Open Technologies Supporting Linked Open Data Publishing: Croatian Population Census Case Study

Karlo Kević, Ana Kuveždić Divjak

University of Zagreb Faculty of Geodesy, 10000 Zagreb, Croatia –
(karlo.kevic, ana.kuvezdic.divjak)@geof.unizg.hr

Keywords: Census, Ontology, Linked Data, Open Data.

Abstract

Population census data in Croatia is provided in spreadsheet and without explicit geometric representation of its spatial units which makes it not directly usable in spatial analyses and data-based decision-making. This poses challenges for data interoperability and limits practical usefulness of census data. To overcome these limitations, this paper aims to propose ontology mapping schema to structure population contingents by age and sex in RDF and publish it as Linked Open Data using open-source data wrangling tool, OpenRefine. In line with Linked Data best practices, population data and spatial units' geometries were modelled separately and linked through spatial unit's URIs. Also, to ensure interoperability and enable broader integration, the schema reuses semantics from RDF Data Cube Vocabulary and GeoSPAQRL while the use of open technologies ensures that the resulting RDF triples are reproducible. This proposed ontology mapping schema represents a foundational step towards the publication of linked open census data in Croatia, paving the way for improved integration and reuse in the future.

1. Introduction

The growing demand for semantically interoperable population data has revealed the limitations of traditional data dissemination formats. In Croatia, census data and its geometric representation are provided in different file formats and by different institutions which poses significant constraints on data integration and automation, and consequently data reuse. Recent paradigm shifts towards data accessible to anyone (open data) further emphasized this problem and the literature proposes semantic web to be a potential solution. Web of data where structured, machine-accessible data enables automated processing and integration by the computer (Hogan, 2020) has the means to overcome these limitations and ensure better future data exploitation.

The population census is one of the most complex statistical tasks of a country, providing harmonized, complete and detailed data on the population, which is essential for the formulation, implementation and monitoring of policies to address the country's economic and societal challenges (United Nations Statistical Commission, 2015). In the European context, the collection of census data has a long tradition and is governed by legal frameworks such as the 2008 EU Regulation on Population and Housing Censuses, which standardize methodology and ensure comparability between countries. However, the way in which census data is disseminated has changed significantly with the adoption of the EU Open Data Directive in 2019, which obliges governments to make such data publicly available for reuse as open data. Not only did it support unrestricted access to the data, but the Directive also classified census data as high-value data. In this way, it emphasized its immense reuse potential to promote social and economic development and urged its provision in machine-readable formats or via APIs (European Commission, 2019).

The publication of census data as open data on the web enhances its accessibility; however, simply making data available does not fully resolve the challenges related to data integration and interoperability. Integrating structures like spreadsheets — commonly used in census data — often requires extensive manual effort (Meroño-Peñuela et al., 2016), which, as noted by Wong et al. (2024), limits the practical usefulness of census data. Linked Open Data (LOD), a concept closely associated with the Semantic Web, offers a solution by providing semantically enriched, standardized, and interlinked datasets. By adhering to core LOD principles — including the use of IRIs, HTTP, RDF, OWL, and links to external datasets — LOD facilitates meaningful connections between datasets, enhancing interoperability, reusability, and enabling deeper analyses. As Fernandez et al. (2011) suggest, LOD is a key enabler for advancing open data initiatives and the development of the Semantic Web.

The creation and provision of LOD by data providers often encounters several challenges, including the need to adapt and map data to appropriate ontologies, as well as a general lack of motivation to ensure ongoing data maintenance (Conde et al., 2022). Fiorelli and Stellato (2021) suggest that existing data wrangling tools can alleviate these issues by facilitating the transformation of tabular data into RDF triples. Furthermore, Conde et al. (2022) emphasise that open-source tools offer essential capabilities to overcome such barriers, pointing to their continuous improvement in functionalities and adaptability to specific user requirements.

At the European Union level, there is an awareness of linked and open census data (e.g. Eurostat). However, only a few countries have published such data at national level (e.g. Fernandez et al., 2011; Petrou et al., 2013; Pabón et al., 2013; Meroño-Peñuela et al., 2016). In Croatia, the national statistical institution provides geocoded census data as open data, mainly in .xlsx format. For many practical applications, e.g., identifying populations exposed to air pollution, this format requires manual preprocessing to link the data with the corresponding geometries. These limitations pose significant barriers to

automated processing, integration with other datasets, and cross-domain analysis. To address these limitations, this research proposes an ontology mapping schema to represent and publish the Croatian census as LOD, utilising the capabilities of OpenRefine, a fully open-source data processing tool. By adhering to standards such as the RDF Data Cube Vocabulary for statistical data and GeoSPARQL for geospatial semantics, the proposed approach ensures interoperability. In addition, the use of open-source tools promotes transparency, reproducibility and adaptability across different datasets and research contexts.

The transformation of the Croatian population census into LOD represents a significant step towards improving the accessibility and usability of data. It will not only facilitate the integration of data from different sources but also enable improved contextualisation and more precise information retrieval through query-based access.

