Conciliating Perspectives from Mapping Agencies and Web of Data on Successful European SDIs: Toward a European Geographic Knowledge Graph

Bénédicte Bucher, Esa Tiainen, Thomas Ellett von Brasch, Paul Janssen, Dimitris Kotzinos, Marjan Čeh, Martijn Rijsdijk, Erwin Folmer, Marie-Dominique van Damme, Mehdi Zhral

International Journal of Geo-Information

MDPI

Article

Conciliating Perspectives from Mapping Agencies and Web of Data on Successful European SDIs: Toward a European Geographic Knowledge Graph

Bénédicte Bucher [1,*], Esa Tiainen [2], Thomas Ellett von Brasch [3], Paul Janssen [4], Dimitris Kotzinos [5], Marjan Čeh [6], Martijn Rijsdijk [7], Erwin Folmer [7], Marie-Dominique Van Damme [1] and Mehdi Zhral [1]

1 LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint Mandé, France; marie-dominique.van-damme@ign.fr (M.-D.V.D.); mehdi.zrhal@ign.fr (M.Z.)
2 Finnish Geospatial Research Institute-FGI, National Land Survey of Finland, Opastinsilta 12 C, 00520 Helsinki, Finland; esa.tiainen@nls.fi
3 Kartverket, Kartverksveien 21, 3511 Hønefoss, Norway; thomas.ellett@kartverket.no
4 Geonovum, Barchman Wuytierslaan 10, 3818 LH Amersfoort, The Netherlands; p.janssen@geonovum.nl
5 ETIS Lab UMR 8051, Paris Seine University, University of Cergy Pontoise, ENSEA, CNRS, 2 Av. A. Chauvin, 95000 Pontoise, France; Dimitrios.Kotzinos@u-cergy.fr
6 University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova Cesta 2, 1000 Ljubljana, Slovenia; Marjan.Ceh@fgg.uni-lj.si
7 Kadaster, Hofstraat 110, 7311 KZ Apeldoorn, The Netherlands; Martijn.Rijsdijk@kadaster.nl (M.R.); erwin.folmer@kadaster.nl (E.F.)
* Correspondence: benedicte.bucher@ign.fr; Tel.: +33-143-988-315

Received: 1 November 2019; Accepted: 19 January 2020; Published: 21 January 2020

Abstract: Spatial Data Infrastructures (SDIs) are a key asset for Europe. This paper concentrates on unsolved issues in SDIs in Europe related to the management of semantic heterogeneities. It studies contributions and competences from two communities in this field: cartographers, authoritative data providers, and geographic information scientists on the one hand, and computer scientists working on the Web of Data on the other. During several workshops organized by the EuroSDR and Eurogeographics organizations, the authors analyzed their complementarity and discovered reasons for the difficult collaboration between these communities. They have different and sometimes conflicting perspectives on what successful SDIs should look like, as well as on priorities. We developed a proposal to integrate both perspectives, which is centered on the elaboration of an open European Geographical Knowledge Graph. Its structure reuses results from the literature on geographical information ontologies. It is associated with a multifaceted roadmap addressing interrelated aspects of SDIs.

Keywords: SDI; open data; metadata; ontologies; linked data; provenance; mediation

1. Introduction

European Spatial Data Infrastructures (SDIs) have been under discussion for many years. We present, in this section, their definition, the issues that are still at stake, and some specific remaining challenges assessed by practitioners. We also present our approach to building a contribution for European SDIs based on a combination of perspectives from mapping agencies and the view brought by the Semantic Web.

1.1. European Spatial Data Infrastructures: Stakes and Challenges

Reusing the same data to monitor local, regional, national, and European environmental activities, as well as regulation and planning, is key for Europe to reach its commonly agreed goals for its territory as a whole. Using common data is essential to achieving consistency and trust between the different levels. It ensures that the European vision on a specific theme, for example, forests, will be compliant with the specific realities in the different member states; for example, Nordic forests are much different from French forests. In 2007, the European Commission issued the INSPIRE directive for an Infrastructure for Spatial Information in Europe, which was defined as "metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use; and coordination and monitoring mechanisms, processes and procedures, established, operated or made available in accordance with this Directive" [1]. SDI is defined as "the technology, policies, standards, human resources and related activities necessary to acquire, process, distribute, use, maintain, and preserve spatial data" [2]. INSPIRE is thus establishing the political decision for a European SDI and also the vision of a technical solution to ground European regulation on meaningful and quantifiable concepts that can be interpreted in national decisions. A major aim is to cope with the variety of national concepts, either in reality (like in the example of forests) or in information systems and regulation. The implementation of INSPIRE mainly relies on legally mandated organizations in member states, in charge of providing spatial datasets and services, and in particular the National Mapping Agencies (NMA).

INSPIRE has much broader goals than simply monitoring European environmental activities. The United Nations 2030 Agenda has defined seventeen goals for a sustainable planet and for eradicating poverty, which are consolidated into a set of indicators. Monitoring these indicators requires reusing existing data, creating new data, and defining integration methods and spatial analysis applied to existing geographical and statistical data [3]. More generally, we witness a growing amount of spatial data stemming from different technologies (data delivered by the administration under the Public Sector Information directive, satellites, air borne sensors, in situ sensors, social networks, etc.), and it is more and more complex for a user to discover data, compare them, and select the best ones in terms of benefit for his application and in terms of cost. Cost includes the following: the money paid as a customer or as a citizen (tax), the time spent to find a solution, the cognitive effort, the investment in expertise, the accepted uncertainties in the result, and the accepted loss of privacy. In a biased infrastructure, the most relevant datasets presented to a user could be selected based on advertisers and not on users need. Considering how many crucial activities rely on spatial data, it is key for Europe to get enough control on the infrastructure that connects people (whatever their activity) to the most relevant spatial data. Besides, we believe spatial data will be an important asset to support the connection between different generic open data, not necessarily just spatial. This makes the control of a European SDI all the more crucial.

More generally than the European level, there have been numerous SDI projects in the last forty years. In the early century, more than 200 countries (had) embarked on some form of SDI initiative [4]. While the first SDIs in the 1980s were prioritizing a better reuse of existing products or the design of specific new products from multiple sources [5], the next generation has shifted priorities to address the added value of SDIs and the quality of decision processes made on these infrastructures [6]. With the evolution of technologies, mainly Web 2.0 and smartphones equipped with GPS, new approaches to design SDIs have emerged where the user is also the data producer.

Parallel to SDI development, the World Wide Web has evolved toward the Semantic Web vision promoted by Sir Tim Berners-Lee [7], which meets many SDI functions. Web-application developers should easily find and reuse data through Web protocols, from whatever domain, just as a user of the Web browses Web documents thanks to search engines and hyperlinks. Twenty years later, these technologies for the Semantic Web have been adopted at an industry scale, noticeably by key players in our digital society, such as Facebook, IBM, Microsoft, and Google, and in particular the technology of knowledge graphs [8].

Different organizations have investigated the application of Semantic Web technology to SDIs. Ordnance Survey UK was an early adopter [9], followed by several research prototypes, like at IGN-France [10,11] and many large implementations by, amongst others, Ireland [12], Belgium, Switzerland, and the Netherlands [13]. However, scaling up these projects is a slow process. In the Netherlands, a community of linked data practitioners and researchers from both private and public sectors, including Kadaster, are running the Platform Linked Data Netherlands (PLDN, [14]). In this platform's projects, it is noticeable that policy makers are struggling in deciding whether they should define policies predicated on linked data or not, especially in the context of e-government. We interpret this as evidence of a knowledge gap in the business rationale behind the implementation of linked data. PLDN events and publications show a large amount of linked data implementations and projects, both in industry and government, but also, in particular, in the spatial domain. This reluctance of NMA to invest in Semantic Web technologies is explained here in the example of IGN-France. It has to face and cope with disruptions in many technological domains in order to pursue its mission of providing the national territory with the best spatial information infrastructures at the lowest possible cost. Additional new technological domains include the growing amount of satellites (like Sentinel but also Planet labs), sensors (like Lidars, Unmanned Aerial Vehicles, and in situ sensors), new information life cycle, and the involvement of citizens and communities in the production process and new information technologies (Deep Learning, Semantic Web, Blockchains, and Digital Twins). In this context, and with a limited amount of resources, it is not possible for an NMA to invest in all technologies, and it must adopt a strategy based on a convincing business rationale or internal intuition and strong belief in the potential for innovation. In this context, many mapping agencies are waiting for Semantic Web success stories before they invest in this technology, while users of authoritative geospatial data published as Linked data are waiting for more data to be available and connected in order to scale up existing local prototypes.

