Transactions in GIS

WILEY

Transactions in GIS

RESEARCH ARTICLE OPEN ACCESS

Integrating Forestry and Local Administrative Data Using Semantic and GIS Technologies

Natalia Crespo-Lera[1,2] | Guillermo Vega-Gorgojo[1,2] | José M. Giménez-García[1,2] | Cristina Mayo Sarmiento[3] | Felipe Bravo[1] | Irene Ruano[1]

[1]SMART Ecosystems Research Group, Instituto de Investigación en Gestión Forestal Sostenible (iuFOR), Universidad de Valladolid, Palencia, Spain | [2]Grupo de Sistemas Inteligentes y Cooperativos (GSIC) ETSI de Telecomunicación, Universidad de Valladolid, Valladolid, Spain | [3]Subdirección General de Tecnología y Despliegue Digital, Dirección General de Catastro, Madrid, Spain

Correspondence: Natalia Crespo-Lera (natalia.crespo@uva.es)

Received: 6 June 2024 | **Revised:** 10 October 2024 | **Accepted:** 20 December 2024

Funding: This work was supported by European Regional Development Fund (Grant PID2023-146692OB-C32). NextGenerationEU (Grant TED2021-130667B-I00). Junta de Castilla y León. Horizon-RIA programme of the European Union (Grant 069/230331).

Keywords: GeoSPARQL | GIS | linked open data | local administrative units | National Forest Inventory data

ABSTRACT

Linking National Forest Inventory (NFI) data with local administrative units (LAUs) unlocks a wealth of benefits for forest management. It enhances precision in assessments, facilitates policy alignment, optimizes resource allocation, and enables effective monitoring and research at the local level. This study presents an automated process to convert official Spanish municipality geometries from the National Geographic Institute's geospatial database into Linked Open Data (LOD) format. This facilitates the assignment of Spanish NFI plots to their corresponding municipalities using GeoSPARQL and their publication in an open LOD repository. Additionally, we compared the results of this assignment with a GIS-based solution. To demonstrate the potential of this spatial integration, we conducted analyses in two case studies. This work highlights the benefits of integrating forest information with other cross-domain data using Semantic Web technologies, which can be further complemented by GIS software for spatial analysis.

1 | Introduction

Municipalities are at the bottom of the local administrative units (LAU) in each country worldwide. The use of this lower-level LAU holds different benefits for forest managers and stakeholders as (1) LAUs cover the whole country of interest with no overlapping geographical polygons, (2) municipalities are nested within upper LAUs and NUTs (Nomenclature of Territorial Units for Statistics), such as counties, provinces and so forth allowing information upscaling, and (3) LAUs nomenclature are standardized with unique identifiers leading to an easier data integration. By making forestry data integrated with local administrative units publicly available, we provide a crucial foundation for the digital and bio-based economy. This increases opportunities for innovation, sustainability, public policy evaluation, and the development of competitive business models (see Rantala et al. (2020) for the case of opening data in the Finnish forest sector).

Municipality or county based forest assessment, and informed decision-making typically rely on National Forest Inventories (NFIs). These inventories serve as the foundation for analyzing and comparing forested areas across various topics, including reforestation (e.g., Marey-Pérez and Rodríguez-Vicente (2009)), biodiversity (e.g., Guadilla-Sáez et al. (2019)), biomass content and carbon capture (e.g., Gil et al. (2011); Lorenzo-Sáez et al. (2022)), or wildfires (e.g., Martínez, Vega-Garcia, and Chuvieco (2009)). The location of NFI plots within municipalities allows local authorities to tailor forest management strategies to their specific needs, as the characteristics

and extent of forest cover can vary significantly within municipal boundaries.

Accurate analysis of forests at the municipal and, subsequently, county level using data from the Spanish National Forest Inventory (SNFI) is currently hindered by data inconsistencies. The latest edition of the SNFI (SNFIv3) assigns municipality codes to plots through two separate tables, one pre-fieldwork and one post-fieldwork. This redundancy leads to challenges: 47% of plots have both codes (with 2% inconsistencies), 34% have only one code, and 19% have no code at all. Integrating SNFI data with the official geospatial dataset of municipalities could address this limitation. By performing a spatial join, we can definitively assign SNFI plots, along with all their associated information, to the correct municipality.

Data integration is a difficult problem that can be solved with Semantic Web technologies (Lan et al. 2022), a set of knowledge representation languages standardized by the World Wide Web Consortium (W3C) that have an inherent ability to formally depict knowledge in a semantically enriched way (Hayes 2003). The Semantic Web (Berners-Lee, Hendler, and Lassila 2001) aims to create a "Web of Data", that is, the extension of the Web with a global data space based on open standards (Heath and Bizer 2022). To achieve this vision, data publishers are encouraged to follow the linked data principles (Pascal 2021). The result is Linked Open Data (LOD), structured data which are interlinked to other data and released under an open license. Louarn et al. (2019); Kamdar et al. (2019); Arslan, Desconnets, and Mougenot (2022) are some examples of using LOD and Semantic Web technologies for data integration.

In the forestry domain, the European project Cross-Forest is a prime example of the use of Semantic Web technologies for data integration (Fierro García et al. 2022). The main outcome of this project is the creation of a freely accessible LOD resource, henceforth the Cross-Forest dataset, seamlessly integrating the most up-to-date NFIs and land cover maps from Spain and Portugal.

In this paper, we aim to support forestry studies and processes at local level in Spain. We use the Cross-Forest dataset as a starting point, since it already includes SNFIv3. We then integrate the official municipality dataset provided by the Spanish National Geographic Institute. In a next step we obtain the assignment of SNFI plots to municipalities with GeoSPARQL (LOD-based) and a GIS. The obtained results are republished as LOD in the Cross-Forest dataset, increasing its value by allowing data reuse and the creation of new applications for diverse purposes (commercial, educational, management, conservation, etc.). Finally, we exploit the enhanced dataset in some application case studies that comprise the analysis of the dominant species per municipality in a Spanish National Park and the forest diversity of 17 selected Spanish counties.

The contributions are thus fourfold:

1. Comparison of the performance of different GeoSPARQL implementations (Virtuoso and Fuseki) and a GIS (QGIS) in employing the "within" spatial predicate.
2. Correct Assignment of SNFI plots to Spanish municipalities.
3. Publication as LOD of the obtained results to facilitate the realization of studies at the local level.
4. Presentation of two case studies illustrating the benefits of LOD for local-based forestry management.

The rest of the paper is organized as follows: Section 2 introduces the Semantic Web technologies used in this research and provides an overview of the source datasets employed. Section 3 details the geospatial processing techniques applied, including plot data preparation, municipality data preparation, plot-to-municipality assignment, cross-validation, and publication processes. Section 4 presents two case studies to demonstrate the practical applications and benefits of integrating LOD for local forest management. Section 5 compares the performance of different GeoSPARQL (LOD-based) implementations and a GIS when performing the plot-to-municipality assignment. It also includes examples of maps and graphs that illustrate the two case studies and demonstrate the effectiveness of the proposed methodologies. Finally, Section 6 concludes the paper by reflecting on the results, discussing the applications of the proposed methodologies, and suggesting possible directions for future research.

2 | Background Knowledge

This section begins with an introduction of Semantic Web technologies and Linked Open Data (LOD), providing the background for semantic data integration. Section 2.2 then details the data sources used in this work: the third Spanish National Forest Inventory (SNFIv3) and the official database of Spanish municipalities.

2.1 | Semantic Web Technologies

For integrating the datasets, we employed the principles and best practices of LOD for publishing and connecting structured data on the Web. This involves using Semantic Web technologies and standards established by the W3C (Berners-Lee, Hendler, and Lassila 2001). These standards include RDF (Resource Description Framework) as the data modeling language, OWL (Web Ontology Language) or RDFS (RDF Schema) for constructing ontologies or vocabularies, and SPARQL (Protocol and RDF Query Language) as the query language for RDF data. Furthermore, SPARQL CONSTRUCT is a type of SPARQL operation that allows the generation of new RDF graphs—a collection of interconnected triples—extracted from a dataset. The adoption of these standards offers new possibilities for analyzing, processing, and modeling interdisciplinary data, facilitating greater insight and pattern recognition in highly complex datasets (Lausch, Schmidt, and Tischendorf 2015).

RDF is a standard model for data interchange on the Web based on "triples". An RDF triple is a tuple of three terms `(subject, predicate, object)`; subjects (the resources being described) and predicates are identified by Internationalized Resource Identifiers (IRIs) (Dürst and Suignard 2005), whereas objects (the values for the properties) can be either other resources or literals (values that do not correspond with resources, such as strings, numbers, or dates). RDF provides a data model

for making statements about resources, but it does not make any assumptions about the meaning of IRIs. In practice, RDF is used in combination with ontologies (OWL and/or RDFS) that define the terminology of a specific domain, such as forestry.

IRIs are effective for identifying resources on the Web but they can be very long to write out in detail. To address this, a simplified IRI abbreviation scheme called QNames (qualified names) (Bray et al. 2009; Allemang and Hendler 2011) is used in this work for examples. A qname representation consists of two parts: a namespace and an identifier, separated by a colon (e.g., `nfi:Species72` for the identifier `Species72`, which is a term defined in a species ontology, in the `nfi` namespace). The Table 1 provides the list of prefixes and namespaces employed in this work.

TABLE 1 | Prefixes and their corresponding namespaces used in this work.

Prefixes	Namespaces
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
owl	http://www.w3.org/2002/07/owl#
geof	http://www.opengis.net/def/function/geosparql/
wkt	http://www.opengis.net/ont/geosparql#
spatialF	http://jena.apache.org/function/spatial#
spo	http://crossforest.eu/position/ontology/
epsg	http://epsg.w3id.org/ontology/
crs	http://crossforest.eu/epsg/data/crs/
axis	http://epsg.w3id.org/ontology/axis/
polygon	https://datos.iepnb.es/recurso/sector-publico/medio-ambiente/ifn/polygon/
territory	https://datos.iepnb.es/recurso/sector-publico/territorio#
province	https://datos.iepnb.es/recurso/sector-publico/territorio/Provincia/
municipality	https://datos.iepnb.es/recurso/sector-publico/medio-ambiente/ifn/municipality/
nfi	https://datos.iepnb.es/def/sector-publico/medio-ambiente/ifn/
plot	https://datos.iepnb.es/recurso/sector-publico/medio-ambiente/ifn/plot/
tree	https://datos.iepnb.es/recurso/sector-publico/medio-ambiente/ifn/tree/

Figure 1 provides an example of two triples with the same subject, a tree with IRI `tree:33-1518-N-N-2`. An example of one of the triples shown in Figure 1 with prefixes is as follows: `tree:33-1518-N-N-2 rdf:type nfi:Species72`. This triple indicates that the tree belongs to nfi:Species72, which represents the species *Castanea sativa* Mill. This class is defined within an ontology of tree species according to the SNFI codes and an ontology of NFIs. The second triple specifies the tree's height in meters, using a property defined in the NFIs ontology (`nfi:hasTotalHeightInMeters`). The object of the triple is a literal value (decimal number).

Semantic data are stored in specialized database systems called triplestores. These systems are designed specifically to manage and query RDF data efficiently. Triplestores offer a variety of features and capabilities, allowing researchers to select the one that best suits their project's needs. Popular examples include Virtuoso, Fuseki, and Graph DB. All of these triplestores provide a SPARQL endpoint, which acts as a query processor for the SPARQL language.