2. Background

2.1 Croatian Population Census

Population data in Croatia is collected every ten years by the National Bureau of Statistics through the Census of Population, Households, and Dwellings. The most recent census was conducted in 2021 and was based on the traditional enumeration method with the novelty of self-enumeration (The Official Gazette of the Republic of Croatia, 2021). The scope of the statistical variables collected, and the methodology are regulated at national level and aligned with European regulations (EC 763/2008 and EC 2017/712). A total of 36 population-related variables — such as name, gender, and date of birth — are collected and disseminated in aggregated form per spatial unit and predefined population profiles (e.g., age contingents, nationality) (The Official Gazette of the Republic of Croatia, 2021). The data is primarily disseminated in tabular form (.xlsx) to the settlement level, but also as vector ESRI shapefiles in a 1km x 1km grid resolution. Although the finest spatial resolution for data aggregation is enumeration district, which is defined as a spatial unit with up to 130 households (The Official Gazette of the Republic of Croatia, 2020), this data is only available upon request due to data protection regulations.

The spatial component of the census, the corresponding spatial units, is the fundamental part of the Register of Spatial Units. Maintained by the State Geodetic Administration, Croatia's national cadastre and mapping authority, the Register comprises various types of spatial units, including administrative, statistical and enumeration units, as well as addresses and cadastral parcels (The Official Gazette of the Republic of Croatia, 2020). These spatial units are crucial for planning and conducting the census and for determining the scale on which the census data is aggregated and disseminated. The census related units follow administrative hierarchy in which subordinate units are nested within the boundaries of a higher-level ones. In addition to the geometric representations, the Register provides attribute data such as unique identifiers that support unit identification and hierarchical linking. The spatial units are provided up to the level of enumeration district, in GML format or via WFS services, using the projected coordinate system and compliant with the INSPIRE data specifications.

The collection and processing of census data in Croatia complies with European statistical regulations, ensuring comparability and compatibility with indicators from other countries. As a result, existing LOD publication approaches adopted by other European countries are well suited for application to the Croatian census.

2.2 Semantic Web Technologies for Linked Open Data

Given the five-star deployment schema by Tim Berners Lee, for data to be LOD, it must be published on the web, in a way that it links to other people's data to provide context. This means different semantic web technologies must be employed to enable description, access to and query of LOD on the web: Uniform Resource Identifier (URI) to uniquely identify pieces of data, Resource Description Framework (RDF) to structure the data, Web Ontology Language (OWL) to capture semantics, Hypertext Transfer Protocol (HTTP) to access the data through descriptions in URIs, and SPARQL Protocol and RDF Query Language (SPARQL) to query data on the web. While HTTP and SPARQL support access, allow retrieval and manipulation of LOD, URIs, RDF and OWL describe the information contained in data. URI uniquely identifies things which enables their linking, RDF integrates HTTP URIs into Subject-Object-Predicate triples and standardises structure and OWL exploits semantics to allow data integration.

A key advantage of LOD is its ability to integrate with other datasets, which is made possible by the semantic richness embedded in RDF triples. To support this integration, the use of well-defined ontologies and vocabularies is essential. Various vocabularies can be used to describe statistical data such as census information. The RDF Data Cube Vocabulary (QB), for instance, is specifically designed for publishing statistical data in RDF format (Cyganiak et al., 2014), while the Core Ontology for Official Statistics (COOS) is intended to represent the full lifecycle of official statistics, including data collection, processing, and dissemination (Rizzolo et al., 2023). The QB vocabulary, which has been endorsed as a W3C standard, is widely adopted for representing census data in LOD. It is inspired by earlier models such as the (now deprecated) Statistics Core Vocabulary (SCOVO) and draws upon the Statistical Data and Metadata eXchange (SDMX) information model. Based on a multidimensional data structure, QB organises data according to three main characteristics: *dimension*, indicating what the observation applies to; *measure*, representing the values observed; and *attributes*, providing additional metadata such as units of measurement. This flexible structure makes QB suitable not only for statistical datasets, but also for other forms of structured data, including survey results.

Several ontologies can be employed to semantically describe geometries within census data, each offering different levels of support for spatial representation and querying. For example, the NeoGeo Geometry Ontology allows for the description of geographical regions but lacks expressiveness for spatial relationships and does not support geometric representation (Salas et al., 2012). Similarly, the iCity Geometry Ontology facilitates the expression of geometries using the Well-Known Text (WKT) format but only partially addresses spatial relationships and is used for geometry representation rather than spatial querying (Katsumi, 2016). In contrast, the GeoSPARQL ontology, a standard developed by the Open Geospatial Consortium (OGC), has gained widespread adoption due to its full support in representing and querying geographic information in RDF. It extends SPARQL with spatial query capabilities and is grounded in established standards such as the General Feature Model (ISO 19109), Simple Features Access, GML, and builds upon SKOS. GeoSPARQL defines two

primary classes: *geo:SpatialObject*, which represents anything with a spatial representation, and *geo:Feature*, which denotes real-world entities associated with at least one geometry (Car et al., 2024). The ontology uses the *geo:hasGeometry* property to associate features with their geometric descriptions, thereby enabling robust spatial data modeling and querying on the Semantic Web.

