a tsn:UnitVersion ;
    tsn:isMemberOf nuts:V2003_L2 ;
    tsn:hasIdentifier "ES63"^^xsd:string ;
    tsn:hasName "Ciudad Autonoma de Ceuta" ;
    tsn:hasSuperFeature nuts:V2003_L1_ES6 ;
    tsnchange:hasNextVersion nuts:V2006_L2_ES63 ;
    tsn:isVersionOf nuts:L2_ES63 ;
    tnschange:outputUnitVersion
        nuts:change_extractionchange_V1999_L2_ES63 ;
    geosparql:hasGeometry nuts:Geometry_3246 ;
    <http://dbpedia.org/ontology/languageCode>
        "es"^^xsd:string ; .
nuts:Geometry_3245 a  geosparql:Geometry ;
     geosparql:asWKT
         "<http://www.opengis.net/def/crs/EPSG/0/4326>
         MULTIPOLYGON (((-2.84 35.2937, -2.913 35.2633,
         -2.9522 35.3492, -2.8959 35.3657,
         -2.893 35.3619, -2.8593 35.3185, -2.84 35.2937)),
         ((-5.2888 35.7974, -5.3748 35.7767,
         -5.4141 35.8383, -5.334 35.8586,
         -5.2888 35.7974)))"^^geosparql:wktLiteral ;
     geosparql:is3D "false"^^xsd:boolean ;
     geosparql:isEmpty "false"^^xsd:boolean ;
     geosparql:isSimple "true"^^xsd:boolean ;
     geosparql:spatialDimension  2 .
nuts:Geometry_3246 a  geosparql:Geometry ;
     geosparql:asWKT
         "<http://www.opengis.net/def/crs/EPSG/0/4326>
         POLYGON (((-5.2888 35.7974, -5.3748 35.7767,
         -5.4141 35.8383,  -5.334 35.8586,
         -5.2888 35.7974)))"^^geosparql:wktLiteral ;
     geosparql:is3D "false"^^xsd:boolean ;
     geosparql:isEmpty "false"^^xsd:boolean ;
     geosparql:isSimple "true"^^xsd:boolean ;
     geosparql:spatialDimension 2 .
```

Listing Code 11.11 – A description of two versions of the TUs ES63 in NUTS 1999 and 2003 using the TSN and the GeoSPARQL ontologies (RDF-Turtle syntax).

Listing 11.12 shows how to define observations on the two versions of the TU ES63 described in Listing 11.11. Here, we describe three observations of the Eurostat indicator called *Total population*. The first observation measures the total population of the TU `nuts:V1999_L2_ES63`, in 2001. The second observation measures the total population of the TU `nuts:V2003_L2_ES63`, in 2001. The third observation measures the total population of the TU `nuts:V2003_L2_ES64`, in 2001. From line 9 to 12, we declare a predicate `st-stat:refArea` as a sub-property of the predicate *qb4st:refArea*, that is the canonical predicate of QB4ST to declare the reference area of a QB observation. This new sub-property allows us to specialize the range that has to be of the type `tsn:UnitVersion` *i.e.*, the reference area of the observation have to be a `tsn:UnitVersion`.

Line 14 to 18, we declare the Eurostat indicator called *Total population* and the structure of this indicator (its space and time dimensions, its source, unit of measure...) is described from line 20 to 28. Then, the 3 observations are described according to this data structure (lines 30-34; 36-40; 42-46). The 3 observations describe the same indicator, in 2001, but the reference area is not the same. The first observation measures a total population in 2001 for Ceuta and Melilla, of 142,000 people (TU version `nuts:V1999_L2_ES63`). The second observation measures a total population in 2001 for Ceuta, of 74,000 people (TU version `nuts:V2003_L2_ES63`). The third observation measures a total population in 2001 for Melilla of 68,000 people (TU version `nuts:V2003_L2_ES64`). We show with this example that one may declare on the LOD Web the observations of an indicator on different TSN versions that is to say, different versions of the geographic divisions can coexist to say different things about the territory (because they actually are different).

```
1 @prefix st-stat: <http://purl.org/steamer/statistics/> .
2 @prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3 @prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
4 @prefix sdmx-dimension:  <http://purl.org/linked-data/sdmx/2009/dimension#> .
5 @prefix sdmx-attribute:  <http://purl.org/linked-data/sdmx/2009/attribute#> .
6 @prefix qb:              <http://purl.org/linked-data/cube#> .
7 @prefix qb4st: <http://www.w3.org/ns/qb4st/>
8
9 st-stat:refArea  a rdf:Property, qb:DimensionProperty;
10    rdfs:label "reference area"@en;
11    rdfs:subPropertyOf qb4st:refArea;
12    rdfs:range tsn:UnitVersion.
13
14 st-stat:dataset-poptot-eurostat a qb:DataSet;
15    rdfs:label "Population, total"@en;
16    rdfs:comment "Total population, both sexes."@en;
17    qb:structure st-stat:dsd-dataset-poptot-eurostat;
18    sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year>.
19
20 st-stat:dsd-dataset-poptot-eurostat a qb:DataStructureDefinition;
21    #The dimensions
22    qb:component [qb:dimension st-stat:refArea];
23    qb:component [qb:dimension sdmx-dimension:refPeriod];
24    #The measures
25    qb:component [qb:measure st-stat:totalPopulation];
26    #The attributes
27    qb:component [qb:attribute sdmx-attribute:UnitMeasure];
28    qb:component [qb:attribute sdmx-attribute:DataSource].
29
30 st-stat:observation-ES63-2001-nuts1999 a qb:Observation;
```

```
31     qb:dataset st-stat:dataset-poptot-eurostat;
32     st-stat:refArea nuts:V1999_L2_ES63;
33     sdmx-dimension:refPeriod [time:year "2001"^^xsd:gYear] ;
34     st-stat:totalPopulation 142000 .
35
36 st-stat:observation-ES63-2001-nuts2003 a qb:Observation;
37     qb:dataset st-stat:dataset-poptot-eurostat;
38     st-stat:refArea nuts:V2003_L2_ES63;
39     sdmx-dimension:refPeriod [time:year "2001"^^xsd:gYear] ;
40     st-stat:totalPopulation 74000 .
41
42 st-stat:observation-ES64-2001-nuts2003 a qb:Observation;
43     qb:dataset st-stat:dataset-poptot-eurostat;
44     st-stat:refArea nuts:V2003_L2_ES64;
45     sdmx-dimension:refPeriod [time:year "2001"^^xsd:gYear] ;
46     st-stat:totalPopulation 68000 .
```

Listing Code 11.12 – A description of the Total population Eurostat indicator - Two observations of the same indicator in different NUTS versions for the ES63 TU (RDF-Turtle syntax).

11.5 The TSN catalogs of areas for statistics

In addition to the triplestore, a second store is available online, and enables users to query data according to OGC standards. Indeed, we publish three OGC services (recognized as GIS standards), using the GeoServer [13] software (at `http://lig-tdcge.imag.fr/geoserver/web/`): using our Catalog Service for the Web (CSW), users discover the list of TUs through their metadata; from our Web Feature Service (WFS), users may download TUs representations (geometry, identifier, name, level...), in XML format for instance; the Web Map Service (WMS) enables the visualization of TUs, in PNG for instance. Thus, the Web services disseminate data and metadata in various forms (image or text file) and formats (e.g. PNG, PDF, JPEG, GeoJSON, CSV, Shapefile). Then, TUs information (name, geometry, etc.) may be downloaded from the WFS layer and loaded from any client software (*e.g.* ArcGIS, QGIS or web-mapping clients Open Layers, Leaflet) that can read data in standard formats such as GML, CSV, GeoJSON, KML or Shapefile. By connecting to the PostGIS spatial database of the framework, the Geoserver software automatically generate these OGC services. We register in the PostGIS database, for each TU Versions, its URI in the triplestore. Thus, we build the bridge between the geomatics and the semantic worlds. For instance, searching for the $ES63$ TU, in version 1999, using the WFS OGC standard request *getFeature*, in Listing 11.13, we obtain the response in JSON displayed in Listing 11.14. By clicking, in the JSON file, line 4, one connects to the triplestore of the framework and obtains a HTML representation of the semantic resource.

13. `geoserver.org/` free and open source software

```
http://lig-tdcge.imag.fr/geoserver/nuts/ows?service=WFS&version
    =1.0.0&request=GetFeature&typeName=nuts:unit_version&featureID=
    V1999_L2_ES63&outputFormat=application%2Fjson
```

Listing Code 11.13 – A WFS GetFeature Request on the OGC Web Services of the Theseus Framework.

```
{"type":"FeatureCollection","totalFeatures":1,"features":
  [{"type":"Feature","id":"unit_version.V1999_L2_ES63","geometry":{"
     type":"MultiPolygon","coordinates":[[[[-2.84,35.2937],
     [-2.913,35.2633],
  [-2.9522,35.3492],...]]]},"geometry_name":"geom", "properties":{
    "uri":"http://purl.org/steamer/nuts/V1999_L2_ES63",
    "id_unit_nomenclature":"ES63",
    "unit_name":"Ceuta y Melilla",
    "idnomenclature_version":"1999",
    "valid_from":"1999-12-31Z",
    "valid_until":"2003-06-29Z",
    "level":"2",
    "supunit":"ES6"}}],
    "crs":{"type":"name","properties":{"name":"urn:ogc:def:crs:EPSG
        ::4326"}}}
```

Listing Code 11.14 – A WFS GetFeature Response on the OGC Web Services of the Theseus Framework.

A GUI has been developed (available online at `http://lig-tdcge.imag.fr/tsn-catalog/`) using the OpenLayers 3 Web Mapping library [14] and standard requests addressed to the WMS of the framework (see Figure 11.4). In this Web mapping GUI, we implement WMS Filter requests (see on top of Figure 11.4) so that the user may filter the version and level of the TSN, s/he wants to visualize. By clicking one TU on the map (here, for instance, the UK TU has been selected) the user accesses a description of the TU on a table (see the right side of Figure 11.4). S/He also visualizes below the table the HTTP request (*GetFeatureInfo* WMS request) run on the Geoserver to get such information. This GUI provides statisticians with a catalog of TU versions that they can visualize on a map. Also, it helps Spatial Data Infrastructure developers accessing our system: they may copy the HTTP request text and provide a link to our semantic catalog of TUs from their system.