Paradoxically, remaining issues in SDIs would benefit from an enhanced management of semantics, which would argue for a collaboration with Semantic Web experts to apply their knowledge and technologies to today's SDIs. Eurogeographics is a European organization gathering national mapping agencies, to provide authoritative spatial reference data at the European level and to exchange experiences and knowledge among its members. EuroSDR, European Spatial Data Research, is a European nonprofit organization established in 1953 and whose core activity is knowledge exchange and joint benchmarking between its participants, national mapping and cadastral agencies, universities, and industries. During workshops organized by Eurogeographics Knowledge Exchange Network on INSPIRE (KEN INSPIRE) and the EuroSDR Commission on information usage, several remaining issues emerged from national presentations. These discussions are documented in presentations and minutes published by Eurogeographics on their website [15] and in the official EuroSDR report [16,17]. Below, we list unsolved issues discussed in these workshops that refer to a need for enhanced management of semantics, particularly in a European context:

- More automation in data integration preserving different semantics, for example, extracting evolution of land use from data at different dates with different conceptual schema.
- More automation in the generation of consistent views at different scales, for example, automatic filtering of the most important place names to show in a map at small scales.
- Designing user-oriented catalogues, for example, supporting a scientist working on urban climate to retrieve every datum required to describe building behavior with respect to heat and moisture flux.
- Documenting the different uncertainties (including semantic ones) of a multisource product, for example, a multisource land-cover product where networks are provided by different authorities, and possibly by non authoritative sources.
- Knowledge management solutions when designing SDIs adapted to thematic applications, for example, to identify if the water-related concepts (lakes, rivers, and wetlands) referenced

by the biodiversity community are the very same as topographic features maintained by mapping agencies.

These are examples, and there could be many more similar issues.

1.2. Objectives and Approach of This Work

Our work targets the management of semantic heterogeneities on European SDIs to meet the stakes of the INSPIRE directive, of the UN Agenda 2030, and also to empower citizens, scientists, and private and public sectors through enabling better access to open data.

Our approach is to co-design solutions integrating insights from the disruptive new technology that is the Semantic Web and the expertise and theories elaborated over many years in national mapping and cadastral agencies and in the domain of geographical information science. One practical aspect that we need to consider is the scalability of our approach to engaging authoritative data providers to invest in this technology enough for the corresponding development community to take off; and, vice versa, we need to engage this development community enough to demonstrate the capacity of these technologies to improve the management of semantic heterogeneities on SDIs and convince data providers to pursue their investment. The primary proposal presented in this paper is to set up the conditions for the creation of a European Geographical Knowledge Graph, the evolution of current Knowledge Graph technologies to address geographical information specificities and the actual documentation of this Knowledge graph and its exploitation.

Section 2 of this paper explains why contributions from two distinct communities, cartographers and Web developers, are needed. It also identifies differences in their perspectives, in particular critical ontological differences, that impact communication and collaboration. Section 3 presents our proposal that extends upon the complementary aspects of these communities. It also reconciles these perspectives by supporting the Open World Assumption (OWA) and the notion of Universe of Discourse into a European Geographical Knowledge Graph. Section 4 proposes a way forward to set up the conditions for the European Geographical Knowledge Graph to be documented and exploited to improve the management of semantics in European SDIs. Finally, Section 5 concludes the paper and provides starting points for discussion on future work.

2. Why Is Collaboration between Both Communities Needed and Yet So Difficult?

Two distinct communities, mapping agencies and geographic information scientists on the one hand, and Semantic Web designers on the other, have developed relevant competencies and capacities for European SDIs. They also have different perspectives, possibly diverging, which must be considered carefully for effective collaboration.

2.1. Expertise and Theories Related to the Provision of Reusable Geographical Data, Grounding Critical Decisions: Mapping Agencies and Geographic Information Science

National mapping agencies (NMAs) are missioned to provide nationwide information products to be used as a basis for decisions related to the surface of earth, especially public action, like simulating flood impacts, finely monitoring a bridge movement, or deciding how much in subsidies a farmer is entitled to. Once seen simply as providers of traditional maps, NMAs have evolved to become data providers which feed user applications. An important mission of NMAs is to design product specifications in order to cover as many as possible critical applications at the best cost for society (digital elevation models, land-use products, topographic data, gazetteers, etc.). Product specifications also ensure that a homogeneous product will be delivered even if the production relies on different operators. NMAs are also engaged in providing metadata supporting the correct interpretation of their data content, early on as pdf documents and lately as machine readable data. A trend over the last decade, data production is not always realized within the NMA; instead, data production has started to be subcontracted, while maintaining commitments on the quality of the produced data. NMAs also engaged in stronger connections with users to co-design the geographical data needed for user

application, for example, locating people in an emergency based on how they describe their location, managing ambulance itineraries, simulating flood, simulating urban heat islands, etc.

In the 1990s, NMAs developed competences in distributed architectures and interoperability with the creation of pan-European products and metadata bases with the MEGRIN network and later Eurogeographics. They also developed competencies in designing national geoportals, allowing citizens to visualize their information products through the Web, as well as locating data from other providers (for example, geological surveys). Moreover, since the INSPIRE directive, NMAs have contributed to the design of national components of a European infrastructure for Spatial Information and to the development of standards, mainly in the Open Geospatial Consortium (OGC) and ISO TC211, and more lately the W3C, in order to enhance the management of semantic heterogeneities in SDIs. In the new national geospatial platform of Finland, each entity in service interface schemas (features, attributes, and attribute values) is to be annotated by (the URIs of) an ontology concept to support its discovery. In this domain, too, in order to monitor the UN Sustainable Development Goals Indicators, for example, "proportion of the rural population who live within 2 km of an all-season road", both statistics and geographic information communities have cooperated to achieve better data integration. In a Finnish prototype, Linked Data have been experimented with to encode alignment between spatial features used in the statistical survey and the corresponding administrative features in the National Land Survey data (http://allusion.spatineo-devops.com/test/). This relies on persistent identifiers for both types of objects in the different databases, on the alignment of informational entities based on points, as well as the use of a common pivot ontology related to real world objects the informational entities are referring to, the general Finnish ontology (YSO, [18]).

Designing reusable vocabularies to describe European geographical reality has been partly addressed at an operational level in INSPIRE, where shared data specifications have been established as a European view on available national reference geodata, so not on real-world entities but rather on information objects. The way forward for the INSPIRE implementation has been more generally a knowledge-management process between contributors. A key experience of the project was to learn about the impracticality of a one-size-fits-all solution for geodata in Europe, as well as to grasp the different concepts developed from one place, one time, and from one domain to another, to represent a located feature. The complexity of the modeling based on INSPIRE led to other efforts, which yielded intermediate schemas that provided enough conceptual power for specific applications. For example, [19] proposes a model for groundwater data management which offers a simpler view on the subject by allowing only for the relevant concepts to be used and extending the necessary ones, while [20] proposes a model for landslide-susceptibility management, again in an effort to facilitate the users usability of the model. These intermediate models are easier to understand than the original INSPIRE model itself, easier to use for building a Knowledge Base, and easier to use for querying and reasoning operations. One important point to mention is that these models remain INSPIRE compatible by providing direct and indirect associations or extensions to the INSPIRE concepts, so exporting data with INSPIRE descriptions remains possible and easy. From the scientific perspective, designing reusable vocabularies about a geographical reality is complex, as spatial concepts are often socially constructed structures with a cross-disciplinary nature. There were several scientific attempts of structuring the knowledge of Geographic space into a general ontology, for instance [21–24], even more, into a Semantic reference system of geographic space [24,25]. Reference [24] treats geographical space simultaneously as the taxonomy of objects and activities. He introduces two main concepts, namely spatial objects (physical and abstract) and human activities (basic and advanced) as the top-level nodes of a hierarchical semantic geospatial datum. In a proposal for a Geographical Taxonomy for Geo-Ontologies, [25] recognizes three types of geo-ontologies, namely spatial geo-ontologies (SGO), physical (or natural) geo-ontologies (PGO), and human (HGO) geo-ontologies. Other authors addressed the alignment of application schemas [26,27].