2.3 Existing Census Linked Open Data Practices

Many countries around the world have embraced the publication of census data in LOD to improve its accessibility and interoperability and to support data-based decision-making processes. Since 2016, the Statistics Bureau of Japan has released census data as part of its broader statistical LOD initiative, with the goal of facilitating use by both domestic and international stakeholders. Interoperability is achieved through the application of established vocabularies such as the RDF Data Cube Vocabulary (QB) and the International Monetary Fund (IMF) vocabulary, which are used to structure the data into RDF triples (Asano et al., 2016). For instance, the "Population by Age and Sex" dataset is modelled using seven QB *dimension* components (area, sex, nationality, age, refArea, timePeriod, and measureType), one *measure* component (population), and four *attribute* components (unitMeasure, unitMult, obsType, and obsStatus). Spatial information in the model is represented using the GeoSPARQL ontology and designed to capture changes in geographic boundaries over time.

In Canada, a comprehensive ontology model for publishing population census data in LOD was introduced in 2024 (Wong et al., 2024). The model integrates both statistical and spatial components of the census and is built upon standards such as ISO/IEC 21972, ISO/IEC 5087-1, and the GeoSPARQL ontology. It introduces a *Characteristic* class, derived from ISO/IEC 21972, to represent census indicators (e.g. population count). These indicators are grouped within a *CensusProfile* class, which aggregates all characteristics associated with a specific geographic area. The actual values and units tied to each characteristic are captured in instances of the *Measures* class, also based on ISO/IEC 21972 and all the individuals being described by a given characteristic are represented through instances of the *Population* class. Spatial component in the model is described using locations and geometry, where the *hasLocation* property (from ISO/IEC 5087-1) connects census data to its location, and geometries are expressed via the *asWKT* property from the GeoSPARQL ontology.

In Europe, various countries have initiated efforts to publish census data as Linked Open Data. One of the earliest is the initiative by Fernandez et al. (2011), who proposed a model for publishing the 2001 Spanish census, emphasizing flexibility, openness, and adherence to linked data principles. Their approach utilizes LOD technologies to capture relationships among individuals, housing, and buildings, supported by a custom ontology to define semantics. While geographic areas are identified using URIs from the GeoNames ontology, the model does not include geometric representations of spatial units.

Petrou et al. (2013) developed a framework for publishing the 2011 Greek census data in LOD. They structured data on population (e.g., sex, marital status), households, and dwellings (e.g., address, type of residence) from .xls files using the QB vocabulary. The entire dataset was mapped to the qb:DataSet class, with columns treated as dimensions linked to relevant concepts in the QB schema: geocodes for regional divisions were mapped to *qb:DimensionProperty*; population count to *qb:MeasureProperty*; and unit of measurement to *qb:AttributeProperty*. The regional division was modelled to reflect administrative unit levels (based on geocodes) using concepts such as *dic:geolevel* from the SKOS vocabulary (*skos:ConceptScheme*), though it did not include geometric representation. Data modeling and publishing were carried out using the GoogleRefine data wrangling tool.

In Italy, Aracri et al. (2014) developed a model for publishing the 15th Italian Population and Housing Census in LOD. The model links census indicators to the corresponding datasets, focusing on aggregated data, similar to the approach taken by Petrou et al. (2013). The model uses the QB vocabulary to structure RDF triples, but unlike Petrou et al. (2013), the authors define their own attributes (e.g., *census:Year*, *census:Sex*) rather than using *qb:AttributeProperty*. Additionally, they employ the SKOS ontology to describe the classification of administrative units. In later adaptations, Aracri et al. (2018) revised the ontology to support the publication of microdata, including data on persons, families, and cohabitations. To achieve this, they mapped data from a relational database to RDF using R2RML, a customized language designed for these purposes (Das et al., 2012).

Meroño-Peñuela et al. (2016) proposed a model for publishing Dutch Historical Census data (1795-1971) in LOD. The authors highlighted that the concept of census, including characteristics, indicators, and their semantics, evolved over time, making the data modeling process a complex task. To represent the data in RDF, they utilized classes and relationships from the QB vocabulary for statistical variables. The spatial component was standardized using URIs from GeoNames and DBpedia, as well as the gemeentegeschiedenis.nl portal, which provides historical toponyms in LOD (Meroño-Peñuela et al., 2016).

Publishing census data as Linked Open Data requires more than just a formal ontology model. It also needs tools that can transform conventional file formats into RDF triples and support its publication on the web.

