TripleGeo-CSW: A Middleware for Exposing Geospatial Catalogue Services on the Semantic Web

Spiros Athanasiou[§] Nikos Georgomanolis[§] Kostas Patroumpas[†,§]
Michalis Alexakis[§] Thodoris Stratiotis[§]
[§]Institute for the Management of Information Systems, "Athena" Research Center, Hellas
[†]School of Electrical & Computer Engineering, National Technical University of Athens, Hellas
{spathan, ngeorgomanolis, kpatro, alexakis, stratiot}@imis.athena-innovation.gr

ABSTRACT

A wealth of data and services are available on the Web, and often have geographical context as well. But the vast quantity of offered geospatial information is rather difficult to explore, and its quality hard to assess, due to lack of sufficient metadata. Hence, the Open Geospatial Consortium has specified application profiles for publishing, accessing, and searching over collections of spatial metadata with standardized *Catalogue Services for the Web* (CSW). Unfortunately, existing spatial metadata remain largely unexploited by Semantic Web technologies. In this paper, we introduce TripleGeo-CSW, a middleware that can be used to discover metadata from existing CSWs through a virtual SPARQL endpoint. Acting as broker between a request (in SPARQL) and catalogue services (in XML/GML), this platform can provide on-the-fly information (as RDF triples) on available geodata according to multiple, user-specified criteria (e.g., area of interest, date of last update, keywords). As a proof of concept, we have set up an instance of this middleware against CSWs from public authorities across Europe, which involve datasets complying with the EU INSPIRE Directive. Our experience testifies that TripleGeo-CSW can assist stakeholders to repurpose existing CSWs with minimal overhead and readily expose spatial metadata on the Semantic Web.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Design, Management, Standardization

Keywords

Catalogue services, geospatial data, metadata, CSW, RDF, GeoSPARQL, INSPIRE.

1. INTRODUCTION

Proliferation of location-aware devices (smartphones, car navigators, etc.) over the past decade has led to an unprecented offering of geospatial information on the Web. Not only maps of the finest detail or satellite imagery of the entire planet, but also geotagged photographs and geolocation hashtags in social networking underscore the importance of geography in our everyday life and activities. Crowdsourcing has also become a valuable means of providing up-to-date geodata for free, thanks to initiatives such as OpenStreetMap [21], GeoNames [10], or Wikimapia [34] that engage thousands of volunteers worldwide.

However, all this geospatial information comes in so many different formats, heterogeneous schemas, proprietary systems, customized services, etc., such that assessing its quality becomes a burden even for experts. For example, choosing an unreliable road network for a routing application may disappoint users despite its friendly interface; having updated locations for points of interest (restaurants, cinemas, bars, etc.) in a digital city guide could be the key to commercial success; and an accurate geological map is indispensable in mineral exploration or when constructing transport infrastructures. With so many geospatial data coming from commercial vendors, governmental agencies, or crowdsourcing, the need for precise *metadata* is indisputable. Such metadata can provide a brief summary about the content, purpose, quality, location of the spatial data, and also report on its creation procedures. Indeed, information about the geographical reference (i.e., its Coordinate Reference System– CRS), resolution (i.e., the map scale used in digitization), date of last update, or textual keywords describing the content of digital maps, can greatly assist users to choose the geospatial features that best suit their needs.

ISO standard 19115:2003 [12] (recently updated to ISO 19115-1:2014 [13]) offers specifications for *standardized metadata* that can support users in effective discovery and retrieval of geodata. With the endorsement of the Open Geospatial Consortium (OGC), this standard establishes a common terminology for metadata elements on geospatial features, properties, and entire collections of geodata. Furthermore, *catalogue services* are important in publishing and searching collections of metadata for geospatial data and related web services. Metadata in catalogues represent resource characteristics that can be queried and presented for evaluation and further processing by both humans and software. OGC standard on *Catalogue Services for the Web* (CSW) [18] specifies a framework and interfaces for defining application profiles of services based on geospatial metadata.

This metadata can be queried in order to return results in well-known content models (metadata schemas) and encodings, e.g., in Geography Markup Language (GML) [19]. For example, returned metadata records may contain information about the title of datasets, their format, geographical extent (i.e., their Bounding Box in latitude and longitude coordinates), the Coordinate Reference System, licensing policies, as well as links to other associated metadata.

Unfortunately, accessing such spatial catalogue services is currently disjoint from the Semantic Web, without any means to repurpose the contents of existing catalogues according to the Linked Data paradigm [3]. Having high-quality linked metadata resources on available geodata could offer great advantages for users and applications. Catalogue contents would become machine reabable and potentially interlinked with information from third parties. Fortunately, the recent OGC GeoSPARQL standard [20] proposes structures for storing RDF geometries, querying them through a SPARQL extension [31] equipped with a variety of spatial operations [2], as well as with support for spatial reasoning on Linked Open Data (LOD). We regard this as a great opportunity to expose spatial metadata from catalogues encoded in RDF [30] and queried through GeoSPARQL.

Yet another development may also act as a catalyst for publishing linked spatial metadata. By 2020, implementation of the INSPIRE Directive (*INfrastructure for SPatial InfoRmation in Europe*) [7] will enable discovery, download, and visualization of geospatial information across the European Union in a common, cross-boundary manner. Paving the way towards geospatial data interoperability and dissemination, several public organizations across Europe have begun publishing metadata in spatial catalogues according to the ISO and OGC specifications. Availability of such official, diachronic, high-quality information can have major benefits to governance, research, and enterpreneurship. In case such metadata were made accessible via SPARQL endpoints, they would certainly offer great perspectives for extracting spatial knowledge and interlinking.