14. `http://openlayers.org/`

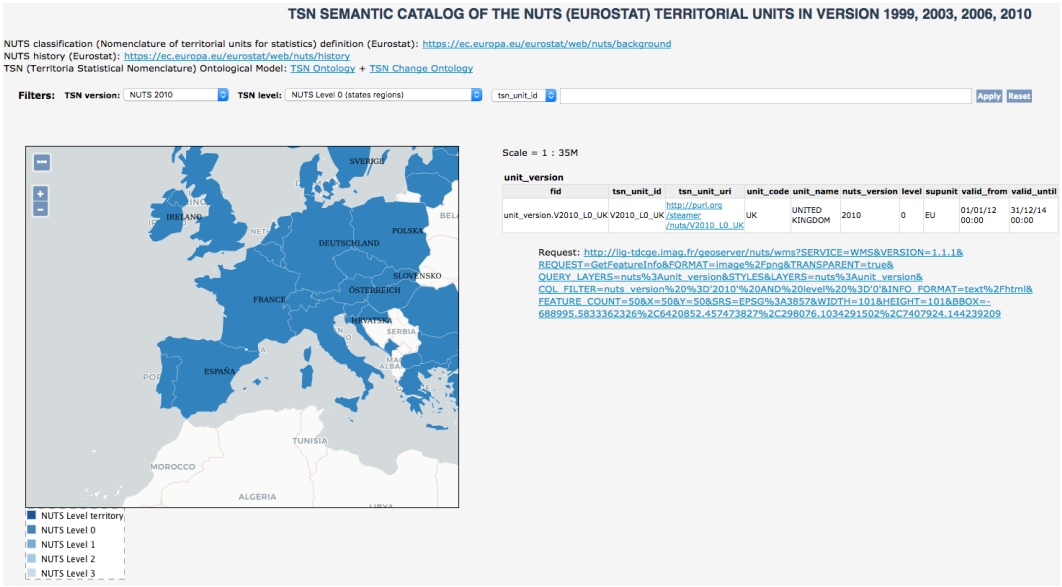


Figure 11.4 – The Theseus Framework Web Mapping GUI using request to the Geoserver WMS (available at `http://lig-tdcge.imag.fr/tsn-catalog/`).

11.6 Conclusion

With respect to the re-usability of the published resources, several online documentations (*e.g.*, project documentation `http://purl.org//steamer/tsndoc/` [15]) and papers (*e.g.*,Bernard et al. [2018b,a]) are available and provide potential users with some examples on how to use the ontologies and the data sets. The *Change graphs* are available online from a SPARQL Endpoint [16]. Thus, one can query the graph using the SPARQL language and its geospatial extension *GeoSPARQL* in order to extract information such as: the number of changes between two versions, the most prevalent change type, the life line of a given TU, the description of one *Change Graph*, etc. from the three data sets processed.

We use the `https://purl.org/` service to provide persistent URIs to our resources and concepts. Also, to make ontologies and data sets accessible by both humans and machines, content negotiation operations are set up on the server's side (they return Web pages in HTML format or RDF files describing the requested resource).

Finally, with regard to the access control of the ontologies, they are both accessible from the Linked Open Vocabularies Portal and registered in GitHub. The two ontologies are under Creative Commons Attribution 4.0 International license [17].

15. *e.g.*, ontologies documentation, TSN catalog of changes in the SAU compared to description provided by the Swiss Federal Statistical Office `http://purl.org//steamer/tsndoc/resources/sau_2017_2018_tsn_change_descriptions.pdf`
16. `http://steamerlod.imag.fr/sparql?&query=`
17. http://creativecommons.org/licenses/by/4.0/

Regarding the three data sets we have transformed in RDF, and the change descriptions we have created, with respect to:
– the Eurostat NUTS data sets (versions 1999, 2003, 2006, 2010): the permission to use the data is granted on condition that: the data will not be used for commercial purposes; the source will be acknowledged (copyright notice, ©European Union, 1995-2019 and ©EuroGeographics for the administrative boundaries). This information is mentioned for each of the NUTS version *e.g.*, `http://purl.org/steamer/nuts/V2003`.
– the SAU data sets: the boundaries are provided by the Federal Office of Topography swisstopo, under the data set named swissBoundaries3D. The terms of use of this data set are: *Open use. Must provide the source. Use for commercial purposes requires permission of the data owner.*
– the ASGS data sets: they are under ©Commonwealth of Australia and Creative Commons Attribution 4.0 International license.
With regard to our RDF description of changes, *e.g.*, based on Australian Bureau of Statistics data, they fall under the license of the geospatial files from which we compute the changes description.

In the following Chapter, we conclude this manuscript by recalling the contributions of this work and listing some of the research and development perspectives it opens up.

Chapter 12

Conclusion

The way the geographic divisions evolve over time is often ignored, and not represented. Most of the time, only the latest revisions of the geographic divisions are shown and are queryable on a map. For instance, Web mapping services such as Google Maps, or Open Street Map show only one border, the very last, and only one toponym for a city, the last one(s) (sometimes two names are shown in contested cases). However, in many cases, it may be useful to access the previous boundaries, the old names of municipalities, and also to search them and locate them on a map. Keeping each version of the geographical divisions and all the information on geographical units and their changes on the open data Web would therefore greatly improve the search engines of geographic features over time. This would allow, for instance, historians to find on a map the location of the *Duché de Savoie* in 1477. In this thesis report, we have studied in particular the evolution of *Territorial Statistical Nomenclatures* that are sets of artifact geographic areas built by Statistical Agencies to observe a given territory at several geographic division levels (*e.g.*, regions, districts, sub-districts levels).

Nowadays, the challenge for SAs is to publish their geo-coded statistics and associated TSNs on the Linked Open Data Web since the more contextualized data are (linked to historical, environmental information, etc.), and linked to each other, in time and space, the more analyzes carried out on the territories using these data will be multi-criteria, relevant and reusable by policies and researchers. This implies a significant change of emphasis from traditional Spatial Data Infrastructures (SDI) by adopting an approach based on general Web standards [Tandy et al., 2017] *i.e.*, adopting the more general and domain independent Linked Open Data approach.

No ontology on the LOD Web is generic enough to enable the description of any TSN in the world though it would foster the exchange of information among SAs. Similarly, no generic ontology on the LOD Web enables the description of the way TSNs change over time, while description of such changes may avoid misinterpretations of statistics, statistical biases and enhance the understanding of territorial evolution, providing statisticians, and general public with descriptions to comprehend the motivations and the impact of changes on geo-coded data (on electoral results for instance).

The fundamental question we have addressed in this manuscript is how to enhance the understanding of territorial dynamics, providing geographers and statisticians with tools to comprehend the impact of territorial changes by means of semantic representations of evolving territorial entities.

12.1 Summary of the contributions

We have presented in this manuscript the *Theseus Framework*, a framework designed for managing TSNs and their evolution in the semantic Web, according to a management process that consists of the following activities: Specify, Model, Generate, Publish, and Exploit. It embeds our contributions, summarized below.

First, **an ontological model have been proposed** in order to describe any TSN hierarchical structure in the world (*i.e.*, structures with more than one territorial level, where each level *covers* the whole study area, and is composed of *non-overlapping* TUs, which corresponds to the structure of most of the administrative divisions in the world), their geographic areas (unambiguously defined in time and space), and their changes over time. This model is composed of two ontologies: the *TSN Ontology* and *TSN-Change Ontology* that are published on the LOD Web. Through this model, we have proposed **an innovative approach for detailed description of the evolution of TSNs over time taking into account the nested nature of such geographic divisions, at the time of describing their changes also**. We decide to describe each change impacting each element of the territorial structure, then to link together the changes when they occur on nested areas. This new approach is based on the combination of two existing approaches: the 4D-fluent approach of [Welty et al., 2006] and the Change Bridge approach of [Kauppinen and Hyvönen, 2007]. Thus, using our approach, one may represent both the life line of the areas (4D object, space-time worm) and the way these areas change from one time-slice (of the space-time worm) to another. The semantics of changes we have proposed, based on [Claramunt and Thériault, 1995; Plumejeaud et al., 2011] works, acts as a preliminary analysis of the territory and helps, among other users, statisticians in estimating rapidly the areas that change and how they change. Indeed, it must be noted that these changes are often much more complex than a simple fusion of regions (because they impact several levels), and are very numerous the more we go down the territorial divisions (municipalities in Switzerland for example). Therefore, these changes are more difficult to see without an assisted visual detection or semantic descriptions, as the ones we propose.

Second, **we have proposed the *TSN Semantic Matching Algorithm* in order to automate the detection and semantic description of similarities and changes** between two consecutive versions of a TSN. This algorithm extends previous works from [Plumejeaud et al., 2011; Plumejeaud, 2011]. It is configurable and expandable in order to adapt to the specifics of each TSN. In that respect, we have designed it in a highly configurable way. The principle of the algorithm is to compare TUs of two successive versions of a TSN and to measure their difference, taking into account a list of their attributes (*e.g.*, name, geometry). An expert of the TSN configures the algorithm, according to h-er/is knowledge on the TSN: s/he constructs a definition of what makes the identity of the geographic objects processed. S/He defines which attributes hold the identity of the TUs in the TSN, and in which proportion they can vary before TUs lose their identity.

Third, we have **tested our whole approach and the *Theseus Framework* on different data sets in order to prove the ability of our framework to**

manage several TSNs, and automate the detection and description of their changes on the LOD Web. The created RDF graphs constitute catalogs of evolving geographic areas that enhance the understanding of territorial dynamics, providing statisticians with descriptions to comprehend the motivations and impact of redistricting. We now plan to use our method on other TSNs such as the Corine Land Cover data set and the Census Tracks of United States. We believe our approach constitutes a step towards knowledge graphs of evolving geo-coded statistics made up of RDF Data Cube data sets, geospatial TSN data sets, event data sets, law data sets, etc., from which one could build intelligent tools for the analysis of the territory dynamics over time.

Let us come back to the representation of the evolution of the number of medals won per country for the Olympic winter games, over a period from 1924 to 2010 (see Subsection 2.1.2). If one wants to show on a map the medals won by country from 1924 to 2010, it is necessary to have a representation of the borders of the countries that varies over time. Therefore, one has to create a data set containing the list of countries with their validity period and their geometry valid for the period of existence of the country (referring to the International Olympic Committee's Country Code Data set [1]). Under these conditions, it would be possible to display on a map the countries according to their period of existence and the number of medals over this period. The user should also be informed that the count of medals is only done over the period of the country's existence, which does not necessarily cover the period 1924-2010. Then, using our approach that could be used in order to draw *Change Bridges* between former States and current one, we offer a solution to navigate through the data set of medal over time and space.