2.4 Data Wrangling Open-source Tools

According to Fiorelli and Stellato (2021), systems to transform tabular data into RDF can be classified as standalone converters, database-to-RDF, knowledge development environments, data wrangling tools and full-fledged ETL (extract, transform and load) solutions. Data wrangling tools, in particular, are typically lightweight, equipped with graphical user interfaces, and well-suited for handling local files or small datasets during exploratory data tasks. Among these, open-source solutions have proven functionalities to create RDFs from tabular data (Conde et al., 2022). For instance, GraphDB OntoRefine offers a user-friendly interface that facilitates predicate and type selection, streamlining the mapping process to an RDF schema. It also supports various input formats (e.g., .xls, .xlsx) and includes cleaning features to preprocess messy data (Ontotext, 2024). Similarly, LODRefine, a web-based application for RDF creation, provides format conversion and data preprocessing functionalities (Fioreli and Stellato, 2021). Both of these are built upon and extend OpenRefine, an open-source tool maintained and run by the community.

OpenRefine, previously named GoogleRefine, is an open-source, community-driven data wrangling tool designed for cleaning, enhancing, and transforming messy datasets (Ham,

2013). It features intuitive graphical interface and requires no previous knowledge of programming or query languages, making it accessible for users to quickly process data. While the core version of OpenRefine does not natively support mapping data to RDF structure, this capability can be added through extensions like RDF Extension or RDF Transform. These add-ons enable features like ontology term auto-completion and, when combined with OpenRefine's own expression language (GREL), support the creation of custom URIs tailored to the dataset.

Even though simple, OpenRefine is a powerful tool for transforming data into RDF triples. Its suitability for the creation of census LOD was showcased in Petrou et al. (2011) where authors used its functionalities to create LOD in a quick and easy way.

3. Methodology

The main goal of this research is to define ontology mapping schema and publish Croatian census in LOD. Based on existing vocabularies and ontologies relevant in the domain, the idea is to propose interoperable data model that will be used to structure census data in RDF triples and publish it in LOD. The proposed methodology is therefore divided into three steps: (1) data identification and preprocessing, (2) ontology schema development and (3) data transformation.

3.1 Data identification and preprocessing

Census data is made publicly available by the National Bureau of Statistics via its official data portal under the Open Government License. While data is offered in multiple formats and for different reference years, for the purposes of this research, the authors selected the most recent, 2021 census. Among its various population profiles, the dataset focusing on population counts by age and sex across spatial units down to the city/municipality level (in .xlsx format) was chosen due to its broad applicability in real-world applications.

Information in the dataset is divided between rows and columns. First five columns hold a hierarchical administrative classification of spatial units (census spatial units) in both Croatian and English language. Each entry includes the spatial unit's name, its administrative type (city or municipality), and its associated county. This classification enables unique identification of each record based on its spatial attributes. The remaining columns present aggregated population counts across 11 predefined age groups: total population; 0–6, 0–14, 0–17, 0–19 years; women of reproductive age (15–49 and 20–29 years); working-age population (15–64 years); and those aged 60+, 65+, and 75+ years. While age information is captured in the columns, sex is recorded in the rows with three entries per spatial unit—for men, women, and total population.

Spatial geometries corresponding to the census units are maintained separately by the State Geodetic Administration and can be accessed through their geoportal, also under the Open Government License. These data are distributed via an ATOM service in GML format, use the projected coordinate system (HTRS96/TM), and cover all levels of national spatial units' classification. Apart from geometry, every spatial unit includes attributes such as national code, localID, name, and type. However, not all attributes from the official Register of Spatial Units (see The Official Gazette of the Republic of Croatia, 2020) are included in the publicly available dataset.

Because the census and spatial data originate from different agencies and are provided in differing formats and structures, they are not inherently linked. Moreover, the current data structures do not support direct mapping to RDF. To address these limitations, several preprocessing steps were required. The census data, originally structured with multi-row entries for each spatial unit, was transformed into a single-row format. The sex classification was added as a refinement to the age dimension, making each row represent one spatial unit-sex-age contingent combination, which aligns better with our proposed RDF modelling approach. For spatial data (geometry), the GML file was filtered using QGIS to retain only spatial units up to the municipality level and was then reprojected into the WGS84 geographic coordinate system to enhance sharing. To ensure modularity and reusability, the census and geometry datasets were modelled independently and linked using URIs of spatial units. Since the provided national code and localID in spatial data (geometries) lacked a consistent hierarchical structure, custom unique IDs were assigned to census spatial units. These five-digit IDs encode administrative levels starting with a digit for the country, followed by couple of numbers representing counties and municipalities; based on alphabetical ordering of names.

The newly created IDs were then added to geometric representation of spatial units using QGIS. In QGIS, a join operation was performed using the spatial unit name as the join key. Since some municipalities share the same name (Otok, Privlaka and Sveta Nedelja) errors emerged from automated matching were fixed manually. Additionally, since the City of Zagreb functions as both a city and a county, its duplicate census entry was removed, and its sub-districts were excluded from modelling. Finally, geometry layer of spatial units was assigned a Well-Known Text (WKT) geometry representation for compatibility with semantic web technologies.