Towards these goals, we introduce TripleGeo-CSW [1], which is essentially an open-source *CSW-to-RDF middleware*. Easily coupled with a web interface so as to constitute a virtual GeoSPARQL enpoint, it allows users to explore the quantity and quality of spatial datasets available from several existing Catalogue Services according to multiple search criteria. With TripleGeo-CSW, GeoSPARQL queries are translated on-the-fly into requests against CSW on remote servers over HTTP protocols. Using RDF mappings for XML/GML encodings of standard geospatial metadata, the server response is suitably transformed via XSL stylesheets [33] into RDF triples that are finally returned as answers. To the best of our knowledge, this is the first attempt to build an abstraction layer on top of the CSW and INSPIRE infrastructures based on GeoSPARQL concepts, thus making spatial catalogues accessible and discoverable as linked metadata with geometries. Our contribution can be summarized as follows:

- We have implemented TripleGeo-CSW, a middleware that enables searching for available geodata through a virtual GeoSPARQL interface for CSW.

- We have specified application profiles that can be used as templates for transforming geospatial XML/GML metadata into RDF by an XSLT parser.

- As a proof of concept, we exposed existing INSPIRE-aligned catalogue services as linked data sources in RDF. With minimal overhead, this web interface enables queries in GeoSPARQL for discovering geospatial resources across Europe.

The remainder of this paper proceeds as follows. In Section 2, we survey related work on standards regarding spatial metadata and catalogue services. In Section 3, we present the architecture of TripleGeo-CSW by examining its components, the processing flow, and its current implementation status. In Section 4, we present a working case study on data discovery over INSPIRE catalogue services. Section 5 concludes the paper.

2. BACKGROUND & RELATED WORK

2.1 Catalogue Services for the Web (CSW)

Catalogue Services for the Web (CSW) is an OGC standard [18] that describes application profiles for publishing, accessing, and searching over collections of metadata on geospatial data, services, and related resources. This metadata must be encoded in XML and the schema of its records is usually conformant to more specific standards (like ISO 19139 [15], Dublin Core [6], or INSPIRE [7]). The spatial extent of a dataset is given with its bounding box encoded in GML [19]. Users may submit a number of different requests (either `GET` or `POST` HTTP methods) to a CSW and the response is encoded in an XML document as well. Typical requests that must be always supported by a CSW are:

- `GetCapabilities` can be used to retrieve metadata describing the type of requests the CSW can accept (e.g., version, acceptable parameters, output formats, etc.).

- `DescribeRecord` returns a description of the metadata records' model, i.e., an XML schema definition (XSD).

- `GetRecords` retrieves actual metadata records that satisfy criteria and filters specified in the request. Figure 4 illustrates one such request to CSW, asking for available geodata within a rectangular area (given in longitude/latitude coordinates) and matching specific textual criteria on the subject of the dataset and its associated keywords.

- `GetRecordsById` returns records matching a list of specific identifiers given as parameters in the request.

Other requests are non-mandatory for CSWs, like:

- `GetDomain` returns the range of values of a metadata record field or a request parameter.

- `Transaction` can be used to create metadata records, as well as to edit or delete existing ones.

- `Harvest` pulls metadata from third-party sources to create new records or update existing ones in the CSW.

2.2 Spatial Metadata as Linked Data Sources

There are mainly two (often complementary) approaches to *cataloguing linked metadata*. Data Catalogue Vocabulary (DCAT) [28] is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the

```
<dct:conformsTo>
  <dct:Standard>
    <dct:title xml:lang='en'>
      <xsl:value-of select='//gmd:report//gmd:title/gco:CharacterString'/>
    </dct:title>
    <dct:issued rdf:datatype='http://www.w3.org/2001/XMLSchema#date'>
      <xsl:value-of select='//gmd:report//gco:Date' />
    </dct:issued>
  </dct:Standard>
</dct:conformsTo>
```

Figure 1: Excerpt of XSL stylesheet for transforming metadata elements into RDF.

Web. The VoID Vocabulary (VoID) [32] makes use of an RDF Schema vocabulary to express metadata about RDF datasets, and aims at data discovery, cataloguing and archiving. However, both approaches make extensive use of terms from other vocabularies, in particular Dublin Core [6].

Based on similar vocabularies, a few initiatives and case studies headed towards *linked metadata on spatial datasets*. Among them, the Mimas Linked Data Project for LandMap Spatial Discovery in the UK has made some preliminary work [17], mostly by identifying vocabularies and defining RDF mappings for a subset of their datasets. An open source prototype for Data Catalogue Vocabulary services based on DCAT is being implemented in GeoNetwork [11], and would eventually provide support to harvest, search and link catalogue contents with other interlinked resources. Public authorities across Europe have also begun publishing spatial metadata through SPARQL enpoints, such as the municipality of Zaragoza in Spain [35].

With regard to the particular task of exposing spatial metadata as linked open data, the crosswalking approach is suggested in [24, 16]. *Metadata crosswalking* involves mappings from popular geospatial metadata schemas to the Dublin Core vocabulary, addition of extra metadata elements, and finally expressing the metadata terms as RDF. The authors in [24] suggest an alternative method for publishing geospatial metadata provisioned by custom catalogue services as linked open metadata. In this case, RDF metadata terms are published directly from the UML representation of the underlying custom schemas.

