applied sciences

MDPI

Article

Smart Environmental Data Infrastructures: Bridging the Gap between Earth Sciences and Citizens

José R.R. Viqueira [1,*], Sebastián Villarroya [2], David Mera [1] and José A. Taboada [1]

1 COGRADE, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 - Santiago de Compostela, Spain; david.mera@usc.es (D.M.); joseangel.taboada@usc.es (J.A.T.)

2 Computer Science & Electrical Engineering, Jacobs University Bremen, 28759 - Bremen, Germany; s.villarroyafernandez@jacobs-university.de

* Correspondence: jrr.viqueira@usc.es;

Received: 20 November 2019; Accepted: 20 January 2020; Published: 25 January 2020

check for updates

Abstract: The monitoring and forecasting of environmental conditions is a task to which much effort and resources are devoted by the scientific community and relevant authorities. Representative examples arise in meteorology, oceanography, and environmental engineering. As a consequence, high volumes of data are generated, which include data generated by earth observation systems and different kinds of models. Specific data models, formats, vocabularies and data access infrastructures have been developed and are currently being used by the scientific community. Due to this, discovering, accessing and analyzing environmental datasets requires very specific skills, which is an important barrier for their reuse in many other application domains. This paper reviews earth science data representation and access standards and technologies, and identifies the main challenges to overcome in order to enable their integration in semantic open data infrastructures. This would allow non-scientific information technology practitioners to devise new end-user solutions for citizen problems in new application domains.

Keywords: smart data; semantic web; environmental data; geospatial data; linked data; semantic integration; open data

1. Introduction

Many applications require detailed knowledge about environmental conditions to incorporate in different decision-making tasks. An example in the field of public health is the prediction of the impact that environmental variables such as sea and air temperature, rainfall and humidity have on the behavior of diseases such as Cholera [1] and Influenza [2]. Other examples include the prediction of landslides from rainfall and soil conditions [3] and the study of the interactions between atmosphere and biosphere, such as heat and hydrological transfer and their impact on agriculture [4]. Apart from the above applications in science and engineering, environmental conditions are very important also in decision making related to the daily activities of citizens. Examples of this are the impact of weather, sea conditions and air pollution in many open-air activities such as running, hiking, yachting, surfing, sightseeing, and sea bathing. More precisely, the air quality information generated in the scope of the TRAFAIR project (http://trafair.eu/) may be used by running and sightseeing applications to help in the elaboration of relevant routes inside the cities. Similarly, the sea surface conditions, including currents and wave heights, generated by the high-frequency radar (HF Radar) infrastructure of the project RADAR-ON-RAIA project (http://radaronraia.eu/) may be incorporated in decision support for yachting and surfing applications. In general however, to leverage the existence of environmental datasets in such applications, currently available open data infrastructures should enable IT practitioners to discover and access them in an effective and efficient manner. Unfortunately, as it is illustrated in Figure 1, the models, standards and technologies involved in the construction of the geospatial and environmental data infrastructures used by scientific and engineering applications, are not aligned with the semantic technologies on which general purpose open data infrastructures are based.

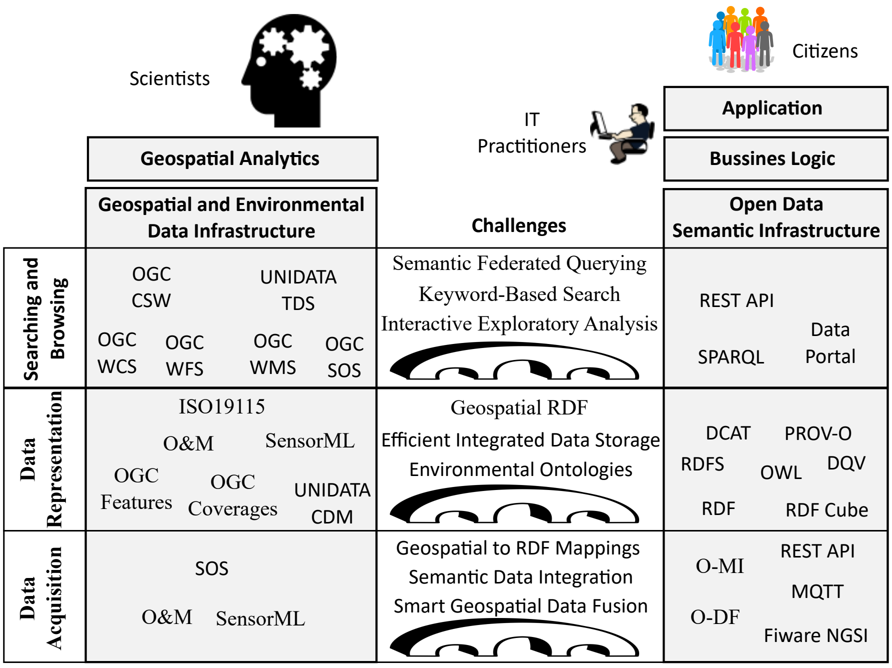


Figure 1. Research challenges to bridge the gap between geospatial and conventional open data infrastructures.

If we focus on the data acquisition layer in Figure 1, both environmental and conventional applications incorporate data generated by sensors, models and people. The main difference among them is that environmental data use to have a spatio-temporal component that is much more complex than that of conventional data. In particular, the data that is required for environmental management tasks is collected and generated at a large scale by different authorities and geoscientific communities. Conventional observation systems include large and heterogeneous networks of static in-situ environmental stations (such as meteorological and air quality stations and sea buoys), mobile in-situ platforms such as radiosondes, gliders, and drifters, static remote sensing systems such as meteorological and coast high-frequency radars, and mobile remote sensing systems on board aircraft and satellites. Besides, environmental models help both in understanding related physical processes and in estimating past, present and future environmental conditions [5]. All the above-advanced data acquisition infrastructures and models generate data with complex spatio-temporal structure and semantics. Due to this, specific data and metadata models and service interfaces have been proposed to achieve integrated data acquisition from such complex and heterogeneous data sources (see the bottom left part of Figure 1). Contrary to the above, conventional data acquisition infrastructures as those used in internet of things (IoT) and smart city applications rely on more simple models and interfaces in which the geospatial component restricts to basic geometries such as points and where provenance metadata is not generally provided (see the bottom right part of Figure 1).

Regarding the data representation layer in Figure 1, many research works in the areas of Geographic Information Systems (GIS) and Spatial Data Management were dedicated to the definition of geospatial data models [6,7]. Research solutions led to the proposal and adoption of standards, first of the Open Geospatial Consortium (OGC) and now also of ISO (see the center-left part of Figure 1). Two main types of geospatial data entities have been identified, namely, features and coverages. A feature [8] is an entity of a given application domain that may be described by properties of conventional and geospatial data types [9], which include points, lines, and surfaces represented with vectors of coordinates (vector geometric representation). Examples of features with geospatial properties are cities, rivers, roads, and municipalities. A coverage [10] is a special type of feature whose conventional properties may vary along its spatial, temporal or spatio-temporal domain. An example of a spatio-temporal coverage is the variation of air temperature, wind speed, wind direction and rainfall through a given geospatial region and during a given period of time. The most common discrete representation of the domain of spatial coverages, called raster representation, is based on a regular square grid or tessellation. In addition to the above geospatial data types, metadata standards were also defined, both at the level of generic geospatial dataset level metadata (ISO 19115) and at the level of provenance environmental observation metadata [11]. Contrary to this, conventional open data infrastructures rely on semantic web data and metadata representation standards (see the center-right part of Figure 1). Despite the existence of some basic standardized solutions for the incorporation of geospatial vector data in semantic web technologies [12], and for the definition of ontologies on the environmental data management domain [13,14], in general, the support for geospatial data representation and environmental data semantics in those technologies are still far from being satisfactory.

Finally, the searching and browsing layer in Figure 1 supports data discovering and access through OGC and ISO standard web services (as it is shown in the top left part of Figure 1). Those standards enable metadata searching [15], feature and coverage data access [16,17] and server-side map rendering [18] for geospatial visualization and browsing. On the other hand, conventional open data infrastructures rely on either simple open data formats published through specific REST APIs or data portals, or on more complex data query solutions based on the use of the SPARQL Protocol and RDF Query Language (SPARQL) [19]. In any case, despite the existence of a geospatial extension of SPARQL [12] that supports only geospatial vector feature data, in general, the complexity of the spatio-temporal structures and semantics found in environmental data sources is not supported by currently available conventional open data infrastructures.

This paper reviews the state of the art related to geospatial and environmental semantic data infrastructures, which include: i) approaches for environmental data acquisition, especially focusing on data integration problems, ii) technologies and approaches for semantic data representation and storage and iii) solutions for searching and browsing semantic geospatial and environmental data. Based on the characteristics of the existing solutions, the main contribution of the paper is the identification of challenges for new research activities required to bride the gap between environmental and semantic infrastructures, which would enable the construction of smart environmental data infrastructures (SEDIs). In general, the existence and appropriate management of high-quality environmental metadata is a key prerequisite to enable such new smart environmental data solutions. Specific research challenges for each software layer are shown in the center column of Figure 1. More precisely, main challenges regarding smart data acquisition and semantic integration are related to the appropriate mapping between geospatial data models and RDF, the automatic discovery of data integration knowledge and the smart fusion of heterogeneous vector-raster datasets. Regarding data representation and storage, new research efforts should be directed to the support of raster coverages in the RDF model, to the efficient integrated storage of vector and raster data and to the formalization of a generic top-level environmental observation and modeling ontology. Finally, smart data searching and browsing demands new research for the development of efficient vector-raster federated query processing, effective and efficient semantic keyword-based search technologies for data discovery and efficient query processing approaches for interactive exploratory analysis workloads.

It is estimated that these new generation SEDIs will leverage semantic technologies and existing metadata to achieve more effective data discovering, access and browsing and to enable better semantic integration of heterogeneous data sources. Besides, it is estimated that the use of general-purpose semantic open data infrastructure technologies will bring SEDIs closer to Information Technology (IT) practitioners, which will have the opportunity to incorporate high-quality environmental data in new added-value products and business models from which citizens will also benefit, as it is illustrated in the top right part of Figure 1.

The remainder of the paper is organized as follows. Section 2 describes a general architecture for SEDIs. State of the art related to the smart representation and storage of geospatial and environmental data is described in Section 3, together with the identification of relevant research challenges. Section 4 reviews classical and modern environmental data sources, describes solutions for semantic data integration and fusion and identifies relevant research challenges towards smart data acquisition and integration. Current technologies and approaches for the discovering, searching and browsing of geospatial and environmental datasets are described in Section 5, and research challenges towards smart data searching and browsing are also identified. The motivation for all the identified research challenges is summarized and discussed in Section 6. Finally, Section 7 concludes the paper.

2. System Architecture

A software architecture organized in three layers (three-tier architecture) is a classic design pattern that is still of common use in current systems [20]. At the bottom of the architecture, a data layer is responsible for all the required data management tasks, which include data and metadata storage, data access, query processing, and transaction management. In the middle of the architecture, an application layer, also called business logic layer, implements all the processes that define the application functionality. The architecture is completed with a presentation layer at the top, which enables the friendly interaction with the users. The adoption of the above architectural design pattern in the area of GIS is also widespread. Thus, for example, the generic architecture for GIS proposed in [21] describes the main structure and functionality that the above three layers should have in GIS data browsing applications.

Another example is the OpenGIS Service Architecture [22] proposed by the OGC, which has also been adopted by ISO (ISO 19119:2016). The above standard classifies geographic services into six categories, namely, human interaction, model/information management, workflow/task, processing, communication, and system management. The standard shows also how the above types of services fit the different layers of a multi-tiered architecture. Thus, the data layer is composed of model/information management services, the application layer includes processing and workflow/task services, the presentation layer has human interaction services, the communication services enable the interaction between the layers and system management services are transversal to the architecture.

To enable the interoperable interchange of geospatial data and metadata in Europe, the INSPIRE directive [23] proposes the use of network services. The most important ones are those related to the data layer, and they include the following. Discovery services that enable the searching of metadata to find both datasets and services. View services that enable the displaying and navigation of geospatial datasets. Download services that enable obtaining parts of geospatial datasets in standard formats.

A specific software architecture for an environmental observatory information system has been proposed in [24]. General components are proposed for observation and communication, i.e., sensor data acquisition, data and metadata storage, data quality and provenance, publication and interoperability and discovery and presentation. The functionality of the above components is mainly related to that of a data Layer, except the discovery and presentation component that deals with human interaction at an application layer.

Based on the context described above, the SEDIs considered in this paper fit completely inside the data layer of a three-tier geospatial software architecture. As it is shown in Figure 2, the architecture is composed of three main components that enable data acquisition from different sources of earth observation and modeling, data storage and data searching and browsing.

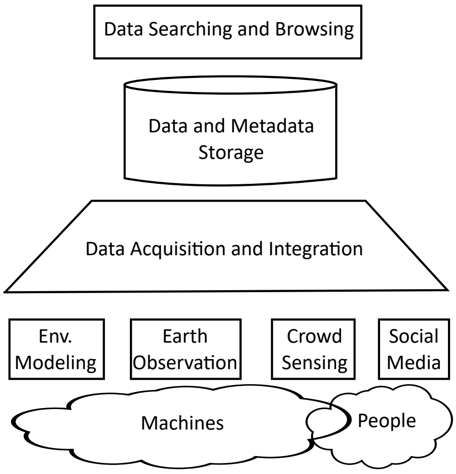


Figure 2. Main components of an environmental data infrastructure.