3.2 Ontology schema development

To foster visibility, interoperability, and potential future integration of RDF triples, the authors chose to reuse semantics from established ontologies and vocabularies when designing the ontology mapping schema for Croatian census data. The RDF Data Cube Vocabulary (QB) was employed to model the statistical (census) data, while GeoSPARQL was used to represent spatial geometries. To support additional semantic aspects beyond the core vocabularies, several complementary ontologies and vocabularies were incorporated, including RDF and RDF Schema, Dublin Core (dc, dcterms), Friend of a Friend (foaf), Simple Knowledge Organization System (skos), Web Ontology Language (owl), OWL-Time (time), XML Schema (xsd), and INSPIRE Administrative Units (au). These resources provided the necessary semantics for expressing metadata, add temporal properties, and administrative classifications.

Before developing the ontology mapping schema, a URI schema for publishing the census LOD was established, following the best practices recommended by the EC ISA Programme (Archer et al., 2012) To ensure URI persistence, two base domains were defined: one for census data at https://linked-census.hr and another for spatial units at https://linked-spatialunits.hr. These base URIs were then extended to distinguish between schema components and actual data.

Schema components, which define the structure of the census data—including specifications for dimensions, measures, and

attributes—are placed under the *{base_URI}/schema/{ComponentType}* path. Actual datasets and their corresponding observations are organized under the *{base_URI}/data/{DataSetName}* and *{base_URI}/data/{DataSetName}/{ObservationName}* paths. For spatial data, URIs are structured as *{base_URI}//unit/{SpatialUnitID}* to represent individual spatial units, and *{base_URI}//geometry/{SpatialUnitID}* to define their geometries.

For the spatial units (Figure 1.), each unit was defined as an INSPIRE Administrative Unit and *geo:Feature* and enriched with multilingual labels, its corresponding level within the national administrative hierarchy, and geometry expressed in WKT format. To represent the administrative structure, the *geo:isWithin* property was used to connect each lower-level unit to its respective higher-level unit. This hierarchical relationship was established by leveraging existing dataset attributes and applying conditional logic through GREL (Google Refine Expression Language).

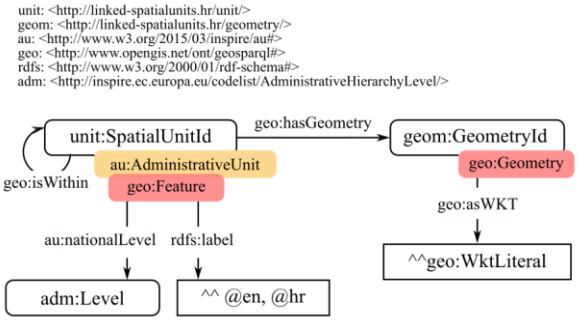

Figure 1. Proposed ontology schema for mapping Croatian spatial units in RDF

In the QB (RDF Data Cube) model (Figure 2.), census data, represented as observations, are grouped within a dataset that adheres to a predefined data structure definition (DSD). Following this premise, we first defined the structure *schema:PopulationContingents2021* with its three *dimensions* (*schema:spatialUnit, schema:contingent, and schema:sex*), one *measure* (*schema:populationCount*), and one *attribute* (*schema:unitOfMeasure*) (see Figure 2).

Next, we created the dataset *data:PopulationContingents2021*, linked it to the defined structure, and assigned with metadata such as a label, descriptive comment, publisher, reference year, and license. To ensure semantic interoperability, the publisher was described using the FOAF vocabulary, indicating the organization type and its name.

Since the unit of measure is consistent across all observations, it was defined at the dataset level to avoid redundancy. Finally, the observation schema was developed to map the dataset's actual content into RDF triples. Within this schema, each observation was assgined *data:PopulationContingents2021* dataset and mapping rules for dimensions and measures. Contingent groups and sex categories (both dimensions) were introduced as schema classes and further described as *skos:Concept* instances with preferred labels, enabling their reuse across different census profiles or other domains applications.