For instance, the following SPARQL query retrieves all *Castanea sativa* trees within the SNFIv3 dataset that have a height of 15 m or more. This query can be executed against the Cross-Forest SPARQL endpoint. To demonstrate the query format, we have included prefixes in this example; in subsequent queries, these prefixes will be omitted as they are specified in Table 1.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf
    -syntax-ns#>
PREFIX nfi: <https://datos.iepnb.es/def/
    sector-publico/medio-ambiente/ifn/>

SELECT DISTINCT ?tree WHERE {
  ?tree rdf:type nfi:Species72;
    nfi:hasTotalHeightInMeters ?height.
  FILTER (?height >=15)
}
```

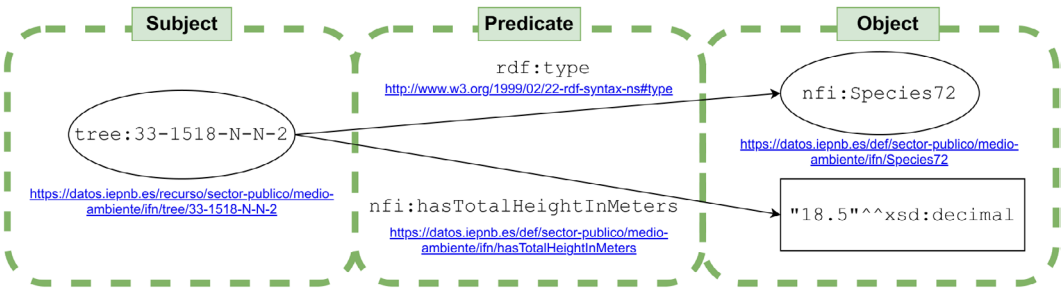

FIGURE 1 | Example of a graph with two RDF triples about a tree with IRI `tree:33-1518-N-N-2`. The first triple specifies that this tree is of type `nfi:Species72` (*Castanea sativa*). The second one indicates its height by using a property from the NFI ontology (`nfi:HasTotalHeightInMeters`) with value 18.5 (a literal).

The `SELECT` part of the query specifies the variables we want to retrieve from the dataset. `DISTINCT` ensures that the results are unique, avoiding duplicate entries. `?tree` is a variable that will hold the results, in this case, the trees with the specified characteristics.

The `WHERE` clause defines the conditions or patterns that the data must match to be included in the results. Within the braces, we indicate that we are looking for a variable named ?tree of type `nfi:Species72`, which means we are searching for trees classified as `Species72` (*Castanea sativa*). Additionally, we specify that each tree must have a recorded height, corresponding to variable `?height` in the query, through property `nfi:hasTotalHeightInMeters`. Finally, the `FILTER` clause applies a condition to the results, ensuring that only trees higher or equal to 15 m will be retrieved.

Furthermore, to simplify querying and handling geospatial data in LOD format, the Open Geospatial Consortium (OGC) has developed the GeoSPARQL standard (OGC 2012). This standard consists of an ontology for describing geospatial RDF data and a set of spatial functions for integrated use with the SPARQL query language for performing spatial analyses. GeoSPARQL supports representing features using various geometries like points, lines or polygons, and spatial functions (OGC 2012); one of these functions is `wkt:sfWithin` that serves to detect if a geometry is entirely contained within another geometry.

Many triplestores, such as Virtuoso or Fuseki, provide support for GeoSPARQL, allowing SPARQL queries on semantically annotated geospatial data. However, the level of GeoSPARQL support may differ across different implementations (Jovanovik, Homburg, and Spasić 2021; Li et al. 2022), which should be taken into account when using different triplestores for geospatial analysis.

2.2 | Data Sources

2.2.1 | National Forest Inventory Data

This research leverages plot locations extracted from the Spanish National Forest Inventory (SNFI) version 3 (SNFIv3). Conducted between 1997 and 2007, SNFIv3 represents the most recent and comprehensive publicly available inventory dataset, accessible through the Ministry for the Ecological Transition and the Demographic Challenge (MITECO). MITECO initially positioned SNFIv3 plots using a systematic sampling approach. This approach relied on three key criteria: (1) a 1 km × 1 km Universal Transversal Mercator (UTM) grid, (2) the Spanish Forest Map at a 1:50,000 scale to identify forest cover, and (3) the locations of plots sampled during the second SNFI edition (1986–1996).

During fieldwork, a total of 99,045 plots (Figure 2) were identified by their central coordinates using the UTM ED50 (European Datum 1950) reference system. However, ED50 is no longer the official datum for mapping and geospatial applications in Europe. Currently, the preferred data are ETRS89 (European Terrestrial Reference System 1989), the official geodetic system in Spain since 2007 (BOE 2007), and WGS84 (World Geodetic System 1984). For any further analysis or integration with other datasets, it is crucial to consider a datum transformation from ED50 to either ETRS89 or WGS84. This ensures data compatibility and accuracy in modern geospatial workflows. While both data are very close, there are slight coordinate variations.

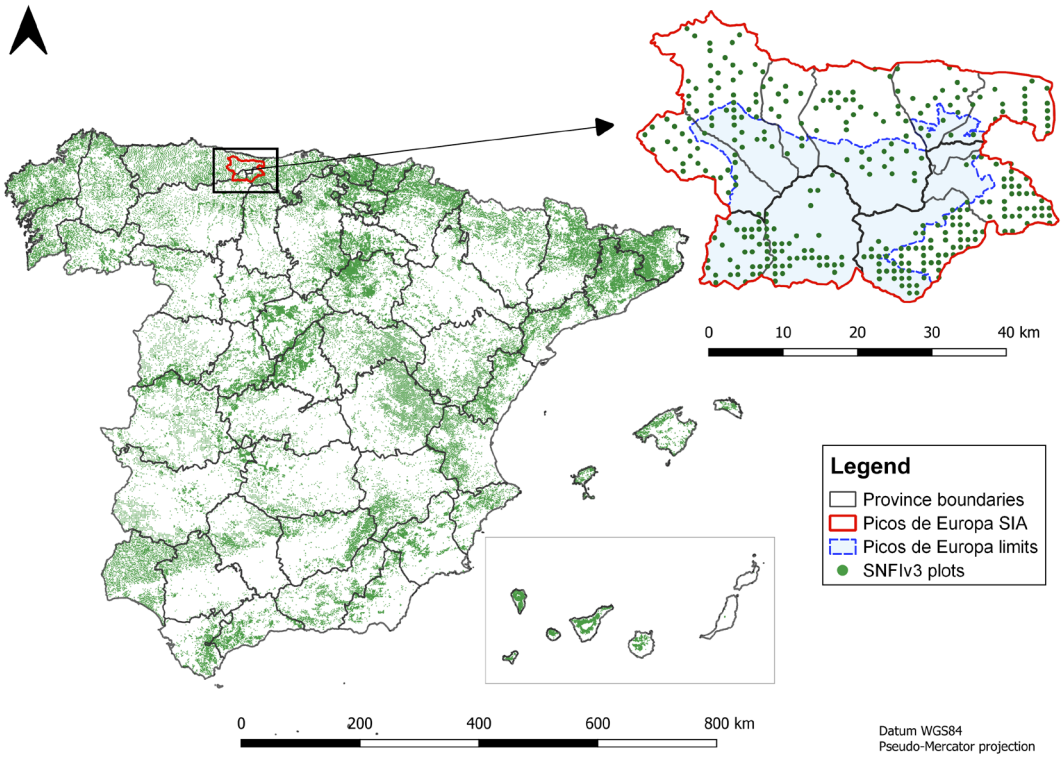

FIGURE 2 | Third National Forest Inventory plots (SNFIv3) distribution within the province division of Spain with special focus on the Picos de Europa National Park Socioeconomic Impact Area (SIA) and limits of the National Park Protected Area.

Notably, WGS84 is used for global positioning and navigation, whereas ETRS89 is specific to Europe.

Information regarding municipalities within SNFIv3 is scattered across two separate database tables. These were created at different times and through distinct processes, leading to inconsistencies in their contents. The first database table was created before the field sampling and contains residual information from the cartographic layer available at that time. This table encompasses municipality codes for approximately 81% of the inventory plots. While its creation aimed to provide an initial reference, its reliance on outdated cartographic data may lead to inaccuracies. The second database table, compiled during fieldwork by forestry technicians, assigns municipality codes to plots. However, these data cover only 47% of the plots, leaving a substantial information gap. Furthermore, a 2% discrepancy exists between the codes when information is present in both database tables in the SNFIv3. Table 2 (see below) provides a comparison of the municipality code data associated with the inventory plots in the SNFIv3 dataset.

TABLE 2 | Municipality information associated with plots in the third edition of the Spanish National Forest Inventory (SNFIv3).

	Before fieldwork	After fieldwork
Plots with municipality code	80,286 (81%)	46,443 (47%)
Plots without municipality code	18,759 (19%)	52,602 (53%)
Same municipality code	45,670 (98%)	
Different municipality code	773 (2%)	

In general, utilizing the SNFIv3 dataset published by MITECO presents several challenges: (1) the data are not consolidated into a single database, making multi-province analysis difficult; (2) the dataset comprises 1.1 k distinct tables and 282.9 k different columns, resulting in a complex schema; (3) there are minor format inconsistencies across the different files; and (4) relies on the old Microsoft Access 2007 version, which is a proprietary database format no longer supported.

2.2.1.1 | **Cross-Forest Dataset.** The Cross-Forest project (Portolés et al. 2021) provides open access to the latest Spanish National Forest Inventory (SNFIv3) data. These data are published in Linked Open Data (LOD) format, allowing for flexible querying and analysis through the Cross-Forest SPARQL endpoint, using the freely available Virtuoso triplestore (version 07.20.3230). Each plot within the Cross-Forest dataset is identified by a special code. This code combines four elements: (1) the province code (e.g., 33 for Asturias); (2) a unique field identifier for the specific plot within the province (e.g., 1518); (3) a class designation indicating whether the plot was surveyed during the second edition of the SNFI (A) or introduced in the third edition (N); and (4) a subclass providing additional details. The Cross-Forest dataset goes beyond SNFIv3 by incorporating additional geospatial information: Spanish Forest Map (1:50,000 scale), Portuguese Land Use and Occupancy Map (2018), and Summarized Portuguese National Forest Inventory (version 6). Moreover, a set of ontologies has been developed to annotate the data, covering aspects such as geographic positions, coordinate reference systems, measures, forest inventories, and land cover maps. The Cross-Forest project is committed to continuous improvement. Future plans include integrating new and existing datasets: the second edition of the Spanish NFI, a

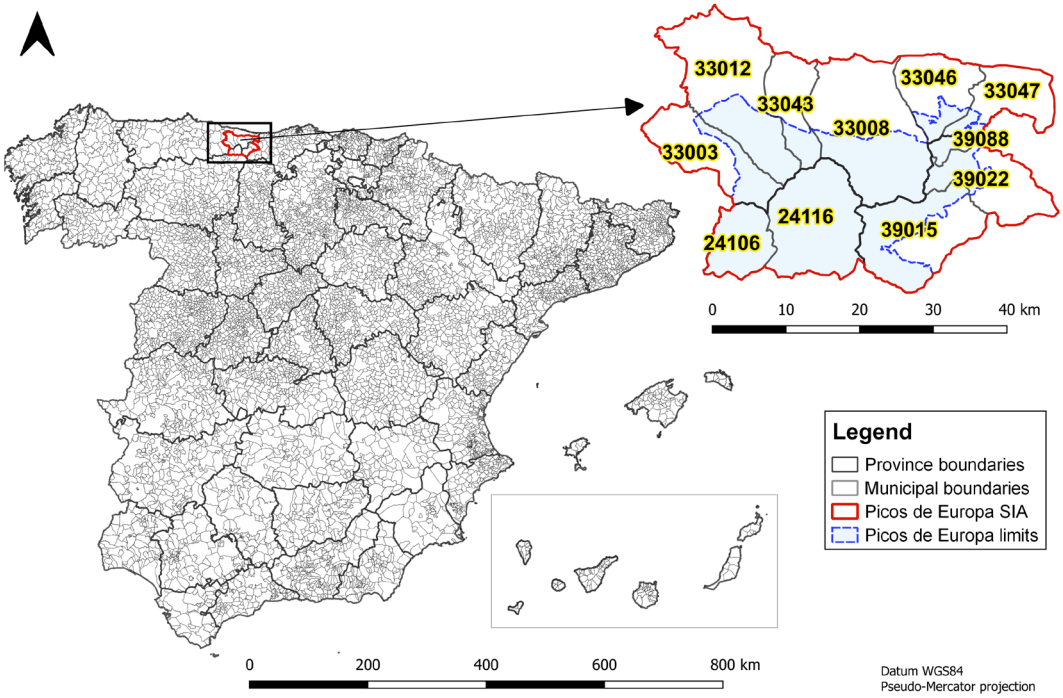

FIGURE 3 | Spanish municipal boundaries within their respective provinces, with a focus on the municipalities of the Picos de Europa National Park Socioeconomic Impact Area (SIA).

higher resolution Spanish Forest Map (1:25,000 scale), LiDAR maps, geographical boundaries, and bioclimate databases. This expansion will further enrich the Cross-Forest dataset, making it an even more powerful tool for a wide range of forest research endeavors.

2.2.2 | Spanish Municipality Data

The National Geographic Institute (NGI) publishes the official territorial boundaries of the Spanish municipalities that are updated every year. The dataset consists of 8131 polygons or multipolygons of varying sizes, providing the delimitations of the municipal boundaries. Figure 3 represents the Spanish municipalities boundaries within each province and highlights their diverse sizes, indicating varying land resources and administrative capacities. The figure focuses on the Picos de Europa National Park, which covers three provinces, León (codes starting with 24), Asturias (codes starting with 33), and Cantabria (codes starting with 39), illustrating the challenges involved in managing such cross-provincial areas.

TABLE 3 | Attribute Table of Amieva Municipal Boundary (inside Picos de Europa National Park), including extracted municipality code (IDmuni) and geometry vertices in WKT Format.

INSPIRED	ES.IGN.BDDAE.34022222131
COUNTRY	COUNTRY: ES
NATLEV	https://inspire.ec.europa.eu/codelist/AdministrativeHierarchyLevel/4thOrder
NATLEVNAME	Municipio
NATCODE	34,033,333,003
NAMEUNIT	Amieva
CODNUT1	ES1
CODNUT2	ES12
CODNUT3	ES120
IDmuni	33,003
GEOMETRY	'Polygon (((−0.43640375 38.68807148, −0.43605047 38.68830699, −0.43525394 38.68872198...)))'

The data are organized into two separate Shapefiles with geometries based on the interpretation of legal titles included in the Central Register of Cartography. One Shapefile covers the Peninsula and the Balearic Islands, while the other is dedicated to the Canary Islands. The two Shapefiles are represented in two distinct datums: ETRS89 for the Peninsula and Balearic Islands, and REGCAN95 (National Geodetic Network by Canary Islands Spatial Techniques 1995) for the Canary Islands. The latter closely aligns with WGS84.

These geometries are linked to an attribute table (see an example in Table 3) that contains information, including the national code (NATCODE) with format CCAAPPMMMMM. CC refers to the country code, AA is the autonomous community code, PP is the province code, and MMMMM is the unique municipality code. For this study, we focused specifically on extracting the last 5 digits, which are related to the municipality code associated with each polygon or multipolygon (IDmuni).

3 | Geospatial Processing, Results Validation, and Publication

The methodology, outlined in this section, can be broken down into three key stages: (1) data preparation, requiring the integration of SNFIv3 plots and Spanish municipalities; (2) assignation of inventory plots to municipalities through two alternative approaches (GIS and GeoSPARQL); and (3)

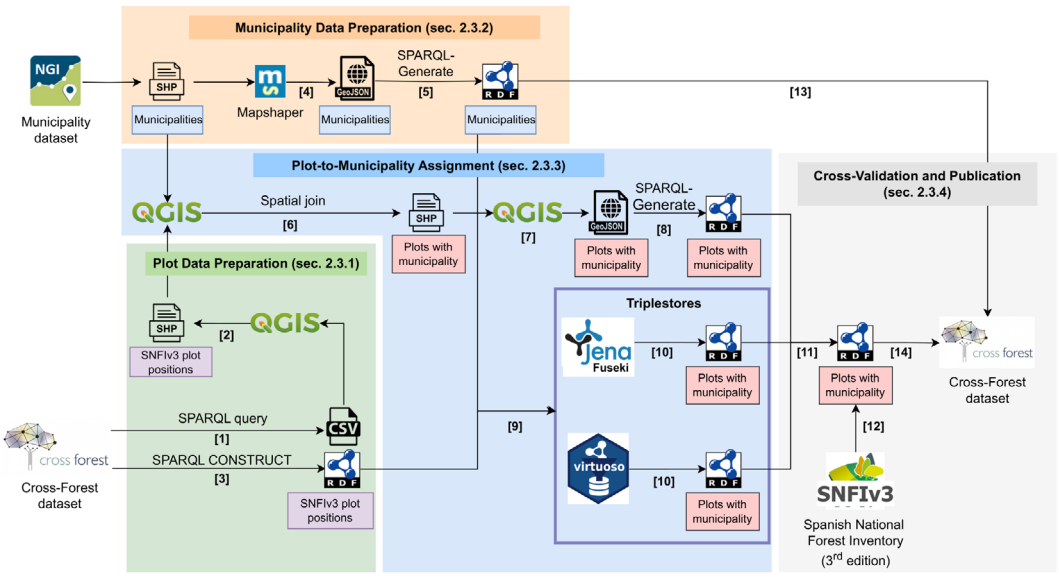

FIGURE 4 | Flowchart summarizing the methodology for the assignment of plots to their municipalities.

validation of the results and publication of the plot to municipalities assignment.

For clarity and organization in our study, we have devised the structured workflow illustrated in Figure 4. This workflow represents the methodology we used to integrate, validate and publish the combined plot and municipality information from SNFIv3. For the integration we used both a semantic solution and a GIS-based approach, which allowed us to evaluate the advantages of each alternative, and to perform a cross-validation.

Initially, we prepared the plot and municipality data for both semantic and GIS-based processing. For this, we define the workflow blocks "Plot Data Preparation" (see Section 3.1) and "Municipality Data Preparation" (see Section 3.2), followed by "Plot-to-Municipality Assignment" (see Section 3.3) comparing the two proposed methods. In the semantic approach, we use two different triplestores due to the irregular support of GeoSPARQL (see Section 2.1). In the "Cross-validation and Publication" phase (see Section 3.4), we validated the assignment generated in this work and compared it with the official municipality information for the SNFIv3 plots published by the Spanish Government. We also published the corrected assignment in the Cross-Forest dataset.

3.1 | Plot Data Preparation

Determining the municipality for each of the 99,045 SNFIv3 plots requires their central coordinates. Originally spread across 50 Microsoft Access files (see Section 2.2.1). These data are now retrieved from the Cross-Forest dataset to avoid extensive preprocessing. The Cross-Forest dataset offers plot locations in both the original ED50 datum and the more widely used WGS84 format, facilitating integration with municipal boundaries. We opted for WGS84 to ensure a standardized framework for analysis.

The Cross-Forest dataset uses several ontologies to describe its data and relationships. These queries use the Position and EPSG ontologies, which define coordinate reference systems and their corresponding EPSG data.

In order to acquire plot locations (latitude and longitude in WGS84) from the Cross-Forest dataset, we employed this SPARQL query:

```
SELECT DISTINCT ?plot ?lat ?lng WHERE {
  ?plot rdf:type nfi:Plot;
    spo:hasPosition ?spo.
  ?spo spo:hasCoordinateReferenceSystem crs
     :4326;
    axis:106 ?lat;
    axis:107 ?lng.
}
```

In this query, we retrieve the latitude and longitude of the plot centers, expressed in the WGS84 coordinate reference system (EPSG:4326). The geodetic latitude and longitude in the Cross-Forest dataset are represented by axes 106 and 107, respectively. These axes correspond to the standardized definitions found in the EPSG registry: axis 106 for latitude (https://epsg.io/106-axis) and axis 107 for longitude (https://epsg.io/107-axis).

The above query allowed us to download the information for all Spanish plots in various formats:

- For the GIS assignment (QGIS version 3.28.6. (QGIS Development Tesam, 2023)): we downloaded the data in CSV format (Figure 4 [1]) and subsequently transformed it into a point-geometry Shapefile (Figure 4 [2]).
- For the GeoSPARQL assignment: we downloaded the data directly in RDF format (on which the Linked Open Data are based) (Figure 4 [3]).

3.2 | Municipality Data Preparation

Spanish municipality data were obtained from the National Geographic Institute in two Shapefiles: one for the Iberian Peninsula and the Balearic Islands (datum ETRS89), and one for the Canary Islands (datum REGCAN95). To ensure compatibility with plot locations (WGS84), we merged these files into a single WGS84 Shapefile using a GIS.

- For the GIS assignment: due to its compatibility with GIS software, the data in Shapefile format can be used without requiring additional transformations.
- For the GeoSPARQL assignment: we transformed the merged municipality Shapefile into Linked Open Data (LOD) format. For this, first we used Mapshaper (Bloch 2023), a simple, free and open-source software tool for working with geospatial data. Mapshaper efficiently transformed the merged Shapefile of all Spanish municipalities into GeoJSON (Figure 4 [4]). This step is necessary because there are currently no known tools for direct Shapefile to RDF conversion. And second, a SPARQL-Generate query (Lefrançois, Zimmermann, and Bakerally 2017) was used to transform the GeoJSON data into RDF format, including municipality geometries and attributes (codes, names, etc.) (Figure 4 [5]). A custom shell script developed for the Cross-Forest project[1] automated this SPARQL-Generate execution.

3.3 | Plot-To-Municipality Assignment

We compared two approaches for assigning plots to their corresponding municipalities: a Geographic Information System (GIS) approach and a Linked Open Data (LOD)-based GeoSPARQL approach.

- For the GIS assignment: we used two Shapefiles, plot locations from the Cross-Forest dataset (see Section 3.1) and the Shapefile containing the geometries of all Spanish municipalities along with their attributes (see Section 3.2).

Within the QGIS environment, the "Join attributes by location" tool (Figure 4 [6]) with the "Within" function efficiently assigned each plot (point) to its containing municipality (polygon).

This generated a single Shapefile with plot-to-municipality assignments for the plots. To integrate this information with the existing inventory data, we transformed the results into RDF and incorporated them into the Cross-Forest dataset. This involved converting the Shapefile to GeoJSON (Figure 4 [7]) and then to RDF (Figure 4 [8]), similar to the municipality data conversion (Section 3.2).

- For the GeoSPARQL assignment, we set up two test environments using open-source GeoSPARQL-compatible triplestores: Virtuoso (version 07.20.3236) and Fuseki (version 4.7.0). These environments contained minimal data: SNFIv3 plot positions and municipality data in RDF format (Figure 4 [9]). We chose these two triplestores because GeoSPARQL implementations can vary. Virtuoso is the triplestore employed by the Cross-Forest dataset, and Fuseki offers an implementation more closely aligned with the OGC GeoSPARQL standard than others (Jovanovik, Homburg, and Spasić 2021).

To assign plots to municipalities, we used a GeoSPARQL query in our testing environments (Figure 4 [10]). This query is customized to utilize the specific geospatial functions of each triplestores.

For Virtuoso, we employed the `st_within` spatial predicate to check if one geometry is entirely within another. It is important to note that Virtuoso uses its own extension of GeoSPARQL (prefixed with `bif`), so adjustments are needed if the same query is used in other triplestores like Fuseki:

```
SELECT ?plot ?muni WHERE {
  ?plot rdf:type nfi:Plot;
    spo:hasPosition ?spo.
  ?spo spo:hasCoordinateReferenceSystem crs
     :4326;
    axis:106 ?lat;
    axis:107 ?lng.
  ?muni spo:hasPolygon ?poly.
  ?poly wkt:asWKT ?wkt_muni.
  BIND(spatialF:convertLatLon(?lat, ?lng) as
       ?point)
  FILTER(geof:sfWithin(?point, ?wkt_muni))
}
```

The GeoSPARQL query above is designed to automatically assign points (`?plot`) to polygons (`?wkt_muni`) using the GeoSPARQL functions implemented in Fuseki. The query specifies that (`?plot`) is of type (`nfi:Plot`) with a position (`?spo`) in WGS84 datum (`crs:4326`). It extracts latitude (`?lat`) and longitude (`?lng`) from the position. Additionally, it defines that (`?muni`) has a polygon (`?poly`) and extracts the Well Known Text (WKT) vector geometry of the polygon (`?wkt_muni`). In Fuseki, the `spatialF:convertLatLon` function, a proprietary extension, converts latitude and longitude into a point in (`?point`). The `geof:sfWithin`[2] function, defined by the OGC, is then used to determine if that point falls within the polygon of the municipality, similar to the `st_within` function in Virtuoso.

The key differences between the two triplestores are

- Fuseki uses GeoSPARQL functions like geof:sfWithin, while Virtuoso uses its own functions such as bif:st_within.
- Fuseki converts latitude and longitude into a geospatial point with spatialF:convertLatLon, whereas Virtuoso creates the point with bif:st_point.

These variations are minor but important for ensuring compatibility in geospatial operations across triplestores.

Through this assignment, we generated RDF triples that associate each plot's unique identifier with the identifier of the municipality it belongs to.

3.4 | Cross-Validation and Publication

Following the plot-to-municipality assignment using both GIS and GeoSPARQL (implemented in Virtuoso and Fuseki) approaches (Figure 4 [11]), we performed a cross-validation of the results (Figure 4 [12]). Additionally, we compared the validated assignments with municipality codes associated with plot locations in the original SNFIv3 data (refer to Section 2.2.1 for details). This comparison focused on two specific tables containing information collected before and after fieldwork.

The validated plot-municipality associations were then transformed into RDF triples and integrated into the Cross-Forest dataset (Figure 4 [13, 14]). This dataset is accessible through a SPARQL endpoint, allowing for analysis, querying, and data download in various formats (RDF, CSV, JSON, GeoJSON ...). This enables filtering or aggregating inventory data by municipality based on plot and municipality identifiers. For transparency and reproducibility, all files used and generated during this project, along with the developed scripts, are openly available on a GitHub repository.

4 | Description of Case Studies for Analysis

Integrating SNFIv3 plot data with municipal information opens doors to calculating various metrics previously difficult to obtain at the local level across Spain. The following sections explore this potential through two case studies. The first case study uses plot-specific information associated with municipalities within a National Park. The second case study utilizes municipality data to perform aggregations, extracting inventory plot information based on counties.

4.1 | Picos de Europa National Park Dominant Species

Picos de Europa National Park encompasses a total of 11 municipalities within its Socioeconomic Impact Area (SIA) in three different provinces (León, Asturias and Cantabria), covering a surface area of 133,845 ha. The SIA refers to an association of

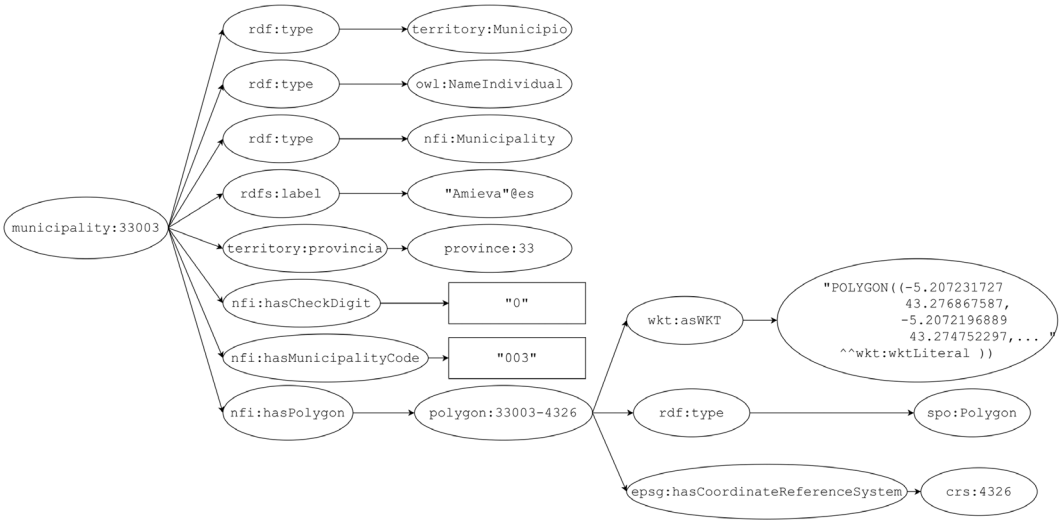

FIGURE 5 | Subject, predicate and object of Amieva municipality properties and geometry in LOD format (only a couple of the vertices forming the polygon are shown).

municipalities that contribute land to a National Park, where public administration will implement active policies for its development (BOE 2014). We aimed to identify the dominant tree species in each municipality within the park. To achieve this, we analyzed the mean basal area (G) expressed in m^2/ha per species in forest inventory plots. Mean basal area is a common metric for evaluating forest composition, reflecting the total tree stem area per unit area (Riofrío, del Río, and Bravo 2016; Hedwall et al. 2019; Canedoli et al. 2020).

With the SNFIv3 inventory data integrated into a single dataset (Cross-Forest) and linked to municipal information, we could efficiently extract all relevant forest inventory data for the municipalities within the Picos de Europa National Park. Basal area serves as an example, but the potential for analysis extends to many other variables within this dataset, such as volume, diameter class, or the number of trees per species.

4.2 | Diversity in Selected Spanish Counties

Operational forest management needs sound data at plot level, but forest tactical and strategic planning is conducted at forest/ownership level and county/landscape level. Counties in Spain are administrative organizations defined following landscape and cultural criteria and, are demarcated by municipal boundaries. Our main result (an automatic procedure to transform the geometries of Spanish municipalities from the geospatial database of the National Geographic Institute to LOD format) allows us to assign plot to municipalities and by aggregation to define counties, and finally to obtain forest plot database by county.

We tested the capability of our solution to address this problem by defining a set of representative counties in Spain (17) covering a wide range of ecological and forestry situations. Once we have selected the counties, we extracted the plots that belong to the municipalities that integrate them (using the information integrated in this work in LOD format). Carrying out SPARQL queries based on basic calculations of SNFI data, we extracted the plots in each county with one main species occupying > 95% of basal area (G) (pure stands), plots with two main species whose sum occupies 95% of G (mixed forest—low diversity plots), and plots with 3 or more species occupying 95% of G (mixed forest—high diversity plots).

5 | Results

5.1 | Municipality Data Processing

After converting municipality data to LOD, we obtained a file with 89,441 triples, 11 per municipality. Figure 5 displays the 11 triples generated of an example municipality located within the Picos de Europa National Park (Amieva).

For defining the IRI that identify the municipality of Amieva we use its municipality code (like an ID number, in this case 33,003). We attach additional information to this entity, such as its geometry corresponding to `polygon:33003-4326`. The IRI of this polygon is formed by combining the municipality code and the WGS84 datum code (defined by the European Petroleum Survey Group (EPSG:4326)) (European Petroleum Survey Group 2020) of the polygon's coordinates. This polygon represents the surface area of the Amieva municipality (11,498 ha) and is further associated with the 769 vertices forming its geometry in WKT format (see Figure 5).

5.2 | Assignment of Plots to Their Municipalities Using Different Procedures

Throughout the automated assignment process, the 99,045 SNFIv3 plots were associated with their respective municipality codes through three separate procedures, obtaining three sets of outcomes.

The results obtained by Fuseki and QGIS in this assignment were the same, as follows:

- 99,026 (99.98%) plots were assigned to a single municipality.
- 19 (0.02%) plots were not assigned to any municipality.

The results obtained using Virtuoso for the assignment were incorrect in the majority of cases. Although all plots were assigned to at least one municipality, 73% of the plots were associated with more than one municipality, resulting in false positives:

- 26,591 (26.85%) plots were associated with 1 municipality.
- 46,668 (47.12%) plots were associated with 2 municipalities.
- 20,869 (21.07%) plots were associated with 3 municipalities.
- 4284 (4.33%) plots were associated with 4 municipalities.
- 560 (0.57%) plots were associated with 5 municipalities.
- 71 (0.06%) plots were associated with 6 municipalities.
- 2 plots were associated with 7 municipalities.

In Figure 6, we present a specific case where Virtuoso assigns the `plot:33-1518-N-N`, located in Asturias province within the Picos de Europa National Park SIA, to multiple municipalities near the position of the plot. The accompanying map highlights the plot's primary location within the municipality with code 33003 (Amieva). However, the GeoSPARQL query executed by Virtuoso has erroneously assigned it to three neighboring municipality codes: 33012 (Cangas de Ons), 33,045 (Parres) and 33,050 (Ponga). We have opened an issue regarding this problem at the Virtuoso GitHub repository available at https://github.com/openlink/virtuoso-opensource/issues/1098, however, the company has not yet proposed a fix at the time of publication.

In contrast, Fuseki and QGIS provides the same assignment to `plot:33- 1518-N-N`, corresponding to `municipality:33003` (Amieva). Visual inspection (Figure 6) is consistent with the proposed assignment.

5.3 | Cross-Validation of the Assignments

After the assignment of plots to their municipalities using the three procedures (SPARQL queries in Virtuoso and Fuseki, and spatial join in QGIS), the results obtained are compared in Table 4. All three procedures were executed on the same system with the following specifications: an Intel(R) Core(TM) i7-1165G7 @2.80GHz, 16.0 GB RAM, 512 GB SSD storage, and the Windows 11 Pro (x64) operating system. Notably, the assignment process took significantly longer in Fuseki (almost 2 h) compared to Virtuoso or QGIS (less than 3 min).

The assignments carried out using Fuseki and QGIS were entirely identical. Due to the issue discussed in Section 5.2, Virtuoso makes a 73% of incorrect assignment of plots to municipalities. To further validate these conclusions, we performed a visual inspection of 30 randomly chosen plots assigned with the three procedures. Besides, the perfect match between Fuseki and QGIS results, achieved through distinct approaches (GIS and GeoSPARQL functions), strengthens our confidence in their reliability.

Furthermore, we identified 19 plots assigned by Virtuoso to municipalities, but not by Fuseki or QGIS. Investigating these discrepancies, we discovered that the central coordinates of these plots fall outside any Spanish municipality's boundaries. Some plots are located at sea, while others lie in France. Since these plots are not within the jurisdiction of any Spanish municipality, they cannot be assigned a valid municipality code.

Comparing the correct assignments with pre-existing SNFIv3 municipality information, we found that 56% of pre-fieldwork assignments were incorrect. However, for plots assigned during fieldwork (even though only 47% of plots had a municipality code assigned in the field inventory), the error rate dropped significantly to just 12%.

5.4 | Applications of Integrated Plot and Municipal Data

Applications The successful allocation of plots to their respective municipalities has greatly facilitated the development of a comprehensive map showcasing the dominant species within each municipality. This map takes into account the basal area and encompasses the entire country's forested areas (Figure 7). Through this map, a holistic understanding of Spain's forest ecosystems is presented, highlighting the primary species within distinct regions of the Peninsula. The map also illustrates species biodiversity at the national level, showing the dominance of 59 different species across Spanish municipalities. It is important to acknowledge that not all municipalities contain forested areas. Out of Spain's 8131 municipalities, only 6056 have data from forest inventory plots (SNFIv3) that provide detailed tree information.

To obtain data on the species with the highest basal area (G) per municipality, the following SPARQL query was executed on the integrated dataset. The first part of the query calculates the average basal area per hectare for each species within each municipality, while the second part identifies the maximum basal area per municipality. By combining the results of these two sub-queries, the species with the highest average basal area in each municipality is identified.