To summarize, NMAs and geographic information scientists have developed practices and skills that are very relevant to the management of semantics on European SDIs; the provision of geographical

data to support different critical applications with meaningful information, the documentation of quality metadata, and the integration of informational geographical objects across different contexts—different products within a given NMA, different data across NMAs in Europe, and different data across authorities in the same country. NMAs have also developed practical competences in the design of shared vocabularies and the preservation and exchange of local semantics. Lastly, geographic information scientists have also advanced theories on spatiotemporal ontologies.

2.2. The Semantic Web and Linked Data

The notion of the Semantic Web (SW) has emerged as the capacity to support better decisions by reusing the growing amount of data coming from various sources, and using the Web as the platform to do so [7]. According to [28], while the Web is an elegant publication platform for documents, it is not possible to search for data at a sub-document level. For example, it is easy to search and retrieve documents about the registration of certain buildings that are regularly published by Kadaster, the Dutch NMA; however, it is not possible to search, e.g., for the oldest building among these documents. Even though the year of construction occurs in the aforementioned registrations, the original Web does not allow this information to be encoded in such a way that it can be uniquely identified. The concept of the SW was envisioned to tackle this exact flaw of the original Web. For the above example, this implies that each data attribute that appears in the registration document is individually recognizable, retrievable, and combinable into aggregate statistics. Besides, a revolutionary aspect of the Semantic Web in this initial vision was the capacity for Web applications to manipulate knowledge about the real world, to better meet a user request, thanks to encoded knowledge describing relationships between entities in the real world.

The three key questions of the Web of Data are listed by [28]:

- "How best to provide access to data so it can be most easily reused?"
- "How to enable the discovery of relevant data within the multitude of available data sets?"
- "How to enable applications to integrate data from large numbers of formerly unknown data sources?"

Basic elements to support data reusability are that data is somehow physically accessible, i.e., that data are on the Web, free of charge, and in a structured format. More advanced aspects are as follows: easy understandability of the data structures by developers and the capacity to crawl the Web of Data thanks to links between data repositories. In order to fill the SW with data that meet these more demanding requirements, the Linked Data (LD) initiative has been launched [29] to promote the use of semantic standards for representing and publishing information on the Web at data level. Linked Data adopts a formal model to describe data, as well as data structures, the Resource Description Framework. Linked Data supports the publication of explicit connections between datasets stemming from different producers based on URIs, which can be seen as external keys within datasets. The foundations of Linked Data also include the usage of description logics and ontologies to represent knowledge on the Web so that this knowledge can be used, for instance, to better interpret and adapt to a user context and propose better results and relevant recommendations.

More than 15 years were needed for Linked Data (LD) and the Semantic Web (SW) technologies to evolve from a mere vision presented in [7] to mature technologies residing in the plateau of production of the Gartner diagram [30]. In Osterwalder's terms [31], the main value propositions of implementing Linked Data (adopted from [32]) are decreased costs and increased flexibility of data integration and management. This leads to improved data quality and gives rise to new services. This can be observed from many inquiries indicating that, depending on the scale and scope of a Linked Data project, the saving potential in the management and reutilization of data can be noteworthy (e.g., [33]).

Beyond interconnected data, important assets from SW technologies are Knowledge Graphs (KG). References [34–37] have identified characteristics for a KG: a graph composed of statements about the real world that has both instances and schema and that covers more than one knowledge domain.

KGs include instance-level statements (e.g., things) and statements about background knowledge (ontologies) needed to understand the meaning of instances. Usually, instance-level information is by several orders of magnitude larger than that of schema level [36]. Reusability of the latter plays a major role in the context of the third property of KG. It is required for integrating, dereferencing, and disambiguating cross-domain knowledge [35]. Companies like Google, Microsoft, eBay, IBM, or Facebook already employ integrated knowledge graphs in their technologies. For instance, Facebook is using a knowledge graph about users, their connections, and the things they are interested in, such as music, films, or celebrities. This KG is used to identify relevant information and new connections to present to the user. Google's Knowledge Graph is probably the best-known use case of knowledge graphs. It allowed the company's traditional search engine to be more efficient in answering search for information, and also to design new solutions, like Google Assistant, which can answer questions expressed in natural language. Moreover, eBay is also developing a Product Knowledge Graph that is meant to represent products, entities, and the semantic relationships between them [38].

To summarize, the Semantic Web and Linked Data communities have developed good practices and new protocols to support the query of Web content as structured data with a concern to leverage as much as possible the silos between domains. They have established links between different databases. They also proposed contributions to the issue of self-learnability of data by an application developer outside the domain of the data. Knowledge graphs (KG) have proven a valuable asset to enhance user interaction with data.

2.3. Different Perspectives on Uncertainty, Authorities, and Context

In order to co-design solutions for European SDIs based on the competences of both communities described in Sections 2.1 and 2.2, durable collaborations are needed between them. However, as highlighted in the introduction, despite specific successful projects, there is little scalability of the results. From 2015 on, a series of technical seminars were organized by EuroSDR in order to address pending issues in SDIs with new technologies from the Web. Participants to these different seminars were practitioners, developers, and scientists from different communities: NMAs, Geographic information scientists, located data providers (geological survey, statistical institute), and the Web of Data domain. In these seminars, participants exchanged their expectations and experiences in engaging in Linked Data activities. These exchanges led the authors to relevant findings regarding gaps that exist between perspectives adopted by these communities on successful SDIs.

A first difference arises from the respective priorities attached to uncertainty management and reusability by both communities. NMAs and cartographers core priority is to leverage uncertainty in user decisions made on the abstraction (maps or data) provided by the NMA. They manage uncertainty by including clear metadata description and the curator's direct relation with published data. More precisely, metadata include the data-specification—spatial and temporal coverages, spatial resolution (usually distinguishing altimetric and planimetric)—features of interest and how they are represented, and quality criteria which correspond to an accuracy threshold the data provider commits to respect, and last but not least, the lineage information and provider. The core priority in the development of the Web of Data is to leverage as much as possible obstacles to data-loading in applications far from the context of production, prior to any processing of these data. Well-designed protocols and good practices aim at a seamless and enhanced circulation of information on the Web. These protocols do not put any condition on uncertainties documentation in the source data. However, both communities are respectively more and more attached to the reusability (for NMAs) and to the achievement of real applications where uncertainties cannot be ignored (for Web developers). These differences that were an obstacle to communication are seen in this paper as complementary approaches that can be reconciled.