Many other data sets like the medals data set, require that the history of the regions (their name, borders, ...) is preserved. This is particularly the case for cultural heritage data. However, a major obstacle is that geographic information describing these spatial units is often not available [2]. One could imagine a crowdsourcing platform in order to collect this information (past geometries, names of units ...), following the example of the collaborative project Open Historical Map [3], for instance. The contributors would be asked to provide information to characterize the historical regions as in the TSN model. Then, our program would automatically computes the life line of the geographic areas. Finally, all these descriptions could be published in the semantic Web and could therefore be linked to DBpedia or Wikidata resources describing the events related to the change of the regions over time.

Thus, the approach we propose in this manuscript may be useful in many areas (socio-economic, archaeological, cultural, linguistic, ...) in order to correctly locate the data in time and space. Indeed, our approach consists in describing very precisely

1. https://raw.githubusercontent.com/askmedia/datalogue/master/olympics/code_ioc.csv
2. However, in France, we notice the *Geo-LARHRA* initiative from a Research Group of historian that have redraw the boundaries of the french cantons from 1884 to 1966, available at http://geo-larhra.ish-lyon.cnrs.fr/?q=atlas-historique
3. https://wiki.openstreetmap.org/wiki/Open_Historical_Map

the regions, both spatially and temporally, for the accuracy of (statistical, cultural) data attached to these regions. Generalize this approach to other spatial data may improve search engines that deal with geographical data, offering the possibility to search by municipalities' obsolete names, for instance, and then visualize the life line of a municipality and its transformation over time. However, in such cases, our approach of describing all the units in each version of a TSN should be adapted in order to describe more one-time changes. We discuss the perspectives to this work in Section 12.2.

12.2 Future research and development directions

12.2.1 Automatic contextualization of changes

One of the first perspectives to this work is to automatically find the reason for the change: the "Why", *e.g.*, automatically explain the evolution of the electoral boundaries of congressional districts in the U.S.A.

In the U.S.A., the borders of districts are based on census data. Using the U.S. Census Bureau data, every State in the U.S.A. defines the boundaries of its districts. This leads sometimes to redistricting in order to establish a political advantage for a particular party (a process called "gerrymandering"). A famous district illustrating this situation is the Illinois's 4th congressional district (see Figure 12.1). This district was re-drawn successively in 1993, 2003 and in 2013 by the Illinois State Legislature with the intention of having a majority of Hispanic people in the Chicago area, based on population information from the census. "*This district combines two geographically separated areas whose populations are mainly Hispanic (74.5%). The western border of the district consists of a portion of Inter-state 294 but little of the surrounding area. This clever use of the interstate ensures that the district is pathwise connected, or contiguous, a legal requirement in most states.*" [Hodge et al., 2010]. On DBpedia, `http://dbpedia.org/page/Illinois's_4th_congressional_district`, this district is described as follows: "*It was featured by The Economist as one of the most strangely drawn and gerrymandered congressional districts in the country and has been nicknamed "earmuffs" due to its shape. It was created to pack two majority Hispanic parts of Chicago into one district*" and preserve the influence of Hispanic people voters on other neighbors districts.

Describing this kind of redistricting and the evolution of such electoral redistricting may be useful to inform people about politics and elections. Contextual information on the Web attached to this kind of change description are of paramount importance in these cases.

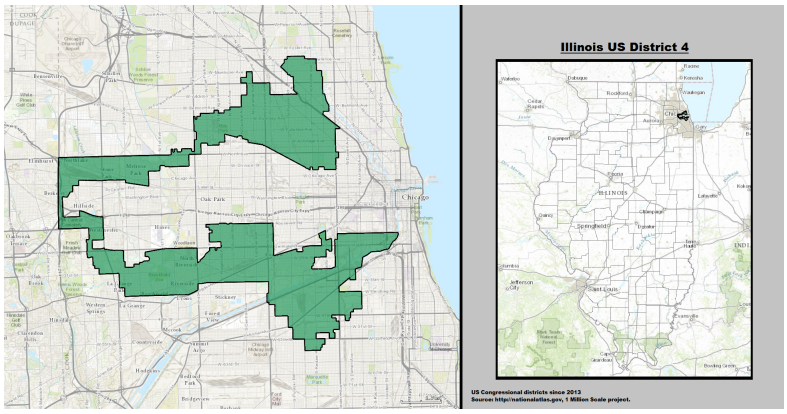


Figure 12.1 – Boundaries for Illinois's 4th United States Federal Congressional District, since 2013 (source: GIS (congressional districts, 2013) shapefile data was created by the United States Department of the Interior. Data was rendered using ArcGIS software by Esri. File developed for use on Wikipedia (Public domain)).

In Section 11.3, we have explained how one could automate the contextualization of such changes: by extracting events from WikiPedia pages and automatically create LOD descriptions of these events (using the LODE ontological model for instance).

This research perspective will probably need further exploration of the existing works and collaboration with experts in the domain of language processing and particularly relationship extraction between entities. The automation of the process will definitely have to be flexible enough to adjust to each TSN characteristics, and must of all to adjust to the way the over triple store database describe the unit or the events that explain the changes.

If one wants to link the TUs of a TSN version with other resources on the LOD Web, then one has to be careful since, for instance, the DBPedia or WikiData pages provide general description of the areas, without precise temporal information. Then, pages such as the DBPedia page describing the Illinois's 4th congressional district should be used for extraction of events (that cause the changes of the district), and less to link the units to their encyclopedic (atemporal) representation, with regard to spatio-temporal accuracy.

12.2.2 Managing other kinds of geographic divisions

In this thesis, we have studied the way TSNs evolve over time. We have focused on a specific type of geographic divisions while multiple others types exist. Because of the standard nature of TSNs, we have avoided multiple problems such as claimed territories: our descriptions only reflect the boundaries chosen by statistical agencies for the collection and restitution of their statistical data.

A perspective to this work is to modify our model in order to describe geopolitical

divisions and report on territorial changes in real time.

Another perspective is to manage the evolution of other kinds of TSN (medico-social areas, cross-border areas, employment areas, cities ...) such as:
– 1) non-covering hierarchies, where some units may have superior unit(s) at a territorial level not immediately above;
– 2) non-strict hierarchies, where a TU may have several super features;
– 3) non-onto hierarchies where the height of the hierarchy is varying;
– 4) nomenclatures where TUs at one level (or at several levels) do not cover the whole study area, such as the *UMZ* urban nomenclature, or other nomenclatures of cities;
– 5) nomenclatures composed of only one territorial level, such as the *UMZ* urban nomenclature.

1) Regarding TSNs with several levels, processing non-covering TSNs (where some units may have superior unit(s) at a territorial level not immediately above) requires no further development and modifications of the Theseus Framework, since each TU is stored in the system with information on its super TU and the level this super TU belongs. At the time of chaining the changes that propagate through the TSN levels, the software program uses this information in order to link the changes of the current processed TU, to the changes of its super TU. Whether this TU is directly above, or not, does not matter.

2) In the same way, for non-strict hierarchies, instead of connecting to the changes of a single super TU, the program will connect to several ones.

3) In TSNs where the height of the hierarchy is varying, some units may have no super TU, or no sub TU(s) while others have both. Our TSN Ontology requires a unit to be always connected to a super unit or to a territory (a specific type of unit which may correspond to the study area of the nomenclature). Then, regarding non-onto TSNs where the height is varying because some TUs don't have super TU, data preprocessing is required before the TUs enters the system. Each TU has to be linked to a super TU that could be the territory it belongs. Regarding non-onto TSNs where the height is varying because some units don't have sub TUs, there is no major obstacle since the program already manages this case, when traversing the lowest level in the hierarchy, composed of units that do not have sub units.

4) Regarding TSNs where TUs at one level (or at several levels) do not cover the whole study area, there is no major obstacles. Indeed, for instance, the program is able to detect merge of TUs (merge of two cities for instance) that are not spatially close, but that adopt a common identity after the change event, under a shared multi-polygon geometry, which covers the two initial polygons.

5) With regard to flat TSNs, composed of only one territorial level, we do not foresee any major obstacles to the processing of such TSNs. The TSN Ontology can be used for the description of such TSNs on the LOD Web. A parameter should be added to the SMA algorithm in order to disable or activate everything related to

the chaining of changes over the hierarchy of territorial levels.

Regarding spatial nomenclatures with only one level, our methodology is applicable to divisions of a territory on the physical characteristics of the landscape elements. For instance, we could extend our framework in order to manage the change of land cover in the Corine Land Cover data sets. Because our framework is configurable, we could add to the list of identity attributes the nature of the land cover. In order to perform the management of data sets such as the different versions of the CLC data set (available online from the Copernicus Web site `https://land.copernicus.eu/pan-european/corine-land-cover`), we propose to:

Figure 12.2 – Corine Land Cover Classes (source: Copernicus Project `https://land.copernicus.eu/Corinelandcoverclasses.eps.75dpi.png/`).

– add a new attribute to the list of TUs attributes: a CLC class describing the land cover (one of the classes of the CLC hierarchy of land cover classes, see Figure 12.2).
Using the SmOD Land Cover Vocabulary and its predicate `lc:corineLandCover`,

one can link a TU (GeoSPARQL feature object) to a CLC class, using the SKOS representation of the CLC taxonomy available at `http://www.w3.org/2015/03/corine` (as proposed in the online documentation of the SmOD Vocabulary `http://www.w3.org/2015/03/inspire/lc#`). Thanks to the configurable nature of the Theseus identity match test (see Equation 9.1), the program can take into account the land cover class of a region when determining if its identity changes over time.

– add a new distance test on this new attribute: according to [Harbelot et al., 2015] a semantic distance test is appropriated here. Using the distance between concepts in the CLC hierarchy of classes (see Figure 12.3), our algorithm could automatically determine if a parcel has changed and how much it has changed (*e.g.*, land cover change from "Coniferous forest" to "Airports").

– add a new weight parameter (attached to this CLC class attribute) to the TSN Semantic matching Algorithm. This weight indicates the importance of this attribute when determining the persistence of the identity of the TU over the versions.