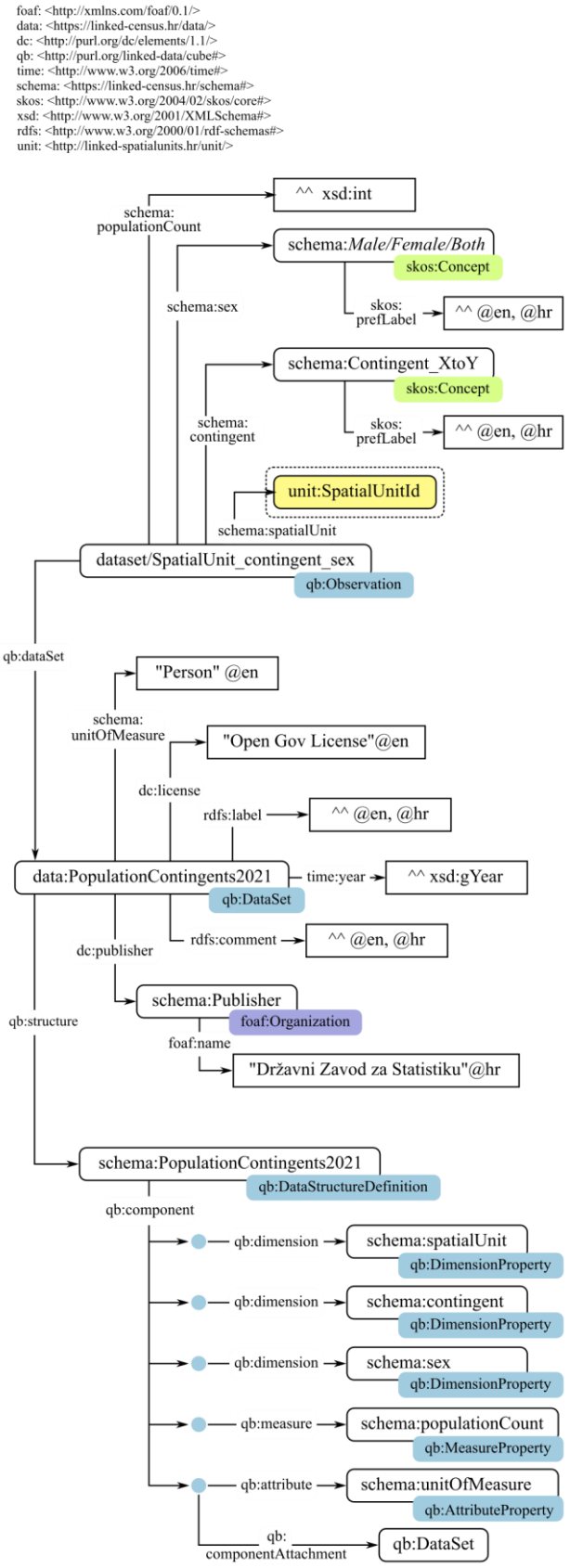

Figure 2. Proposed ontology schema for mapping Croatian population census age contingents in RDF

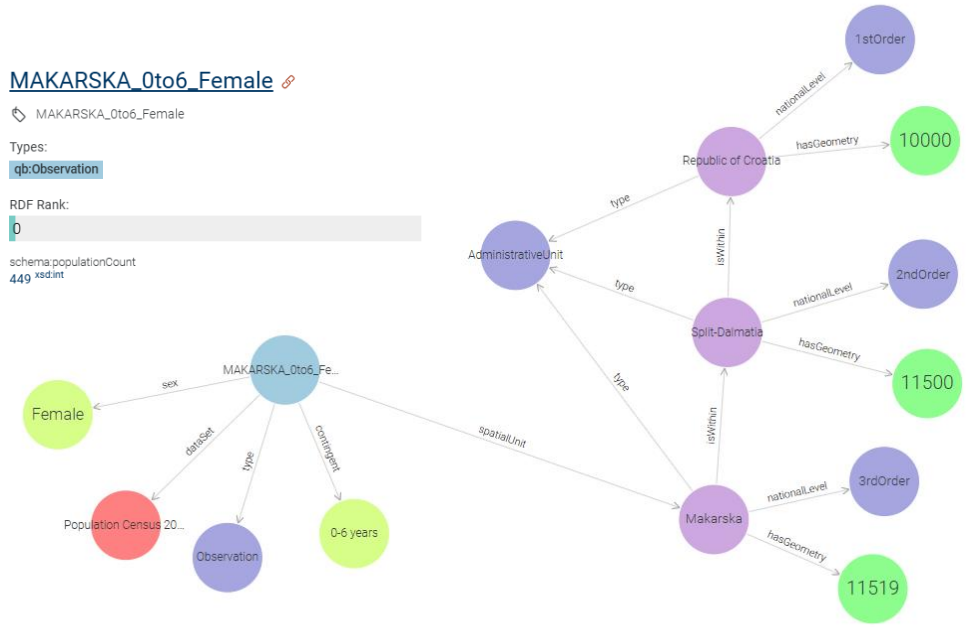

Figure 3. An RDF representation of female sex group aged 0-6 years for town of Makarska, Split-Dalmatia County

3.3 Data transformation

The transformation of data from .xlsx format into RDF triples was carried out using OpenRefine tool. To represent geometries of spatial units in RDF, we utilized OpenRefine's GREL language for conditional labeling. Each spatial unit was assigned a unique URI based on its corresponding 5-digit identifier. Since spatial unit names were distributed across three columns, GREL was employed to filter these values and to generate multilingual labels in Croatian and English (rdfs:label). Additionally, the administrative hierarchy (au:nationalLevel property) was inferred from the column sourcing the name, classifying units as first-order (country), second-order (county), or third-order (local level: city/municipality) depending on the source column. Geometry column was assigned *geo:wktLiteral* property, and spatial containment relationships were established via *geo:isWithin*, determined through GREL-based conditional logic.