Data acquisition and integration (Section 4). Nowadays, real-time data of what is happening on the Earth's surface may be obtained by infrastructures that are based on machines and/or people. Since long ago, different types of sensing and computing infrastructures have been used to perform environmental observation and modeling. Models are available in many scientific domains such as hydrology, meteorology, oceanography, ecology, etc. They generate large amounts of numerical estimations of many variables related to the past, present and future of different physical systems. Many Earth observation infrastructures are currently deployed along the surface of our planet, including many networks of in-situ stations with different types of sensors and many remote sensing complex systems and campaigns. Recently, the broad proliferation of the use of mobile devices connected to the Internet has opened new opportunities to complement the above Earth observation means by people supported sensing. Generally, this kind of new sensing may have the form of either crowdsensing applications that make use of sensors already installed or connected to the mobile devices or virtual sensors that extract useful information from social media. The main challenge to be overcome by this component is the effective integration and fusion of the above highly heterogeneous types of data sources, taking into account both syntactic and semantic conflicts.

Data and metadata storage (Section 3). Various data modeling and representation approaches have been proposed for geospatial and environmental data in the areas of GIS and spatial and environmental data management. As a consequence, well-established solutions exist for both vector features and coverages, and relevant standards are implemented by current tools. General models for the representation of geospatial metadata at different levels of the dataset structures have also been proposed, including specific models for sensor observation semantics. The adoption of semantic web technologies for geospatial and environmental datasets have also been investigated, leading to some general-purpose solutions for representation and storage. More specific ontologies and vocabularies have also been designed that foster interoperability between disconnected data silos. The main research challenges in the scope of this component are related to the effective and efficient incorporation of raster coverage representation and storage within semantic web ecosystems.

Data searching and browsing (Section 5). Efficient geospatial data querying has been a topic of research in the area of spatial data management since more than 30 years ago and it is supported in current spatial databases. Additionally, relevant mature OGC web service interfaces exist for filtering over datasets of either vector features or raster coverages. Catalog services enable geospatial data discovery based on standard metadata models. Web map services enable the server-side rendering of maps for geospatial dataset browsing. Regarding semantic web technologies, federated querying is supported for vector entities through GeoSPARQL endpoints.

Main open issues and challenges for research are related to the integration of very large datasets of vector and raster data during efficient federated querying, effective and efficient keyword-based searching and efficient query processing for interactive exploratory analysis workloads.

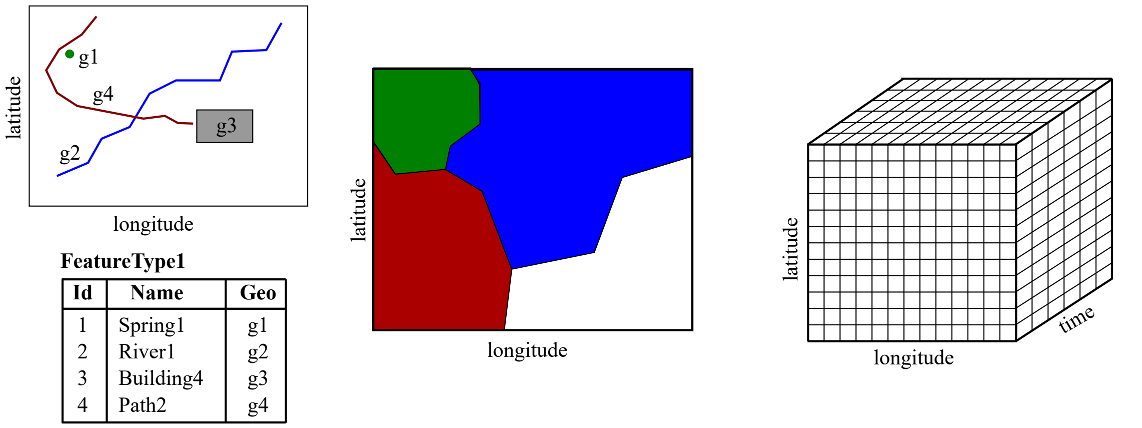
3. Data and Metadata Representation and Storage

First geospatial and environmental data representation and storage solutions are described and next related semantic solutions are discussed. Based on the above descriptions, relevant challenges for future research are identified.

3.1. Geospatial and Environmental Data Modeling and Storage

The representation of geospatial data has been a topic to which much attention was paid in the areas of GIS and spatial data management [6,7]. Two main data modeling approaches were proposed. A first approach considers space as populated by objects or entities. Such geospatial entities may have both conventional (alphanumerical) and geospatial properties, which define their location and shape. Geospatial properties are represented with geometric approximations based on vectors of 2D or 3D pairs of coordinates (vector representations). The second approach represents directly the properties of each point of space, by considering different kinds of regular or irregular discrete spatial grids. Given that the above approaches fit better different types of geospatial datasets, both are currently supported by the data modeling framework defined by OGC and adopted by ISO.

The reference model defined by OGC in [8] (also ISO 19101) provides metamodel artifacts to represent geospatial entities (called features in OGC notation), entity types (feature types) and entity properties (attributes). Rules for the creation of application models following the above reference framework are given in the ISO 19109 standard. Figure 3a depicts the geometric representation and the values of the conventional properties of four features of a feature type named "FeatureType1". Properties of space are represented with coverages in the OGC modeling framework [10] (also ISO 19123). A coverage is a special type of feature whose property values vary along its spatial, temporal or spatio-temporal extension. More precisely, a coverage is defined as a collection of functions (one for each conventional property) with a common domain (spatial, temporal or spatio-temporal) and with ranges of conventional data types. Different types of coverages, depending on the representation of its domain are supported by the above standard. Discrete Coverages has a domain composed of discrete zones where the values of the properties are constant. An example of a Discrete Spatial Coverage with four zones is depicted in Figure 3b. On the other hand, the progressive variation of a property (such as temperature, elevation above sea level, etc.) along space, time or both is represented with a continuous coverage. Five different types of continuous spatial coverages are considered, depending on the discrete approximation used for its continuous domain, namely, Thiessen polygon coverages, triangulated irregular network (TIN) coverages, square grid coverages, hexagonal grid coverages and curve segment coverages (its domain is a line feature). By far, the most common continuous representation used for coverages is the one based on square grids, which are called raster coverages. Figure 3c illustrates a raster coverage whose domain is composed of two spatial dimensions and a temporal dimension. Recalling the above discussion on the two data modeling approaches, the same spatial phenomena may be represented either as a collection of features or as a discrete coverage. This is the case for example of a discrete coverage of the soil type, which may also be represented as a collection of features (soil type zones). Beyond that, depending on the properties of interest, a municipality may be represented either as a feature with a property representing its total population at a given moment, or as a continuous coverage that records the continuous variation of its population density along its spatial extension.



(a) Geographic Features (b) Discrete Spatial Coverage (c) Spatio-temporal Raster Coverage

Figure 3. Illustration of geospatial features and coverages.

The OGC model that defines the representations for geospatial properties of Features [9] (also ISO 19107) contains vector-based geometric and topological primitives, which may be combined to construct complex structures. Geometric primitives include points, lines, polygons, and 3D solids. Homogeneous and heterogeneous collections of primitives, called aggregates, are also supported. Geometric complex geometries are collections of primitives, but contrary to the aggregates they form a single connected geometry. Topological primitives represent nodes, arcs, and faces of geospatial networks. Besides the representation of spatial network topologies, they also enable the representation of the zones (faces) of discrete coverages. The OGC Geography Markup Language (GML) [25] (also ISO 19136) provides an XML encoding for collections of Features that implements the above data modeling framework.

An important profile of the above standard due to its widespread use in different types of geospatial tools is the simple feature access (SFA) implementation specification [26]. As it is shown in Figure 4, this profile defines three types of geometric primitives, namely, Point, LineString, and Polygon. A Point is represented with a tuple of real coordinates. A LineString is an approximation of a curve represented with an ordered sequence of points, using linear interpolation between consecutive vertices. Polygons are represented with an exterior LinearRing (circular LineString whose start and end points are identical) and possibly with a collection of interior holes, represented also with LinearRings. In addition to the basic primitives, SFA supports also collections of geometries (GeometryCollection). Heterogeneous collections are direct instances of type GeometryCollection, whereas specific subtypes (MultiPoint, MultiLineString and MultiPolygon) are defined for homogeneous collections of either of the three basic types. Finally, an abstract type Geometry enables the representation of any of the above types in heterogeneous Feature collections like the one in Figure 3a. The above standard defines two encodings for geometric objects, a text encoding called well-known text (WKT) encoding and a binary encoding called well-known binary (WKB) encoding. Besides, SFA is also the ground for the GeoJSON encoding [27] of geospatial features and for most of the current implementations of spatial DBMSs and toolkits.

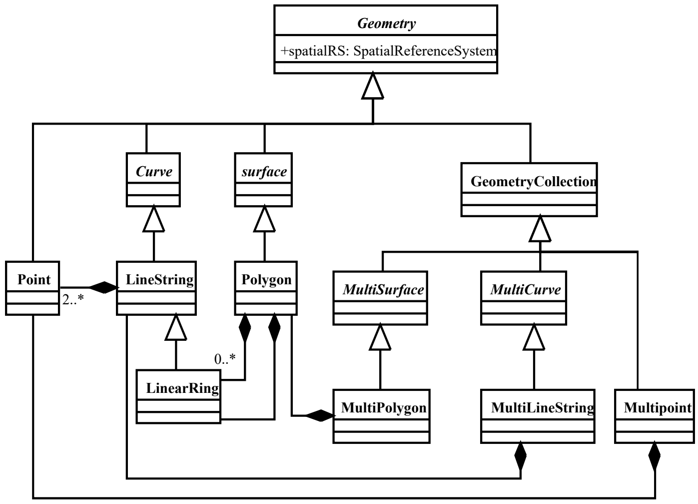


Figure 4. Geometric types in Open Geospatial Consortium (OGC)-simple feature access (SFA) implementation specification.

Geographic Coverages may also be encoded in GML, but as it is also the case for Features, a simplification of the model that considers other encodings has also been defined in the context of OGC activities [28]. The most important data encodings are the GeoTIFF format, widely used in the geospatial domain and the NetCDF format, which is very well known among environmental experts. While GeoTIFF restricts to 2D spatial coverages, the data model supported by NetCDF goes much further. NetCDF and its data model, called common data model (CDM), have been proposed by UNIDATA (https://www.unidata.ucar.edu/), a community of research and education entities whose objective is the sharing of geoscientific data and accessing and visualization tools. As it is shown in Figure 5, a dataset of the CDM has a tree of groups. Each group may have a collection of variables. Each Variable is a multidimensional geoscientific array of a specific data type. The dimensions of each variable are also defined in the same group and they may be shared by multiple variables. A group might also contain enumerations that may be used as data types of variables and attributes that are used to record metadata (pairs key-value) at the level of group or variable. One dimensional coordinate variables may be used to record the values of dimensions. It is noticed that the multidimensional CDM model of UNIDATA resembles the conventional multidimensional models used in data warehouses. A main difference is the specific support of CDM for sampling dimensions over space and time.

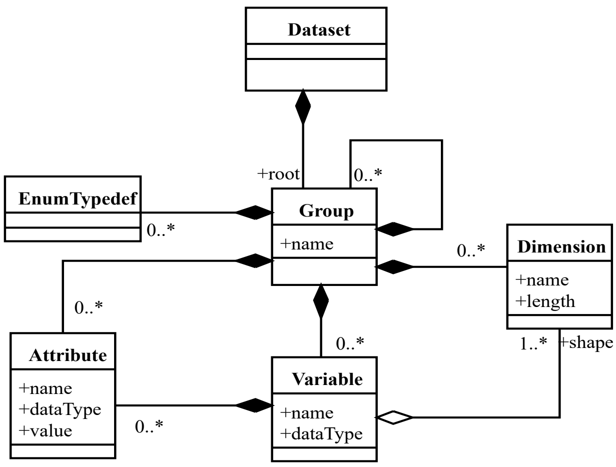


Figure 5. Main structures of the UNIDATA common data model (CDM).

Regarding data storage, although some general-purpose scientific array data managers have already been developed [29,30] their use is still not widespread among the scientific communities, which still rely on the direct recoding of coverages in files with formats such as GeoTIFF and NetCDF.

As it was already stated in the introduction, to achieve smart datasets and smart data infrastructures, a key issue is the availability of high-quality metadata. The representation of metadata for geospatial datasets and services is also standardized through the ISO 19115 Geographic metadata standard. The metadata specified by this standard includes descriptions of the identification, spatial and temporal extension, data quality and provenance (lineage), other spatial and temporal characteristics, contents, visualization, distribution, etc. Identification metadata includes an identifier, title, abstract, date, organization or person responsible, geographic and temporal extents, languages used, information about scale and resolution, URL for on-line access, keywords, access restrictions, etc. The standard includes both mandatory and optional metadata elements, enabling the specification of a minimum set of metadata for a broad range of application domains, but also providing an extension mechanism to overcome specific requirements.

Metadata related to provenance and observed features and properties are particularly important to enable the correct interpretation of environmental sensor observation data. These metadata go beyond the dataset level metadata represented by ISO 19115, thus specific data models are required. Therefore, various designs have been proposed in the environmental data management research domain. As an example, the observations data model (ODM) [31] has been developed as part of the data layer of the hydrological information system (HIS) of the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) [32]. This model enables the recording of hydrological observed data values together with metadata that includes the sites where they were measured, the variables that were measured and the methods used to perform the measurement. The VOEIS (Virtual Observatory and Ecological Information System) observation data model (VODM) is based on ODM and it has been designed to represent observation data in the VOEIS Data Hub [33]. VODM extends ODM with data streams and dataset level metadata. ODM has also been extended and generalized in [34] to deal with discrete Earth observations. In [35], the authors review more than 40 systems for water management and propose a new water management data model (WaMDaM).