A second difference is related to cost-benefit-authority frameworks. Both communities target the design of reusable frameworks to increase the benefits brought by geospatial data and lower the costs. These frameworks comprehend the recognition of authorities, yet they have neither the same

benefits nor the same costs in mind, and this impacts on the authorities' assessments. Geospatial data infrastructures are justified based on applications reusing cross-sector data contents, hence multiplying the value of data according to diverse and versatile user needs. Benefit here is the capacity to define and conduct a public action based on data in different public domains, for instance, risk management, urban planning, and land management. Legal authority mainly comes from legal-mission, law-empowerment of public organizations that commit to maintain data products. Knowledge authority comes from mastering all relevant techniques of the GI domain, ranging from geodetic survey to spatial analysis and visual rendering. The costs are mainly covered by society, rather than by the user, except for the investment in competences to utilize information served by the mapping agencies. In the Semantic Web, the authority is strongly related to usages, and the costs involved are normally borne by the developer of an application: How much money, effort, and time does one need to invest and spend to access and reuse the data? The developers considered here are Web developers mastering state of the art protocols and practices, to publish and exchange data, whatever the domain, on the Web. The notion of authority here is strongly related to user adoption of the provider data services. Both notions of authority are conflicting as soon as the predominantly used data are not the data which would be the most relevant from a technical and legal perspective. The notions of costs also are conflicting with respect to developer investment in learning to use the data. For NMAs, this cost must be "paid" by developers, and they design tutorials and forums to facilitate this investment. For Web developers, they do not want to invest in specific skills needed to use geospatial data.

A third difference identified during the workshops was the definition of context. Context is a recurring notion since both communities aim at representing the context of entities of interest for an improved understanding of these entities. However, there are different kinds of contexts. For cartographers, the context of an entity is made up of its geographical context, including proximity and topological neighborhood. More precisely, two views are needed to contextualize an entity: the focus view that represents its components and close neighborhood, and a larger view that represents its position in a larger environment. For example, Paris will be typically described with two views: a map of Paris with its main topography (la Seine, the avenues and boulevards), its monuments, its and boundaries, and a map of France and neighboring countries with main cities and main roads. Spatial relationships that compose a context are seldom explicitly represented in geodatabases and are rather carried by coordinates and evaluated on demand either by applying geo-computational algorithms to coordinates or, for most relationships, by visualization. Semantic Web developers initially aimed at building a knowledge base to store connections that exist in the real life between entities, for example, between someone and the place she was born in, in order to better manage these entities on the Web. However, it has evolved toward describing what exists on the Web and on making connections between Web entities. Context is then more understood as an informational context, connecting different perspectives on the same feature. These different perspectives on context are conflicting because they lead to different data structures. NMAs data are structured as groups of objects with few explicit links between them. Even when transformed into the Resource Description Framework (RDF) promoted by the Semantic Web, the structure is not really a navigable graph. Visual rendering is needed to reveal context. Vice versa, it is difficult for cartographers to derive their views of interest from Web data, as the notions of scales, as well as of rendering styles associated to geographical feature types, are lacking.

2.4. Closed-World Assumption vs. Open-World Assumption

Regarding semantic modeling, a fundamental difference exists between both communities. Where cartographers design datasets related to a specific contained universe of discourse, the Semantic Web approach is to define data according to an OWA. Both are legitimate but are also conflicting.

In cartography, since the late 1990s, object-oriented modeling (OO) has become the main approach in how the digital representation of the real world is established. The object, or feature, is an abstraction of a real-world phenomenon (ISO 19101:2014), and a feature type is a class of features having common characteristics (ISO 19156:2011). Phenomena are expressions that are part of, and defined by, the

view of the real or hypothetical world that includes everything of interest, the universe of discourse. Within each view, only the relevant information is modeled. An example of view of interest is a roadmap used to find directions and to drive. It includes neither paths nor every building, but only landmark buildings that can be used to become oriented. Application models are according to context-related specifications, depending on specific views of reality. This approach is generally referred to as the Closed World Assumption (CWA); what is not known to be true, is false, so that absence of information is interpreted as negative information [8]. If a flood map contains inhabited buildings in its specifications, the absence of buildings in the neighborhood of a given river on the map can be interpreted as the absence of buildings in the reality. Another aspect is that, in general, OO semantics are defined in and dedicated to specific use cases, which leads to semantics that are closely related to the vocabulary of a domain. The CWA can be associated with quality metadata describing if data are compliant with the given universe of discourse and what is the gap, the percentage of missing roads on a road map, for instance, while it is a never-ending task to describe the gap between a dataset and reality in the absolute. The downside of CWA is the aspect of harmonization and data interoperability between domains. If the universes of discourse are distinct when the basic concepts are established, it can be difficult to reunite them. An example is the use of detailed topographical data in asset-management processes and systems engineering, with these being more related to Building Information Models (BIM) and having different requirements related to construction and maintenance on features such as roads, road segments, building, and building parts. Existing topographical data do not often meet these requirements and are therefore not easily reused or adapted for BIM purposes without transformation to a new schema, in a different universe of discourse. We explain this difficult harmonization by the fact that historically the corresponding universes of discourse were disjoint. Topographical data, or Geographic Information System (GIS) models, have been designed in a Universe of Discourse where buildings were viewed only through their spatial footprint. Later on, with the advance of new sensors and 3D models, the domain has gradually integrated more details on buildings, like roof shape, facade textures, number of floors, etc. Historically, the BIM domain was concentrated on the building only, but gradually it has embraced the building neighborhood. Today, both universes of discourse have a strong overlap and need to align their concepts.

As opposed to this, LD follows the Open World Assumption (OWA): What is not known to be true or false is interpreted as unknown information [39]. Linked Data does use the Web as its publication platform, not in the way that a traditional service-oriented architecture (SOA) does by the publish, find, and bind mechanism, but by publishing data as data-entities in the Web of Data: the Semantic Web. In this Web world, anybody can say anything about anything, which is referred to as Anybody Can Make Statements About Any Resource (https://www.w3.org/TR/rdf-concepts/). This flexible data management is made operational by the concept of triples (https://www.w3.org/TR/rdf11-primer/). This triple concept allows for publishing data that can be extended by an infinite number of predicates published by anybody at any Web location. A single subject can be linked to predicates and objects defined in multiple ontologies covering several domains. Strong points from this perspective are the easy accessibility of data, as they are based on mainstream Web technology and related browsers. Moreover, the links between contents favor the discoverability of new content, just like we browse the Web of documents thanks to hyperlinks. The open aspect leads to a growing knowledge base. Knowledge models based on ontologies can be combined and used to generate information that was never part of any singular data publication. The downside in this case is that validation and securing data quality is much weaker and sometimes not addressed. However, this validation aspect is becoming more elaborate with the introduction of the SHACL vocabulary in Linked Data [40] (https://www.w3.org/TR/shacl/). The vision is that ontologies should be easily linkable in any context. This tends to create a very holistic vocabulary that does not favor recognition in a particular application domain. Whereas the OO approach provides a mechanism for a basis of standardized and structured data within domains, Linked Data provides an open mechanism for sharing and combining data [41].

3. Advancing Ontologies and Knowledge Graph Technologies for European SDIs

This section presents our proposal to co-design ontologies needed for an improved management of semantic heterogeneities on European SDIs, based on the complementary aspects between national mapping agencies' traditional assets and the SW technologies.

3.1. A Referencing Framework to Align and Compare Geographical Data: The Ontology Geographical Space Context (OGSC)

A first item is a referencing framework where digital assets of interest for an SDI can be described in such a way that they are easier to discover, compare, integrate, and reuse. This proposal extends upon the work in [23,42], to design ontologies that support dataset discovery and interconnection on traditional SDIs. It extends also on the proposal in [13] to provide a standards-based approach to add context to single datasets so that KGs may be built automatically.

Digital assets of interest for an SDI are, for instance, ontologies, gazetteers, a corpus of regulation, city models, scientific papers, or any other piece of information that may need to be shared and reused based on its spatial characteristics. They range from written sentences in natural language within a scientific paper to models of the Earth based on technologically advanced observations, measurements, calculations, cartographic mapping, and capturing of spatial data in contemporary structured digital databases.