The Joint Research Centre (JRC) of the European Commission has begun an exploratory investigation [23] regarding geospatial metadata on the Semantic Web. Of course, they mainly focus on alignment with the EU INSPIRE Directive [7] towards a LOD-enabled INSPIRE prototype, by creating a corpus of RDF metadata exposed via a SPARQL endpoint. Still, their preliminary version of RDF mappings for INSPIRE metadata elements offers a concrete RDF representation [9] for spatial metadata based on DCAT-AP and other relevant vocabularies (such as DCT, SKOS, vCard, etc.). In this work, we take advantage of such mappings and we offer generic stylesheets in XSL (EXtensible Stylesheet Language) [33], which can be used to transform XML files with OGC-compliant spatial metadata into an equivalent RDF representation. To the best of our knowledge, ours is the first attempt to offer application profiles in RDF for standardized geospatial metadata through catalogue services.

3. MIDDLEWARE ARCHITECTURE

In this Section, we present TripleGeo-CSW, an open-source middleware for data discovery from geospatial catalogue services on the Semantic Web. We first describe the way that spatial metadata elements can be translated into RDF triples. Then, we analyze the processing flow in TripleGeo-CSW, as well as its capabilities of searching against CSW with multiple criteria via a virtual GeoSPARQL endpoint.

Table 1: Some RDF mappings for spatial metadata.

Metadata element	RDF mapping of attribute
Resource title	`dct:title`
Resource language	`dct:language`
Keyword	`dcat:keyword`
Geographic Bounding Box	`dct:spatial`
Responsible organization – Owner	`dct:rightsHolder`

3.1 Metadata Application Profiles in RDF

Although still a work-in-progress, the RDF mappings suggested by the JRC [9] offer a valuable basis to develop a methodology for transforming spatial metadata elements into RDF. Our goal is to facilitate such transformations on-the-fly, such that contents from existing CSWs can be made accessible through (Geo)SPARQL. We are mostly interested in exposing the spatial coverage of data, as well as the temporal range of their lifecycle (i.e., when data was created, published or modified). However, many more metadata elements are important as well, such as descriptions (e.g., title, abstract, subject, keywords), content assessments (like quality, provenance, or conformity), as well as their legal status (owner, license, point of contact, etc.). A few RDF mappings from indicative metadata elements to vocabularies such as DCAT and DCT are shown in Table 1. Once this metadata gets exposed on the Semantic Web, it may be potentially interlinked with other features, such as terms in code lists or multilingual thesauri [23].

In practice, we manually created an application profile for such metadata as a set of templates employed in XSLT transformation [33]. Our custom XSL stylesheet[1] accepts an XML file with metadata records obtained as a response from a request to a CSW. Once invoked with an XSLT parser, the stylesheet turns metadata elements into suitable RDF statements according to the mapping; the result is an RDF/XML representation of original OGC-compliant metadata records. The XSL stylesheet is generic, covers all elements, and can be reused against any metadata conforming to OGC/ISO specifications. The excerpt shown in Figure 1 refers to handling of elements related with dataset conformity. Regarding the geographical coverage, its bounding box can be suitably expressed either as a GeoSPARQL polygon

[1] Stylesheet `Metadata2RDF.xsl` is included in the source code [1]; it has been also integrated into our TripleGeo tool [22] for directly transforming locally stored metadata files from XML into RDF.

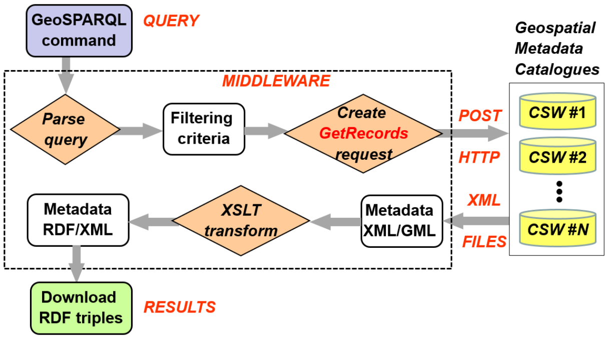


Figure 2: Flow diagram for processing GeoSPARQL queries in the TripleGeo-CSW middleware.

or a 2-dimensional rectangle BOX2D.

Our design adheres to reusing existing URIs as much as possible, especially in statements concerning spatial, temporal, and identification elements. However, blank nodes exist in the resulting triples, since these RDF mappings are based on the abstract schema of metadata elements. Such blank nodes are used as locally-scoped artifacts, which need not be explicitly labelled. Provided that the user is aware of the underlying schema (ISO/OGC), formulating (Geo)SPARQL queries against such metadata is straightforward.

3.2 Processing GeoSPARQL Queries over CSW

We assume that a list of catalogue services (CSW) exists, and each service is operational and accepts HTTP requests. In order to facilitate discovery of matadata from such CSWs through (Geo)SPARQL, we have implemented the TripleGeo-CSW middleware. The processing flow of this middleware is illustrated in Figure 2. It is triggered by a (Geo)SPARQL query, where the user can specify one or more conditions according to the spatial metadata ontology, as explained in Section 3.1. We developed a parser, which identifies several types of such conditions, including spatial predicates as documented next. The OGC standard defines a specific model for CSW requests (POST/GET HTTP protocols), which covers several cases. However, our major concern here is the CSW <Filter> element, which controls whether metadata should be retained according to specific criteria. Hence, the user-specified GeoSPARQL conditions must be internally rewritten and then integrated into the <Filter> element of a GetRecords request for CSW. Thanks to the OGC standard for CSW [18], an identical such request will be submitted simultaneously via POST HTTP protocol against each of the listed catalogues. Once each CSW provides its response as a collection of qualifying metadata records in a separate XML file. With the XSL stylesheet described in Section3.1, these metadata records (conforming to ISO 19115) are finally converted into XML/RDF triples and are available for download by the user.