In parallel with the above research initiatives, and aligned with them, the OGC has been defining also its own general-purpose data model for environmental observation data. The general structure of this model, called Observations and Measurements (O&M) [11] (also ISO 19156), is depicted in the UML class diagram of Figure 6. As it is shown in the figure, the central class of the model is the OM_Observation, which represents and observation produced at a given resultTime by a given observation process (OM_Process), generally based on some sensor system. The observation provides a value (result) for an observed property (GF_PropertyType) of a given feature of interest (GF_Feature) that applies at a given phenomenonTime. Besides, an observation may also have parameters (key-value pairs that provide additional context information), general metadata and specific data quality metadata. A specific subtype of features is defined by the standard to represent sampling features (SF_SamplingFeature), i.e., intermediate features related to the sampling method from which data has to be recorded. When the location of this sampling feature is important then what we have is a spatial sampling feature (SF_SpatialSamplingFeature). Examples of spatial sampling features are the elements of a network of sampling stations, which are platforms where the processes (sensors) are hosted. The sampling feature is related to the ultimate feature that is being sampled, in our example of stations, the whole geographic area where the stations are located. The model supports also sampling methods based on the collection of specimens (SF_Specimen) that are next observed in a laboratory. The observations are classified with respect to the data type of their result value, including simple data types such as real, string, boolean and integer, complex data types constructed from records of simple types and even complex Coverage observations, including temporal and spatial coverages. It is finally noticed that the model is general enough to support any kind of observation processes that observe any kind of application-dependent feature of interest. In particular, they include the results of in-situ observations such as the one in the above example of stations and remote sensing that generates spatial coverages. The sensing devices may be hosted onboard static and mobile platforms as will be discussed in the next section.

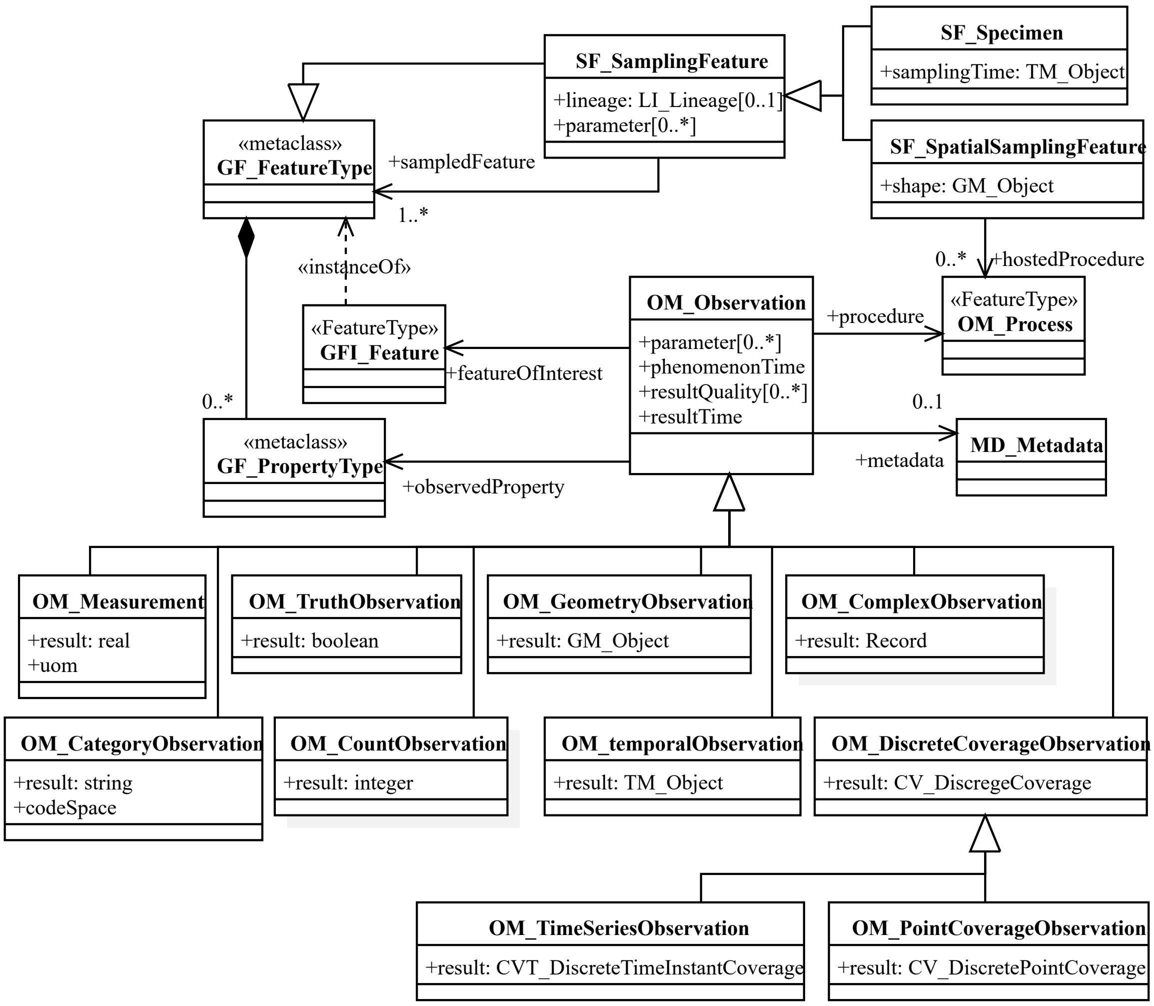


Figure 6. Main structures of the OGC observations and measurements (O&M) data model.

3.2. Geospatial Smart Representations

The core of the smart data representations used by semantic technologies is the resource description framework (RDF) model [36]. Resources in RDF are identified with universal resource identifiers (URIs) and are described in terms of simple properties and values, which are also identified and represented with URIs. RDF expressions are triples of the form (subject, predicate, object), used to represent things (subject), properties of things (predicate) and values of those properties (object). Such values may also be identifiers of other resources, enabling this way the construction of complex knowledge graphs where things and values are nodes and properties are arcs. The use of standard text encodings of RDF graphs such as RDF XML and RDF Turtle eases the open interchange of RDF between systems.

The vocabularies used in RDF statements (object and subject URIs) may be defined in RDF with RDF Schema primitives. Those primitives enable the definition of classes of objects using RDF Schema properties rdf:type (to state that an object is an instance of a specific class) and rdfs:subClassOf (to define hierarchies of classes). The domain and range class (or data type) of each property may also be defined in RDF Schema, which also enables the definition of hierarchies of properties and sub-properties. The Ontology Web Language (OWL) [37] extends RDF Schema with advanced class relationships that include property restrictions expressed with universal and existential quantifiers and advanced property relationships such as inverseof.

RDF data and metadata (ontologies) may be accessed directly through the web via HTTP and web services. However, to enable the declarative querying of RDF graphs in RDF stores, W3C has defined a set of specifications of languages and protocols called SPARQL Protocol and RDF Query Language (SPARQL) [19]. The GeoSPARQL [12] geographic extension of SPARQL provides an RDFS/OWL vocabulary that is based on the OGC SFA data model and it contains the following elements: (i) A class geo:Feature to represent OGC Features with geometry, (ii) a data type geo:wktLiteral to represent geometric literals encoded in OGC WKT, (iii) a class geo:Geometry to represent feature geometries, (iv) standard properties geo:hasGeometry and geo:hasDefaultGeometry to link each Feature to its geometric property values and (v) standard property geo:asWkT to link a geometry with its WKT encoding of geo:wktLiteral data type. Support for continuous raster coverages with semantic technologies has not still been developed and incorporated into relevant standards.

Many semantic data storage technologies have been developed to support the recording of RDF graphs. More specifically, there are implementations that support GeoSPARQL RDF graphs [38]. Metadata related to semantic dataset schemes are naturally represented with OWL ontologies. Other dataset level metadata may be represented in RDF using the Data Catalog Vocabulary (DCAT) [39] and its geospatial extension GeoDCAT-AP. Specific ontologies have already been proposed by the W3C to represent data quality and data provenance metadata. The W3C Data Quality Vocabulary (DQV) [40] represents five different types of quality information of a dataset: (i) a quality annotation that gives feedback and quality certificates about the dataset, (ii) a data quality standard to which the dataset conforms, (iii) a data quality policy or agreement, (iv) a data quality measure that may be either qualitative or quantitative and (v) an entity involved in the dataset provenance and distribution. Data quality metrics may be given for different data quality dimensions (for example its availability). Provenance data describes the entities involved in the generation of a dataset, which may give insight about its quality and reliability that must be taken into account during data fusion and analytic processes. The W3C PROV Ontology (PROV-O) [41], at a high level of abstraction, represents how different agents may be involved in the generation of entities through the execution of different activities. Some research activities already report the use of PROV-O in the scope of environmental data management. Thus, for example, in [42] PROV-O is used to design an information model that represents the provenance information related to a climate assessment report, in the scope of climate change analysis. In [43], the gaps between ISO 19115 lineage model and PROV-O are analyzed to undertake a PROV-O extension with ISO 19115 based semantics, enabling this way PROV-O to be effectively used in geospatial application domains.

Beyond the above and other general-purpose top-level ontologies, more specific ontologies that may be used as top-level models and vocabularies for environmental applications have already been designed. One such ontology is the semantic sensor network ontology (SSN) [13,44], which includes elements to describe sensing systems and their observations. The lightweight core of SSN is the sensor observation, sample, and actuator (SOSA) ontology. This ontology may be used to represent datasets that follow simple OGC O&M models, but it may be extended to support its whole expressiveness [45]. Additionally, SOSA enables the representation of complex sensor systems (see Section 4) and provides concepts to model the actuation mechanisms that are needed in IoT application domains.

Many vocabularies have been designed in the scope of different geoscience application domains. A first example is the Semantic Web for Earth and Environmental Terminology (SWEET) [14], which is a collection of ontologies constructed in NASA with the broad scope of Earth science. It consists of 11 primary ontologies that contain a total of several thousand concepts related to natural phenomena, human activities, physical processes and properties, units, etc. Efforts to design and integrate vocabularies in more specific application domains have been developed. Some examples may be found in meteorology (http://cfconventions.org/index.html), oceanography [46], environment (https://www.eionet.europa.eu/gemet/), geology [47] and mineral exploration [48].

One step beyond the definition of simple vocabularies and terminologies is the incorporation of complex constraints and knowledge rules that may be expressed in the Semantic Web Rule Language (SWRL) [49], a rule language that combines the representation power of OWL with the reasoning power of RuleML. Such knowledge may next be used to perform inferences to achieve for example complex data validation [50] and advanced decision support tools [51].

To summarize, the smart representation of geospatial vector features in RDF is currently supported by the GeoSPARQL ontology [12]. However, such ontology is not designed to support efficient spatial and spatio-temporal raster coverages. Only the metadata of those coverages may be efficiently managed [52]. An initial attempt to support array data representation and querying with semantic technologies is reported in [53], where arrays are represented with a new data type as RDF collections of numbers. This is somehow similar to the use of raster data types in current spatial DBMSs. The main problem of this approach is that it assumes a complex nested model where arrays are represented as data elements of the RDF model. Therefore, to be able to manipulate array elements, the user must use specific array operators of the relevant data type. Contrary to the above approach, the multidimensional model defined in [54] supports vector feature collections and spatial and spatio-temporal raster coverages with a unique and simple primitive data structure (a data cube called MappingSet). The RDF Cube Vocabulary [55] enables the representation of multidimensional data cubes in RDF, however it does not support the sampling spatial and temporal dimensions required to achieve efficient representations for spatio-temporal coverages. Therefore, research is still needed to reach efficient integrated smart data representation and storage technologies for vector and raster geospatial data warehouses. Furthermore, in spite of the existence of general purpose ontologies like SSN [13,44], a complete ontology that supports the representation of any kind of environmental observation (including in-situ and remote sensing) and modeling processes has not still been designed.

Based on the above, the following research challenges, which are graphically illustrated in the information architecture of Figure 7, should be overcame to achieve effective smart geospatial and environmental data representations and efficient relevant storage technologies.

Challenge 1: geospatial RDF representation. An effective RDF representation that enables the integrated modeling of vector features and raster coverages in heterogeneous geospatial data warehouses has to be defined. Such representation should treat vector features and raster coverage cells as first class citizens, to ease the integrated querying of both types of structures.

Challenge 2: efficient integrated semantic vector-raster storage. Separate data storage and access technologies already exist for vector features and raster coverages, and many datasets are currently stored with those technologies. However, the efficient combined access to vector and raster geospatial data may only be achieved if new integrated data storage and access technologies are developed.

Challenge 3: environmental observation and modeling ontology. A general top-level ontology must be designed to support the representation of all the data and metadata involved in any type of environmental data generation process, which include in-situ and remote environmental sensing and also environmental modeling. That ontology should be aligned to top-level ontologies like SSN [13,44] and should enable the incorporation of existing vocabularies such as the climate and forecast (CF) conventions.

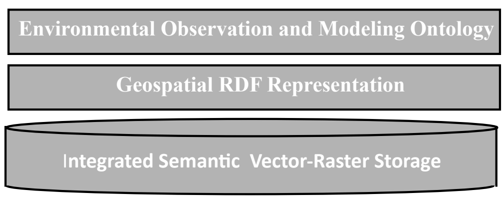


Figure 7. Illustration of challenges towards smart environmental data and metadata representation and storage.

4. Data Acquisition and Integration

This section analyzes currently available environmental data sources, including both traditional observation and modeling systems and modern crowdsourcing approaches, and discusses current solutions and future challenges to achieve smart environmental data integration.

4.1. Earth Observation and Modeling Data Sources

Two main types of sources of data about our environment exist in relevant applications, namely, observation data sources and modeling data sources. Observation data sources are based on different kinds of sensing processes that obtain values of relevant properties over entities of interest. A data model [11] and a related ontology [13] have already been described in Section 3.