– extend the TSN-Change typology of changes with new tag(s) describing the change of this new attribute. For instance, a new concept such as "LandCoverClassChange" could be added to the TSN Change Typology. Or, we could propose more precise terms to describe the change process. In [Harbelot et al., 2015], the authors propose to describe three kinds of derivation, by aligning with the 3 levels in the CLC nomenclature of land cover classes (see Figure 12.3). Based on these degrees of derivation, we can propose more precise tags to describe the CLC changes.

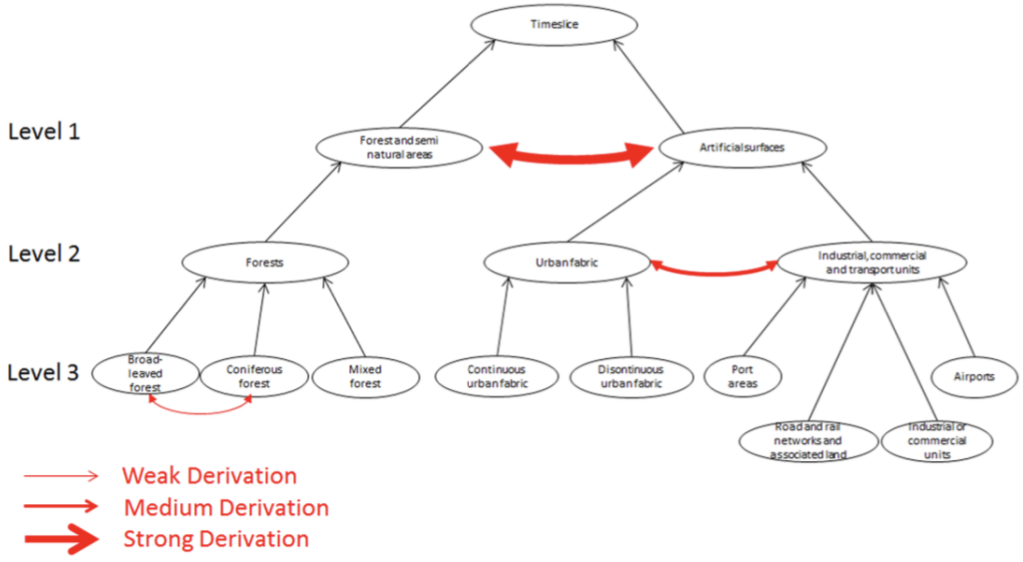


Figure 12.3 – An example of three kinds of derivation in the CLC data set: weak, medium and strong derivation [Harbelot et al., 2015].

To conclude, regarding the CLC data sets, one could set the identity match test with two parameters (as shown on the Equation 12.1): the geometry of the parcel (TU), and its CLC class. Thus, our algorithm could automatically construct the lineage of the land-parcels through time, describing both the change of the geometries of the TUs and the change of the land cover.

$$\begin{aligned} &F(\alpha_1, geom), (\alpha_2, class) \qquad where \qquad \alpha_1 = 0.5, \alpha_2 = 0.5 \\ &(u', u'') \mapsto 0.5 * (\left|u'.geom - u''.geom\right|) \\ &+0.5 * (\left|u'.class - u''.class\right|) \end{aligned} \tag{12.1}$$

12.2.3 Create bridges between TSNs

The matching of TUs of different TSN versions is of paramount importance when statisticians have various geo-coded statistical data sets that use different versions of a geographic support for the collection of data. In the same way, matching between TUs that belong to different TSNs is of paramount importance when statisticians have various geo-coded statistical data sets that are based on different TSN supports. Indeed, the correspondences between units of different nomenclatures would allow statisticians to transfer data from one TSN to another and this would make it easier to compare statistical data collected using different TSNs.

The Change Bridge approach we have used in order to chain former TSN version to new ones could also be used in order to link different TSNs together. The correspondence between TUs of different TSNs could be made only if the two TSNs cover the same territory/ies and if such correspondence between TUs seems feasible (*e.g.*, the TUs have almost the same surface area size or the correspondence is officially declared by the SAs, as for the NUTS TUs at level 3 that correspond to the SAU TUs at level 0 (see Section 10.1).

12.2.4 Describe other kinds of changes

In this manuscript, we have started from the assumption that what constitutes the identity of a TU is a list of attributes, with different weights. This list of attributes contains only attributes that describe the TUs adopting a geographic nomenclature point of view: this list contains geometric information (*e.g.*, area geometry, area surface, or location of the capital), spatial structure information, toponymic information. The Theseus Framework de-correlates the geographic information from all other thematic information associated with a TU, such as: the number of inhabitants, the elected representatives, etc. This kind of thematic information is considered so far as information that has to be described elsewhere, using another ontology than the TSN Ontology. For instance, the number of inhabitants should be described using the RDF Data Cube ontology and linked to the UnitVersion resource it describes by specifying that this TU is the spatial dimension of the observation (as described in Section 11.4). However, in order to describe how the values of these observations evolved over time, we have identified two different

approaches:

– (1) Consider that the observation takes part in the identity of the TU.

– (2) De-correlate the observation from the TU and describe its changes elsewhere than in the TSN.

Adopting the approach (1) or (2) depends on the identity criteria defined by the expert of the TSN. Thus, the two approaches are both perspectives to this work.

12.2.4.1 Approach 1 – Consider that the observation takes part in the identity of the TU

In the first approach, the expert of the TSN adds new attribute(s) (such as the "number of inhabitants") to the list of attributes that take part in the identity of the TUs. Thus, the first approach implies to extend the TSN and TSN-Change Ontologies and to make the Theseus Framework more configurable. Indeed, in order to consider new attributes in the identity definition and to describe their changes with specific tags, we propose to build more domain specific ontologies that extend the TSN ontological model. This consists in extending the TSN ontological model by: (1) adding new attribute(s) to the list of characteristics of a TU, that is to say defining a new predicate that apply to a UnitVersion (*e.g.*, `tsn:numberOfInhabitants` where `tsn:UnitVersion` is the domain of this predicate). (2) adding new change sub-concepts to the `tsnchange:Change` concept, in order to characterize the changes of the new attribute(s) (*e.g.*, `tsnchange:increase`, `tsnchange:decrease`).

For future work, we propose to make more generic the Theseus Framework by applying the following methodology's steps:
– (a) the expert of the TSN sets the list of TUs' attributes considered for matching, including variables such as the "number of inhabitants".
– (b) the expert defines the TSN and TSN-Change terms in order to describe the new attributes and their changes.
– (c) according to the change terms defined in (b) for the new attribute(s), the expert and the Theseus developers define and write the code of new distance test(s) on the new attribute(s). This new distance test forms part of the distance tests library of the Theseus Framework.
– (d) the expert sets the thresholds and weights.

In this perspective to further develop the framework, we would like to make it easier to use by experts, by creating several GUIs to:
– select the list of attributes (in the existing list of attributes already processed by the framework) or record a new attribute (and the change vocabulary associated with this new attribute). The recording of a new attribute will generate new domain sub-ontologies to the TSN Ontology (a new property is associated to the UnitVersion concept, defining that a UnitVersion has a new attribute, for instance

`tsn:hasLandCoverType a rdf:Property ; tsn:hasLandCover rdfs:domain tsn:UnitVersion.`) and TSN-Change (new change types are added such as `tsn:LandCoverTypeChange a owl:Class.`) ;
– select the distance tests to apply to each attribute, in the existing list of distance tests of the framework library, or create a new distance test that will be added to the library of tests of the framework;
– select the weight of each attribute;
– draw a small bounding box on a map that displays one version of the TSN: only the TUs in this bounding box should be used for tests. Thus, the program could be run several times, in few minutes, in order to well adjust manually the weights and thresholds to the TSN characteristics;
– visualize the changes descriptions created by the program (see Subsection 12.2.5).

12.2.4.2 Approach 2 – De-correlate the observation from the TU and describe its changes elsewhere than in the TSN

This second approach consists in describing the variation of the observation using an appropriate vocabulary of changes. This could be done by extending the vocabularies for observation descriptions on the LOD (*e.g.*, RDF Data Cube, O&M, SSN).

We believe that the work presented in this manuscript may help people in better understanding the way their territory evolves and in quickly identifying which area and how the areas change over time. In the same way, one of the perspectives for future work is to help people in quickly identify which and how the values of observations made on these areas change over time. By creating a vocabulary of Change terms for observations, made of few tags such as *increase, decrease*, one could describe with semantics the evolution of the values of an indicator over time. This semantic information may be useful for analysts, it may act as a pre-reading of the indicator values, a pre-analysis of its trends. This is particularly important when one wants to analyze many indicators, measuring different environmental, social, economical variables, etc. on a territory. For instance, this may help analysts in rapidly identifying influence of the human action on the local flora and fauna, whenever such environmental indicators are available.

12.2.5 GUI for territorial changes visualization

Another perspective work is to give to see the TSN changes on a map. Indeed, the RDF graphs are difficult to read (even if we provide HTML representation of the graphs). A Graphical User Interface may considerably help the users in their analyses of the territory's dynamics over time. Such Web mapping GUI may use the color codes proposed by GitHub in order to color the TUs whether they have change (orange), or not (green), or disappear (red) (see Figure 11.3) and, in the background, the SPARQL requests proposed in Section 11.2.3 in order to select the TUs to show on the map. The GUI may also focuses on one specific TU that has changed. For instance, Figure 12.4 is a first proposal of geo-visualization of

the change of the district 4th in Illinois: two maps are shown (on the left the congressional districts in 1983 to 1992, and on the right the congressional districts in 1993 to 2002). This territorial change has occurred during the transition between the 102nd U.S. congress (January 3, 1991 to October 9, 1992) and the 103rd U.S. congress (January 5, 1993 to December 1, 1994). The congressional district 4th of Illinois in the version 1993 is highlighted in yellow because it has been selected by the user. Under the maps, a part of the TSN RDF graph describing this redistricting could be displayed. This kind of geo-visualization raises multiple issues such as the amount of information shown on the map, and zoom level choice to ensure information readability.