Note that integrity and consistency of metadata information is a responsibility of the owners (governments, organizations, etc.), so each metadata element is supposed to come from a single CSW. Thus, resolving conflicts is not employed when compiling the resulting RDF triples from multiple sources. Of course, the final output is OGC-compliant metadata, since the XSLT transformation uses templates that map each metadata element into DCAT elements.

This open-source software has been developed in Python 2.7.3, and it makes use of several additional libraries:

- urllib2[2], a Python module for fetching URLs (Uniform Resource Locators) and posing requests;
- re[3] provides Perl-style regular expression pattern matching and is used for parsing such expressions in users' SPARQL requests;
- etree[4] performs XML parsing using the concepts of the ElementTree API for Python.

In its current release, TripleGeo-CSW can support user requests to discover whether there are any available, updated spatial datasets according to criteria that may involve a given geographical area, a certain thematic category (e.g., transport, hydrography), or particular keywords (e.g., "water", "rail"). More specifically, a (Geo)SPARQL query that can be handled by this middleware consists of two parts:

- a SELECT or CONSTRUCT clause identifies the attributes that will appear in the query results, and
- a WHERE clause provides the basic graph pattern to match against the metadata, as well as FILTER criteria.

Typically, a graph pattern in a WHERE clause consists of a triple with subject, predicate and object; this pattern is checked for matching against the metadata records. A pattern is formatted as ?s ?p ?o, where ?s is the sought element and ?o is either a specific value (e.g., a string literal like "Environment") or a binding variable. Hence, search involves only triples satisfying match patterns ?s ?p ?o (*Case 1*) or ?s ?p "literal" (*Case 2*). In order to handle the matching candidates, we make use of a List and a Dictionary structure. The list is used for handling all triples with a variable as their object (*Case 1*). The dictionary is a set of <key:value> pairs with the requirement that keys are unique; so, it actually contains <element:value> pairs with unique metadata elements (*Case 2*). In the evaluation, triple patterns are checked one by one for matches, since multiple such criteria may be present in a query. Currently, no query optimization or check for syntax errors is performed; we defer dealing with such issues in future releases.

Concerning filtering, TripleGeo-CSW supports GeoSPARQL queries that may include any of the following criteria:

- *Matching regular expressions* (REGEX) against string literals. Through FILTER conditions in SPARQL, the user can check wildcard pattern matchings of string values (e.g., "^water*") with keywords, titles, subjects, and other textual properties in the metadata.
- *Date comparisons* make use of typical operators (>, <, <=, >=) and a constant date value as an argument, in order to identify datasets issued, modified or published before or after that particular date.
- *Spatial filtering.* OGC-compliant metadata include the BoundingBox of the geographical extent for each dataset

[2] https://docs.python.org/2/library/urllib2.html
[3] https://docs.python.org/2/library/re.html
[4] https://docs.python.org/2/library/xml.etree.elementtree.html

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/geosparql/function/>
CONSTRUCT { ?m dcat:keyword ?k .      ?s dc:subject ?sub .      ?f geo:hasGeometry ?fWKT }
WHERE { ?m dcat:keyword ?k .
        ?s dc:subject "Environment" .
        ?f geo:hasGeometry ?fWKT .
        FILTER (REGEX(str(?k),"^water*") && geof:sfWithin(?fWKT, "BOX2D(-8.24 54.02,-5.18 55.32)"^^geo:wktLiteral)) };
```

Figure 3: Example GeoSPARQL query against spatial metadata exposed in CSW.

as an indication of its coverage area. Hence, it makes much sense to allow users discover availability of data in their region of interest, specified as a 2-dimensional rectangle (`BOX2D`) with four geographical coordinates. The parser recognizes typical GeoSPARQL topological predicates [20] like **`sfWithin()`**, **`sfContains()`**, **`sfIntersects()`**, **`sfOverlaps()`**, etc., which can be translated into equivalent CSW spatial filters over rectangles. For instance, operator **`sfWithin()`** checks whether a user-specified `BOX2D` is totally within the coverage of a dataset, **`sfIntersects()`** identifies whether a given `BOX2D` intersects the coverage of a dataset, etc.

Logical operators for conjunction (`&&`) and disjunction (`||`) can be used to combine filtering criteria, whereas a `UNION` clause can bind statements specifying alternative patterns (e.g., searching for datasets characterized by subjects like "road" or "rail"). For instance, issuing the GeoSPARQL query in Figure 3 will retrieve spatial metadata (in any of the listed CSWs) for environmental features related to water and within the given rectangular area. Note that no RDF graph is specified, as neither do we make use of a physically stored semantic repository nor any RDF triples get materialized or permanently retained. The TripleGeo-CSW middleware automatically rewrites this query into an equivalent `GetRecords` request (shown in Figure 4), which may be submitted to each of the available CSWs. Consequently, all RDF results are generated on-the-fly by XSLT transformation of the XML response received from these CSWs.

3.3 Implementation Status

TripleGeo-CSW is free software and its current version 1.0 is publicly available [1], including the source code in Python and several query examples. TripleGeo-CSW can be redistributed and modified under the terms of the GNU General Public License. The software can work in standalone mode, but it can also be coupled with a web interface.

A basic web interface consists of a client-side JavaScript application that allows the user to edit a query, specify the format for results, and download the qualifying RDF triples. In addition, a server-side PHP application acts both as an API proxy and an abstraction layer. Once it receives a query, this proxy validates it, then builds the properly formulated HTTP request against the list of available CSWs, and finally sends back the results to the user.

Adding or removing a CSW simply involves editing a configuration file in TripleGeo-CSW, i.e., adding or deleting the related URL of that catalogue service. In essence, all processing components are wrapped under this virtual GeoSPARQL endpoint, thus offering flexibility to interact directly with any number of remote catalogues and repurpose