```
SELECT ?muni ?species ?maxG WHERE {
  {
    SELECT ?muni (MAX(?meanG) AS ?maxG)
WHERE
       {
     SELECT ?muni ?species (AVG(?G) AS?
        meanG) WHERE {
       ?plot a nfi:Plot;
          nfi:containsSpeciesPlot?
            infoSpeciesPlot;
          nfi:isInMunicipality ?muni.
       ?infoSpeciesPlot nfi:
          hasBasalAreaInM2byHA ?G;
          nfi:hasSpecies ?species.
       }
     GROUP BY ?muni ?species
  }
    GROUP BY ?muni
}
```

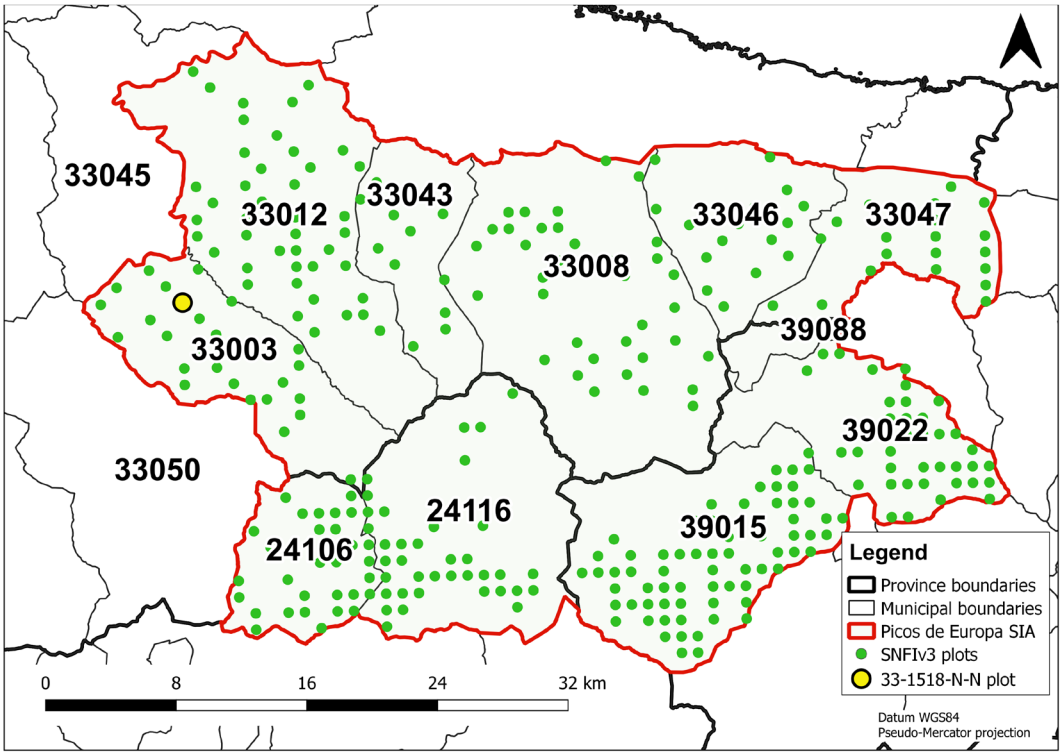

FIGURE 6 | SNFIv3 plots and municipalities inside Picos de Europa National park. `plot:33-1518-N-N` is highlighted in yellow. QGIS and Fuseki correctly assigns this plot to municipality 33,003. In contrast, Virtuoso assigns it to municipalities 33,033, 33,012, 33,045, and 33,050.

TABLE 4 | This table compares the effectiveness of different procedures (Virtuoso, Fuseki, and QGIS) in assigning SNFIv3 forest inventory plots to their corresponding municipalities within Spain.

	Virtuoso	Fuseki	QGIS
Plots assigned	99,045 (100.00%)	99,026 (99.98%)	99,026 (99.98%)
Correct assignments	26,591 (26.85%)	99,026 (99.98%)	99,026 (99.98%)
Incorrect assignments	72,454 (73.15%)	0	0
Plots not assigned	0	19 (0.02%)	19 (0.02%)

```
  # Join the maximum back to retrieve spe-
cies
  {
    SELECT ?muni ?species (AVG(?G) AS
?meanG)
       WHERE {
       ?plot a nfi:Plot;
           nfi:containsSpeciesPlot?
             infoSpeciesPlot;
           nfi:isInMunicipality ?muni.
       ?infoSpeciesPlot nfi:
           hasBasalAreaInM2byHA ?G;
           nfi:hasSpecies ?species.
       }
     GROUP BY ?muni ?species
  }
  FILTER(?meanG = ?maxG)
}
ORDER BY ?muni
```

Our analysis identified the top five tree species dominating a significant portion of Spanish municipalities: the holm oak (*Quercus ilex* L.), which dominates in 959 municipalities, followed by the aleppo pine (*Pinus halepensis* Mill.) (954 municipalities), the maritime pine (*Pinus pinaster* Aiton.) (721), the scots pine (*Pinus sylvestris* L.) (500), and the black pine (*Pinus nigra* Arnold) (371). Collectively, these five species reign supreme in over 57.88% of Spanish municipalities with forest inventory data.

Although detailed municipal-level data are not feasible at this map scale, it highlights the prevalence of certain species. For instance, along the Mediterranean coast, *Pinus halepensis* thrives, while the interior Mediterranean zone is characterized by *Quercus ilex*, whereas *Pinus pinaster* and *Eucalyptus spp.* L'Hér. dominates the northern and northwestern areas. Furthermore, the map reveals the peninsula mountainous areas characterized by the dominance of *Pinus sylvestris* or *Fagus sylvatica* L.

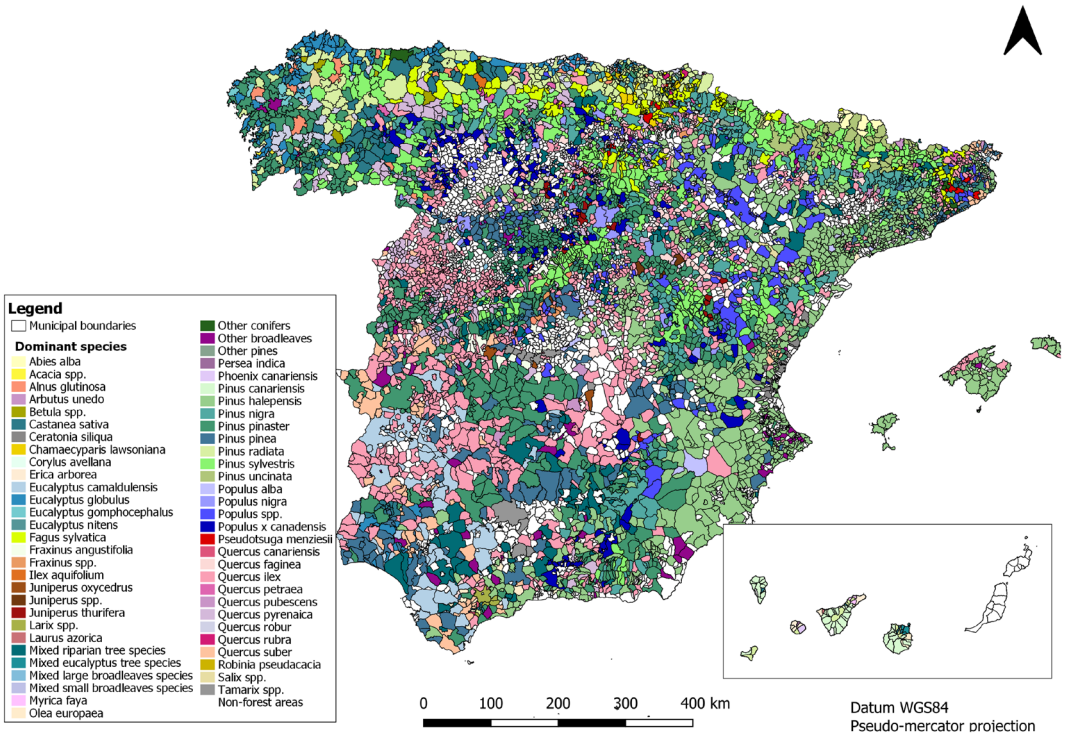

FIGURE 7 | Dominant species per Spanish municipality map based on basal area (m^2/ha).

5.4.1 | Picos de Europa National Park Dominant Species

To continue along the lines of our case study, Figure 8 shows the species dominance in the 10 municipalities of the Picos de Europa National Park illustrated through a map based on their mean G (m^2/ha) per each species.

The most dominant species in Picos de Europa municipalities is the chestnut (*Castanea sativa*) that, according to the G mean, is predominant in four of the 10 municipalities studied. The second more dominant is the European beech (*Fagus sylvatica*), followed by radiata pine (*Pinus radiata* D.Don) plantations and the holm oak (*Quercus ilex*).

To analyze the number of plots dominated by a single species (pure plots) or by multiple species (mixed plots) within the Picos de Europa municipalities, specific SPARQL queries with adapted filter conditions were developed. To identify plots where a single species dominates with 70% or more of the basal area (G), the query utilized the following filter expression: FILTER((100 *?maxSpBasalArea / ?totalBasalArea)>=70). This condition ensures that only those plots where the maximum basal area of a single species reaches or exceeds 70% of the total basal area are classified as pure plots. Conversely, to detect mixed plots, where no single species occupies 70% or more of the basal area, the query was modified to include the filter: FILTER(?maxSpBasalArea < (0.7 * ?totalBasalArea)). This adjustment extracts plots where the maximum basal area of any species is less than 70%, indicating that the plots have multiple species with more balanced distributions of basal area. Users can modify this percentage threshold to suit their specific needs.

Below is an example query that calculates the percentage of pure plots per municipality in Picos de Europa National Park.