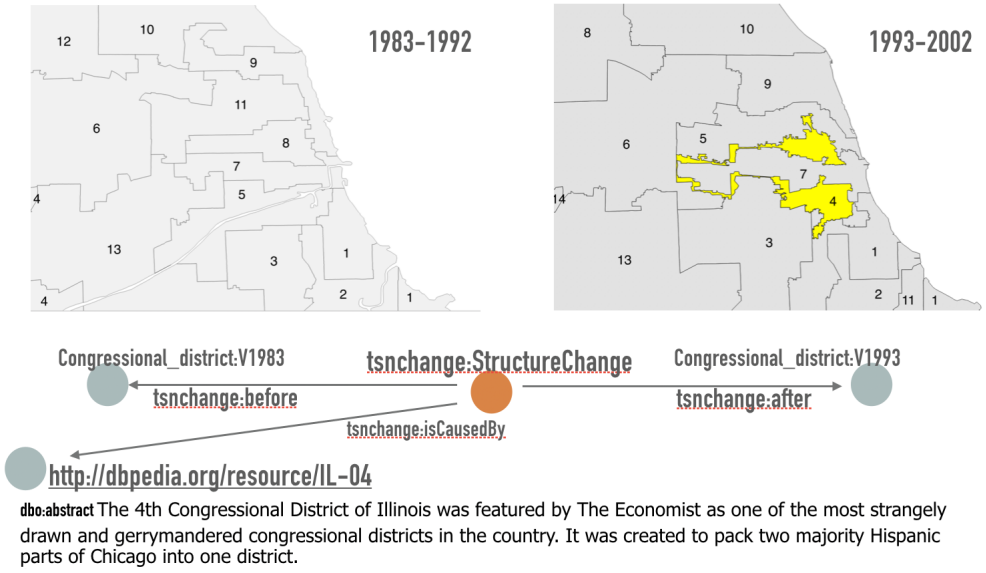


Figure 12.4 – GUI proposition – Geo-visualization of the change of the district 4th in Illinois: two maps are shown (on the left the congressional districts in 1983, and on the right the congressional districts in 1993). (Source: Jeffrey B. Lewis, Brandon DeVine, Lincoln Pitcher, and Kenneth C. Martis. (2013) Digital Boundary Definitions of United States Congressional Districts, 1789-2012. Retrieved from http://cdmaps.polisci.ucla.edu on 2019-11-06.)

This work was supported by the French region Auvergne-Rhône-Alpes [grant number REGION 2015-DRH-0367].

References

Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, and Joel Saltz. Hadoop gis: a high performance spatial data warehousing system over mapreduce. *Proceedings of the VLDB Endowment*, 6(11):1009–1020, 2013.

James F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983. URL http://dl.acm.org/citation.cfm?id=358434.

James F. Allen and George Ferguson. *Actions and Events in Interval Temporal Logic*, pages 205–245. Springer Netherlands, Dordrecht, 1997. ISBN 978-0-585-28322-7. doi: 10.1007/978-0-585-28322-7_7. URL https://doi.org/10.1007/978-0-585-28322-7_7.

Clive W Anderson, Vic Barnett, Philip C Chatwin, and Abdel H El-Shaarawi. *Quantitative Methods for Current Environmental Issues*. Springer Science & Business Media, 2012.

Schwering Angela. Approaches to semantic similarity measurement for geospatial data: A survey. *Transactions in GIS*, 12(1):5–29, 2008. doi: 10.1111/j.1467-9671.2008.01084.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2008.01084.x.

Rob Atkinson. QB4ST: RDF Data Cube extensions for spatio-temporal components, 2017. URL https://www.w3.org/TR/qb4st/.

Bruno Bachimont, Antoine Isaac, and Raphaël Troncy. Semantic commitment for designing ontologies: a proposal. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 114–121. Springer, 2002. URL http://link.springer.com/chapter/10.1007/3-540-45810-7_14.

Ben Balter. Diffable, more customizable maps, 2014. URL https://blog.github.com/2014-02-05-diffable-more-customizable-maps/.

Sandipto Banerjee and Karen C Davis. Modeling data warehouse schema evolution over extended hierarchy semantics. In *Journal on Data Semantics XIII*, pages 72–96. Springer, 2009.

Evdoxios Baratis, Euripides GM Petrakis, Sotiris Batsakis, Nikolaos Maris, and Nikolaos Papadakis. TOQL: Temporal ontology querying language. In *International Symposium on Spatial and Temporal Databases*, pages 338–354. Springer, 2009. URL http://link.springer.com/chapter/10.1007/978-3-642-02982-0_22.

Sotiris Batsakis and Euripides GM Petrakis. SOWL: a framework for handling spatio-temporal information in OWL 2.0. In *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, pages 242–249. Springer, 2011.

URL http://link.springer.com/10.1007%2F978-3-642-22546-8_19.

Sotiris Batsakis, Euripides G.M. Petrakis, Ilias Tachmazidis, and Grigoris Antoniou. Temporal representation and reasoning in OWL 2. *Semantic Web*, 8(6):981–1000, 2017. doi: 10.3233/SW-160248. URL http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-160248.

Atef Bel Adj Ali and François Vauglin. Geometric Matching of Polygons in GISs and assessment of Geometrical Quality of Polygons. In Michael Goodchild Wenzhong Shi and Peter Fisher, editors, *International Symposium on Spatial Data Quality'99*, pages 33–43, Hong Kong Polytechnic University, 1999.

Aaron Beller. Spatial/temporal events in a GIS. In *Proceedings of GIS/LIS*, volume 91, page 4, 1991.

Mark de Berg, editor. *Computational geometry: algorithms and applications*. Springer, Berlin, 3rd ed edition, 2008. ISBN 978-3-540-77973-5.

Camille Bernard, Marlène Villanova-Oliver, Jérôme Gensel, and Benoit Le Rubrus. Spatio-Temporal evolutive Data Infrastructure: a Spatial Data Infrastructure for managing data flows of Territorial Statistical Information. *International Journal of Digital Earth*, 10(3):257–283, March 2017. ISSN 1753-8947, 1753-8955. URL https://www.tandfonline.com/doi/full/10.1080/17538947.2016.1222003.

Camille Bernard, Christine Plumejeaud-Perreau, Marlène Villanova-Oliver, Jérôme Gensel, and Hy Dao. An ontology-based algorithm for managing the evolution of multi-level territorial partitions. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '18, pages 456–459, New York, NY, USA, 2018a. ACM. ISBN 978-1-4503-5889-7. doi: 10.1145/3274895.3274944. URL http://doi.acm.org/10.1145/3274895.3274944.

Camille Bernard, Marlène Villanova-Oliver, Jérôme Gensel, and Hy Dao. Modeling changes in territorial partitions over time: Ontologies tsn and tsn-change. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, pages 866–875. ACM, 2018b. ISBN 978-1-4503-5191-1. doi: 10.1145/3167132.3167227. URL http://doi.acm.org/10.1145/3167132.3167227.

Tim Berners-Lee, James Hendler, Ora Lassila, and others. The semantic web. *Scientific american*, 284(5):28–37, 2001.

Boundless. GeoGit in Action: Distributed Versioning Geospatial Data - Boundless, March 2014a. URL https://boundlessgeo.com/2014/03/geogit-distributed-versioning/.

Boundless. Introducing Versio: Version Control for Spatial Data - Boundless, October 2014b. URL https://boundlessgeo.com/2014/10/introducing-versio/.

U.S. Census Bureau Census Tracts Geographic Products Branch. Census Tracts, 2013. URL https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf.

Tao Cheng and Monsuru Adepeju. Modifiable Temporal Unit Problem (MTUP) and Its Effect on Space-Time Cluster Detection. *PLoS ONE*, 9(6):e100465, June 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0100465. URL https://dx.plos.org/10.1371/journal.pone.0100465.

Paolo Ciccarese, Stian Soiland-Reyes, Khalid Belhajjame, Alasdair JG Gray, Carole Goble, and Tim Clark. Pav ontology: provenance, authoring and versioning. *Journal of biomedical semantics*, 4(1):37, 2013.

Christophe Claramunt and Marius Thériault. Managing Time in GIS An Event-Oriented Approach. In C. J. van Rijsbergen, James Clifford, and Alexander Tuzhilin, editors, *Recent Advances in Temporal Databases*, pages 23–42. Springer London, London, 1995. ISBN 978-3-540-19945-8 978-1-4471-3033-8. URL http://link.springer.com/10.1007/978-1-4471-3033-8_2. DOI: 10.1007/978-1-4471-3033-8_2.

Arzu Çöltekin, Stefano De Sabbata, Corina Willi, Irene Vontobel, Sebastian Pfister, Matthias Kuhn, and Martin Lacayo. Modifiable temporal unit problem. In *ISPRS/ICA workshop. Persistent problems in geographic visualization (ICC2011), Paris, France*, volume 2, 2011.

Reidar Conradi and Bernhard Westfechtel. Version models for software configuration management. *ACM Computing Surveys (CSUR)*, 30(2):232–282, 1998. URL http://dl.acm.org/citation.cfm?id=280280.

Gianluca Correndo and Nigel Shadbolt. Linked nomenclature of territorial units for statistics. *Semantic Web*, 4(3):251–256, 2013. URL http://content.iospress.com/articles/semantic-web/sw079.

Gianluca Correndo, Alberto Granzotto, Manuel Salvadores, Wendy Hall, and Nigel Shadbolt. A linked data representation of the nomenclature of territorial units for statistics. 2010. URL http://eprints.soton.ac.uk/271876/.

Franck Cotton, Richard Cyganiak, RTAM Grim, Daniel W Gillman, Yves Jaques, and Wendy Thomas. Xkos: An skos extension for statistical classifications. In *Proceedings of the 59th World Statistics Congress of the International Statistical Institute The Hague, The Netherlands.* Citeseer, 2013. URL http://2013.isiproceedings.org/Files/CPS203-P32-S.pdf.

Simon Cox and Chris Little. Time Ontology in OWL - W3C Recommendation 19 October 2017, 2017. URL https://www.w3.org/TR/owl-time/.

Richard Cyganiak and Dave Reynolds. The RDF Data Cube Vocabulary, 2014. URL https://www.w3.org/TR/vocab-data-cube/.

Richard Cyganiak, David Wood, and Markus Lanthaler. RDF 1.1 Concepts and Abstract Syntax, 2014. URL https://www.w3.org/TR/rdf11-concepts/.