```
<?xml version='1.0' encoding='utf-8'?>
<GetRecords
     xmlns="http://www.opengis.net/cat/csw/2.0.2"
     xmlns:csw="http://www.opengis.net/cat/csw/2.0.2"
     xmlns:ogc="http://www.opengis.net/ogc"
     xmlns:ows="http://www.opengis.net/ows"
     xmlns:dcat="http://www.w3.org/ns/dcat#"
     xmlns:dc="http://purl.org/dc/terms/"
     xmlns:gml="http://www.opengis.net/gml"
     xmlns:gmd="http://www.isotc211.org/2005/gmd"
     xmlns:apiso="http://www.opengis.net/cat/csw/apiso/1.0"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     service="CSW"
     version="2.0.2"
     startPosition="1"
     resultType="results"
     maxRecords="100"
     outputFormat="application/xml"
     outputSchema="http://www.isotc211.org/2005/gmd"
     xsi:schemaLocation="http://www.opengis.net/cat/csw/2.0.2
http://schemas.opengis.net/csw/2.0.2/CSW-discovery.xsd">
  <Query typeNames="gmd:MD_Metadata">
    <ElementSetName typeNames="gmd:MD_Metadata">full
    </ElementSetName>
    <Constraint version="1.1.0">
      <ogc:Filter>
        <ogc:And>
          <ogc:PropertyIsEqualTo>
            <ogc:PropertyName>dc:subject</ogc:PropertyName>
            <ogc:Literal>Environment</ogc:Literal>
          </ogc:PropertyIsEqualTo>
          <ogc:PropertyIsLike wildCard="^" singleChar="_">
            <ogc:PropertyName>dcat:keyword</ogc:PropertyName>
            <ogc:Literal>^water*</ogc:Literal>
          </ogc:PropertyIsLike>
          <ogc:Within>
            <ogc:PropertyName>ows:BoundingBox</ogc:PropertyName>
            <gml:Envelope>
              <gml:lowerCorner>-8.24 54.02</gml:lowerCorner>
              <gml:upperCorner>-5.18 55.32</gml:upperCorner>
            </gml:Envelope>
          </ogc:Within>
        </ogc:And>
      </ogc:Filter>
    </Constraint>
  </Query>
</GetRecords>
```

Figure 4: The `GetRecords` request to CSW corresponding to the GeoSPARQL query in Figure 3. Note that the spatial condition is translated into an equivalent enclosure within a `gml:Envelope` specified with the given geographical coordinates. The graph pattern and regular expression matching criteria in the query are respectively transformed into equivalent conditions `ogc:PropertyIsEqualTo` and `ogc:PropertyIsLike`, recognizable by CSW services.

Table 2: INSPIRE-aligned metadata available through several CSWs across Europe.

INSPIRE Discovery Service in the Czech Republic:	`http://geoportal.cuzk.cz/SDIProCSW/service.svc/get?request=GetCapabilities&service=CSW`
Estonian National Geoportal:	`http://inspire.maaamet.ee/geoportal/csw/discovery?request=GetCapabilities&Service=csw&language=eng`
Irish Spatial Data Exchange:	`http://catalogue.isde.ie/geonetwork/srv/en/csw?request=GetCapabilities&service=CSW`
National CSW for Norway:	`http://www.geonorge.no/geonetwork/srv/nor/csw-inspire?service=CSW&request=GetCapabilities`
Discovery Service for the UK Location catalogue:	`http://csw.data.gov.uk/geonetwork/srv/en/csw?request=GetCapabilities&service=CSW`
Spanish National Geographic Institute:	`http://www.ign.es/csw-inspire/srv/eng/csw?Service=CSW&Request=GetCapabilities`
Metadata Catalogue of the SDI for Spain:	`http://www.idee.es/csw-inspire-idee/srv/eng/csw?request=GetCapabilities&service=CSW`

their spatial metadata. The only prerequisite for such catalogues is that they must be compatible with the OGC standard for CSW [18] and thus support the related requests, as discussed in Section 2.1.

4. A USE CASE: DISCOVERING INSPIRE THROUGH GEOSPARQL QUERIES

In this Section, we present a use case where TripleGeo-CSW has been applied in practice. This validation of the middleware concerns discovery of INSPIRE-aligned spatial datasets from catalogue services across Europe through a virtual GeoSPARQL endpoint.

4.1 INSPIRE as a Source for Linked Data

The INSPIRE Directive 2007/2/EC [7] sets a unified framework for Spatial Data Infrastructures (SDI) across the EU, so that by 2020 spatial information can be shared among European public authorities in order to assist in environmental policies. Its foundations include technical interoperability standards for geospatial metadata, data and online services, as well as uniform legal rules for data interchange and reuse. Towards establishing such a pan-European SDI, INSPIRE specifications prescribe catalogues of available resources using metadata, common access policies and standards, as well as network services for discovery, viewing, downloading, transformation, etc. for spatial datasets.