For the census dataset, the OpenRefine project was divided into three root nodes: *schema:PopulationContingents2021* (data structure definition), *data:PopulationContingents2021* (dataset), and the observation entries. While the structure and dataset elements were defined conceptually (following QB definition) to support data organization and semantic clarity, each spreadsheet row was treated as a *qb:Observation* in accordance with the RDF Data Cube Vocabulary. As each row contained information covering multiple characteristics for a single geographic area, it was decomposed into 29 distinct observations, each representing a specific combination of age group and sex category. Each observation was linked to its corresponding spatial unit (geometry) via the *schema:spatialUnit* property, and observed value, i.e., the population count, through the *schema:populationCount* property. Age groups (e.g., schema:Contingent_0to6) and sex categories (e.g., schema:Female) were predefined as *skos:Concept* and referenced in observations through the *schema:contingent* and *schema:sex* properties, respectively, and for all these classes, preferred labels in English were provided using skos:prefLabel.

Once mapped to the defined ontology mapping schema, the RDF data was exported in Turtle (Terse RDF Triple Language) format and loaded into the graph database (Ontotext GraphDB) for visualization and validation. Table 1. shows RDF classes and number and direction (incoming/outgoing) of links for each class while Figure 3. illustrates an RDF visualization of the observation representing the female population aged 0–6 years in the town of Makarska. The graph highlights the relationships among class instances, with instance-specific details, such as population counts and WKT geometries, contained within the corresponding graph nodes. All RDF files can be found at our Github repository: https://github.com/kkevic/linked-statistics.git.

Class name	Number of links	Direction
qb:Observation	94 K	outgoing
skos:Concept	61 K	incoming
au:AdministrativeUnit	33 K	both
qb:DataSet	2 K	both
geo:Feature	574	incoming
qb:DimensionProperty	4	both
qb:DataStructureDefinition	1	incoming
foaf:Organization	1	incoming
qb:MeasureProperty	2	both
qb:DimensionComponent	2	both
qb:AttribtueProperty	2	both

Table 1. RDF classes and number of established links

4. Conclusion

This paper presented a case study on publishing publicly available Croatian population census data as Linked Open Data (LOD). In Croatia, census data is currently disseminated in .xlsx

format without accompanying geometric references, which presents significant challenges for reuse and integration. As part of this research, we developed an ontology mapping schema for the 2021 Croatian population census—focusing on sex and age contingency profiles—and incorporated spatial geometries into the LOD framework.

To ensure interoperability and broader applicability, the schema leverages established ontologies and vocabularies. In particular, we adopted the RDF Data Cube Vocabulary to model statistical data and GeoSPARQL for the representation of spatial geometries. Also, by using OpenRefine we ensured replicability of the process and showcased its suitability for publication of census data in the proposed way.

The proposed ontology mapping schema provides a solid foundation for publishing census data in Croatia as LOD. In accordance with LOD principles, census information and spatial geometries were modeled separately and interconnected through URIs. This approach promotes the reuse of spatial geometries across various (non-)statistical datasets and applications. Additionally, incorporating a multilevel hierarchy of spatial units enhances semantic precision and supports inferencing. Modeling census indicators such as age groups and sex categories as reusable concepts rather than literals allows their further refinements and linking with other relevant vocabularies in the domain.

Nevertheless, some limitations remain. The linkage between census data and spatial geometries relies on ID created by the authors which imposes limitations for wider URI reuse. Also, names of spatial units can be linked to external sources such as GeoNames to provide more semantics. For census data, observations that share the same dimension could be grouped under slices of data cube for better data integrity.

Overall, OpenRefine proved to be a capable and user-friendly tool for modeling and publishing census data in RDF. Its intuitive graphical interface features easy-to-use options needed for simple transformation of data into RDF. This way, it is possible not only to map the data to proposed schema, but to perform simple alterations in the original data source.

While the proposed schema marks an important step towards better data integration, accessibility, and insight generation, it currently addresses only one population profile. Future work could therefore involve extending the schema to include additional census profiles, such as ethnicity or religion, and linking population data with housing data, given their shared role in the comprehensive Population and Housing Census.

References

Aracri, R.M., De Francisci, S., Pagano, A., Scannapieco, M., Tosco, L., Valentino, L., 2014. Publishing the 15th Italian Population and Housing Census as Linked Open Data. In *Proceedings of the 2nd International Workshop on Semantic Statistics co-located with 13th International Semantic Web Conference* (SemStats@ISWC 2014). Riva del Garda, Italy Article 3.

Aracri, R.M., Radini, R., Scannapieco, M., Tosco, L, 2018. Using Ontologies for Official Statistics: The Istat Experience, In *Current Trends in Web Engineering*. (ICWE 2017). Lecture Notes in Computer Science, vol 10544, Rome, Italy, 166-172. https://doi.org/10.1007/978-3-319-74433-9_15.

Archer, P., Goedertier, S., Loutas, N., 2012. D7.1.3 – Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC, Interoperability Solutions for European Public Administrations, Available at: https://interoperable-europe.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf.