The Sensor Model Language (SensorML) [56], defined by the OGC, provides a model and an XML encoding of metadata to describe environmental sensor systems (process in OGC notation). The metadata supported for a process includes descriptions of its inputs and outputs. The key element of a simple process is its process method, which provides a description of the methodology used by the process to generate the outputs from the inputs. For example, the process method of an air quality sensor would provide a description of the electrochemical device that generates currents and the algorithms that transform currents in gas concentrations. Aggregate processes are defined as complex interconnections of components that may be in turn simple or aggregate processes. A specific type of process is defined for simple real processing devices, called a physical component. A physical system is an aggregate process that models a complex real device that may be composed of physical and software components.

If we restrict to physical components and systems, they may be classified as either in-situ or remote, depending on whether the observed entity (feature of interest) is located in the surroundings of the sensor or at some distance. Alternatively, they may also be classified as either static or mobile, depending on whether they are installed on static or mobile platforms. Examples of observations obtained from physical systems of different types are illustrated in Figure 8. Examples of static platforms with in-situ sensors are the networks of environmental stations. At the top left of the figure, the locations of a collection of stations are shown, together with the time series of values of a given property in one of the stations. An example of a mobile platform with in-situ sensors is a radiosonde, whose values at each location of its trajectory are illustrated at the top right part of the figure. It is noticed that the geospatial data generated by in-situ sensors are vector feature data. On the other hand, remote sensors use to retrieve raster coverages as values of their observations. If the platform is static, usually, all the coverages have the same spatial extent, as it is illustrated in the bottom-left part of the figure. As an example, a meteorological radar generates a time series of precipitation spatial coverages, all of them with the same spatial extent. On the other hand, spatial coverages generated by mobile remote sensors use to provide some kind of mosaic over a specific area called swath, as it is shown in the bottom right part of the figure. Examples of mobile remote sensors are those installed on-board of satellites.

Apart from the above physical devices, traditional environmental observation is many times supported by field campaigns undertaken by experts. With the proliferation of the use of smartphones, such human-based sensing started to be supported by mobile applications [57]. One step beyond this practice is to open the use of the data collection mobile applications to the citizens, moving towards a people-centric sensing approach [58], i.e., making use of mobile devices to gather information obtained from both lightweight sensors (either equipped in the device or connected to it) and people senses. Mobile sensing approaches were classified as in [59] as either personal or community. The former use personal data for personal applications such as walking and running support, whereas the latter try to monitor large-scale phenomena measured by many people. Community sensing is further subdivided into participatory and opportunistic sensing. The former requiring the active involvement of the people and the latter being more autonomous and automatic. The term mobile crowdsensing is coined in [59] to denote both participatory and opportunistic sensing approaches. Mobile crowdsensing applications are also classified in [59] according to the phenomena that they try to measure into environmental, infrastructure and social. Examples of phenomena that have already been measured in the environmental domain include air quality [60,61] and noise [62].

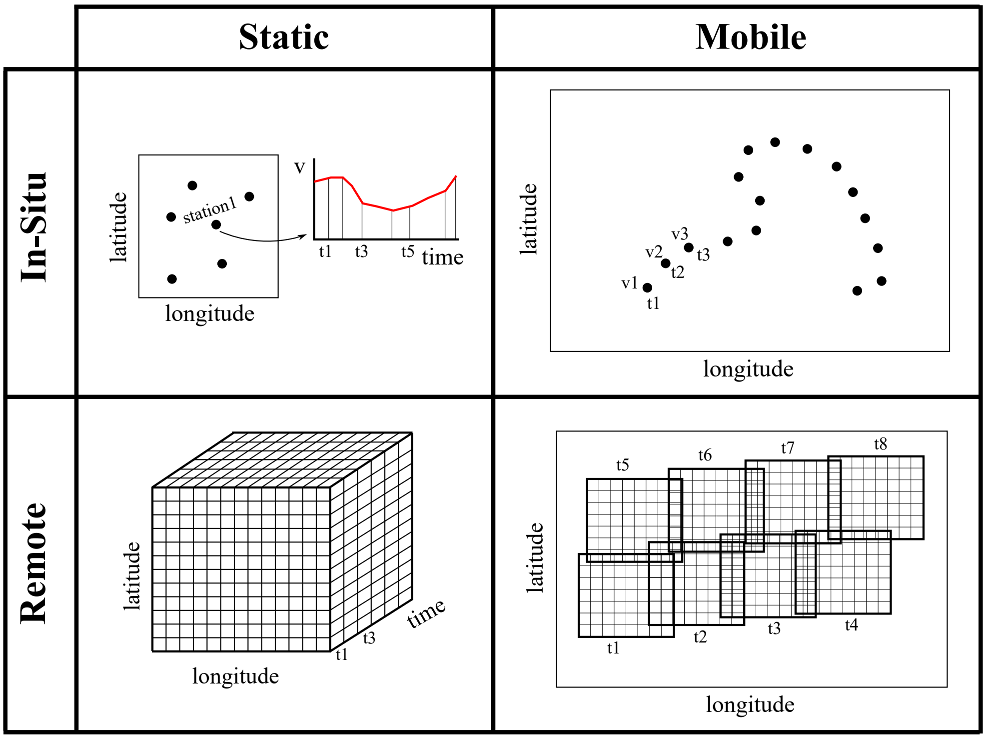


Figure 8. Observations obtained from environmental sensors of different types.

An important kind of mobile crowdsensing applications are those supported by social media analytics, due to the huge amount of data that they are able to generate on a daily basis. Text content generated by citizens, such as tweets, is analyzed with the help of natural language processing tools to extract relevant data. Such data use to include specific keywords related to the application that enables the filtering. It is also common to extract information about the feelings of the user (sentiment data). In environmental applications is very important to collect reference geographic data. Some applications like Tweeter may already provide a location for each message. However, geographic references to local and also remote locations may also be present in the text that may be extracted with specific technologies already developed in the area of geographic information retrieval [63,64], to provide appropriate geographic context related to observed entities.

Many examples of environmental mobile crowdsensing applications based on social media analysis have been reported in the literature, including sentiment analysis to estimate the impact of hurricanes [65], the construction of earthquake reporting systems [66] and the monitoring and prediction of floodings [67–69]. A survey on these kinds of environmental data generation applications classifies the approaches based on the type of data they produce [58]. Some of them produce data about the interaction between people and nature. Examples of these are those in the scope of tourism and recreation such as hiking, walking, etc. Some other applications directly generate data on nature, such as land cover description, water, and air quality monitoring, etc. A final type of applications collect data useful for planning and governance. They contribute to solve problems such as hazard preparedness, urban planning, etc. It is pointed out in [58] that social media offer huge volumes of real time data, and that current large scale data analytics technologies are ready to process them in real time, however, important challenges have to be faced. Such challenges include data heterogeneity, data quality, ethics in data collection and use and uncertainty about future data availability. One of the critical areas of work is the integration of these crowdsensing data with other more traditional sources.

Environmental modeling enables on the one hand a better understanding of environmental processes and on the other hand to generate predictions on the impacts that changes of some variables will produce on some other. Examples of the above are weather monitoring and forecast, human impact in climate change or land degradation, flooding monitoring and prediction, air quality monitoring and prediction and ecosystem analysis. Model outputs may be used to predict future conditions (forecasting), to understand better past situations (reanalysis) and also to have a better coverage of what is happing now (nowcasting). Environmental models may be classified as either mathematical or physical [5]. Physical or hardware models are versions of real world systems at an scale that enables their implementation. Examples are wind tunnels and channel flumes, and they are useful when mathematical models are not appropriate due to high complexity and/or uncertainty, or due to the lack of conceptual knowledge. Mathematical models try to express the relationships between the system variables in terms of mathematical rules. Those models may range from completely empirical models that generate mathematical functions from observed values of the variables (for example statistical or machine learning models)[70] to physical-based models that implement the mathematical rules using physical knowledge of the real world system (for example the simulation models most commonly used in weather (https://www.mmm.ucar.edu/weather-research-and-forecasting-model) and air quality [71] forecasting). The results of environmental models use to have the form of, spatial, temporal or spatio-temporal coverages.

4.2. Smart Environmental Data Integration

The main challenge to be faced to achieve the effective development of data acquisition systems for environmental applications is the very high heterogeneity of the different sources of data described in the previous subsection. The specification and use of standard models and interfaces ease the data integration processes, and therefore, efforts in that direction have been done in the environmental domain, with the O&M data model [11], the SensorML language [72] and the transactional part of the sensor observation service (SOS) [73], among others. The same problem arises also in IoT [74] and Smart City [75] applications. Therefore, similar efforts have been undertaken in these areas to reach standard interfaces and models. Representative examples are the open messaging interface (O-MI) and open data format (O-DF), proposed by The Open Group, the MQ Telemetry Transport (MQTT), which is a standard of both OASIS and ISO, and the NGSI RESTful API of the FIWARE Context Broker.

The simplest data integration solutions provide only syntactic integration [76] of the data sources, enabling uniform access to heterogeneous model paradigms and formats. Thus, it is common to generate RDF data from stored relational data, since many datasets are already available in relational format. This may be done to query directly relational datasets with SPARQL or to implement ETL processes that load relational data into RDF data stores. To ease these tasks, W3C has defined a direct mapping between relational data to RDF [77], where each tuple of each relation is automatically mapped to a set of RDF triples using the database catalog to obtain names for classes, properties, and individuals. In case a target RDF vocabulary is already available, the direct mapping cannot be used. For those cases, Relational Database to RDF Mapping Language (R2RML) [78] enables the definition of highly customized RDF views over relational data through the specification of user defined mappings. The resolution of semantic conflicts is the responsibility of the programmer of either ETL processes in data warehouse architectures [79] or data access wrappers in federated architectures [80] implemented with the classical mediator/wrapper pattern [81].

The semantic integration of data sources has already been identified as a challenging key functionality to be supported in advanced data infrastructures [82], including those in the scope of environmental data [83], and those in the scope of IoT [74]. To achieve this, three main tasks have to be undertaken. First, top-level general and standard ontologies must be developed to provide spaces for the integration process, and data sources must be semantically enriched. Next, semantic mappings between data source and general ontologies must be discovered, to construct the glue knowledge that enables the model based integration process [84]. Finally, reasoning mechanisms must be implemented over the integrated knowledge to solve global data access. During the last decade, the research on semantic technologies applied to e-science mostly centered on the development and alignment of ontologies and on the development of frameworks and use cases [85]. Those solutions have to be now leveraged to construct modern semantic data integration approaches. Regarding geospatial data, according to [86], new semantic data integration solutions should pay attention to the special characteristics of spatial data both during data representation and integration. Various semantic data integration solutions have already been proposed in various environmental application domains, including crop modeling in agriculture [87], solar terrestrial observation [88], water management [51] and geology [89]. Approaches defined from a more general purpose perspective have also been proposed in the scope of OGC data models and services, focusing on geospatial data [90] and more specifically on environmental observation data [45,91]. All the above approaches rely on the existence of data integration knowledge in the form of mappings between local and global ontologies, which has to be provided by experts of the application domains.

The above semantic data integration solutions focus on the resolution of semantic conflicts between data source metadata during global data access. However, other challenges related to the effective fusion of various data sources that refer to the same variables have to be faced, to achieve more complete datasets that are also more precise. Those challenges include issues related to data imperfection, data correlation, data inconsistency and data disparateness [92]. The data fusion problem in the scope of geospatial data sources has already been formally specified [93]. The spatial attributes of geospatial features from various datasets have to be appropriately combined to achieve data concatenation with possible duplicate elimination, geometric and/or temporal correction, feature enrichment and feature update and difference. It is noticed that the above approach restricts to only vector features. A specific solution that combines meteorological and air quality data has been described in [94]. The data sources include numerical environmental data sources and also text documents extracted from the web. The fusion algorithm is based on a statistical approach and it uses expert information to determine source reliability and to transform linguistic values to numeric values. A similar approach is the fusion of observed and modeled air quality data undertaken in [95] for the generation of real-time interpolated data. Notice that the above solutions are application specific, thus, to the best of these authors' knowledge, general purpose solutions for data fusion between vector features and raster coverages in the environmental data management context are not available.

In summary, much work has been done related to the semantic integration of environmental metadata using ontologies, however, the approaches available so far rely on the existence of data integration knowledge provided by experts. Regarding data fusion, the problem has been formalized for vector features, but only application-specific solutions exist with raster coverages of environmental data. Based on the above the two following challenging problems are identified in this area (see Figure 9).

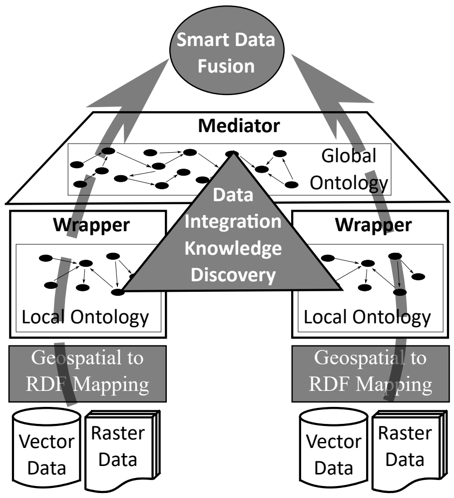


Figure 9. Illustration of challenges towards smart environmental data integration.

Challenge 4: geospatial to RDF mappings. To support the syntactic integration of geospatial data sources into the semantic data representation storage technologies described in the previous section, mapping solutions similar to those between the relational model and RDF [77,78] have to be developed between geospatial data models and RDF. At a first glance, it seems that mappings between relational data and RDF may directly be applied to vector feature datasets, however, this is not the case of geospatial coverages, whose underlying data model is not based on entities.