Our referencing system is built by associating each digital asset with a geographical space context. A geographical space context is the conceptual space of objects and events in which a description of Earth is created and published on the Web. It is organized into two main components: geographical space of interest and information technology used to represent and encode the space. Geographical space context can be explained through the rough metaphor of a spatiotemporal window of interest with a filtering glass through which the selected entities can be seen. The geographical space is the scope of reality that is visible from the spatiotemporal window, and the information technology is the filtering glass. It includes the activity related to the observation, for example, surveying or hiking. Geographical space context supports a better understanding of a digital asset by describing relevant contextual information for the user. It supports the interconnection between digital assets based on their respective geographical spaces. It reconciles OWA and Universe of Discourse. Anyone can express a description. However, in order to facilitate the reuse of this description, it is important to document its context and detail what reality the author of the description targeted and how. The Geographical space context is also adapted to document precisely a given Universe of Discourse, which is often too implicit.

Figure 1 summarizes the categories of information used to reference a geographical digital asset in our proposal. For example, a GPS track of a jogger in a park is a specific digital asset that can be associated to a geographical space context composed of a a geographical space –the park- and an activity (jogging) and an information technology (the smartphone and app used to produce the track). Another example is the French topographic database BDTopo®. This Digital Asset can be associated to a geographical space context composed of a geographical space that is a Universe of Discourse (the set of physical features at the surface of Earth on France at the granularity of the building and a metric resolution), an activity that is the professional activity of topographic survey and an information technology that is the database implementation specifications, as well as sources used for the acquisition, including aerial imagery.

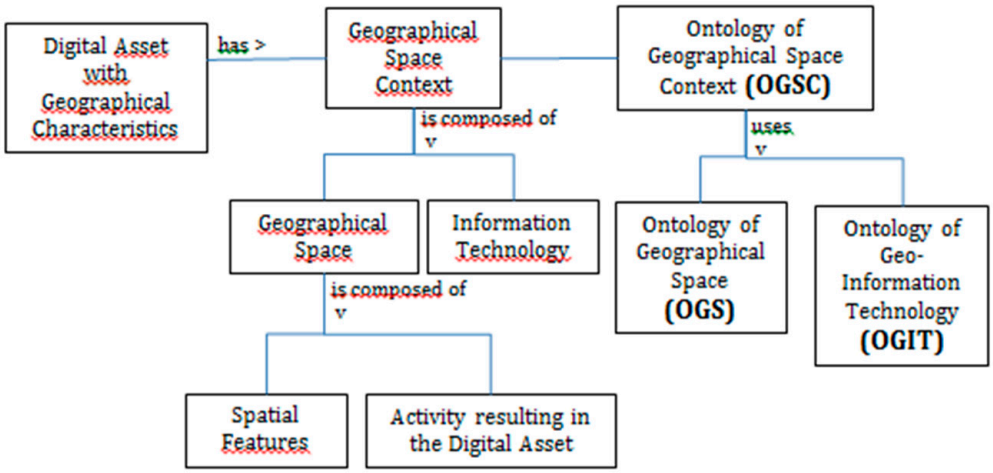

Figure 1. Our proposed referencing framework for improved management of semantics of digital assets with geographical characteristics, either classical SDI components or any other asset, and the needed ontologies. The spatial features are used to specify a CWO or an OWA.

We illustrate the relevance of this referencing framework for the interconnection of non spatial data with spatial data on SDIs through place names. To improve current text-to-space technologies, there is a need to better document digital resources and their scope [16]. We propose to describe this scope as a specific geographical space context: What is the geographical space in which a gazetteer is valid and complete? What is the informational structure of it? The scope of a document should also be described through a geographical-space context: What is the geographical space covered by the author of the document and what is the information structure of the document? The added value of an Ontology of Geographical Space Context (OGSC) to current geo-tagging solutions can be described on the GeoRef framework. It is a joint initiative between the Finnish National Land Survey, several national research institutes, the national thesaurus and ontology service, memory organizations, universities, large cities, private companies and YLE, the national broadcasting company, and the private sector. In current GeoRef architecture, illustrated in Figure 2, documents can be annotated thanks to ontologies (vocabularies), which are domain-specific; for example, the geo-ontology (http://dev.finto.fi/geo/fi/) is used to annotate a corpus of scientific paper in geology. Documents can also be registered thanks to place names provided; they use the labels from the place name register, shown on Figure 2 also, in a consistent manner. In current stage, the prototype is backed by the authoritative place-name register, which manages synonymy and homonymy expressions of place and which covers official and, to some extent, vernacular place names from 1996 and later. The correct selection of the relevant place-name register is key to handle synonymy and homonymy. Identification of the domain, scope of the documents, and appropriate ontology and place names register can be more or less automated through their metadata or natural language processing (NLP) tests. However, the introduction of the proposed shared ontology, OGSC, to describe the scope of SDI's digital assets like place names registers, as well as the scope of digital assets that have a geographical component like research reports, could automate the identification of the correct ontology and the correct place-name register to geo-reference research reports and the scalability of frameworks like GeoRef.

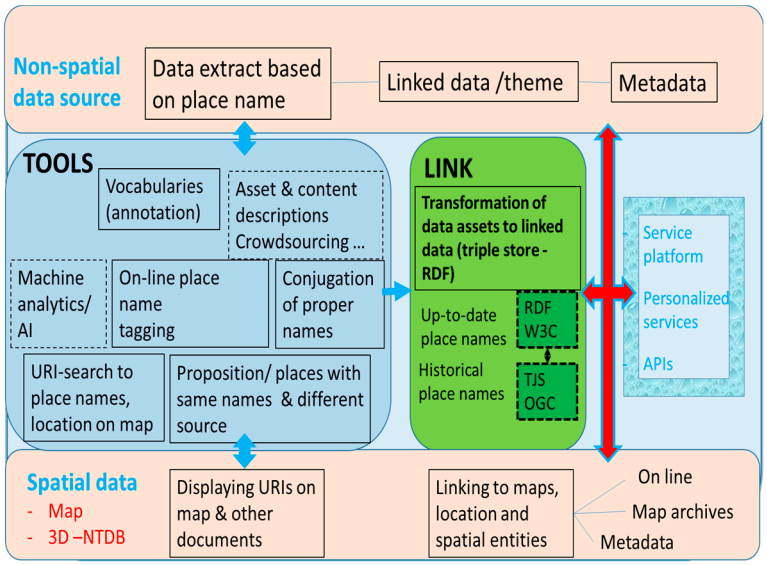

Figure 2. Internal structure of the current GeoRef prototype which connects nonspatial sources with spatial data based on place names automatically extracted and analyzed in these documents [43]. The introduction of the OGSC to reference nonspatial data sources, spatial data, and vocabularies shall improve the capacities of the TOOLS and the framework scalability.

An especially challenging item of our proposed referencing framework is the item used to describe the geographic space of interest the digital asset is referring to. This item is called Ontology of Geographic Space (OGS). As introduced in Section 3.1, in the geographic space, numerous spatially materialized objects with human origin exist, i.e., artificial objects. Besides them, there are also spatial schemes within the geographical space, either with consensually defined flat boundaries (for instance, municipality boundaries) or observed spatial formations of natural (continuous) phenomena and social phenomena. Reference [23] proposed a geographic-space ontology adapted to the assets managed within the Slovenian national SDI, supporting local, regional, and state community activities, including economic activities of private sectors. The simultaneous treatment of human activities and spatial objects was applied to build a conceptual graph of geographic space context in natural language. Conceptual graph theory was used to build an initial taxonomy of five levels of abstraction, from upper ontology to subordinated lower ontologies in Slovenian and English language. Reference [23] enriched the Slovenian version by ontology modeling in OWL and analyzed the content of the Digital Topography Database of the Slovenian NMA with Inductive Logic Programing language. Semantic suitability of this ontology as a semantic reference system was estimated manually against data catalogues of six different Slovenian databases (Land use—cadastre, Land use—agriculture, Topographic map scale 1:500, Topographic map scale 1:5000, CORINE Land Cover, and Digital Topographic database scheme, with semantic matching rate from 69% to 99%).