Asano, Y., Takeyoshi, Y., Matsuda, J., Nishimura, S., 2016. Publication of Statistical Linked Open Data in Japan, In *Proceedings of the 4th International Workshop on Semantic Statistics co-located with 15th International Semantic Web Conference* (ISWC 2016). Kobe, Japan, Article 1.

Car, N.J., Homburg, T., Perry, M., Knibbe,F., Cox, S.J.D., Abhayaratna, J., Bonduel, M., Cripps, P.J., Janowicz, K., 2024. OGC GeoSPARQL – A Geographic Query Language for RDF Data, OGC Standard, http://www.opengis.net/doc/IS/geosparql/1.1 (21.3.2025).

Conde, J., Munoz-Arcentales, A., Choque, J., Huecas, G., Alonso, Á., 2022. Overcoming the Barriers of Using Linked Open Data in Smart City Applications. *Computer*, 55(12), 109-118. doi: 10.1109/MC.2022.3206144.

Cyganiak, R., Reynolds, D., 2014. The RDF Data Cube Vocabulary, W3C Recommendation. http://www.w3.org/TR/vocab-data-cube/ (20.3.2025).

Das, S., Sundara, S., Cyganiak, R., 2012. R2RML: RDB to RDF Mapping Language, W3C Recommendation, https://www.w3.org/TR/r2rml/

European Commission, 2019. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). *Official Journal of The European Union* L 172/56, Available at: https://eurlex.europa.eu/eli/dir/2019/1024/oj#document1 (25.3.2025)

Fernández, J. D., Martínez-Prieto, M. A., Gutiérrez, C., 2011. Publishing open statistical data: the Spanish census. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* (dg.o '11). Association for Computing Machinery, New York, NY, USA, 20–25. https://doi.org/10.1145/2037556.2037560.

Fiorelli, M., Stellato, A., 2021. Lifting Tabular Data to RDF: A Survey. In Metadata and Semantic Research (MTSR 2020). In *Communications in Computer and Information Science*, vol 1355, Madrin, Spain, 85-96. https://doi.org/10.1007/978-3-030-71903-6_9.

Ham, K., 2013.: OpenRefine (version 2.5). http://openrefine.org. Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association*, 101(3), 233-234, doi: 10.3163/1536-5050.101.3.020.

Hogan, A., 2020. *Web of Data*. Springer, Cham. https://doi.org/10.1007/978-3-030-51580-5_2

Katsumi, M., 2016. iCity Geometry Ontology, http://ontology.eil.utoronto.ca/icity/Geom/ (22.4.2025)

Meroño-Peñuela, A., Ashkpour, A., Guéret, C., Schlobach, S., 2016. CEDAR: The Dutch historical censuses as Linked Open Data. *Semantic Web*, 8(2), 297-310. doi:10.3233/SW-160233.

The Official Gazette of the Republic of Croatia, 2021. Zakon o popisu stanovništva, kućanstava i stanova u Republici Hrvatskoj 2021. godine, 34, in Croatian, Available at: https://www.zakon.hr/z/2501/zakon-o-popisu-stanovnistva%2C-kucanstava-i-stanova-u-republici-hrvatskoj-2021.-godine#google_vignette (16.3.2025).

The Official Gazette of the Republic of Croatia, 2020. Pravilnik o registru prostornih jedinica, 37, in Croatian. Available at: https://narodne-novine.nn.hr/clanci/sluzbeni/2020_03_37_808.html (15.3.2025).

United Nations Statistical Commission, 2015. The 2020 World Population and Housing Census Programme, Economic and Social Council, Official Records 2015 Supplement No. 4, Available at: https://unstats.un.org/unsd/statcom/46th-session/documents/statcom-2015-46th-report-E.pdf (24.3.2025)

Ontotext, 2024. GraphDB EE Documentation, Release 9.11.0. Available at: https://graphdb.ontotext.com/documentation/9.11/enterprise/loading-data-using-ontorefine.html (20.3.2025).

Pabón, G., Gutiérrez, C., Fernández, J.D., Martínez-Prieto, M.A., 2013. Linked Open Data Technologies for Publication of Census Microdata. *Journal of the American Society for Information Science and Technology*, 64(9), 1802-1814. https://doi.org/10.1002/asi.22876.

Petrou, I., Papastefanatos, G., Dalamagas, T., 2013. Publishing census as linked open data: a case study. In *Proceedings of the 2nd International Workshop on Open Data* (WOD '13). Association for Computing Machinery, New York, NY, USA, Article 4, 1–3. https://doi.org/10.1145/2500410.2500412.

Rizzolo, F., Gillman, D., Vucko, F., 2023. A Core Ontology for Official Statistics, https://linked-statistics.github.io/COOS/coos.html (22.4.2025).

Salas, M.J., Harth, A., 2012. NeoGeo Geometry Ontology, http://geovocab.org/geometry (22.4.2025).

Wong, A., Fox, M., Katsumi, M., 2024. Semantically interoperable census data: unlocking the semantics of census data using ontologies and linked data. *International Journal of Population Data Science*, 9(1):2378. doi: 10.23889/ijpds.v9i1.2378.