Challenge 5: geospatial data integration knowledge discovery. Data integration knowledge has the form of semantic relationships between classes of local and global ontologies. Most commonly, those relationships define class hierarchies and equivalences. The automatic discovering of these kinds of relationships is the main objective of ontology mapping solutions [96]. They may either define mappings between a global ontology and various local ontologies, or directly between local ontologies. The key issue in this challenge is to take advantage of both the special characteristics of geospatial features and coverages and the already existing environmental data models and ontologies to improve the effectiveness of the solutions.

Challenge 6: smart geospatial fusion of heterogeneous vector-raster data. The data integration knowledge discovered by the approaches of the previous challenge, enables the identification of which geospatial vector features and raster coverages may be candidates to undergo a further data fusion process. The required functionality will range from simple detection of duplicates [97] to more advanced smart data fusion solutions. Already existing geospatial and environmental specific data fusion solutions should serve as use cases to be generalized towards tools of more general purpose. Those tools will take advantage of currently available statistical and intelligent data analysis techniques and will incorporate geospatial relationships (topological, directional and metric) between the data elements during the fusion process. Specific environmental provenance information defined in existing data models and ontologies will help in the estimation of the reliability and precision of the data sources, which are key characteristics to be considered during the integration.

5. Data Searching and Browsing

First, current solutions in the areas of geospatial and environmental data management are revised and next, it is shown how current semantic technologies are being applied to this problem and which are the challenges to be faced towards the construction of relevant smart solutions for environmental data searching and browsing.

5.1. Geospatial Data Searching and Browsing

Geospatial data searching and browsing over vector features are generally supported by the spatial query capabilities of spatial databases [6,7]. Spatial extensions of relational and object-relational DBMSs are based on OGC [26] and ISO [98] standards. Efficiency is achieved through spatial indexing structures and spatial query processing techniques. Declarative querying over raster coverages is not supported in standard SQL. In spite of this, some spatial database implementations such as the PostGIS (https://postgis.net/) extension of PostgreSQL incorporate already a raster data type. However, on the one hand, raster queries are complex in the sense that raster arrays are nested in tuples and the user must combine raster functions with SQL, and on the other hand the efficiency of the implementation is still not appropriate. Specific managers for array datasets are also available [29,30] that incorporate efficient declarative query capabilities. In spite of all the above solutions, effective and efficient integrated querying of raster and vector datasets has still not been achieved by any solid implementation [54].

The implementation of geospatial data discovering and access infrastructures on the web are now supported by OGC standard service interfaces and INSPIRE recommendations [99,100]. As it is illustrated in the UML components diagram of Figure 10, geospatial searching, and browsing client applications may discover the existence of geospatial datasets using a Catalog Server that implements the OGC catalog service (CSW) standard interface [15]. This service enables filtering (including spatial and temporal conditions) over sets of geospatial metadata, usually modeled according to ISO 19115. These metadata may be harvested from Geospatial Data Servers, which provide access to datasets of vector features and raster coverages. Structured data searching over vector features and raster coverages are provided through OGC Web Feature Service (WFS) [16] and Web Coverage Service (WCS) [17] standard interfaces, respectively. Browsing geospatial datasets has to be done through map based interfaces. Map rendering may be done at the client side using the data retrieved through WFS and WCS, however, server-side map rendering is also generally supported by Geospatial Data Server implementations through OGC Web Map Service (WMS) [18] standard interface. Client-side rendering of maps enables more flexible and interactive interfaces, but on the other hand server-side map generation enables the browsing of very large datasets with light-weight clients, as those usually available in mobile devices.

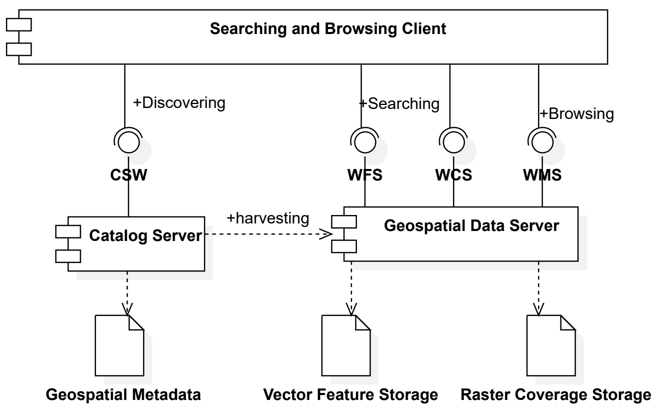


Figure 10. Illustration of geospatial data discovering, search and browsing in OGC Spatial Data Infrastructures.

If we focus now on environmental observation and model data, various infrastructures and platforms have already been constructed. A first example is the Global Earth Observation System of Systems (GEOSS) Common Infrastructure [83,101], which provides services to discover, search and access Earth Observation data sources. A key element of this infrastructure is its Discovery and Access Broker (DAB) framework, which includes brokering solutions for data discovery, data access, and semantic interoperability. The taxonomy proposed in [90] classifies web services to ease their incorporation in the GEOSS Component and Service Registry. The main types of services according to this taxonomy are: (i) Data Access Services, (ii) Catalog/Registry Services, (iii) Portrayal and Display Services and (iv) Data Transformation Services. Notice that except Data Transformation Services, which deal with data processing, the other three types of services are needed for data discovering, access and browsing. Two more examples of environmental data access infrastructures for environmental observation data related to water resources and ecological research, respectively, are the Hydrologic Information System (HIS) of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) [32], and the Data Hub for the Virtual Observatory and Ecological Information System (VOEIS Data Hub) [33].

The platform most commonly used to distribute geoscientific data is the one based on the tools provided by UNIDATA. In particular, the THREDDS Data Server (TDS) (https://www.unidata.ucar.edu/software/tds/current/) provides metadata and data access for geoscientific arrays, which conform to the Common Data Model already described in Section 3, using different interfaces and protocols. Those interfaces include the OGC WCS and WMS standard interfaces, but also the NetCDF Subset Service (NCSS) that provides filtering functionality over CDM dimensions.

The OGC has also a data access service specific for sensor observation data called Sensor Observation Service (SOS) [73], which provides searching functionality for datasets that conform to the OGC O&M data model already described in Section 3. Despite of spread spectrum of observation types supported by the O&M model, current SOS implementations (https://52north.org/software/software-projects/sos/; https://sourceforge.net/projects/pysos/) are based on spatial databases and mainly restrict to sampling methods based on in-situ observation. Therefore, remote sensing, which demands the recording of geospatial coverages in the result of the observations, is still completely ignored in SOS implementations. An approach that solves remote sensing data publishing based on OGC interfaces is described in [102]. More precisely, this solution combines a CSW catalog to discover SOS datasets, an SOS to publish observation metadata and a WCS to provide access to the result raster coverages.

5.2. Smart Data Searching and Browsing

Smart data searching and browsing solutions should leverage semantic geospatial and environmental data representations and combine them with existing semantic technologies to produce a better user experience. Some research approaches have already tried to achieve that by incorporating semantic technologies in OGC web services. As an example, RDF is used in [103] to improve data discovery in geographic data portals. Data discovery through CSW services is also improved in [104] with the use of ontologies. In [72], SensorML descriptions of sensor systems are semantically annotated to enable their registration in SOS services. A semantic SOS (SemSOS) implementation is described in [105]. Semantically annotated O&M and SensorML documents are generated from semantically annotated observation data and metadata using SPARQL. Finally, the SOS-based semantic mediation approach described in [45] provides also enhanced searching supported by basic reasoning over integrated global and local ontologies.

A different approach is the incorporation of geospatial data management capabilities in semantic web data management technologies. As it was already mentioned in Section 3, declarative querying of RDF datasets in the semantic web is supported with SPARQL [19]. This is a set of standard specifications that include a query language, an update language, a protocol to submit queries and update request to SPARQL services (SPARQL end points), a method for discovering and a vocabulary for describing SPARQL services, definitions of JSON, XML, CSV and TSV encodings for query results, and extension for federated queries and the definition of semantics of SPARQL queries under different entailment regimes. SPARQL has already been extended to support geospatial queries over vector feature datasets. Such an extension is called GeoSPARQL [12]. The RDF/OWL vocabulary that enables the representation of vector features in GeoSPARQL was already briefly described in Section 3. Beyond gespatial data representation, GeoSPARQL includes also primitive geospatial predicates and functions, based on OGC SFA [26], to support geospatial query formulations. Implementation of GeoSPARQL are also available [38]. Raster coverages are not supported in GeoSPARQL, as they are also not supported in standard SQL. To the best of these authors' knowledge, only the SciSPARQL solution [53] has attempted to incorporate raster coverage management in SPARQL. However, this approach suffers from similar problems to those already reported above for spatial databases with raster data types, i.e., high complexity in query specification due to the nested model and low performance.

Discovering semantic datasets is achieved by querying metadata catalogs, that may be represented using the DCAT standard vocabulary [39]. An application profile of DCAT, called DCAT-AP has been defined for European Data Portals, and an extension of DCAT-AP, called GeoDCAT-AP incorporates specific metadata of geospatial datasets.

Based on the above review of currently available technologies and approaches for semantic searching and browsing over geospatial data, the following challenges are identified that should be undertaken to achieve really effective and efficient smart geospatial searching and browsing (see gray boxes in Figure 11).

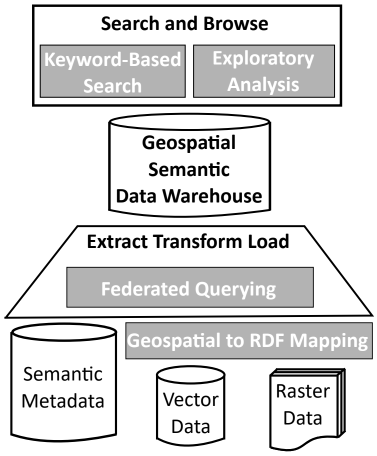


Figure 11. Illustration of challenges towards smart geospatial searching and browsing.

Challenge 7: semantic federated querying over vector and raster datasets. Most of the available geospatial datasets are not stored with semantic technologies, on the other hand, as it is illustrated at the bottom of Figure 11, to construct integrated datasets, ETL task may need to access semantic stores, generally to retrieve metadata, spatial databases or files to retrieve vector features and array files or managers to retrieve raster coverages. The above technologies are heterogeneous not only in their storage formats, but also in their query capabilities. The definition of model mapping solutions from geospatial vector and raster models to RDF have already been identified as a challenge (challenge 4) in Section 4. Efficient federated querying should support geospatial predicates and functions and should implement distributed query processing techniques that delegate parts of the query plan in the data sources based on their query capabilities.

Challenge 8: keyword-based search over geospatial data sources. General purpose dataset discovery is supported with SPARQL queries over metadata catalogs, which may incorporate full-text search capabilities to enable keyword-based search. Similarly, discovering geospatial datasets may be achieved with GeoSPARQL queries over GeoDCAT-AP metadata collections. Beyond dataset discovery, supporting keyword-based queries directly on the contents of the datasets may be of great importance [106]. Thus, for example, scientists researching on Cholera outbreaks are interested in downloading data from areas near the seaside with low elevation (accessible for bathing) and high temperature and rainfall. Such keyword-based queries may not be directly translated to GeoSPARQL and additionally, they cannot be replied using only dataset metadata, as they require accessing vector features and raster coverages.

An initial basic approximation to the above problem has been proposed in [107], where it is shown that the challenge is complex and that it requires the combination of geospatial and text indexing methods with linguistic values defined with fuzzy sets to incorporate user and expert subjectivity.

Challenge 9: interactive exploratory analysis of very large geospatial datasets. Once the required parts of the datasets of interest have been identified, and before advanced data mining and analytic techniques are applied, it is usually required to perform some exploratory analysis of the data. Such descriptive analytical workloads in data warehouses are supported by the evaluation of online analytical processing (OLAP) queries. Integrated geospatial vector-raster OLAP requires the efficient implementation of joins between vector features and raster coverages, and the efficient evaluation of aggregate functions over spatial and temporal dimensions. Achieving interactive response times over very large datasets is an elusive objective, even if parallel processing is implemented. However, given that during browsing the required precision for the results does not have to be high, approximate query processing techniques [108,109] may be adapted to geospatial data. Those techniques will combine geospatial indexing structures with dataset synopsis such as samplings and/or sketches [110].

6. Wrap-up Discussion on New Research Challenges

This section describes a framework that defines dimensions to evaluate the progress on the achievement of the proposed nine challenges. Besides, it discusses an initial evaluation of currently available semantic approaches according to the proposed framework, and it summarizes the related geospatial technologies which might be related to future solutions for each challenge. The evaluation of the currently available semantic technologies has been synthesized in Table 1. Further details are given below for each challenge.

Challenge 1: Geospatial RDF Representation

A geospatial RDF representation should support both vector features and raster coverages. Besides, those types of geospatial data should be supported in an integrated manner, i.e., avoiding the need for two separate model components or data structures. Notice that having separate representations for vector and raster data disables for example: (i) the representation of a raster cell (i.e., a spatial point with a specific spatial resolution) as the value of an attribute of a vector feature and (ii) the use of geometric types as data types for coverage attributes. Separate data models have been defined for vector features [8,9,26] and raster coverages [10] by the OGC, and such an approach is followed in general by the currently available geospatial data management technologies. In [54], an integrated data model was proposed, but a relevant implementation is not available. Regarding current semantic technologies, RDF [36] and RDF Cube [55] support semantic conventional and general purpose multidimensional data, however, they do not incorporate geospatial data representation capabilities. GeoSPARQL [12] supports only vector features and SciSPARQL [53] supports only raster coverages.