Implementing Rules [8] for INSPIRE-compliant metadata propose a schema for describing datasets, dataset series, services and thematic layers across Europe. This schema is designed according to ISO standards [12, 13] and contains metadata elements for data regarding its identification, topic, quality, geographical and temporal extent, as well as points of contact with the responsible parties. In addition, ISO-19119 [14] defines a framework for developing services that can be used to access and process geospatial data. This framework supports access to different data sources through a generic, platform-neutral application interface. INSPIRE metadata should not violate these ISO standards, but since the latter require many more elements (e.g., points of contact, restrictions) these have to be provisioned as well. On the other hand, metadata published according to the ISO-19115 core is not guaranteed to conform with the INSPIRE ontology, so an alignment is necessary.

Unfortunately, no complete INSPIRE ontology in RDF/OWL [29] currently exists. This reflects the difficulty of bridging the "closed world" assumption of UML models in INSPIRE with the "open world" view of RDF. Admittedly, this limitation refers not only to INSPIRE, since exposing geospatial information as open linked data is a relatively new research topic. Especially for INSPIRE SDIs, some prominent opportunities of utilizing linked open data have been highlighted [25] by the Joint Research Centre of the European Commission, along with the requirements for achieving it. Exposing INSPIRE datasets as linked data has attracted some research interest. The proposed approaches either translate INSPIRE-compliant GML data models as semantic OWL ontologies [26], or generate an ontology model mixing a number of different existing ontologies and vocabularies along with tools for RDF extraction and interlinking [27], or even deriving linked data from GML data and reusing existing concepts from vocabularies [5].

In contrast to the aforementioned approaches on spatial data, there has not been any attempt to expose INSPIRE *metadata* from existing catalogues according to the GeoSPARQL standard [20], as we present next. Our TripleGeo-CSW suite for the Semantic Web can not only be used by stakeholders that wish to make their SDI contents accessible in RDF, but also for discovering available third-party data via GeoSPARQL requests against CSWs.

4.2 Data Discovery from INSPIRE CSWs

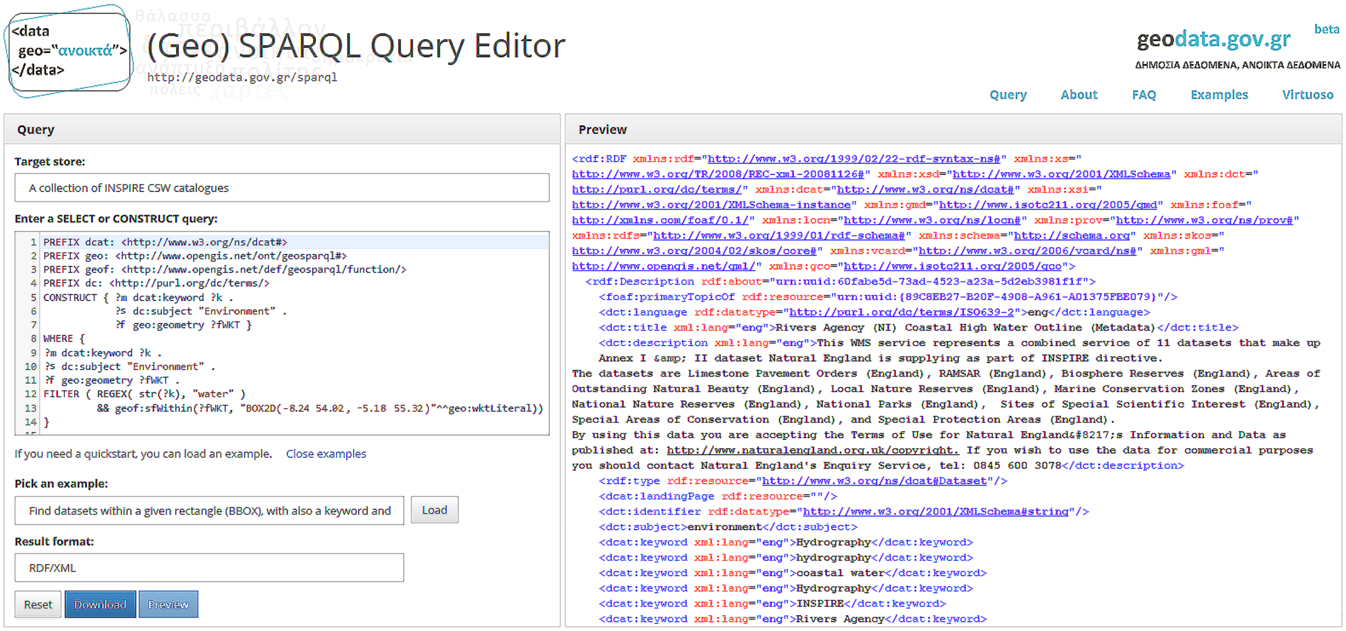
Catalogue services for INSPIRE-compliant metadata have become already available in various European countries, even in non-EU member states like Norway, as indicated in Table 2. Our work is focused on exposing such CSWs on the Semantic Web through our CSW-to-RDF middleware TripleGeo-CSW. In short, we wish to enable GeoSPARQL queries with user-specified criteria against the contents of such catalogues, so as to facilitate INSPIRE data discovery. In this case, TripleGeo-CSW acts as a broker between a virtual GeoSPARQL endpoint and a list of INSPIRE-compliant CSWs, and undertakes to request any available information from the CSWs, collect the partial XML results, and finally return any qualifying metadata as RDF triples.

Towards this goal, we have made use of INSPIRE metadata from CSWs across Europe (Table 2). We stress that this is just an indicative list of currently operating CSWs. Of course, this list may be extended as more INSPIRE-compliant such services become available, without necessitating absolutely any change in our existing framework. It only requires including any additional CSW into the list of such services, i.e., editing the respective configuration file that is accessible by the middleware.

In order to provide a simple and uniform interface to end-users, we have implemented a web application that offers the ability to issue GeoSPARQL queries against CSWs and receive response in a variety of formats (RDF/XML, CSV, HTML, etc.). This web interface is publicly available at:

`http://geodata.gov.gr/sparql/`

Users wishing to explore available INSPIRE geodata across Europe must choose *"A collection of INSPIRE CSW catalogues"* as their (virtual) target store. We stress that no triple store is used to physically retain any RDF metadata received from such CSW services. Instead, qualifying metadata records are collected in XML and transformed on-the-fly into a RDF serialization.



Target store: A collection of INSPIRE CSW catalogues

Enter a SELECT or CONSTRUCT query:

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/geosparql/function/>
PREFIX dc: <http://purl.org/dc/terms/>
CONSTRUCT { ?m dcat:keyword ?k .
            ?s dc:subject "Environment" .
            ?f geo:geometry ?fWKT }
WHERE {
?m dcat:keyword ?k .
?s dc:subject "Environment" .
?f geo:geometry ?fWKT .
FILTER ( REGEX( str(?k), "water" )
         && geof:sfWithin(?fWKT, "BOX2D(-8.24 54.02, -5.18 55.32)"^^geo:wktLiteral))
}
```