Christophe Debruyne, Alan Meehan, Éamonn Clinton, Lorraine McNerney, Atul Nautiyal, Peter Lavin, and Declan O'Sullivan. Ireland's authoritative geospatial linked data. In *International Semantic Web Conference*, pages 66–74. Springer, 2017.

Géraldine Del Mondo, John G Stell, Christophe Claramunt, and Rémy Thibaud. A graph model for spatio-temporal evolution. *J. UCS*, 16(11):1452–1477, 2010.

Géraldine Del Mondo, M.A. Rodríguez, C. Claramunt, L. Bravo, and R. Thibaud. Modeling consistency of spatio-temporal graphs. *Data & Knowledge Engineering*, 84:59–80, March 2013. ISSN 0169023X. doi: 10.1016/j.datak.2012.12.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0169023X12001188.

Thomas Devogele. A new merging process for data integration based on the discrete fréchet distance. In *Advances in spatial data handling*, pages 167–181. Springer, 2002.

Martin Doerr. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75, 2003.

Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48, 2016.

ESRI. United States census geography—Related Concepts | ArcGIS, 2017. URL https://learn.arcgis.com/en/related-concepts/united-states-census-geography.htm.

Lorena Etcheverry and Alejandro A Vaisman. QB4OLAP: a new vocabulary for OLAP cubes on the semantic web. *Proceedings of COLD*, 2012.

European Joint Research Centre. D2.8.III.10 INSPIRE Data Specification on Population Distribution - Technical Guidelines, 2013a. URL https://inspire.ec.europa.eu/id/document/tg/pd.

European Joint Research Centre. D2.8.III.1 INSPIRE Data Specification on Statistical Units - Technical Guidelines, 2013b. URL https://inspire.ec.europa.eu/id/document/tg/su.

Eurostat. *Manual of concepts on land cover and land use information systems*. Office for Official Publications of the European Communities, Luxembourg, 2001. ISBN 978-92-894-0432-7. OCLC: 464871883.

Rihab Fahd and Arwa Yousuf. Linked Open Data Life Cycle. *International Journal of Computer Applications*, 175(3):34–39, October 2017. ISSN 09758887. doi: 10.5120/ijca2017915489. URL http://www.ijcaonline.org/archives/volume175/number3/almutawa-2017-ijca-915489.pdf.

Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. Data on the Web Best Practices - W3C Recommendation, January 2017. URL https://www.w3.org/TR/dwbp/.

James D Fearon. What is identity (as we now use the word)? *Unpublished manuscript, Stanford University, Stanford, Calif*, 1999.

Thore Fechner, Dennis Wilhelm, and Christian Kray. Ethermap: Real-time collaborative map editing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3583–3592, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702536. URL http://doi.acm.org/10.1145/2702123.2702536.

Robin Flowerdew. Data integration: statistical methods for transferring data between zonal systems. *In Handling geographical information: methodology and potential applications*, pages 39–54, 1991.

Andrew Frank, Jonathan Raper, and Jean-Paul Cheylan. *Life and Motion of Socio-Economic Units: GISDATA*, volume 8. CRC Press, 2003.

Alessandro Furieri. Optimizing SQL access to really complex polygons, 2011. URL http://www.gaia-gis.it/spatialite-3.0.0-BETA1/WorldBorders.pdf.

Francisco Javier Gallego. A population density grid of the European Union. *Population and Environment*, 31(6):460–473, July 2010. ISSN 0199-0039, 1573-7810. doi: 10.1007/s11111-010-0108-y. URL http://link.springer.com/10.1007/s11111-010-0108-y.

Felix Gantner. *A Spatiotemporal Ontology for the Administrative Units of Switzerland*. Theses, University of Zurich - Switzerland , April 2011.

A. E. Gelfand. On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45, March 2001. ISSN 14654644, 14684357. doi: 10.1093/biostatistics/2.1.31. URL https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/2.1.31.

Michael F Goodchild, Nina Siu Ngan Lam, and University of Western Ontario. Dept. of Geography. *Areal interpolation: a variant of the traditional spatial problem*. London, Ont.: Department of Geography, University of Western Ontario, 1980.

Carol A Gotway Crawford and LJ Young. Change of support: an inter-disciplinary challenge. *Geostatistics for environmental applications*, pages 1–13, 2005.

August Gotzfried and Marco Pellegrino. The Euro-SDMX Metadata structure and quality indicators, 2008. URL https://unstats.un.org/unsd/ccsa/cdqio-2008/Ses1-Pap4.pdf.

Claude Grasland and Malika Madelin. *Modifiable Area Unit Problem*. ESPON, 2006. URL https://halshs.archives-ouvertes.fr/halshs-00174241.

Pierre Grenon and Barry Smith. SNAP and SPAN: Towards Dynamic Spatial Ontology. *Spatial Cognition & Computation*, 4(1):69–104, March 2004. ISSN 1387-5868, 1542-7633. doi: 10.1207/s15427633scc0401_5. URL http://www.tandfonline.com/doi/abs/10.1207/s15427633scc0401_5.

Q. Guo, B. Palanisamy, and H. A. Karimi. A mapreduce algorithm for polygon retrieval in geospatial analysis. In *2015 IEEE 8th International Conference on Cloud Computing*, pages 901–908, June 2015. doi: 10.1109/CLOUD.2015.123.

Willem Robert van Hage, Véronique Malaisé, Roxane H. Segers, Laura Hollink, and Guus Schreiber. Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2), 2011. ISSN 1570-8268. URL http://www.websemanticsjournal.org/index.php/ps/article/view/190.

Benjamin Harbelot, Helbert Arenas, and Christophe Cruz. Continuum: A spatiotemporal data model to represent and qualify filiation relationships. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, pages 76–85. ACM, 2013. URL http://dl.acm.org/citation.cfm?id=2534312.

Benjamin Harbelot, Helbert Arenas, and Christophe Cruz. LC3: A spatio-temporal and semantic model for knowledge discovery from geospatial datasets. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:3–24, 2015. URL http://www.sciencedirect.com/science/article/pii/S1570826815000840.

Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language, 2013. URL https://www.w3.org/TR/sparql11-query.

J. T. Hastings. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22(10):1109–1127, October 2008. ISSN 1365-8816. doi: 10.1080/13658810701851453. URL https://doi.org/10.1080/13658810701851453.

Michael Hausenblas. 5-star Open Data, January 2012. URL https://5stardata.info/en/.

Daniel Hienert and Francesco Luciano. Extraction of Historical Events from Wikipedia. In Elena Simperl, Barry Norton, Dunja Mladenic, Emanuele Della Valle, Irini Fundulaki, Alexandre Passant, and Raphaël Troncy, editors, *The Semantic Web: ESWC 2012 Satellite Events*, volume 7540, pages 16–28. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. ISBN 978-3-662-46640-7 978-3-662-46641-4. doi: 10.1007/978-3-662-46641-4_2. URL http://link.springer.com/10.1007/978-3-662-46641-4_2.

Jonathan K Hodge, Emily Marshall, and Geoff Patterson. Gerrymandering and convexity. *The College Mathematics Journal*, 41(4):312–324, 2010.

Kathleen Hornsby and Max J. Egenhofer. Identity-based change: a foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science*, 14(3):207–224, April 2000. ISSN 1365-8816, 1362-3087. doi: 10.1080/136588100240813. URL http://www.tandfonline.com/doi/abs/10.1080/136588100240813.

Erick Howenstine. Measuring demographic change: The split tract problem. *Professional Geographer*, 45(4):425, November 1993. ISSN 00330124.

Wang Huibing, Tang Xinming, Lei Bing, Yang Ping, and Chu Haifeng. Modeling spatial-temporal data in version-difference model. page 4, 2005.

Bernadette Hyland, Ghislain Auguste Atemezing, and Boris Villazón-Terrazas. Best Practices for Publishing Linked Data, January 2014. URL https://www.w3.org/TR/ld-bp/.

Evangelos Kalampokis, Bill Roberts, Areti Karamanou, Efthimios Tambouris, and Konstantinos A Tarabanis. Challenges on Developing Tools for Exploiting Linked Open Data Cubes. In *SemStats@ ISWC*, 2015.

Tomi Kauppinen and Eero Hyvönen. Modeling and reasoning about changes in ontology time series. In *Ontologies*, pages 319–338. Springer, 2007. URL http://link.springer.com/chapter/10.1007/978-0-387-37022-4_11.

Tomi Kauppinen, Jari Väätäinen, and Eero Hyvönen. *Creating and using geospatial ontology time series in a semantic cultural heritage portal*. Springer, 2008. URL http://link.springer.com/chapter/10.1007/978-3-540-68234-9_11.

Setrag N Khoshafian and George P Copeland. *Object identity*, volume 21. ACM, 1986.

Eva Klien and Florian Probst. Requirements for geospatial ontology engineering. In *8th conference on geographic information science (AGILE 2005)*, pages 251–260, 2005. URL http://www.agile-online.org/Conference_Paper/CDs/agile_2005/papers/79_Eva%20Klien.pdf.

Javier Lacasta, Francisco Javier Lopez-Pellicer, Aneta Florczyk, Francisco Javier Zarazaga-Soria, and Javier Nogueras-Iso. Population of a spatio-temporal knowledge base for jurisdictional domains. *International Journal of Geographical Information Science*, 28(9):1964–1987, September 2014. ISSN 1365-8816, 1362-3087. URL http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.911412.

Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Linna Li and Michael F. Goodchild. An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, 2(4):309–328, December 2011. ISSN 1947-9832, 1947-9824. doi: 10.1080/19479832.2011.577458. URL http://www.tandfonline.com/doi/abs/10.1080/19479832.2011.577458.

LocationTech. GeoGig, April 2018. URL http://geogig.org/.

Martin Loidl, Gudrun Wallentin, Robin Wendel, and Bernhard Zagel. Mapping bicycle crash risk patterns on the local scale. *Safety*, 2(3), 2016. ISSN 2313-576X. doi: 10.3390/safety2030017. URL http://www.mdpi.com/2313-576X/2/3/17.

Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind. *Geographic information systems and science*. John Wiley & Sons, 2005.

Francisco J. López-Pellicer, Aneta J. Florczyk, Javier Lacasta, Francisco Javier Zarazaga-Soria, and Pedro R. Muro-Medrano. Administrative units, an ontological perspective. In *Advances in Conceptual Modeling–Challenges and Opportunities*, pages 354–363. Springer, 2008. URL http://link.springer.com/chapter/10.1007/978-3-540-87991-6_42.