Challenge 2: Efficient Integrated Semantic Vector-Raster Storage

Implementations of the above geospatial RDF representation in both efficient data storage and open data transfer technologies should be achieved. Those implementations should achieve a reasonable storage size efficiency for both vector features and raster coverages, i.e., similar to the one achieved by current geospatial data storage formats. Besides, open text vector and raster encodings should be provided to foster interoperability. Finally, efficiency should also be achieved in terms of data access time for both vector and raster data filtering, but also for jointly accessing vector and raster data in queries. Currently, available geospatial storage formats and systems achieve a reasonable performance (both in storage size and access time) for both vector (Spatial DBMSs, shapefile, etc.) and raster (Array Managers, GeoTIFF, UNIDATA-NetCDF, etc.) data, and they also provide text encodings [25,27]. Efficient vector-raster join implementations are still missing [54]. Regarding semantic technologies, RDF text encodings such as RDF-XML and RDF-Turtle do not incorporate geospatial features. GeoSPARQL [12] provides support only for vector data, whereas SciSPARQL [53] has been designed only for raster data.

Challenge 3: Environmental Observation and Modeling Ontology

A generic top-level ontology that may be used as a global model for integrated knowledge representation in environmental applications has to be designed. The ontology should enable the representation of dataset level geospatial metadata, but also record level environmental metadata, which includes specific application and sampling feature metadata, provenance metadata related to either observation processes or environmental models and data quality metadata. Furthermore, the ontology should support the representation of simple observation and model results, but also of complex multidimensional coverages, including possibly spatial and temporal dimensions. Dataset level metadata is supported by both ISO 19115 and GeoDCAT-AP. ISO 19115 provides also means for the modeling of geospatial data quality. Available observation data models [11,31,33,35,56] provide constructs for the representation of observation processes and features of interest, including sampling features. Complex coverage results and data quality metadata are also supported. The main limitation is that the generic properties supported for processes and application features do not enable still the automatic generation of efficient data and metadata storage schemata, and therefore, existing generic data storage implementations lack the required data access performance. Similar characteristics are also present in the SSN ontology [13,44]. The combination of the record level sensor metadata representation of SSN with the GeoDCAT-AP dataset level metadata would be an interesting starting point. Such a combined ontology should be extended with necessary classes and properties to enable the representation of environmental model metadata and to support the automatic generation of efficient data storage structures. Those structures should enable the efficient integrated storage (in terms of space) and the efficient access (in terms of response time) to vector and raster observation and modeling results with all the required dataset level and record level metadata. Finally, the semantic integration of already existing vocabularies such as SWEET [14], CF, GEMET, and other [46–48] should be enabled by the system, to foster semantic interoperability between data sources and applications.

Challenge 4: Geospatial to RDF Mappings

The first step towards data integration is to achieve syntactic homogeneity among all the datasets within the same data representation framework. Therefore, to implement SEDIs, appropriate mappings are required from existing geospatial data models to the geospatial RDF representation defined in challenge 1. Those mappings should support both vector feature and raster coverage sources and, in a way similar to that of the relational to RDF mappings, they should include both direct [77] and user defined [78] mappings. In particular, relational to RDF mappings should be extended to have full support of vector feature sources, and to add raster coverage support.

Challenge 5: Geospatial Data Integration Knowledge Discovery

Once syntactic integration is achieved, semantic conflicts between the data sources have to be identified and solved. To achieve this, first a global data model is generally assumed. The characteristics of such a model has already been described as part of challenge 3. Next, data integration knowledge that maps local concepts of each data source to global concepts, has to be generated. Currently available solutions assume in general that such knowledge is already available [45,51,87–91]. However, defining such knowledge is a tedious process that requires the implication of application domain experts. Therefore, automatic and semi-automatic data integration knowledge extraction mechanisms should be devised to alleviate the expert's work. Those knowledge extraction processes will leverage already existing automatic ontology mapping approaches [96], which should be extended to incorporate the geospatial dimension of the data and metadata, to achieve better effectiveness (precision and recall) in the knowledge discovery process and more usable graphical user interfaces for the interaction with the application domain experts.

Challenge 6: Smart Geospatial Fusion of Heterogeneous Vector-Raster Data

Once local and global data structures have been matched, both syntactically and semantically, data element level integration, i.e., data fusion must be undertaken. Ad-hoc approaches that combine different types of environmental data, including text and geospatial formats [94], and observation and model data [95], have already been implemented. Those ad-hoc approaches obtain, in general, integrated datasets with an average error that is lower than that of the input datasets. Ad-hoc solutions may be designed to combine vector and raster datasets. However, every new data fusion task requires a new algorithm to be designed and implemented, needing the participation of both data science and application domain experts. Generalizing those ad-hoc approaches maintaining at the same time the performance in terms of error is the main objective of this challenge. The general data fusion problem has been formalized for vector features [93], and it includes dataset concatenation with possible duplicate elimination [97], data accumulation, feature correction, feature enrichment, and feature update and difference. However, the solution does not consider raster sources and of course, the performance in terms of error may still be improved.

Challenge 7: Semantic Federated Querying over Vector and Raster Datasets

Geospatial data querying functionality is needed to implement data searching and browsing. Additionally, federated querying may also be of great importance during the implementation of Extract Transform and Load (ETL) processes in the data integration stage, after the RDF mappings of Challenge 4 have been applied. Both vector and raster data sources should be supported by a semantic geospatial federated query engine. To achieve and efficient implementation in terms of query response time, it is also very important to take into account the data processing capabilities of the data sources during the construction of the federated query plans, trying to place geospatial processing as close to the data as it is possible. In particular, the federated query engine should leverage vector data processing capabilities present in spatial DBMSs and raster data processing capabilities provided by array data managers [29,30]. Although semantic federated querying is already provided by SPARQL [19] implementations, they do not support integrated geospatial querying of vector and raster data. Some available GeoSPARQL [12] implementations may federate various vector data sources, however, they do not leverage different data processing capabilities of heterogeneous geospatial data sources. Finally, SciSPARQL [53] supports raster querying, but does not allow the federation of various data sources.

Challenge 8: Keyword-Based Search over Geospatial Data Sources

Keyword-based searching is a key piece of functionality required to implement data discovering, which is a key process to bring datasets closer to IT practitioners, and therefore to citizen applications. To achieve really effective search, in terms of the precision and recall of the system response, the implementations should enable the searching over both conventional and geospatial metadata elements and over both text and numeric properties. Besides, fuzzy spatial and temporal relationships between data elements should be incorporated in a simple and usable keyword-based query language. The Full Text Search (FTS) capabilities of SPARQL may only be used to query text properties, however, the user must have knowledge about the schema of the dataset. The OGC CSW standard [15] enables dataset discovering using keyword-based and spatial search over geospatial metadata. However, it may not be used to query the dataset contents with keywords. Similar functionality is also available in open data portals that use the GeoDCAT-AP data model. Relational keyword-based search approaches support the query over text properties [106] of relational datasets, however, in general they do not leverage the existence of geospatial metadata and they do not support either the querying of numeric properties with keywords, using fuzzy linguistic values, nor the incorporation of spatial and temporal relationships such as "after", "nearby", "inside", etc. Finally, an initial approach towards generic keyword-based search over environmental datasets, called Keywordterm, has already been described [107]. However, much work is still required related to this challenging problem to achieve solutions with good performance, both in terms of effectiveness (precision and recall of the result) and efficiency (size of the indexing structures, index construction and update times and search response time).

Challenge 9: Interactive Exploratory Analysis of Very Large Geospatial Datasets

Once the data has been discovered and before it may be incorporated into complex decision-making mechanisms, it has to be analytically explored. Exploring geospatial and environmental data is usually done through map based graphical user interfaces. However, required system functionality does not restrict to simple map rendering, since statistical analysis has to be evaluated during data exploration. Supported data types must include both vector features and raster coverages, but two other important requirements related to the system performance must be included. In particular, the response time must be short enough to guarantee interactive exploration of the analysis results. To achieve that, approximate results that have some controlled error may be generated. Current OGC WMS service implementations support both vector and raster data, but only partially since they only provide map rendering and do not support user-defined statistical analysis. Besides, no control over the resulting error is given to the user. Response times may be improved, despite the good performance achieved when dataset (vector or raster) tiling and map overviews are used. Current GeoSPARQL [12] implementations incorporate geospatial indexing techniques that are used for the querying of vector data, however, raster data is not supported and query approximation is not implemented. Therefore, interactive response times may not be achieved when large datasets are analyzed. Finally, conventional approximate query processing techniques [108,109] exploit the use of dataset synopsis [110] to achieve interactive query response times with controlled result error. It is estimated that the combination of those synopses with geospatial approximate representations within geospatial data structures would enable the construction of efficient geospatial approximate query processing engines, which would support integrated exploratory analytics over vector and raster datasets.

Table 1. Evaluation of current semantic technologies in the context of each research challenge.

Semantic Approach	**Evaluation Dimensions**						
1. Geospatial RDF Representation		**Vector**		**Raster**		**Integrated**	
RDF		N		N		N	
GeoSPARQL		Y		N		N	
RDF Cube		N		N		N	
SciSPARQL		N		Y		N	
2. Efficient Integrated Semantic Vector-Raster Storage	**Efficient Vector Encoding**	**Efficient Raster Encoding**	**Text Vector Encoding**	**Text Raster Encoding**	**Efficient Vector Access**	**Efficient Raster Access**	**Efficient Vector-Raster Join**
RDF Text Encodings	N	N	N	N	N	N	N
GeoSPARQL Implementations	Y	N	Y	N	Y	N	N
SciSPARQL	N	Y	N	Y	N	Y	N
3. Environmental Observation and Modeling Ontology	**Dataset Level**	**Observation Processes**	**Modelling Process**	**Application FOIs**	**Sampling Features**	**Coverage Results**	**Data Quality**
ISO 19115	Y	N	N	N	N	N	Y
GeoDCAT-AP	Y	N	N	N	N	N	N
Observation Models (O&M, ODM, etc.)	N	P	N	P	P	Y	Y
SSN	N	P	N	P	P	Y	Y
4. Geospatial to RDF Mappings		**Feature Sources**		**Coverage Sources**		**Direct and User Defined Mappings**	
Relational to RDF Mappings		P		N		Y	

Table 1. *Cont.*

Semantic Approach	Evaluation Dimensions			
5. Geospatial Data Integration Knowledge Discovery	**Precision-Recall**		**Usability of expert GUI**	
Ontology mapping approaches	P		P	
6. Smart Geospatial Fusion of Heterogeneous Vector-Raster Data	**Vector Sources**	**Raster Sources**	**Vector and Raster Sources**	**Error Measure**
Ad-hoc approaches	P	P	P	P
Vector feature fusion	Y	N	N	P
7. Semantic Federated Querying over Vector and Raster Datasets	**Vector sources**	**Raster Sources**	**Leverage Vector Processing**	**Leverage Raster Processing**
SPARQL	N	N	N	N
GeoSPARQL Implementations	Y	N	N	N
SciSPARQL	N	N	N	N
8. Keyword-Based Search over Geospatial Data Sources	**Leverage Geo Metadata**	**Text Properties**	**Numeric Properties**	**Fuzzy Spatial and Temporal relationships**
SPARQL Full Text Search	N	P	N	N
OGC Catalog Service (CSW)	Y	N	N	N
GeoDCAT-AP	Y	N	N	N
Relational Keyword-based Search	N	Y	N	N
Keywordterm	P	P	P	N
9. Interactive Exploratory Analysis of Very Large Geospatial Datasets	**Vector Features**	**Raster Coverages**	**Result Error**	**Response Time**
OGC Map Service (WMS)	P	P	N	P
GeoSPARQL Implementations	Y	N	N	N
Approximate Processing	N	N	P	P

7. Conclusions

Environmental datasets, generated from observation and modeling processes, are important sources of highly valuable information for decision making in many application domains. Those applications include different domains of scientific research, environmental and infrastructure management, agriculture, and other applications more directly in contact with citizens such as smart cities and touristic and leisure activities. However, and in spite of some attempts to apply semantic technologies in the geospatial and environmental domain, environmental data discovering and access infrastructures are still targeted to experts in the areas of GIS and environmental data management. Due to this, in general, IT practitioners are either not aware of the existence of these infrastructures or do not have the required skills to get the required profit from them. As a consequence, many opportunities to reuse these expensive and high-quality datasets in many other application domains and to open new interesting business models from them are lost.

SEDIs are proposed in this paper as the solution to the above problems. Their design and implementation will be based on the effective extension of semantic web technologies with geospatial data management capabilities. Research challenges to achieve the above objective focus mainly in (i) the incorporation of raster data in semantic representations and ontologies, (ii) the automatic discovering of data integration knowledge and the effective fusion of vector and raster datasets, (iii) the efficient federated querying of vector and raster datasets, (iv) the effective and efficient semantic keyword-based search and (v) the implementation of efficient exploratory analysis.