```
SELECT? muni
  ?totalPlots
  ?numPurePlots
  (?numPurePlots * 100 / ?totalPlots AS
    ?percPurePlots)
WHERE {
  VALUES ?muni {
    municipality:24106
    municipality:24116
    municipality:33003
    municipality:33008
    municipality:33012
    municipality:33043
    municipality:33046
    municipality:33047
    municipality:39015
    municipality:39022
    municipality:39088
  }
  # Total number of plots in the municipality
  {
    SELECT ?muni (COUNT(DISTINCT ?plot) AS?
       totalPlots) WHERE {
      ?plot a nfi:Plot;
        nfi:containsSpeciesPlot?
          infoSpeciesPlot;
        nfi:isInMunicipality ?muni.
      ?infoSpeciesPlot nfi:
          hasBasalAreaInM2byHA ?G.
    }
    GROUP BY ?muni
  }
  # Pure plots with species >=70% basal area
  {
```

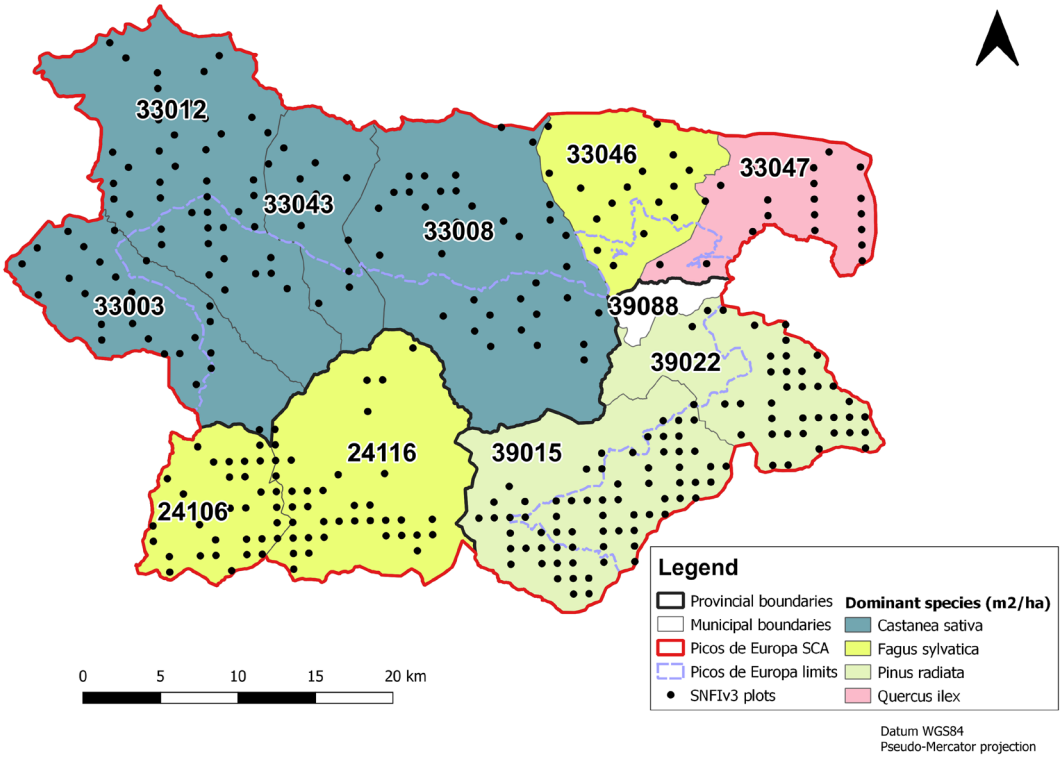

FIGURE 8 | Map showing dominant species by municipality based on basal area (m^2/ha) and SNFIv3 plots in Picos de Europa National Park SIA (Socioeconomic Impact Area).