If you need a quickstart, you can load an example. Close examples

Pick an example: Find datasets within a given rectangle (BBOX), with also a keyword and

Result format: RDF/XML

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:xs="
http://www.w3.org/TR/2008/REC-xml-20081126#" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:dct="
http://purl.org/dc/terms/" xmlns:dcat="http://www.w3.org/ns/dcat#" xmlns:xsi="
http://www.w3.org/2001/XMLSchema-instance" xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:foaf="
http://xmlns.com/foaf/0.1/" xmlns:locn="http://www.w3.org/ns/locn#" xmlns:prov="http://www.w3.org/ns/prov#"
xmlns:rdfs="http://www.w3.org/1999/01/rdf-schema#" xmlns:schema="http://schema.org" xmlns:skos="
http://www.w3.org/2004/02/skos/core#" xmlns:vcard="http://www.w3.org/2006/vcard/ns#" xmlns:gml="
http://www.opengis.net/gml/" xmlns:gco="http://www.isotc211.org/2005/gco">
  <rdf:Description rdf:about="urn:uuid:60fabe5d-73ad-4523-a23a-5d2eb3981f1f">
    <foaf:primaryTopicOf rdf:resource="urn:uuid:{89C8EB27-B20F-4908-A961-A01375FBE079}"/>
    <dct:language rdf:datatype="http://purl.org/dc/terms/ISO639-2">eng</dct:language>
    <dct:title xml:lang="eng">Rivers Agency (NI) Coastal High Water Outline (Metadata)</dct:title>
    <dct:description xml:lang="eng">This WMS service represents a combined service of 11 datasets that make up
    Annex I &amp; II dataset Natural England is supplying as part of INSPIRE directive.
The datasets are Limestone Pavement Orders (England), RAMSAR (England), Biosphere Reserves (England), Areas of
Outstanding Natural Beauty (England), Local Nature Reserves (England), Marine Conservation Zones (England),
National Nature Reserves (England), National Parks (England),  Sites of Special Scientific Interest (England),
Special Areas of Conservation (England), and Special Protection Areas (England).
By using this data you are accepting the Terms of Use for Natural England&#8217;s Information and Data as
published at: http://www.naturalengland.org.uk/copyright. If you wish to use the data for commercial purposes
you should contact Natural England's Enquiry Service, tel: 0845 600 3078</dct:description>
    <rdf:type rdf:resource="http://www.w3.org/ns/dcat#Dataset"/>
    <dcat:landingPage rdf:resource=""/>
    <dct:identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string"/>
    <dct:subject>environment</dct:subject>
    <dcat:keyword xml:lang="eng">Hydrography</dcat:keyword>
    <dcat:keyword xml:lang="eng">hydrography</dcat:keyword>
    <dcat:keyword xml:lang="eng">coastal water</dcat:keyword>
    <dcat:keyword xml:lang="eng">Hydrography</dcat:keyword>
    <dcat:keyword xml:lang="eng">INSPIRE</dcat:keyword>
    <dcat:keyword xml:lang="eng">Rivers Agency</dcat:keyword>
```

Figure 5: The virtual GeoSPARQL endpoint at `http://geodata.gov.gr/sparql/` over INSPIRE CSWs.

This web interface (illustrated in Figure 5) includes a few predefined (Geo)SPARQL query examples against these CSW services. We have employed `CONSTRUCT` queries in order to receive results as RDF triples, and also verify the robustness of our middleware and validate its functionality. These queries explore a wide range of metadata features, e.g., keywords, subjects, titles, as well as the geographical area covered by the INSPIRE datasets referenced in the CSWs. Indicatively, users can:

- Search for datasets tagged with a given *keyword* (e.g., "administrative");
- Find available datasets that specify the given *subject* (like "Environment") in the metadata;
- Find datasets with spatial coverage inside a given rectangle (i.e., *Bounding Box*);
- Identify datasets on a given *subject* (e.g., "Environment") and whose *title* includes a particular term (e.g., "network").

Users may submit such queries "as is", modify them to reflect their specific search criteria, and of course, write their own queries in order to discover INSPIRE-compliant datasets offered by the available CSWs.

5. SUMMARY

In this paper, we introduced an open-source software that can be used to repurpose existing catalogue services (CSW) on geospatial metadata as high quality Linked Data sources. In effect, TripleGeo-CSW acts as a CSW-to-RDF middleware, which translates a given GeoSPARQL query into an equivalent request for available metadata records against multiple CSWs. As soon as the response is collected, the original XML metadata elements are transformed on-the-fly into RDF triples and returned as answers.