3.2. *A European Geographical Knowledge Graph*

Our core proposal is an open European Geographical Knowledge Graph (EGKG) that will support semantics management within European SDIs. The EGKG will host references to different digital assets of shared interest in European SDIs and associate them with their respective Geographic Space Context, described based on the ontology presented in the previous section. For example, it will host identifiers of specific gazetteers and associate them with a Geographic space context: scope in the real world (Geographic space) and schema. It will host also identifiers for authoritative or standard representation of key features, for instance, the Eiffel Tower, which will also be associated with a Geographic Space and a description of the corresponding technology (3D model, RDF, or other). In other words, this European Knowledge Graph will support the open-world assumption but also be

able to host representations with a specific and well-identified Universe of Discourse. To set up the EGKG, a first step is to derive a Web implementation of the Ontology of Geographic Space Context and ensure its adoption. We propose to align it with commonly used vocabularies on the Web and on SDI. An important vocabulary is the General Multilingual Environmental Thesaurus (GEMET) established by the European Environment Agency as a key element for an SDI.

This section details specific items of the EGKG that are references to entities of common interest from the real world and their informational counterparts.

Traditional topographical maps have traditionally been used to interconnect different domains, e.g., a given disease occurrence and a given crop, by drawing their entities of interest on the same map. This drawing is supported by direct location (assessing geometric attribute to an entity and coordinates) or indirect location (relating the information to a topographic entity visible on the map, depending on the map scale and the theme: a crossroad, a building, a city, a lake, a mountain, or a parcel). Authorities were legally mandated and committed to maintain these references with a strong connection with user requirements to design relevant referencing frameworks that would meet all needs at the best cost for society. In a similar way, URIs are needed for such entities in the real world, as well as the capacity for anyone to use these URIs to locate information in thematic databases. These URIs will be hosted in the EGKG.

To illustrate the value of these URIs, let us imagine that Kadaster registered an object, a building of the Saint Catharine church erected in 1900 in Eindhoven, with a certain registration ID. In Figure 3, this information is shown in the initial database and translated into Linked Data, a graph where the arbitrary wording is replaced by standardized notions and their URIs [28]. As only resources with URIs can be linked, use of URIs is key to allow linking data items between datasets. The Saint Catharine church in this example is an outstanding building and appears in many datasets, classified differently (as an addressable object and as a church). By linking these informational objects together, [30] proposes to infer additional knowledge, e.g., that a church is an addressable object. In this way, previously disconnected datasets can be linked together.

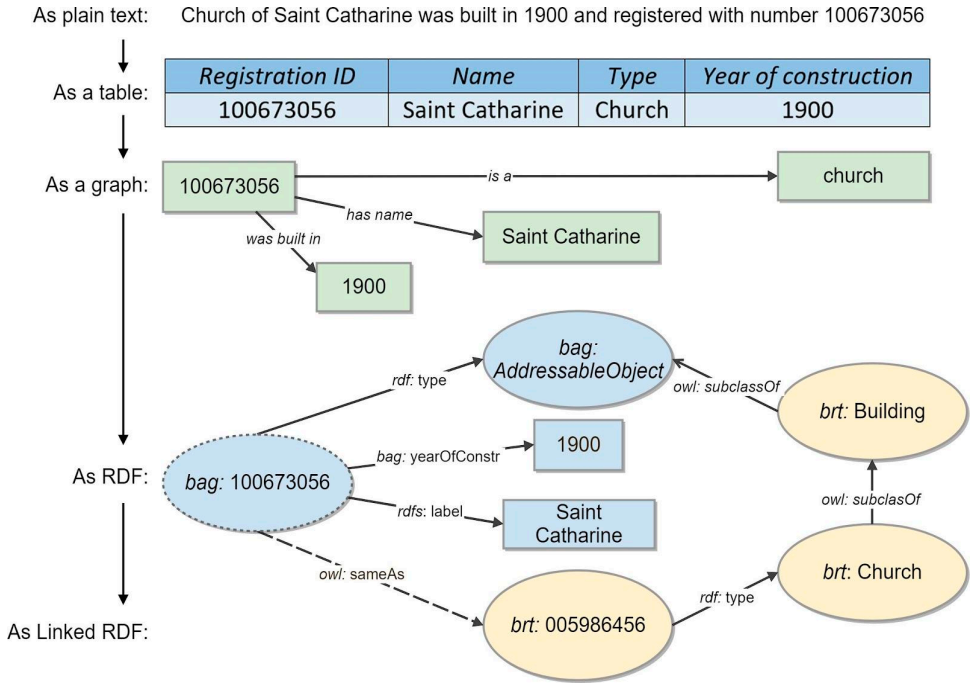

Figure 3. The population of a Knowledge Graph with statements from different registries. The data are firstly translated in graph version and then turned into Linked Resource Description Framework (RDF). After [30].

Another example is the Finnish implementation of connections between datasets from different authorities. In their proposal, different authorities' datasets describe real-world objects, for example, the lake Saimaa described by the National Land Survey and the city of Lappeenranta described by the Statistical Survey. Both institutes use official URIs for these real-world objects in the national platform, for example, http://paikkatiedot.fi/id/100700/abc123 for the lake Saimaa, as shown in Figure 4.

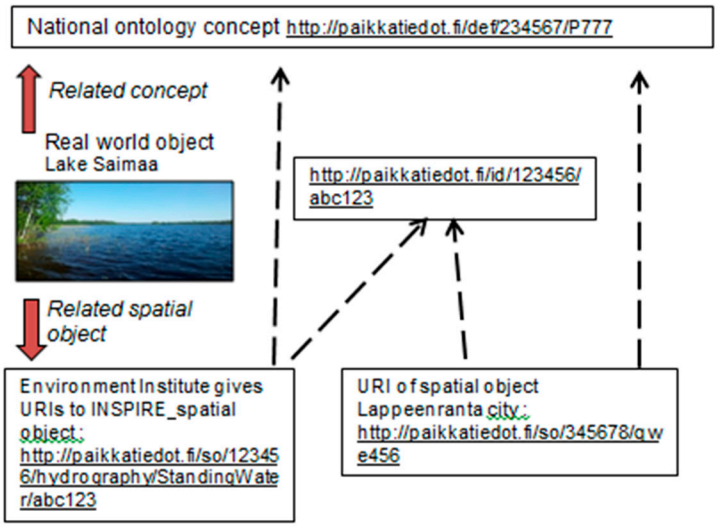

Figure 4. Implementation of URI references for real-world entities in the new national geospatial platform in Finland, according to National recommendation for unique identifiers of spatial data.

There are many questions attached to the design of such pieces for a Geographical Knowledge Graph: how to agree on what entities to put in the graph; how to define the boundaries of a real-world entity, in space and in time; and how to engage authorities in charge of informational entities of common interest to define URIs for these entities, as well as how to make connections with the represented real world entity and how to ensure user adoption of these URIs.

On the one hand, governments own and control systems of legal definitions. These systems are often hierarchical, and their structure can be traced top-down to identify the precise meaning of relations between concepts on the ontological level. Legally mandated authorities can contribute to such a graph. To do so, a straightforward solution experimented at Kadaster is to derive ontologies from the existing UML models of their three base registers (BAG, BRT, and BRK) and consequently linking them. On the other hand, this approach does not help in defining instance-level relations because their number and complexity grow very fast with every instance added to a KG. The network effect makes it difficult to foresee and formalize all possible relations.

This is why, complementary to authorities' publishing URIs and connections, we recommend a use-case driven approach to let entities of interest emerge, and possibly even new authorities. A lesson learned from the Web of Data is, in our opinion, to concentrate on the question of user adoption, even if a product is not flawless and to iterate a process where users give feedback on product specifications. Hence, we target involving users from the beginning of the creation of such a Knowledge graph. When it comes to serving URIs for objects in the real world, user adoption can be dependent on the consistency of data with current de facto authorities. Wikipedia is, in our opinion, necessary to align to. It is not a specifically geographic content, yet it is at the center of the cloud of linked data, meaning that it is the content most referenced. Wikipedia and Wikidata are widely used to provide URIs to features, for example, https://en.wikipedia.org/wiki/Eiffel_Tower for the URI of the Eiffel Tower in the French Wikipedia and https://www.wikidata.org/wiki/Q243 for the URI of the Eiffel Tower in Wikidata. Whereas Wikipedia URIs are more human readable—hence, also language specific—Wikidata is a knowledge graph and provides machine-readable statements about resources. Aligning to Wikidata

URIs is especially interesting for a European Knowledge Graph. Besides, we recommend that the Knowledge Graph be associated with possibilities to evaluate spatial and temporal relationships between user information and entities of interest. Indeed, location is very often a key dimension for interrelating data.