Francisco J. Lopez-Pellicer, Javier Lacasta, Aneta Florczyk, Javier Nogueras-Iso, and F. Javier Zarazaga-Soria. An ontology for the representation of spatiotemporal jurisdictional domains in information retrieval systems. *International Journal of Geographical Information Science*, 26(4):579–597, April 2012. ISSN 1365-8816, 1362-3087. URL http://www.tandfonline.com/doi/abs/10.1080/13658816.2011.599811.

Romain Louvet, Didier Josselin, Jagannath Aryal, and Cyrille Genre-Grandpierre. R as a GIS: Illustrating Scale and Aggregation Problems with Forest Fire Data. In *Proceedings of Spatial Statistics 2015 - Procedia Environmental Sciences*, volume 27, pages 66–69. Elsevier - Science Direct , June 2015. URL https://hal.archives-ouvertes.fr/hal-01251316.

Alexander Maedche and Steffen Staab. Learning ontologies for the semantic web. In *Proceedings of the Second International Conference on Semantic Web-Volume 40*, pages 51–60. CEUR-WS. org, 2001. URL http://dl.acm.org/citation.cfm?id=2889973.

Hélène Mathian and Lena Sanders. *Spatio-temporal approaches: Geographic objects and change process*. John Wiley & Sons, 2014.

Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, 41(2), March 2014. ISSN 1523-0406, 1545-0465. doi: 10.1080/15230406.2014.880327. URL http://www.tandfonline.com/doi/abs/10.1080/15230406.2014.880327.

Vereinte Nationen, editor. *International merchandise trade statistics: compilers manual*. Number 87 in [Studies in methods] / United Nations, Department of Economic and Social Affairs, Statistics Division Ser. F. United Nations, New York, 2004. ISBN 978-92-1-161454-1. OCLC: 250039369.

Marco Negretti. Operation-based revision control for geospatial data sets. page 15, 2015.

Natalya F. Noy and Michel Klein. Ontology Evolution: Not the Same as Schema Evolution. *Knowledge and Information Systems*, 6(4):428–440, July 2004. ISSN

0219-1377, 0219-3116. doi: 10.1007/s10115-003-0137-2. URL http://link.springer.com/10.1007/s10115-003-0137-2.

OECD. *OECD glossary of statistical terms*. 2008. URL https://ec.europa.eu/eurostat/ramon/coded_files/OECD_glossary_stat_terms.pdf.

Ontotext. What is a Knowledge Graph?, 2018. URL https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/.

Stan Openshaw. The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*, 38, 1984. ISSN 0 86094 134 5. URL https://ci.nii.ac.jp/naid/10024464407/en/.

Stan Openshaw and Peter J Taylor. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, 21:127–144, 1979.

Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

Torben Bach Pedersen, Christian S Jensen, and Curtis E Dyreson. Extending practical pre-aggregation in on-line analytical processing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, pages 663–674. Morgan Kaufmann Publishers Inc., 1999.

Matthew Perry and John Herring. OGC GeoSPARQL - A Geographic Query Language for RDF Data. page 75, 2012.

Christine Plumejeaud. *Modèles et méthodes pour l'information spatio-temporelle évolutive*. PhD thesis, Université de Grenoble, 2011.

Christine Plumejeaud, Julie Prud'homme, Paule-Annick Davoine, and Jérôme Gensel. Transferring Indicators into Different Partitions of Geographic Space. In David Taniar, Osvaldo Gervasi, Beniamino Murgante, Eric Pardede, and Bernady O. Apduhan, editors, *Computational Science and Its Applications ICCSA 2010*, volume 6016 of *Lecture Notes in Computer Science*, pages 445–460, Heidelberg, 2010. Springer. ISBN 978-3-642-12155-5.

Christine Plumejeaud, Hélène Mathian, Jérôme Gensel, and Claude Grasland. Spatio-temporal analysis of territorial changes from a multi-scale perspective. *International Journal of Geographical Information Science*, 25(10):1597–1612, 2011. URL http://www.tandfonline.com/doi/abs/10.1080/13658816.2010.534658.

Mgr Petr Přidal and Petr Žabička. Tiles as an approach to on-line publishing of scanned old maps, vedute and other historical documents. *e-Perimetron*, 3(1): 10–21, 2008.

Yves Raimond, Samer A Abdallah, Mark B Sandler, and Frederick Giasson. The music ontology, 2007.

Gabriel Recchia and Max Louwerse. A Comparison of String Similarity Measures for Toponym Matching. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP '13*, pages 54–61, Orlando FL, USA, 2013. ACM Press. ISBN 978-1-4503-2535-6. doi: 10.1145/2534848.2534850. URL http://dl.acm.org/citation.cfm?doid=2534848.2534850.

Timothy Redmond, Michael Smith, Nick Drummond, and Tania Tudorache. Managing Change: An Ontology Version Control System. page 10, 2008.

Agnar Renolen. History graphs: Conceptual modeling of spatio-temporal data. *Gis frontiers in business and science*, 2:46, 1996.

Agnar Renolen. Conceptual modelling and spatiotemporal information systems: how to model the real world. In *ScanGIS*, volume 97, pages 1–22, 1997.

Margaret Rouse. What is versioning?, 2007. URL https://searchsoftwarequality.techtarget.com/definition/versioning.

Juan J. Ruiz, F. Javier Ariza, Manuel A. Ureña, and Elidia B. Blázquez. Digital map conflation: a review of the process and a proposal for classification. *International Journal of Geographical Information Science*, 25(9):1439–1466, September 2011. ISSN 1365-8816. doi: 10.1080/13658816.2010.519707. URL https://doi.org/10.1080/13658816.2010.519707.

Tuukka Ruotsalo and Eero Hyvönen. An event-based approach for semantic metadata interoperability. In *The Semantic Web*, pages 409–422. Springer, 2007.

Nayan B Ruparelia. The history of version control. *ACM SIGSOFT Software Engineering Notes*, 35(1):5–9, 2010.

Robert D. Sack. Human Territoriality: A Theory. *Annals of the Association of American Geographers*, 73(1):55–74, March 1983. ISSN 0004-5608, 1467-8306. doi: 10.1111/j.1467-8306.1983.tb01396.x. URL http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.1983.tb01396.x.

Percy E Rivera Salas, Michael Martin, Fernando Maia Da Mota, Sören Auer, Karin Breitman, and Marco A Casanova. Publishing statistical data on the web. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 285–292. IEEE, 2012.

Rui Santos, Patricia Murrieta-Flores, and Bruno Martins. Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth*, 11(9):913–938, September 2018. ISSN 1753-8947, 1753-8955. doi: 10.1080/17538947.2017.1371253. URL https://www.tandfonline.com/doi/full/10.1080/17538947.2017.1371253.

Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. F–a model of events based on the foundational ontology dolce+ dns ultralight. In *Proceedings of the fifth international conference on Knowledge capture*, pages 137–144. ACM, 2009.

SDMX. SDMX 2.1 User Guides | SDMX – Statistical Data and Metadata eXchange, 2012. URL https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf.

Ryan Shaw, Raphaël Troncy, and Lynda Hardman. LODE: Linking Open Descriptions of Events. *ASWC*, 9:153–167, 2009.

Shashi Shekhar and Hui Xiong. *Encyclopedia of GIS*. Springer Science & Business Media, 2007.

Theodore Sider. *Four-dimensionalism: An ontology of persistence and time*. Oxford University Press on Demand, 2001.

Evren Sirin. A Path of Our Own - Stardog, 2017. URL https://www.stardog.com/blog/a-path-of-our-own/.

Barry Smith and Achille C Varzi. Fiat and bona fide boundaries. *Philosophy and phenomenological research*, 60(2):401–420, 2000.

Office fédéral de la statistique. Liste historisée des communes de la Suisse - Explication et utilisation | Publication. Publication do-f-00.04-hgv-01, Office fédéral de la statistique, December 2017. URL https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/publications.assetdetail.4062823.html.

Kostas Stefanidis, Ioannis Chrysakis, and Giorgos Flouris. On Designing Archiving Policies for Evolving RDF Datasets on the Web. In Eric Yu, Gillian Dobbie, Matthias Jarke, and Sandeep Purao, editors, *Conceptual Modeling: 33rd International Conference, ER 2014, Atlanta, GA, USA, October 27-29, 2014. Proceedings*, pages 43–56. Springer International Publishing, Cham, 2014. ISBN 978-3-319-12206-9. URL http://dx.doi.org/10.1007/978-3-319-12206-9_4. DOI: 10.1007/978-3-319-12206-9_4.

Margaret-Anne D Storey, Davor Čubranić, and Daniel M German. On the use of visualization to support awareness of human activities in software development: a survey and a framework. In *Proceedings of the 2005 ACM symposium on Software visualization*, pages 193–202. ACM, 2005.

Chengyu Sun, Divyakant Agrawal, and Amr El Abbadi. Hardware acceleration for spatial selections and joins. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 455–466. ACM, 2003.

Jeremy Tandy, Linda Van Den Brink, and Payam Barnaghi. Spatial data on the Web Best Practices - OGC and W3C Recommendation, January 2017. URL https://www.w3.org/TR/sdw-bp.

Anton Telechev and Benoit Le Rubrus. Nomenclatures Support, 2013. URL http://database.espon.eu/db2/jsf/NomenclatureSupport/NomenclatureSupport_onehtml/index.html.

Walter Tichy. *Software Configuration Management Overview*. Citeseer, 1988. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.8965&rep=rep1&type=pdf.

Ba-Huy Tran, Christine Plumejeaud-Perreau, Alain Bouju, and Vincent Bretagnolle. A semantic mediator for handling heterogeneity of spatio-temporal environment data. In *Research Conference on Metadata and Semantics Research*, pages 381–392. Springer, 2015. URL http://link.springer.com/chapter/10.1007/978-3-319-24129-6_33.

A. Vandecasteele and R. Devillers. Improving volunteered geographic data quality using semantic similarity measurements. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-2/W1: 143–148, May 2013. ISSN 1682-1777. doi: 10.5194/isprsarchives-XL-2-W1-143-2013. URL http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-2-W1/143/2013/.