```
  SELECT ?muni (COUNT(DISTINCT ?plot) AS?
     numPurePlots) WHERE {
    ?plot a nfi:Plot;
     nfi:containsSpeciesPlot?
       infoSpeciesPlot;
     nfi:isInMunicipality ?muni.
    ?infoSpeciesPlot nfi:
       hasBasalAreaInM2byHA ?G.
    # Calculate the total basal area and
       species basal area
  {
    SELECT ?plot (SUM(?spBasalArea) AS?
         totalBasalArea)
           (MAX(?spBasalArea) AS?
            maxSpBasalArea) WHERE {
       ?plot a nfi:Plot;
         nfi:containsSpeciesPlot?
           infoSpeciesPlot.
         ?infoSpeciesPlot nfi:
           hasBasalAreaInM2byHA?
           spBasalArea;
         nfi:hasSpecies ?species.
      }
      GROUP BY ?plot
    }
    # Filter for pure plots (>=70% of basal
      area dominated by one species)
    FILTER((100 * ?maxSpBasalArea/?
      totalBasalArea) >=70)
  }
  GROUP BY ?muni
}
}
ORDER BY ?muni
```

Pure stands are more prevalent in municipalities dominated by radiata pine plantations (a non-native species) and European beech (which often excludes other species due to ecological factors). Within the mixed plots in the Picos de Europa Socioeconomic Impact Area (SIA), 36 plots exhibit dominance by two species, 23 plots with three dominant species, 10 plots with four dominant species, and one plot with five dominant species. This highlights a variation in species composition within the mixed plots, ranging from moderately diverse to highly diverse ecosystems.

Table 5 provides an example of the information that can be obtained for each municipality within the Picos de Europa National Park. This table includes information such as municipal area, the number of SNFIv3 plots within the Socioeconomic Impact Area (SIA) and the park's strict boundaries, and the three dominant tree species based on their basal area. Additionally, we could analyze the proportion of pure and mixed stands within each municipality. These integrated data allow for comprehensive analysis of forest species biodiversity. However, the potential applications extend far beyond this example. It is also possible to calculate biomass per municipality, accumulated CO_2, or compare tree growth at the local level, providing valuable insights for forest management across Spain.

5.4.2 | Diversity in Selected Spanish Counties

To assess the number of pure and mixed plots with different levels of species diversity, we adapted the queries described in

Section 5.4. To identify pure plots, we applied a filter that required a single species to account for more than 95% of the basal area. For plots classified as high or low diversity (mixed plots), we adjusted the filter values to capture plots where species represented between 95% and 5% of the basal area. After extracting the percent basal area of each species in these mixed plots, we counted the number of species present per plot. Plots with more than two species were classified as high diversity, while those with two species were classified as low diversity.

Data aggregated at the municipality level reveal variations in forest diversity across 17 Spanish counties (Figure 9). La Garrotxa, in Catalonia's Girona province, stands out for its prevalence of highly diverse mixed forests (containing three or more species). Conversely, O Carballiño (Galicia) boasts the highest proportion of mixed forests with low diversity (two dominant species).

Several counties exhibit a dominance of pure forests. Sayago (Zamora), Noroeste Murciano (Murcia), and Villuercas-Ibores-Jara (Extremadura) all have a significant percentage of pure forest stands. Notably, Sayago and Villuercas-Ibores-Jara are dominated by pure *Quercus ilex* forests, while Noroeste Murciano is characterized by pure *Pinus halepensis* stands.

6 | Discussion and Conclusions

This study presents a novel approach for analyzing national forest inventory information at the local level. By leveraging Linked Open Data (LOD), researchers and stakeholders can rapidly acquire specific data, like diversity information, from the third edition of the Spanish National Forest Inventory (SNFIv3) using simple queries. This approach opens doors for various applications, including comprehensive biodiversity

TABLE 5 | Details for municipalities within the Picos de Europa National Park SIA, extracted from the integrated dataset.

Municipality	Surface area (ha)	# of SNFI plots inside SIA	# of SNFI plots inside park	Three most dominant species	Mean basal area (%)	# of pure plots	# of mixed plots
Oseja de Sajambre (24106)	7178	28	28	*Fagus sylvatica* *Castanea sativa* *Quercus petraea*	43.37 21.15 14.51	22 (79%)	6 (21%)
Posada de Valdeón (24116)	16,406	32	32	*Fagus sylvatica* *Quercus pyrenaica* *Quercus petraea*	43.24 22.68 15.58	27 (84%)	5 (16%)
Amieva-running example-(33003)	11,498	23	5	*Castanea sativa* *Quercus pyrenaica* *Fagus sylvatica*	21.29 15.49 13.49	14 (61%)	9 (39%)
Cabrales (33008)	23,883	23	14	*Castanea sativa* *Quercus ilex* *Fagus sylvatica*	35.84 21.02 15.11	13 (57%)	10 (43%)
Cangas de onís (33012)	21,324	25	15	*Castanea sativa* *Fagus sylvatica* *Betula spp.*	15.38 13.81 10.25	15 (60%)	10 (40%)
Onís (33043)	7488	7	2	*Castanea sativa* *Fagus sylvatica* *Ilex aquifolium*	20.49 16.49 15.71	3 (43%)	4 (57%)
Peñamellera baja (33046)	9219	10	3	*Fagus sylvatica* *Quercus robur* *Corylus avellana*	38.71 19.61 10.13	6 (60%)	4 (40%)
Peñamellera alta (33047)	8406	10	2	*Quercus ilex* *Other conifers* *Castanea sativa*	22.17 12.77 9.88	6 (60%)	4 (40%)
Comaleño (39015)	16,189	51	27	*Pinus radiata* *Quercus petraea* *Fagus sylvatica*	25.43 15.31 14.61	39 (76%)	12 (24%)
Cillorigo de Liébana (39022)	10,658	21	2	*Pinus radiata* *Fagus sylvatica* *Quercus pyrenaica*	24.18 18.53 17.29	15 (71%)	6 (29%)
Tesviso (39088)	1596	0		—	—	—	—
Total	**133,845**	**230**	**130**			**160**	**70**

index calculations at the local level, similar to the work done by Alberdi, Cañellas, and Condés (2014) in the province of Ávila, or analysis of forest complexity (Bravo et al. 2021) knowing the mixture degree of forest of a given area. Mixed forests are considered as one of the key tools to cope with climate change and new knowledge is under development on temperate forests (Osei et al. 2021).

Our approach allows easy scalability of spatial analysis from the national level down to counties. While no previous studies in Spain have used National Forest Inventory data at the county level, our work not only underscores data availability but also demonstrates its accessibility for accurate local-scale filtering. This is significant because forest management decisions are often made at forest/ownership and county/landscape levels.

Some forestry research relies on municipalities as a unit of analysis, such as Gil et al. (2011), who used GIS to assess carbon storage in municipalities in Castilla y León, or Martínez, Vega-Garcia, and Chuvieco (2009) who studied the causality of forest fires at the municipal level, among others (Rescia et al. 2008; Ameztegui et al. 2021; Sánchez-García et al. 2015; Marey-Pérez and Rodríguez-Vicente 2009). However, the lack of integrated data from local administrative units and forest databases often forces researchers to manually merge separate geospatial datasets into local GIS software, leading to time consumption and potential errors. This study highlights the value of LOD principles for forest data integration. LOD promotes standardized data formats that facilitate seamless information exchange, complementing existing GIS functionalities.

In this study, we compared two approaches to assign the SNFIv3 plots to their municipalities: LOD-based GeoSPARQL and GIS. We leveraged LOD (in RDF format) to integrate plot and municipality data, enabling us to query it using GeoSPARQL. Besides, the Cross-Forest LOD dataset allowed to easily prepare a Shapefile with all plots for GIS processing. Our evaluation revealed that GIS is better suited for locating the coordinates of the plot centers within municipal boundaries. GIS offers fast and accurate results due to its robust spatial analysis capabilities. When using GeoSPARQL, we found problems with specific triplestore implementations: Fuseki performs spatial operations correctly, but it is quite slow; Virtuoso is fast, but obtained results are unreliable due to false positives.

Previous research by Jovanovik, Homburg, and Spasić (2021) has shown that GeoSPARQL support varies considerably across triplestores. Fuseki, for example, provides complete support for GML and WKT formats, encompassing all GeoSPARQL extensions defined by the OGC. Conversely, Jovanovik, Homburg, and Spasić (2021) reported errors in Virtuoso's geospatial queries, primarily due to its internal processing of WKT literals. In our study, we specifically identified issues with Virtuoso's implementation of the `geo:sfWithin` function (implemented as `bif:st_within`), as it incorrectly assigned containment between geometries (polygon containing a point). Although we checked that Virtuoso at least included the correct assignment, we consider the output of Virtuoso unusable due to false positives. Additionally, other Virtuoso users have reported similar issues with GeoSPARQL functions, as evidenced by the 26 open issues currently associated with Virtuoso's GeoSPARQL

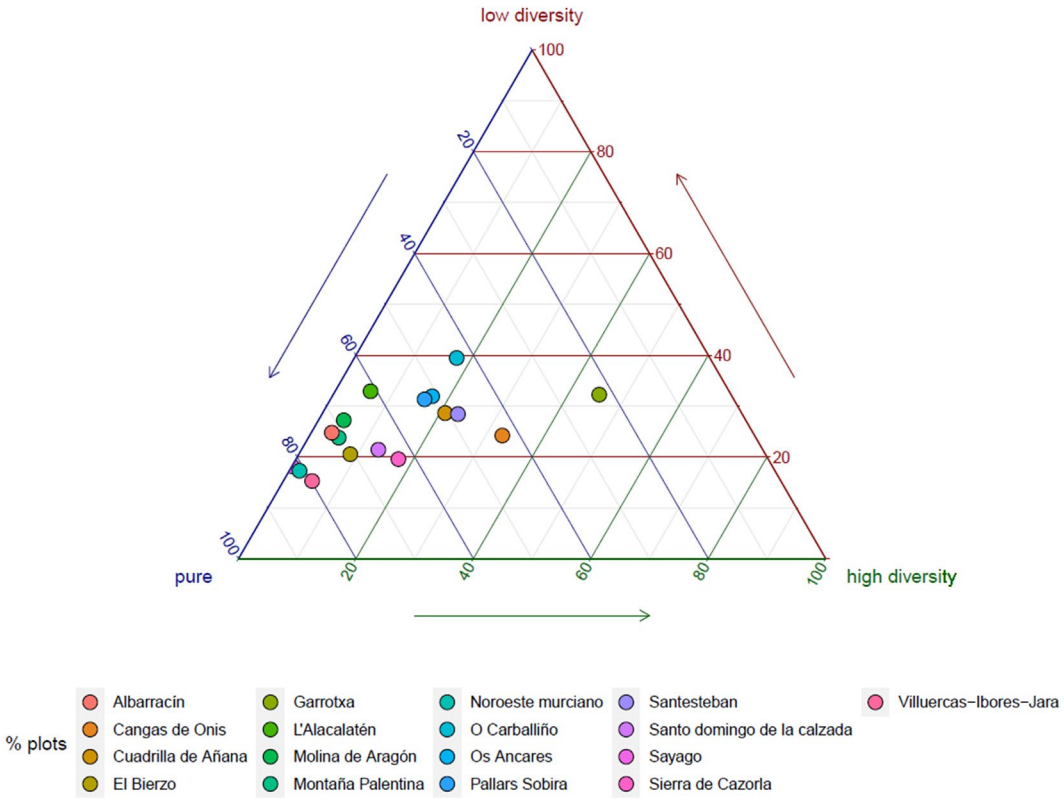

FIGURE 9 | Ternary diagrams (ggtern package Hamilton and Ferry (2018)) representing the percentage of species mixture in SNFI plots of 17 Spanish counties in terms of basal area (m^2/ha).

implementation. Notably, we could not find existing benchmarks that specifically evaluated the functionality or correctness of Virtuoso's `geo:sfWithin`function.

While GIS excels in spatial assignment tasks, dataset integration (100 SNFI relational databases and municipal Shapefiles) benefits significantly from Semantic Web technologies. These technologies have widely demonstrated their effectiveness in overcoming the structural complexities and diverse formats of data sources, enabling a coherent and meaningful connection between heterogeneous sets of information (Pasquier 2008; Lan et al. 2022; Auer 2022). GIS tools manage relational databases with geospatial relationships. However, relational data are quite inflexible: changes in its schema, due to its evolution or to integration with other data, require changes in the structure of tables. These changes are usually non-trivial and can lead to problems if not correctly normalized or if inconsistencies are introduced. This inflexibility, combined with the already complex and inconsistent schema of the multiple database files of the SNFI, makes the integration of the SNFI and municipalities challenging.

LOD uses a graph model with a self-descriptive schema that is easy to evolve and trivial to combine, although graph data can be less human-readable than tabular data and it requires expertise on ontologies; further, there are scarce initiatives on the use of LOD in the forestry domain. Fortunately, we were able to reuse our previous work[34] (Portolés et al. 2021) for converting SNFI into LOD. Thus, we only needed to convert the municipality Shapefiles into LOD, requiring a modest effort. In using Semantic technologies for forest data integration, we discovered significant advantages in terms of machine-readable data and models.

These technologies facilitate automatic understanding of meanings and enhanced reasoning capabilities. This stands in contrast to black-box models, providing improved interpretability, as highlighted by Lan et al. (2022). Additionally, studies such as those conducted by Harbelot, Arenas, and Cruz (2013), Nikolaou et al. (2015), Bernard, Villanova-Oliver, and Gensel (2022), Charles, Aussenac-Gilles, and Hernandez (2023), and Wu, Wang, and Uz Tansel (2024) underscore the challenges and advantages of semantic technologies for integrating historical data, including evolving territorial boundaries such as municipalities. Unlike traditional GIS, semantic technologies excel at modeling and analyzing temporal information through the use of ontologies, thus facilitating more accurate and flexible spatio-temporal data integration and reasoning.

In terms of data access, endpoints equipped with standard Semantic Web query languages act as web services that process queries without local data storage. This approach offers several advantages, including the ability to respond to SPARQL queries, integrate HTTP documents for data interaction, and support flexible querying via HTTP document URLs. It also facilitates accessibility for other user agents or services that encourage the creation of mashups and generates query schema files in various formats like HTML, JSON, CSV, RDF-Turtle, and RDF-N-Triples. These features make semantic technologies accessible and versatile for improved integration and utilization of forest data as also demonstrated by the Cross-Forest project.

While the obtained assignment is only applicable to the Spanish context, we do believe that the proposed procedure for LOD-based integration and plot assignment can be extrapolated to other settings, for example, harmonization of national forest inventories (Mauri, Strona, and San-Miguel-Ayanz 2017) or enrichment with new information, such as territorial units. By adhering to the principles of linking and semantic standardization, LOD provided an effective solution to overcome interoperability barriers, enabling the effective connection of heterogeneous datasets and their open publication.