4. Evolving toward Open SDIs

In order to address capacity building and technology adoption, a first element is to collect the low-hanging fruits from LD technologies, as well as from a European Geographical Knowledge Graph (Section 4.1). A second element is to establish connections with third parties who design today the technologies of future Big Data infrastructures (Section 4.2). The last part is to establish connections between users and metadata (Section 4.3).

4.1. Fostering Adoption of SW and of EGKG from NMAs

To foster adoption of SW technologies by NMAs, as well as to evaluate capacities of the EGKG to improve semantic heterogeneities management, several "low-hanging fruits" have been identified. These are solutions that can be achieved with a limited amount of effort and bring a clear benefit.

4.1.1. Annotating NMA Registries with Linked Data and Shared Ontologies

The first fruit is making NMA registries more visible for SW developers and, more generally, for people using the Web to search for data. In order to do so, the investment is to publish dataset descriptions, using schema.org or DCAT standards. These metadata can be added directly into the web page hosting each registry and should be written in JSON-LD, RDF, or Microdata formats [38]. This will also contribute to the EGKG, as these metadata can be used to reference these datasets in it.

4.1.2. Supporting Feature-Centered Discovery and Exploration of NMAs Datasets

Another fruit is the capacity for a user to discover authoritative datasets based on entities of interest, for example, his house, a specific lake, or the Eiffel Tower. To support this, we make the hypothesis that the user can identify this entity in at least one referenced digital asset in our EGKG, including Wikipedia. Then, based on the links established between informational entities, the user can discover existing datasets that contain data about its feature of interest. This implies a shift from dataset to feature-level search, which could enhance the usability and take up of SDI data also outside the traditional geospatial domain. This is also a direct application of the EGKG, since the graph can be used to retrieve all digital assets associated with a geographic space containing the feature of interest.

Once datasets providing data related to the feature of interest are discovered, our proposal is to transform every retrieved data into Linked Data and to associate a facet browser to these pieces gathered together. Faceted browsers have been developed to present a user with different characteristics of a feature of interest from different sources on the Web; for example, when a customer wants to buy a television, many online stores allow customers to search for a television based on various properties, such as minimum rating, price, weight, screen resolution, and screen size. Customers are able to express a relatively complicated SQL query by interacting with various widgets (checkboxes and sliders) within the Web–user Interface. Generally speaking, creating a faceted browser is a relatively expensive and time-consuming process, since it requires non-trivial development effort for each database. With Linked Data, though, the properties in the database are described in semantic terms, specifying the domain and range types for each property. FacetCheck is a specific prototype developed at Kadaster that maps semantic descriptions onto UI widgets and underlying SPARQL subqueries. As shown in Figure 5, the FacetCheck UI consists of two components: the left-hand side of the screen containing the various widgets, while the right-hand side of the screen displays the entities that match the specified filters. When making selections within the FacetCheck UI, a SPARQL query is automatically assembled out of the subqueries associated with a widget. The entities that adhere to the specified query are retrieved and displayed on the right side of the screen for the user. An instance is

also displayed by a compositional widget. The components of an entity widget are determined by the direct properties that the corresponding entity has in the database. Since FacetCheck allows for the automatic generation of selection and entity widgets, it is relatively easy to create a FacetCheck browser over a specific Linked Dataset. An example configuration of FacetCheck can be used online (https://labs.kadaster.nl/browsers/). In Figure 5, the left-hand side of the screen shows the filters that are based on the properties in the dataset. By scrolling, over 100 data properties can be selected through a map, a slider, or a checkbox list. The right-hand side shows the widgets for the filtered results.

Figure 5. The prototype of FacetCheck which supports visual exploration of Linked Data.

4.1.3. Generating Geographical Data Cards

A last "low-hanging fruit" is the usage of data cards to serve a human readable representation of spatial data. These cards could open up SDI geospatial data to non experts and potentially introduce them to further datasets, thanks to links established between entities in different datasets. To collect this fruit, it is necessary to associate a given entity of interest to one or more maps that can describe it. This technique can benefit from the EGKG when it comes to data whose spatial characteristics are not well detailed. Knowing the represented features from reality, it is, in theory, possible to explore the EGKG and looks for other data representing the same feature but with a more detailed geometry that can be used to draw a more informative map.

4.2. Designing New Technology Spheres for Spatial Data Infrastructures

In order to experiment and evaluate a scalable EGKG, efficient infrastructures are needed. This section presents important third parties who provide this infrastructure and the functional requirements that must be shared with them.

The transformation of the Web from a web of documents to a Web of Data can have an important effect on the description, handling, and storage of geospatial and spatiotemporal datasets. The size and the complexity of such datasets increase, and therefore processing and storage become important issues. In that respect, cloud computing plays a critical role in establishing a highly responsive and scalable infrastructure that can adequately handle the challenges of geospatial-data processing today. Cloud can provide various desirable features for handling geospatial and spatiotemporal data, such as scalability and high-performance computing. On the other hand, it introduces new challenges such as the need to divide up the input geospatial data and the need to support distributed querying on them,

something inherently difficult in highly connected geospatial datasets. There are various efforts in this area, both research and industry led, that provide diverse solutions to choose from, e.g., the data-cubes technology based on "geocube" found in the portal http://www.geoportti.fi/.

There is a cluster of more generic efforts that are implementing generic RDF data management on the cloud, using and extending existing cloud storage solutions to store and query geospatial data, like BigQuery or Amazon S3 and SimpleDB, HBase and Google AppEngine, or Cassandra (http://hbase.apache.org/). Other efforts are using Hadoop and the BigData Warehousing framework to handle data storage and processing, combining MapReduce tasks with analytical query processing. Another notable and quite unique effort in its implementation is Strabon, even though it was not built with cloud-processing in mind. There is also a set of proprietary approaches that support RDF cloud-based data management, like BigData (http://www.systap.com/bigdata.htm) and OWLIM (http://www.ontotext.com/owlim$\delimiter"026E30F$&www.ontotext.com/owlim/geo-spatial). In addition, there are also commercial efforts usually integrated into well-known RDBMS software offerings, like Oracle or native RDF stores like Virtuoso [44]. To summarize, geodata processing on the cloud is required to support three main axes [45]:

- Data distribution: Distributing graph modeled data, like Linked Open Data (LOD), on the cloud is a hard problem, due to interdependencies among the different parts of the dataset.
- RDF/Semantic Data processing and querying: It usually requires large amounts of memory in order to materialize transitive closures for the whole or parts of the graph. This is not always available in cloud computing nodes, so additional distribution might be necessary to support this.
- Geodata analytics: These are highly based on using spatial correlations while data distribution on the cloud is usually oblivious to the actual geo-location of the data, so more elaborate methods should be used.

Cloud offers interesting capabilities for hosting SDI based solutions and supporting NMA activities, since it promises high availability, high performance, and virtually unlimited storage. The EGKG could be hosted on the cloud and reference data also hosted on the cloud or not, as shown in Figure 6. If our geospatial data includes personal (or related) information, despite the recent advances, people raise concerns about the privacy and security of various hosting offerings. Privately managed and hosted clouds or shared resources are necessary for organizations to tackle the benefits of the cloud, while, at the same time, tackling issues of security and privacy.