Gaurav Vazirani. Metaphysics: Ship of Theseus, 2014. URL https://www.khanacademy.org/partner-content/wi-phi/wiphi-metaphysics-epistemology/wiphi-metaphysics/v/ship-of-theseus.

Boris Villazón-Terrazas, Luis. M. Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data. In David Wood, editor, *Linking Government Data*, pages 27–49. Springer New York, New York, NY, 2011. ISBN 978-1-4614-1767-5. doi: 10.1007/978-1-4614-1767-5_2. URL https://doi.org/10.1007/978-1-4614-1767-5_2.

Max Völkel and Tudor Groza. SemVersion: an RDF-based ontology versioning system. page 9, March 2012.

M. Wachowicz. *Object-Oriented Design for Temporal GIS*. Research Monographs in GIS. Taylor & Francis, 2003. ISBN 9780203212394.

Fahui Wang. *Quantitative methods and socio-economic applications in GIS*. CRC Press, 2014.

Chris Welty, Richard Fikes, and Selene Makarios. A reusable ontology for fluents in OWL. In *FOIS*, volume 150, pages 226–236, 2006.

Stefan Wiemann and Lars Bernard. Spatial data fusion in Spatial Data Infrastructures using Linked Data. *International Journal of Geographical Information Science*, 30(4):613–636, April 2016. ISSN 1365-8816, 1362-3087. doi: 10.1080/13658816.2015.1084420. URL http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1084420.

Dan Yamamoto, Akira Ioku, Yoko Seki, Akie Mizutani, Junichi Matsuda, Hideaki Takeda2 Ikki Ohmukai, Fumihiro Kato, Seiji Koide, and Shoki Nishimura. Geographic Area Representations in Statistical Linked Open Data of Japan. In *HybridSemStats@ISWC*, 2017.

Ronan Ysebaert, Isabelle Salmon, Benoit Le Rubrus, and Camille Bernard. Recueil, traçabilité et restitution des données territoriales du programme ESPON. page 21, April 2017.

C. Zhang, W. Li, and T. Zhao. Geospatial data sharing based on geospatial semantic web technologies. *Journal of Spatial Science*, 52(2):35–49, December 2007. ISSN 1449-8596, 1836-5655. doi: 10.1080/14498596.2007.9635121. URL `http://www.tandfonline.com/doi/abs/10.1080/14498596.2007.9635121`.

Appendix

Appendix A

The Theseus data life cycle

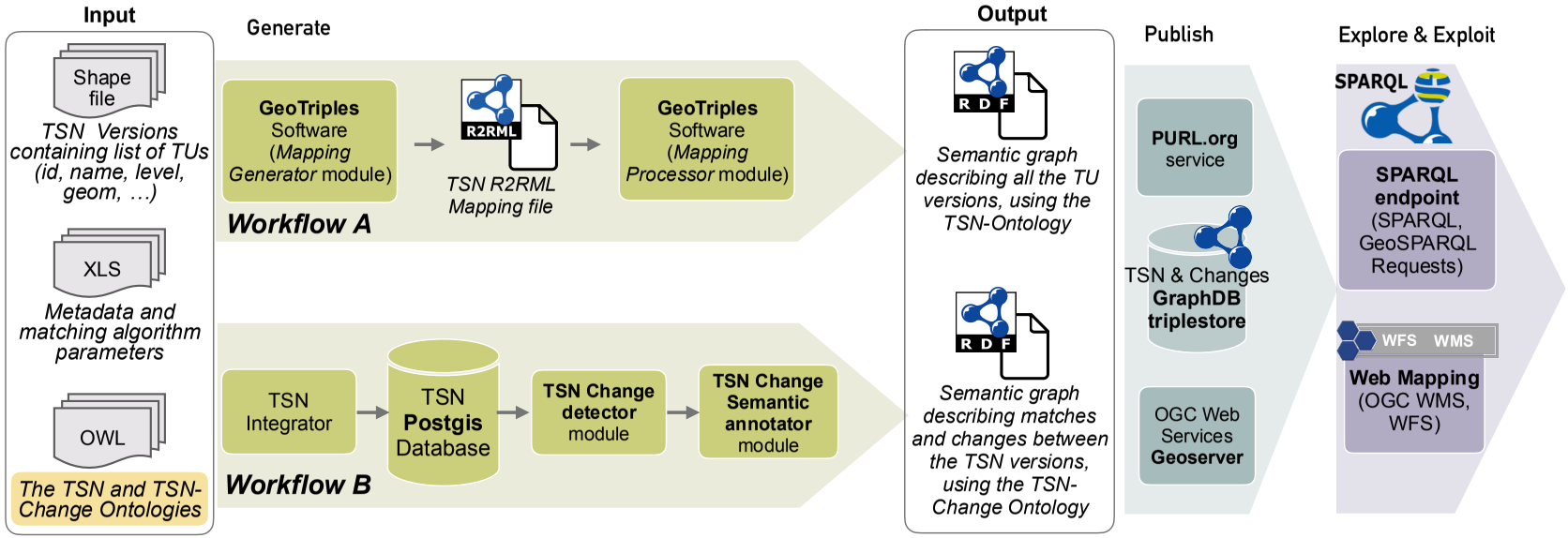


Figure 1.1 – The Theseus Framework data life cycle

Appendix B

The TSN R2RML Mapping File example

In order to populate the TSN Ontology, we implement in the Theseus Framework a workflow that transforms each element composing the TSN versions (each version being described in a shapefile) into RDF triples. It consists first of the alignment of the columns of the input shapefiles with the *TSN Ontology* concepts using a "mapping" file written in R2RML [1] which matches each column of the input shapefile with one concept of the TSN Ontology. This mapping file has been generated using the *Mapping Generator* module of the GeoTriples Software. "*R2RML is a language for expressing customized mappings from relational databases to RDF datasets*" (source: `https://www.w3.org/TR/r2rml/`). The GeoTriples software "*supports the mapping languages R2RML and RML and extends them for modeling the transformation of geospatial data into RDF graphs.*" (source: `http://geotriples.di.uoa.gr/`).

```
@prefix geof: <http://www.opengis.net/def/function/geosparql/> .
@prefix map: <http://purl.org/steamer/nuts/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rrx: <http://www.w3.org/ns/r2rml-ext#> .
@prefix rrxf: <http://www.w3.org/ns/r2rml-ext/functions/def/> .
@prefix strdf: <http://strdf.di.uoa.gr/ontology#> .
@prefix tsn: <http://purl.org/net/tsn#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

map:nuts
    rr:logicalTable [ rr:tableName "`nuts`"; ];
    rr:subjectMap [ rr:class tsn:UnitVersion; rr:template "http://purl.
        org/steamer/nuts/V{`idnomenclature`}_L{`level`}_{`idunit`}/gid/{`
        gid`}"; ];
    rr:predicateObjectMap [
        rr:predicate tsn:isMemberOf;
        rr:objectMap [
          rr:termType rr:IRI;
          rr:template "http://purl.org/steamer/nuts/V{`idnomenclature`}_L
              {`level`}";
        ];
    ];
```

1. `https://www.w3.org/TR/r2rml/`

```
    rr:predicateObjectMap [
        rr:predicate tsn:hasSuperFeature;
        rr:objectMap [
        rr:termType rr:IRI;
        rr:template "http://purl.org/steamer/nuts/V{`idnomenclature`}_L{`
            suplevel`}_{`supunit`}";
        ];
    ];

    rr:predicateObjectMap [
        rr:predicate tsn:hasIdentifier;
        rr:objectMap [
            rr:datatype xsd:string;
            rr:column "`idunit`";
        ];
    ];
    rr:predicateObjectMap [
        rr:predicate tsn:hasName;
        rr:objectMap [
            rr:column "`unit_name`";
        ];
    ];
    rr:predicateObjectMap [
        rr:predicate <http://dbpedia.org/ontology/languageCode>;
        rr:objectMap [
            rr:datatype xsd:string;
            rr:column "`lang_code`";
        ];
    ];

    rr:predicateObjectMap [
        rr:predicate geo:hasGeometry;
        rr:objectMap [
            rr:termType rr:IRI;
            rr:template "http://purl.org/steamer/nuts/Geometry_{`gid`}";
        ];
    ];
    rr:predicateObjectMap [
        rr:predicate tsn:isVersionOf;
        rr:objectMap [
          rr:termType rr:IRI;
          rr:template "http://purl.org/steamer/nuts/L{`level`}_{`idunit
              `}";
        ];
    ];
    .

map:nuts_geometry
    rr:logicalTable [ rr:tableName "`nuts`"; ];
    rr:subjectMap [ rr:class geo:Geometry; rr:template "http://purl.org/
        steamer/nuts/Geometry_{`gid`}"; ];

    rr:predicateObjectMap [
        rr:predicate geo:asWKT;
        rr:objectMap [
            rr:datatype geo:wktLiteral;
            rrx:function rrxf:asWKT;
            rrx:argumentMap
```

```
78              (
79                   [ rr:column "`the_geom`"; ]
80              )
81         ];
82    ];
83    rr:predicateObjectMap [
84         rr:predicate geo:spatialDimension;
85         rr:objectMap [
86              rr:datatype xsd:integer;
87              rrx:function rrxf:spatialDimension;
88              rrx:argumentMap
89              (
90                   [ rr:column "`the_geom`"; ]
91              )
92         ];
93    ];
94    rr:predicateObjectMap [
95         rr:predicate geo:isEmpty;
96         rr:objectMap [
97              rr:datatype xsd:boolean;
98              rrx:function rrxf:isEmpty;
99              rrx:argumentMap
100             (
101                  [ rr:column "`the_geom`"; ]
102             )
103        ];
104   ];
105   rr:predicateObjectMap [
106        rr:predicate geo:is3D;
107        rr:objectMap [
108             rr:datatype xsd:boolean;
109             rrx:function rrxf:is3D;
110             rrx:argumentMap
111             (
112                  [ rr:column "`the_geom`"; ]
113             )
114        ];
115   ];
116   rr:predicateObjectMap [
117        rr:predicate geo:isSimple;
118        rr:objectMap [
119             rr:datatype xsd:boolean;
120             rrx:function rrxf:isSimple;
121             rrx:argumentMap
122             (
123                  [ rr:column "`the_geom`"; ]
124             )
125        ];
126   ];
127   .
```

Listing Code B.1 – The Theseus R2RML Mapping File between an input TSN shapefile and the TSN Ontology Concepts