One of the objectives of this work was to emphasize the versatility of integrated data when carrying out studies or management plans at the local level. The integration of local administrative units and forest inventory data improves public decision-making by providing more precise information that can be used to carry out forecasts and predictions that lead to better outcomes. Furthermore, creating and publishing this dataset as open data will enable its reuse and the development of new applications that can exploit it for various purposes. The automation process devised in our study for associating plots with municipalities significantly enhances the workflow's value. The algorithms we designed and implemented ease synergistic use of geospatial datasets, making it adaptable to various analogous use cases, thus emphasizing its potential for adaptability and scalability.

Adapting our methodology could benefit areas such as natural resource management, urban planning, and environmental monitoring where global scope and local precision are essential. For instance, Koubarakis (2023) discusses in his book how geospatial RDF datasets like the Corine Land Cover Map or the Leaf Area Index (LAI) can be used to extract valuable environmental data within administrative boundaries, as demonstrated in the case of the municipality of Athens.

Future research could explore the seamless integration of historical SNFI data (providing temporal ground-level insights into forest resources) with local administrative unit data and diverse datasets like bioclimatic information, remote sensing imagery, and LiDAR (Light Detection and Ranging) data. The open availability of all of this information in Spain lessens hurdles for such integration, paving the way for a comprehensive approach that can significantly enrich forest ecosystem analysis. By combining these datasets, researchers and stakeholders could develop more robust empirical forest dynamic models and tools to support data-driven forest management decisions.

Acknowledgments

We would like to thank the Spanish National Geographic Institute, the Ministry for Ecological Transition and Demographic Challenge (MITECO), and the TRAGSATEC group for providing the necessary resources and support to conduct this research. We extend our special thanks to the forest technicians who collected the field data for the forest inventories; without their invaluable efforts, this work would not have been possible.

This work has been partially funded by NextGenerationEU and MCIN/AEI through the LOD.For.Trees project (TED2021-130667B-I00), by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF) under the GenieLearn project (PID2023-146692OB-C32), by the Horizon-RIA programme of the European

Union through the Small4Good project (069/230331), and by the Junta de Castilla y León through the CLU-2019-01 and CL-EI-2021-05 initiatives, under the iuFOR Institute Unit of Excellence at the University of Valladolid, co-financed by the European Regional Development Fund (ERDF, "Europe drives our growth") and Ministry of Education of Council of Castilla y León (ORDEN EDU/1009/2024).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Cross-Forest at https://github.com/Cross-Forest.

Endnotes

[1] https://github.com/Cross-Forest/scripts/blob/master/sparql-generate.sh.

[2] https://jena.apache.org/documentation/geosparql/.

[3] https://crossforest.eu/results/data/.

[4] https://datos.gob.es/en/noticia/cross-forest-project-harmonisation-and-promotion-open-forest-data.

References

Alberdi, I., I. Cañellas, and S. Condés. 2014. "A Long-Scale Biodiversity Monitoring Methodology for Spanish National Forest Inventory. Application to Álava Region." *Instituto de Ciencias Forestales (ICIFOR)* 23: 93–110. https://doi.org/10.5424/fs/2014231-04238.

Allemang, D., and J. Hendler. 2011. *Rdf—The Basis of the Semantic Web*, 27–50. Burlington, MA: Elsevier. https://doi.org/10.1016/B978-0-12-385965-5.10003-2.

Ameztegui, A., A. Morán-Ordóñez, A. Márquez, et al. 2021. "Forest Expansion in Mountain Protected Areas: Trends and Consequences for the Landscape." *Landscape and Urban Planning* 216: 104240. https://doi.org/10.1016/j.landurbplan.2021.104240.

Arslan, M., J.-C. Desconnets, and I. Mougenot. 2022. "Environmental and Life Sciences Observations in Knowledge Graphs Using Nlp Techniques to Support Multidisciplinary Studies." *Procedia Computer Science* 201: 543–550. https://doi.org/10.1016/j.procs.2022.03.070.

Auer, S. 2022. "Semantic Integration and Interoperability." In *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*, 195–210. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-93975-512.

Bernard, C., M. Villanova-Oliver, and J. Gensel. 2022. "Theseus: A Framework for Managing Knowledge Graphs About Geographical Divisions and Their Evolution." *Transactions in GIS* 26: 3202–3224. https://doi.org/10.1111/tgis.12988.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. "The Semantic Web." *Scientific American* 284, no. 5: 34–43.

Bloch, M. 2023. *Mapshaper v. 0.6.41*. August 2023. https://github.com/mbloch/mapshaper.

BOE. 2007. *Real decreto 1071/2007, de 27 de julio, por el que se regula el sistema geodésico de referencia oficial en españa*. June 2023. https://www.boe.es/eli/es/rd/2007/07/27/1071.

BOE. 2014. *Ley 30/2014, de 3 de diciembre, de parques nacionales*. Madrid, España: Boletín Oficial del Estado. September 2023. https://www.boe.es/buscar/act.php?id=BOE-A-2014-12588.

Bravo, F., A. Martín Ariza, N. Dugarsuren, and C. Ordóñez. 2021. "Disentangling the Relationship Between Tree Biomass Yield and Tree Diversity in Mediterranean Mixed Forests." *Forests* 12, no. 7: 848. https://doi.org/10.3390/f12070848.

Bray, T., D. Hollander, A. Layman, R. Tobin, and H. Thompson. 2009. *Namespaces in XML 1.0 (Third Edition) (W3C Recommendation)*. Cambridge, MA: World Wide Web Consortium (W3C). http://www.w3.org/TR/xml-names.

Canedoli, C., C. Ferrè, D. Abu El Khair, et al. 2020. "Evaluation of Ecosystem Services in a Protected Mountain Area: Soil Organic Carbon Stock and Biodiversity in Alpine Forests and Grasslands." *Ecosystem Services* 44: 101135. https://doi.org/10.1016/j.ecoser.2020.101135.

Charles, W., N. Aussenac-Gilles, and N. Hernandez. 2023. "Hht: An Approach for Representing Temporally-Evolving Historical Territories." In *The Semantic Web. Eswc 2023. Lecture Notes in Computer Science*, edited by C. Pesquita, J. McCusker, E. Jimenez-Ruiz, et al., vol. 13870. Cham: Springer. https://doi.org/10.1007/978-3-031-33455-925.

Dürst, M., and M. Suignard. 2005. *Internationalized Resource Identifiers (IRIs) (Standards Track RFC 3987)*. Fremont, CA: Internet Engineering Task Force (IETF). September 2023. https://datatracker.ietf.org/doc/html/rfc3987.

European Petroleum Survey Group. 2020. *Database of Coordinate Reference Systems and Coordinate Transformations*. September 2023. https://epsg.io/4326.

Fierro García, B., A. Roldán Zamarrón, M. Lerner Cuzzi, et al. 2022. "Cross-forest, un proyecto de datos forestales abiertos de España y Portugal." In *Actas de la 8o congreso forestal español* (p. 13). Lérida, España.

Gil, M. V., D. Blanco, M. T. Carballo, and L. F. Calvo. 2011. "Carbon Stock Estimates for Forests in the Castilla y león Region, Spain. A Gis Based Method for Evaluating Spatial Distribution of Residual Biomass for Bio-Energy." *Biomass and Bioenergy* 35, no. 1: 243–252. https://doi.org/10.1016/j.biombioe.2010.08.004.

Guadilla-Sáez, S., M. Pardo-de Santayana, R.-G. Victoria, and J.-C. Svenning. 2019. "Biodiversity Conservation Effectiveness Provided by a Protection Status in Temperate Forest Commons of North Spain." *Forest Ecology and Management* 433: 656–666. https://doi.org/10.1016/j.foreco.2018.11.040.

Hamilton, N. E., and M. Ferry. 2018. "Ggtern: Ternary Diagrams Using ggplot2." *Journal of Statistical Software, Code Snippets* 87, no. 3: 1–17. https://doi.org/10.18637/jss.v087.c03.

Harbelot, B., H. Arenas, and C. Cruz. 2013. "Continuum: A Spatiotemporal Data Model to Represent and Qualify Filiation Relationships." In *Proceedings of the 4th Acm Sigspatial International Workshop on Geostreaming, Iwgs 2013*. https://doi.org/10.1145/2534303.2534312.

Hayes, P. J. 2003. "Knowledge Representation." In *Encyclopedia of Computer Science*, 947–951. Chichester, West Sussex, England: John Wiley and Sons Ltd.

Heath, T., and C. Bizer. 2022. *Linked Data: Evolving the Web Into a Global Data Space*. Cham, Switzerland: Springer Nature. https://doi.org/10.1007/978-3-031-79432-2.

Hedwall, P.-O., E. Holmström, M. Lindbladh, and A. Felton. 2019. "Concealed by Darkness: How Stand Density Can Override the Biodiversity Benefits of Mixed Forests." *Ecosphere* 10, no. 8: e02835. https://doi.org/10.1002/ecs2.2835.

Jovanovik, M., T. Homburg, and M. Spasić. 2021. "A Geosparql Compliance Benchmark." *ISPRS International Journal of Geo-Information* 10, no. 7: 487. https://doi.org/10.3390/ijgi10070487.

Kamdar, M. R., J. D. Fernández, A. Polleres, T. Tudorache, and M. A. Musen. 2019. "Enabling Web-Scale Data Integration in Biomedicine Through Linked Open Data." *npj Digital Medicine* 2: 90. https://doi.org/10.1038/s41746-019-0162-5.

Koubarakis, M. 2023. "Geospatial Data Science: A Handson Approach for Building Geospatial Applications Using Linked Data Technologies." *Association for Computing Machinery* 51: 336pp. https://doi.org/10.1145/3581906.

Lan, G., T. Liu, X. Wang, X. Pan, and Z. Huang. 2022. "A Semantic Web Technology Index." *Scientific Reports* 12: 3672. https://doi.org/10.1038/s41598-022-07615-4.

Lausch, A., A. Schmidt, and L. Tischendorf. 2015. "Data Mining and Linked Open Data – New Perspectives for Data Analysis in Environmental Research." *Ecological Modelling* 295: 5–17. https://doi.org/10.1016/j.ecolmodel.2014.09.018.

Lefrançois, M., A. Zimmermann, and N. Bakerally. 2017. "A SPARQL Extension for Generating RDF From Heterogeneous Formats." In *Proceedings of Extended Semantic Web Conference (ESWC'17)*, 35–50. Portoroz, Slovenia: Springer Nature. http://www.maxime-lefrancois.info/docs/LefrancoisZimmermannBakerally-ESWC2017-Generate.pdf.

Li, W., S. Wang, S. Wu, Z. Gu, and Y. Tian. 2022. "Performance Benchmark on Semantic Web Repositories for Spatially Explicit Knowledge Graph Applications." *Computers, Environment and Urban Systems* 98: 101884. https://doi.org/10.1016/j.compenvurbsys.2022.101884.

Lorenzo-Sáez, E., J.-V. Oliver-Villanueva, L.-G. Lemus-Zúñiga, J. F. Urchueguía, and V. Lerma-Arce. 2022. "Development of Sectorial and Territorial Information System to Monitor Ghg Emissions as Local and Regional Climate Governance Tool: Case Study in Valencia (Spain)." *Urban Climate* 42: 101125. https://doi.org/10.1016/j.uclim.2022.101125.

Louarn, M., F. Chatonnet, X. Garnier, T. Fest, A. Siegel, and O. Dameron. 2019. "Increasing Life Science Resources Reusability Using Semantic Web Technologies." In *2019 15th International Conference on Escience (Escience)* (pp. 217-225). https://doi.org/10.1109/eScience.2019.00031.

Marey-Pérez, M., and V. Rodríguez-Vicente. 2009. "Forest Transition in Northern Spain: Local Responses on Large-Scale Programmes of Field-Afforestation." *Ly Use Policy* 26, no. 1: 139–156. https://doi.org/10.1016/j.lyusepol.2008.02.004.

Martínez, J., C. Vega-Garcia, and E. Chuvieco. 2009. "Human-Caused Wildfire Risk Rating for Prevention Planning in Spain." *Journal of Environmental Management* 90, no. 2: 1241–1252. https://doi.org/10.1016/j.jenvman.2008.07.005.

Mauri, A., G. Strona, and J. San-Miguel-Ayanz. 2017. "Eu-Forest, a High-Resolution Tree Occurrence Dataset for Europe." *Scientific Data* 4: 160123. https://doi.org/10.1038/sdata.2016.123.

Nikolaou, C., K. Dogani, K. Bereta, et al. 2015. "Sextant: Visualizing Time-Evolving Linked Geospatial Data." *Journal of Web Semantics* 35: 35–52. https://doi.org/10.1016/j.websem.2015.09.004.

OGC. 2012. *Ogc geosparql: A Geographic Query Language for rdf Data, Version 1.0*. June 2023. http://www.opengis.net/doc/IS/geosparql/1.0.

Osei, R., H. Titeux, K. Bielak, et al. 2021. "Tree Species Identity Drives Soil Organic Carbon Storage More Than Species Mixing in Major Two-Species Mixtures (Pine, Oak, Beech) in Europe." *Forest Ecology and Management* 481: 118752. https://doi.org/10.1016/j.foreco.2020.118752.

Pascal, H. 2021. "A Review of the Semantic Web Field." *Communications of the ACM* 64, no. 2: 78–83. https://doi.org/10.1145/3397512.

Pasquier, C. 2008. "Biological Data Integration Using Semantic Web Technologies." *Biochimie* 90, no. 4: 584–594. https://doi.org/10.1016/j.biochi.2008.02.007.

Portolés, D., G. Vega-Gorgojo, J. M. Giménez-García, B. Fierro, and T. Jurado. 2021. Data Exportation and Publication – Final Report [Computer Software Manual]. Project Cross-Forest. Deliverable 2.3.

Rantala, S., B. Swallow, R. Paloniemi, and E. Raitanen. 2020. "Governance of Forests and Governance of Forest Information: Interlinkages in the Age of Open and Digital Data." *Forest Policy and Economics* 113: 102123. https://doi.org/10.1016/j.forpol.2020.102123.

Rescia, A. J., A. Pons, I. Lomba, C. Esteban, and J. W. Dover. 2008. "Reformulating the Social–Ecological System in a Cultural Rural Mountain Landscape in the Picos de Europa Region (Northern Spain)." *Landscape and Urban Planning* 88, no. 1: 23–33. https://doi.org/10.1016/j.landurbplan.2008.08.001.

Riofrío, J., M. del Río, and F. Bravo. 2016. "Mixing Effects on Growth Efficiency in Mixed Pine Forests." *Forestry: An International Journal of Forest Research* 90, no. 3: 381–392. https://doi.org/10.1093/forestry/cpw056.

Sánchez-García, S., E. Canga, E. Tolosana, and J. Majada. 2015. "A Spatial Analysis of Woodfuel Based on Wisdom Gis Methodology: Multiscale Approach in Northern Spain." *Applied Energy* 144: 193–203. https://doi.org/10.1016/j.apenergy.2015.01.099.

Wu, D., H.-T. Wang, and A. Uz Tansel. 2024. "A Survey for Managing Temporal Data in Rdf." *Information Systems* 122: 102368. https://doi.org/10.1016/j.is.2024.102368.