Author Contributions: state of the art supervision, challenge identification and paper writing, J.R.R.V.; state of the art review S.V. and D.M.; state of the art supervision and writing review, J.A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was co-funded by (i) the TRAFAIR project (2017-EU-IA-0167), co-financed by the Connecting Europe Facility of the European Union, (ii) the RADAR-ON-RAIA project (0461_RADAR_ON_RAIA_1_E) co-financed by the European Regional Development Fund (ERDF) through the Iterreg V-A Spain-Portugal program (POCTEP) 2014-2020, and (iii) the Consellería de Educación, Universidade e Formación Profesional of the regional government of Galicia (Spain), through the support for research groups with growth potential (ED431B 2018/28).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application programming interface
CDM	Common data model
CF	Climate and forecast
CUAHSI	Consortium of Universities for the Advanced of Hydrologic Science
CSW	Catalog Service
DBMS	Database management system
DAB	Discovery and access broker
DCAT	Data Catalog Vocabulary
DCAT-AP	DCAT application profile
DQV	Data Quality Vocabulary
ETL	Extract transform and load
GEOSS	Global Earth Observation System of Systems
GIS	Geographic information systems
GML	Geography Markup Language
HIS	Hidrological information system
HTTP	Hypertext Transfer Protocol
INSPIRE	Infrastructure for Spatial Information in Europe
IoT	Internet of things
ISO	International Organization for Standardization
IT	Information technology
JSON	JavaScript object notation
MQTT	MQ telemetry transport
NCSS	NetCDF subset service
NGSI	Next generation service interface
OASIS	Organization for the Advancement of Structured Information Standards
ODM	Observations data model
OGC	Open Geospatial Consortium
OLAP	Online analytical processing
O&M	Observations and measurements
O-MI	Open messaging interface
O-DF	Open data format
OWL	Ontology web language
PROV-O	Provenance (PROV) Ontology
R2RML	Relational database to RDF mapping
RDF	Resource description framework
RDFS	RDF schema
REST	Representational state transfer
SEDI	Smart environmental data infrastructure
SensorML	Sensor Model Language
SFA	Simple feature access
SOS	Sensor observation service
SOSA	Sensor observation, sample actuator
SPARQL	SPARQL Protocol and RDF Query Language
SSN	Semantic sensor network

SWEET	Semantic web for Earth and environmental terminology
SWRL	Semantic web rule language
TDS	THREDDS Data Server
TIN	Triangulated irregular network
UML	Unified Modeling Language
URL	Uniform resource locator
URI	Uniform resource identifier
VOEIS	Virtual observatory and ecological information system
VODM	VOEIS observation data model
W3C	World Wide Web Consortium
WaMDaM	Water management data model
WCS	Web coverage service
WFS	Web feature service
WKB	Well known binary
WKT	Well known text
WMS	Web map service
XML	Extensible markup language

References

1. Baker-Austin, C.; Trinanes, J.A.; Taylor, N.G.; Hartnell, R.; Siitonen, A.; Martinez-Urtaza, J. Emerging Vibrio risk at high latitudes in response to ocean warming. *Nat. Clim. Chang.* **2013**, *3*, 73–77.
2. Lowen, A.C.; Mubareka, S.; Steel, J.; Palese, P. Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. *PLoS Pathog.* **2007**, *3*, e151, doi:10.1371/journal.ppat.0030151.
3. Utomo, D.; Chen, S.F.; Hsiung, P.A. Landslide Prediction with Model Switching. *Appl. Sci.* **2019**, *9*, 1839, doi:10.3390/app9091839.
4. Cassardo, C.; Andreoli, V. On the Representativeness of UTOPIA Land Surface Model for Creating a Database of Surface Layer, Vegetation and Soil Variables in Piedmont Vineyards, Italy. *Appl. Sci.* **2019**, *9*, 3880, doi:10.3390/app9183880.
5. Wainwright, J.; Mulligan, M. *Environmental Modellin: Finding Simplicity in Complexity*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2013.
6. Rigaux, P.; Scholl, M.; Voisard, A. *Spatial Databases with Application to GIS*; Morgan Kaufmann: San Francisco, CA, USA, 2001.
7. Zeiler, M. *Modeling Our World: The ESRI Guide to Geodatabase Concepts*; Esri Press: Redlands, CA, USA, 2010.
8. Kottman, C.; Reed, C. Topic 5: Features. In *The Opengis Abstract Specification*; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2009.
9. Herring, J. Topic 1: Feature Geometry. In *The Opengis Abstract Specification*; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2001.
10. OGC. Topic 6: Schema for coverage geometry and functions. Version 7.0.0. In *The Opengis Abstract Specification*; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2006.
11. Cox, S. Observations and Measurements. Verion 2.0. In *The Opengis Abstract Specification*; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2013.
12. Perry, M.; Herring, J. *OGC GeoSPARQL—A Geographic Query Language for RDF Data*; Ogc Implementation Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2012.
13. Haller, A.; Janowicz, K.; Cox, S.; Phuoc, D.L.; Taylor, K.; Lefrançois, M. *Semantic Sensor Network Ontology*; W3C Recommendation; World Wide Web Consortium: 2017.
14. Raskin, R.G.; Pan, M.J. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Comput. Geosci.* **2005**, *31*, 1119–1125. doi:10.1016/j.cageo.2004.12.004.
15. Nebert, D.; Voges, U.; Bigagli, L. *OGC Catalogue Services 3.0—General Model*; Ogc Implementation Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2016.
16. Vretanos, P.A. *OpenGIS Web Feature Service 2.0 Interface Standard*; Ogc Implementation Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2010.
17. Baumann, P. *OGC WCS 2.0 Interface Standard—Core: Corrigendum*; Ogc Interface Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2012.
18. de la Beaujardiere, J. *OpenGIS Web Map Server Implementation Specification*; Opengis Implementation Specification; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2006.

19. Harris, S.; Seaborne, A. *SPARQL 1.1 Query Language*; W3C Recommendation; World Wide Web Consortium: 2013.
20. Chelliah, P.R.; Subramanian, H.; Murali, A. *Architectural Patterns: Uncover Essential Patterns in the Most Indispensable Realm of Enterprise Architecture*; Packt Publishing Ltd.: Birmingham, UK, 2017.
21. Luaces, M.R.; Brisaboa, N.R.; Paramá, J.R.; Viqueira, J.R. A Generic Framework for GIS Applications. In *International Workshop on Web and Wireless Geographical Information Systems*; Kwon, Y.J., Bouju, A., Claramunt, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 94–109.
22. Percivall, G. *Topic 12: OpenGIS Service Architecture. Version 4.3*; The Opengis Abstract Specification; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2002.
23. European Parliament; Council of the European Union. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Off. J. Eur. Union* **2007**, *108*, 1–14.
24. Horsburgh, J.S.; Tarboton, D.G.; Maidment, D.R.; Zaslavsky, I. Components of an environmental observatory information system. *Comput. Geosci.* **2011**, *37*, 207–218, doi:10.1016/j.cageo.2010.07.003.
25. Portele, C. *OpenGIS Geography Markup Language (GML) Encoding Standard. Version 3.2.2*; Opengis Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2016.
26. Herring, J. *Simple Feature Access—Part 1: Common Architecture*; Opengis Implementation Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2011.
27. Butler, H.; Daly, M.; Doyle, A.; Gillies, S.; Hagen, S.; Schaub, T. *The GeoJSON Format*; Standards Track; Internet Engineering Task Force (IETF): 2016.
28. Baumann, P.; Hirschorn, E.; Masó, J. *OGC Coverage Implementation Schema*; Ogc Implementation Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2017.
29. Baumann, P. Management of Multidimensional Discrete Data. *VLDB J.* **1994**, *3*, 401–444.
30. Brown, P.G. Overview of sciDB: Large Scale Array Storage, Processing and Analysis. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, SIGMOD '10, Indianapolis, IN, USA, 6–11 June 2010; ACM: New York, NY, USA, 2010; pp. 963–968. doi:10.1145/1807167.1807271.
31. Horsburgh, J.S.; Tarboton, D.G.; Maidment, D.R.; Zaslavsky, I. A relational model for environmental and water resources data. *Water Resour. Res.* **2008**, *44*, 1-12, doi:10.1029/2007WR006392.
32. Horsburgh, J.S.; Tarboton, D.G.; Piasecki, M.; Maidment, D.R.; Zaslavsky, I.; Valentine, D.; Whitenack, T. An integrated system for publishing environmental observations data. *Environ. Model. Softw.* **2009**, *24*, 879–888, doi:10.1016/j.envsoft.2009.01.002.
33. Mason, S.J.; Cleveland, S.B.; Llovet, P.; Izurieta, C.; Poole, G.C. A centralized tool for managing, archiving, and serving point-in-time data in ecological research laboratories. *Environ. Model. Softw.* **2014**, *51*, 59–69, doi:10.1016/j.envsoft.2013.09.008.
34. Horsburgh, J.S.; Aufdenkampe, A.K.; Mayorga, E.; Lehnert, K.A.; Hsu, L.; Song, L.; Jones, A.S.; Damiano, S.G.; Tarboton, D.G.; Valentine, D.; et al. Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environ. Model. Softw.* **2016**, *79*, 55–74, doi:10.1016/j.envsoft.2016.01.010.
35. Abdallah, A.M.; Rosenberg, D.E. A data model to manage data for water resources systems modeling. *Environ. Model. Softw.* **2019**, *115*, 113–127, doi:10.1016/j.envsoft.2019.02.005.
36. Manola, F.; Miller, E. *RDF Primer*; W3C Recommendation; World Wide Web Consortium: 2004.
37. Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P.F.; Rudolph, S. *OWL2 Web Ontology Language Primer*, 2nd ed.; W3C Recommendation; World Wide Web Consortium: 2012.
38. Kyzirakos, K.; Karpathiotakis, M.; Koubarakis, M. Strabon: A Semantic Geospatial DBMS. In *The Semantic Web—ISWC 2012*; Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., et al., Eds.; Springer: Berlin/Heidelberg, 2012; pp. 295–311.
39. Albertoni, R.; Browning, D.; Cox, S.; Beltran, A.G.; Perego, A.; Winstanley, P. *Data Catalog Vocabulary (DCAT)—Version 2*; W3C Candidate Recommendation; World Wide Web Consortium: 2019.
40. Albertoni, R.; Isaac, A. *Data on the Web Best Practices: Data Quality Vocabulary*; W3C Working Group Note; World Wide Web Consortium: 2016.
41. Lebo, T.; Sahoo, S.; McGuinness, D. *PROV-O: The PROV Ontology*; W3C Recommendation; World Wide Web Consortium: 2013.

42. Ma, X.; Zheng, J.G.; Goldstein, J.C.; Zednik, S.; Fu, L.; Duggan, B.; Aulenbach, S.M.; West, P.; Tilmes, C.; Fox, P. Ontology engineering in provenance enablement for the National Climate Assessment. *Environ. Model. Softw.* **2014**, *61*, 191–205, doi:10.1016/j.envsoft.2014.08.002.
43. Jiang, L.; Yue, P.; Kuhn, W.; Zhang, C.; Yu, C.; Guo, X. Advancing interoperability of geospatial data provenance on the web: Gap analysis and strategies. *Comput. Geosci.* **2018**, *117*, 21–31, doi:10.1016/j.cageo.2018.05.001.
44. Compton, M.; Barnaghi, P.; Bermudez, L.; García-Castro, R.; Corcho, O.; Cox, S.; Graybeal, J.; Hauswirth, M.; Henson, C.; Herzog, A.; et al. The SSN ontology of the W3C semantic sensor network incubator group. *J. Web Semant.* **2012**, *17*, 25–32, doi:10.1016/j.websem.2012.05.003.
45. Regueiro, M.A.; Viqueira, J.R.; Stasch, C.; Taboada, J.A. Semantic mediation of observation datasets through Sensor Observation Services. *Future Gener. Comput. Syst.* **2017**, *67*, 47–56, doi:10.1016/j.future.2016.08.013.
46. Graybeal, J.; Isenor, A.W.; Rueda, C. Semantic mediation of vocabularies for ocean observing systems. *Comput. Geosci.* **2012**, *40*, 120–131, doi:10.1016/j.cageo.2011.08.002.
47. Ma, X.; Carranza, E.J.M.; Wu, C.; van der Meer, F.D.; Liu, G. A SKOS-based multilingual thesaurus of geological time scale for interoperability of online geological maps. *Comput. Geosci.* **2011**, *37*, 1602–1615, doi:10.1016/j.cageo.2011.02.011.
48. Ma, X.; Wu, C.; Carranza, E.J.M.; Schetselaar, E.M.; van der Meer, F.D.; Liu, G.; Wang, X.; Zhang, X. Development of a controlled vocabulary for semantic interoperability of mineral exploration geodata for mining projects. *Comput. Geosci.* **2010**, *36*, 1512–1522, doi:10.1016/j.cageo.2010.05.014.
49. Horrocks, I.; Patel-Schneider, P.F.; Boley, H.; Tabet, S.; Grosof, B.; Dean, M. *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*; W3c Member Submission; World Wide Web Consortium: 2004.
50. Shu, Y.; Liu, Q.; Taylor, K. Semantic validation of environmental observations data. *Environ. Model. Softw.* **2016**, *79*, 10– 21, doi:10.1016/j.envsoft.2016.01.004.
51. Howell, S.; Rezgui, Y.; Beach, T. Water utility decision support through the semantic web of things. *Environ. Model. Softw.* **2018**, *102*, 94–114, doi:10.1016/j.envsoft.2018.01.006.
52. Koubarakis, M.; Datcu, M.; Kontoes, C.; Di Giammatteo, U.; Manegold, S.; Klien, E. TELEIOS: A Database-powered Virtual Earth Observatory. *Proc. VLDB Endow.* **2012**, *5*, 2010–2013, doi:10.14778/2367502.2367560.
53. Andrejev, A.; Risch, T. Scientific SPARQL: Semantic Web Queries over Scientific Data. In Proceedings of the 2012 IEEE 28th International Conference on Data Engineering Workshops, Arlington, VA, USA, 1–5 April 2012; pp. 5–10, doi:10.1109/ICDEW.2012.67.
54. Villarroya, S.; Viqueira, J.R.R.; Regueiro, M.A.; Taboada, J.A.; Cotos, J.M. SODA: A framework for spatial observation data analysis. *Distrib. Parallel Databases* **2016**, *34*, 65–99, doi:10.1007/s10619-014-7165-7.
55. Cyganiak, R.; Reynolds, D. *The RDF Data Cube Vocabulary*; W3C Recommendation; World Wide Web Consortium: 2014.
56. Botts, M.; Robin, A. *OGC SensorML: Model and XML Encoding Standard. Verion 2.0.0*; Ogc Encoding Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2013.
57. Sege, J.; Ghanem, M.; Ahmad, W.; Bader, H.; Rubin, Y. Distributed data collection and web-based integration for more efficient and informative groundwater pollution risk assessment. *Environ. Model. Softw.* **2018**, *100*, 278–290, doi:10.1016/j.envsoft.2017.11.027.
58. Ghermandi, A.; Sinclair, M. Passive crowdsourcing of social media in environmental research: A systematic map. *Glob. Environ. Chang.* **2019**, *55*, 36–47, doi:10.1016/j.gloenvcha.2019.02.003.
59. Ganti, R.K.; Ye, F.; Lei, H. Mobile crowdsensing: Current state and future challenges. *IEEE Commun. Mag.* **2011**, *49*, 32–39, doi:10.1109/MCOM.2011.6069707.
60. Dutta, P.; Aoki, P.M.; Kumar, N.; Mainwaring, A.; Myers, C.; Willett, W.; Woodruff, A. Common Sense: Participatory Urban Sensing Using a Network of Handheld Air Quality Monitors. In Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, SenSys ’09, Berkeley, CA, USA, 4–6 November 2009; ACM: New York, NY, USA, 2009; pp. 349–350, doi:10.1145/1644038.1644095.
61. Dutta, J.; Gazi, F.; Roy, S.; Chowdhury, C. AirSense: Opportunistic crowd-sensing based air quality monitoring system for smart city. In Proceedings of the 2016 IEEE SENSORS, Orlando, FL, USA, 30 October–3 November 2016; pp. 1–3. doi:10.1109/ICSENS.2016.7808730.