As a proof of concept, we have successfully enabled users to search for INSPIRE datasets from remote CSW services across Europe, by providing a virtual GeoSPARQL interface on top of TripleGeo-CSW. This ensures that INSPIRE Catalogue Services are accessible with Semantic Web technologies and thus INSPIRE data are discoverable with negligible overhead from stakeholders.

6. ACKNOWLEDGEMENTS

This work was partially supported by the European Commission under FP7-ICT-2011-8 grant #318159 *"GeoKnow – Making the Web an Exploratory Place for Geospatial Knowledge"* and FP7-ICT-2013-SME-DCA grant #609608 *"PublicaMundi – Scalable and Reusable Open Geospatial Data"*. We also wish to thank Andrea Perego and Michael Lutz (European Commission – JRC) for helpful discussions regarding RDF mappings of INSPIRE-aligned metadata.

7. REFERENCES

[1] Athena R.C. TripleGeo-CSW open source middleware. `https://github.com/GeoKnow/TripleGeo-CSW`

[2] R. Battle and D. Kolas. GeoSPARQL: Enabling a Geospatial Semantic Web. *Semantic Web Journal*, 3(4): 355-370, 2012.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data – The Story So Far. *IJSWIS*, 5(3): 1-22, 2009.

[4] DBpedia. `http://dbpedia.org`

[5] L. van den Brink, P. Janssen, W. Quak, and J. Stoter. Linking spatial data: semi-automated conversion of geo-information models and GML data to RDF. *IJSDIR*, 9: 59-85, 2014.

[6] Dublin Core Metadata Initiative. Dublin Core Metadata element set, Version 1.1. July 1999. `http://dublincore.org/documents/dcmi-terms/`

[7] European Commission (EC). INSPIRE Directive –

Infrastructure for Spatial Information in the European Community. http://inspire.jrc.ec.europa.eu/

[8] EC. INSPIRE Implementing Rules. http://inspire.ec.europa.eu/index.cfm/pageid/47

[9] EC. Alignment of INSPIRE metadata with DCAT-AP. https://ies-svn.jrc.ec.europa.eu/projects/metadata/wiki/Alignment_of_INSPIRE_metadata_with_DCAT-AP

[10] GeoNames database. http://www.geonames.org/

[11] GeoNetwork Data Catalog Vocabulary services. http://trac.osgeo.org/geonetwork/wiki/proposals/DCATandRDFServices

[12] ISO 19115:2003. Geographic information – Metadata. http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

[13] ISO 19115-1:2014. Geographic information – Metadata – Part 1: Fundamentals. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798

[14] ISO 19119:2005. Geographic Information – Services. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39890

[15] ISO/TS 19139:2007. Geographic Information – Metadata – XML schema implementation. http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557

[16] F.J. Lopez-Pellicer, A.J. Florczyk, J. Nogueras-Iso, P.R. Muro-Medrano and F. J. Zarazaga-Soria. Exposing CSW Catalogues as Linked Data. In *Geospatial Thinking*, pp. 183-200, 2010.

[17] Mimas Linked Data Project (UK). http://mimasld.wordpress.com/

[18] Open Geospatial Consortium (OGC). Catalogue Service. http://www.opengeospatial.org/standards/cat

[19] OGC Geography Markup Language Encoding Standard, 2007. http://portal.opengeospatial.org/files/?artifact_id=20509

[20] OGC GeoSPARQL Standard - A Geographic Query Language for RDF Data, 2012. https://portal.opengeospatial.org/files/?artifact_id=47664

[21] OpenStreetMap project. http://www.openstreetmap.org/

[22] K. Patroumpas, M. Alexakis, G. Giannopoulos, and S. Athanasiou. TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples. In *LWDM*, pp. 275-278, 2014.

[23] A. Perego. Inspiring Data? Cross-domain Interoperability for EU Spatial Data. In *Using Open Data Workshop*, Brussels, Belgium, June 2012.

[24] J. Reid, W. Waites, and B. Butchart. An Infrastructure for Publishing Geospatial Metadata as Open Linked Metadata. In *AGILE*, 2012.

[25] S. Schade and M. Lutz. Opportunities and Challenges for using Linked Data in INSPIRE. In *Workshop on Linked Spatiotemporal Data*, 2010.

[26] S. Tschirner, A. Scherp, and S. Staab. Semantic access to INSPIRE – How to publish and query advanced GML data. In *Terra Cognita*, pp. 75-87, 2011.

[27] L.M. Vilches-Blázquez, B. Villazón-Terrazas, V. Saquicela, A. de León, O. Corcho, and A. Gómez-Pérez. GeoLinked Data and INSPIRE through an Application Case. In *ACM SIGSPATIAL GIS*, pp. 446-449, November 2010.

[28] W3C. Data Catalog Vocabulary (DCAT). http://www.w3.org/TR/vocab-dcat/

[29] W3C. OWL Web Ontology Language Overview. http://www.w3.org/TR/owl2-overview/

[30] W3C. Resource Description Framework 1.1. http://www.w3.org/TR/rdf11-new/

[31] W3C. SPARQL 1.1 Query Language for RDF. http://www.w3.org/TR/sparql11-query/

[32] W3C. VoID Vocabulary (3/3/2011). http://www.w3.org/TR/void/

[33] W3C. XSL Transformations (XSLT). http://www.w3.org/TR/xslt

[34] Wikimapia. http://wikimapia.org

[35] Zaragoza municipality SPARQL endpoint. http://www.zaragoza.es/datosabiertos/sparql