62. Maisonneuve, N.; Stevens, M.; Niessen, M.E.; Steels, L. NoiseTube: Measuring and mapping noise pollution with mobile phones. In *Information Technologies in Environmental Engineering*; Athanasiadis, I.N., Rizzoli, A.E., Mitkas, P.A., Gómez, J.M., Eds.; Springer: Berlin/Heidelberg, 2009; pp. 215–228.
63. Jones, C.B.; Purves, R.S. Geographical information retrieval. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 219–228, doi:10.1080/13658810701626343.
64. Vasardani, M.; Winter, S.; Richter, K.F. Locating place names from place descriptions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2509–2532, doi:10.1080/13658816.2013.785550.
65. Dong, H.; Halem, M.; Zhou, S. Social Media Data Analytics Applied to Hurricane Sandy. In Proceedings of the 2013 International Conference on Social Computing, Alexandria, VA, USA, 8–14 September 2013; pp. 963–966. doi:10.1109/SocialCom.2013.152.
66. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In Proceedings of the 19th International Conference on World Wide Web, WWW '10, Raleigh, NC, USA, 26–30 April 2010; ACM: New York, NY, USA, 2010; pp. 851–860. doi:10.1145/1772690.1772777.
67. Restrepo-Estrada, C.; de Andrade, S.C.; Abe, N.; Fava, M.C.; Mendiondo, E.M.; de Albuquerque, J.P. Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring. *Comput. Geosci.* **2018**, *111*, 148–158, doi:10.1016/j.cageo.2017.10.010.
68. Wang, R.Q.; Mao, H.; Wang, Y.; Rae, C.; Shaw, W. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Comput. Geosci.* **2018**, *111*, 139–147, doi:10.1016/j.cageo.2017.11.008.
69. Barker, J.; Macleod, C. Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities. *Environ. Model. Softw.* **2019**, *115*, 213–227, doi:10.1016/j.envsoft.2018.11.013.
70. Azeez, O.S.; Pradhan, B.; Shafri, H.Z.M.; Shukla, N.; Lee, C.W.; Rizeei, H.M. Modeling of CO Emissions from Traffic Vehicles Using Artificial Neural Networks. *Appl. Sci.* **2019**, *9*, 313, doi:10.3390/app9020313.
71. Oettl, D. A multiscale modelling methodology applicable for regulatory purposes taking into account effects of complex terrain and buildings on pollutant dispersion: A case study for an inner Alpine basin. *Environ. Sci. Pollut. Res.* **2015**, *22*, 17860–17875, doi:10.1007/s11356-015-4966-9.
72. Bröring, A.; Maué, P.; Janowicz, K.; Nüst, D.; Malewski, C. Semantically-Enabled Sensor Plug & Play for the Sensor Web. *Sensors* **2011**, *11*, 7568–7605, doi:10.3390/s110807568.
73. Bröring, A.; Stasch, C.; Echterhoff, J. *OGC Sensor Observation Service Interface Standard. Version 2.0*; Opengis Implementation Standard; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2012.
74. Yoo, M.J.; Grozel, C.; Kiritsis, D. Closed-Loop Lifecycle Management of Service and Product in the Internet of Things: Semantic Framework for Knowledge Integration. *Sensors* **2016**, *16*, 1053, doi:10.3390/s16071053.
75. Robert, J.; Kubler, S.; Kolbe, N.; Cerioni, A.; Gastaud, E.; Främling, K. Open IoT Ecosystem for Enhanced Interoperability in Smart Cities—Example of Métropole De Lyon. *Sensors* **2017**, *17*, 2849, doi:10.3390/s17122849.
76. Samourkasidis, A.; Papoutsoglou, E.; Athanasiadis, I.N. A template framework for environmental timeseries data acquisition. *Environ. Model. Softw.* **2019**, *117*, 237–249, doi:10.1016/j.envsoft.2018.10.009.
77. Arenas, M.; Bertails, A.; Prud'hommeaux, E.; Sequeda, J. *A Direct Mapping of Relational Data to RDF*; W3C Recommendation; World Wide Web Consortium: 2012.
78. Das, S.; Sundara, S.; Cyganiak, R. *R2RML: RDB to RDF Mapping Language*; W3C Recommendation; World Wide Web Consortium: 2012.
79. Inmon, W.H. *Building the Data Warehouse*, 4th ed.; John Wiley & Sons, Ltd.: NJ, USA, 2005.
80. Sheth, A.P.; Larson, J.A. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Comput. Surv.* **1990**, *22*, 183–236, doi:10.1145/96602.96604.
81. Wiederhold, G. Mediators in the architecture of future information systems. *Computer* **1992**, *25*, 38–49, doi:10.1109/2.121508.
82. Wöhrer, A.; Brezany, P.; Min Tjoa, A. Novel Mediator Architectures for Grid Information Systems. *Future Gener. Comput. Syst.* **2005**, *21*, 107–114, doi:10.1016/j.future.2004.09.018.
83. Nativi, S.; Craglia, M.; Pearlman, J. Earth Science Infrastructures Interoperability: The Brokering Approach. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 1118–1129, doi:10.1109/JSTARS.2013.2243113.
84. Ludascher, B.; Gupta, A.; Martone, M.E. Model-based mediation with domain maps. In Proceedings 17th International Conference on Data Engineering, 2001, Heidelberg, Germany, 2–6 April 2001; pp. 81–90. doi:10.1109/ICDE.2001.914816.

85. Narock, T.; Wimmer, H. Linked data scientometrics in semantic e-Science. *Comput. Geosci.* **2017**, *100*, 87–93, doi:10.1016/j.cageo.2016.12.008.
86. Buccella, A.; Cechich, A.; Fillottrani, P. Ontology-driven geographic information integration: A survey of current approaches. *Comput. Geosci.* **2009**, *35*, 710–723. doi:10.1016/j.cageo.2008.02.033.
87. Porter, C.H.; Villalobos, C.; Holzworth, D.; Nelson, R.; White, J.W.; Athanasiadis, I.N.; Janssen, S.; Ripoche, D.; Cufi, J.; Raes, D.; et al. Harmonization and translation of crop modeling data to ensure interoperability. *Environ. Model. Softw.* **2014**, *62*, 495–508, doi:10.1016/j.envsoft.2014.09.004.
88. Fox, P.; McGuinness, D.L.; Cinquini, L.; West, P.; Garcia, J.; Benedict, J.L.; Middleton, D. Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience. *Comput. Geosci.* **2009**, *35*, 724–738. doi:10.1016/j.cageo.2007.12.019.
89. Wang, C.; Ma, X.; Chen, J. Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. *Comput. Geosci.* **2018**, *115*, 12–19, doi:10.1016/j.cageo.2018.03.004.
90. Lutz, M.; Sprado, J.; Klien, E.; Schubert, C.; Christ, I. Overcoming semantic heterogeneity in spatial data infrastructures. *Comput. Geosci.* **2009**, *35*, 739–752. doi:10.1016/j.cageo.2007.09.017.
91. Regueiro, M.A.; Viqueira, J.R.; Taboada, J.A.; Cotos, J.M. Virtual integration of sensor observation data. *Comput. Geosci.* **2015**, *81*, 12–19, doi:10.1016/j.cageo.2015.04.006.
92. Khaleghi, B.; Khamis, A.; Karray, F.O.; Razavi, S.N. Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion* **2013**, *14*, 28–44, doi:10.1016/j.inffus.2011.08.001.
93. Wiemann, S. Formalization and web-based implementation of spatial data fusion. *Comput. Geosci.* **2017**, *99*, 107–115, doi:10.1016/j.cageo.2016.10.014.
94. Johansson, L.; Epitropou, V.; Karatzas, K.; Karppinen, A.; Wanner, L.; Vrochidis, S.; Bassoukos, A.; Kukkonen, J.; Kompatsiaris, I. Fusion of meteorological and air quality data extracted from the web for personalized environmental information services. *Environ. Model. Softw.* **2015**, *64*, 143–155, doi:10.1016/j.envsoft.2014.11.021.
95. Schneider, P.; Castell, N.; Vogt, M.; Dauge, F.R.; Lahoz, W.A.; Bartonova, A. Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environ. Int.* **2017**, *106*, 234–247, doi:10.1016/j.envint.2017.05.005.
96. Choi, N.; Song, I.Y.; Han, H. A Survey on Ontology Mapping. *SIGMOD Rec.* **2006**, *35*, 34–41, doi:10.1145/1168092.1168097.
97. Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 1–16, doi:10.1109/TKDE.2007.250581.
98. ISO/IEC JTC 1/SC 32 Data Management and Interchange. *ISO/IEC 13249-3:2016. Information Technology—Database Languages—SQL Multimedia and Application Packages—Part 3: Spatial*; ISO Standard; International Organization for Standardization (ISO): Geneva, Switzerland, 2016.
99. Granell, C.; Díaz, L.; Gould, M. Service-oriented applications for environmental models: Reusable geospatial services. *Environ. Model. Softw.* **2010**, *25*, 182–198, doi:10.1016/j.envsoft.2009.08.005.
100. Ángel Latre, M.; Lopez-Pellicer, F.J.; Nogueras-Iso, J.; Béjar, R.; Zarazaga-Soria, F.J.; Muro-Medrano, P.R. Spatial Data Infrastructures for environmental e-government services: The case of water abstractions authorisations. *Environ. Model. Softw.* **2013**, *48*, 81–92, doi:10.1016/j.envsoft.2013.06.005.
101. Nativi, S.; Mazzetti, P.; Santoro, M.; Papeschi, F.; Craglia, M.; Ochiai, O. Big Data challenges in building the Global Earth Observation System of Systems. *Environ. Model. Softw.* **2015**, *68*, 1–26, doi:10.1016/j.envsoft.2015.01.017.
102. Chen, N.; Di, L.; Yu, G.; Gong, J.; Wei, Y. Use of ebRIM-based CSW with sensor observation services for registry and discovery of remote-sensing observations. *Comput. Geosci.* **2009**, *35*, 360–372, doi:10.1016/j.cageo.2008.08.003.
103. Athanasis, N.; Kalabokidis, K.; Vaitis, M.; Soulakellis, N. Towards a semantics-based approach in the development of geographic portals. *Comput. Geosci.* **2009**, *35*, 301–308, doi:10.1016/j.cageo.2008.01.014.
104. Stock, K.; Stojanovic, T.; Reitsma, F.; Ou, Y.; Bishr, M.; Ortmann, J.; Robertson, A. To ontologise or not to ontologise: An information model for a geospatial knowledge infrastructure. *Comput. Geosci.* **2012**, *45*, 98–108, doi:10.1016/j.cageo.2011.10.021.
105. Henson, C.A.; Pschorr, J.K.; Sheth, A.P.; Thirunarayan, K. SemSOS: Semantic sensor Observation Service. In Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems, Baltimore, MD, USA, 18–22 May 2009; pp. 44–53. doi:10.1109/CTS.2009.5067461.

106. Yu, J.X.; Qin, L.; Chang, L. Keyword Search in Relational Databases: A Survey. *IEEE Data Eng. Bull.* **2010**, *33*, 67–78.
107. Álvarez-Castro, D.; Viqueira, J.R.R.; Bugarín, A. Towards Keyword-Based Search over Environmental Data Sources. In *Semantic Keyword-Based Search on Structured Data Sources*; Szymański, J., Velegrakis, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 99–110.
108. Chaudhuri, S.; Ding, B.; Kandula, S. Approximate Query Processing: No Silver Bullet. In Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD '17, Chicago, IL USA, 14–19 May 2017; ACM: New York, NY, USA, 2017; pp. 511–519. doi:10.1145/3035918.3056097.
109. Mozafari, B. Approximate Query Engines: Commercial Challenges and Research Opportunities. In Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD '17, Chicago, IL USA, 14–19 May 2017; ACM: New York, NY, USA, 2017; pp. 521–524. doi:10.1145/3035918.3056098.
110. Cormode, G.; Garofalakis, M.; Haas, P.J.; Jermaine, C. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *Found. Trends Databases* **2012**, *4*, 1–294, doi:10.1561/1900000004.