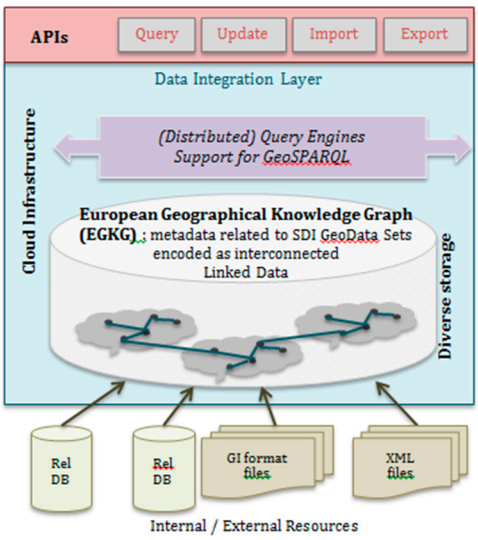

Figure 6. Implementation of the European Geographic Knowledge Graph (EGKG) on the cloud, reusing the general cloud-based infrastructure overview from [42].

To manage geosemantics, it is also important to investigate and experiment on the feasibility of different or combined metadata models (ISO, OGC, W3C: DCAT, Prov-O) in different and emerging data representations (RDF, Json, graph models, etc.) and their serializations. The OGC Geosemantic DWG is targeting stronger semantics for spatial data in order to increase awareness of the benefits of graph data and semantics within the OGC community and to gather support for eventually working on standards that will allow spatial data to be represented and queried effectively as graph data. That process should provide a forum for convergence of different parties, and eventually to start a convergence program, including OGC, W3C, research communities, and other relevant partners (like EuroSDR, JRC, Eurogeographics, EFGS, UNECE, and OECD, which has newly started elaboration of data integration and UN SDGs), and consider the aforementioned aspects.

From a technical standpoint, a substantial number of proprietary and open-source software is available to develop and maintain a full LOD stack (from ontology development to ETL processes, to graph storage and data delivery). However, best practices for choosing and implementing a LOD infrastructure are less readily available, and the knowledge thresholds are still prohibitively high for many. This situation is rapidly changing, however, not least due to international implementation projects delivering best practice or practical experience documents [46] and, even in some cases, full self-contained implementations [45]. Whilst more and more pilot and best practice projects are being conducted within the geospatial domain, one specific area of research that has been somewhat neglected is the use and efficiency of spatial queries over RDF utilizing the GeoSPARQL specification. A workshop at ISWC2017 in Vienna [47] presented findings from research into spatially enabled triple stores, with some queries taking milliseconds, but some (especially joins) taking minutes or hours. In addition, support for the full GeoSPARQL specification was only found in one triple store, with every other implementing only part of the specification. The lack of full GeoSPARQL support and relevant research material hints at a fundamental question for linked geospatial data: "What resolution of spatial data and spatial data analysis is appropriate for linked data?" Is it enough to use simple features geometries with only basic topological query support, for example, or should NMAs be exposing high-resolution complex geometry with full geospatial query and analysis support?

4.3. Collaborative Metadata Curation: Infolabs

The design of an open KG supposes the capacity of users to read metadata associated with datasets and to contribute to these metadata, or to informal descriptions of the datasets. In general, metadata are rarely documented. With INSPIRE, metadata documentation was achieved thanks to law enforcement and the application of penalties to countries where legal data providers would not provide metadata. We wish to improve the capacity of users to contribute to metadata, to provide a feed back to the data provider, or to share with other users how they use available data.

The concept of infolab has been introduced by the FING, Foundation Internet Nouvelle Génération, to accompany the growth of open data. In 2003 and during three years, the association analyzed, in three different regions and different themes (urban mobility, ecological footprints, and life-long orientation), problems encountered by users in exploiting available open data in their projects and activities. In 2013, they coined, as an element of solution, the concept of infolab, inspired by the success of hackerspaces, where people with common interests can collaborate around the usage of a complex resource (like a 3D printer). The FING extrapolated the concept of hackerspace to the reuse of open data and called such solutions infolabs. An infolab can be a web portal, a punctual event, a place, or even an association. The infolab is a solution to support the collaboration of different people around the exploitation of data. In the domain of SDIs, an infolab can be a platform for users to formalize the geographic information needed in their application, with commonsense words or with domain specific words, to share experiences in using existing digital assets, to push a question to the infolab community. Contributors to the infolab can also be data experts publishing references to digital assets possibly less visible for the users.

Two years ago, EuroSDR launched the geometadatalabs.eu project. It is a hub ready to host application specific infolabs. An application specific infolab is a simple wiki dedicated to discussions about usability of geospatial data for the given application. A first such application-specific infolab has been prototyped in the context of the European project URCLIM. In this project, climate scientists need data about specific big cities in Europe to study climate change [48]. In URCLIM infolab, a first phase targeted the specification of user requirements in terms of input data. This need is analyzed in [49] and is summarized on a wiki page. Gathering and sharing references to relevant resources to meet these needs is performed either automatically based on catalogue querying, or manually based on developers publishing their own experiences on the wiki. For example, Finnish participants from the meteorological institute derived a fine digital elevation model thanks to point-clouds data served nationwide by the National Land Survey of Finland and PostGIS tools. Researchers from CNRS in France developed a software program to produce urban indicators based on topographic and statistical data served by national institutes. First iteration of URCLIM infolab has led to more user requirements on the infolab itself: the capacity to load commonly used schemas and taxonomies in the wiki to support text-based search of attributes and looking up the definitions. It has also led to the identification of a complex choreography involving multiple actors from multiple countries to adapt a derivation process involving experiments on local data experimented on one city to another city in another country.

5. Conclusions and Perspectives

Achieving successful European SDIs requires an improved management of semantic heterogeneities. Although Semantic Web communities have developed promising technologies in this field, their adoption in SDIs beyond local or national prototyping is difficult, probably because of a knowledge gap and of difficult communication between these communities.

A first contribution is our analysis of existing competences and assets from NMAs and geographic information scientists on the one hand, and from SW and LD communities on the other, to improve European SDIs, and of differences in their perspectives on SDIs that can lead to conflicts and hinder collaboration.

A roadmap to contribute to successful European SDIs based on the assets of both communities has been elaborated jointly by researchers and practitioners from both communities. It embraces different inter-related aspects of a successful European SDI, from data and metadata to institution adoption, infrastructures and user adoption.

From a data- and knowledge-modeling perspective, considering the literature on spatial ontologies, as well as key assets of successful SW projects, we propose to build an open European Geographic Knowledge Graph (EGKG). It is structured according to the principles of an Ontology of Geographic Space Context and with Web technologies, reusing standard vocabularies on the Web. It will support better alignment, interconnection, and comparison between digital assets on the Web having spatial characteristics. This relies on providing any asset with the explicit description of the "spatio-temporal window of interests and filtering glass" that were used to build this asset. Providing such an EGKG, identifying its top nodes, and defining similarity measures on it require consensus and cooperation at an international level. This kind of engagement requires a long-term commitment from NMAs, to ensure a body of work that delivers the benefits already mentioned. The EGKG would need to be populated with datasets—complex digital assets—but also with more atomic items, features of interest. We recommend that this population is not a top-down process but, instead, driven by use cases. The definition of similarity measures on the Ontology of Geographic Space Context (OGSC) must consider existing similarity measures within an Ontology of Geographic Space (OGS), where spatial concepts are defined, and similarity measures within an Ontology of Geographic Information Technologies (OGIT) where information technology concepts are defined.

From an institutions' adoption perspective, it is necessary to illustrate benefits brought by the SW to contribute to unsolved issues in SDIs, to foster EGKG population and exploitation. The first

element is to collect a list of low-hanging fruits from Linked Data technologies (see Section 4.1), which are based on prototypes already developed, which we think can be useful to populate the EGKG and to assess its value in the context of SDI. From an infrastructure perspective, dialogue with information technology (IT) specialists designing next-generation infrastructures is essential as it will have a strong impact on Linked Data benefits. Functional requirements expressed from a European SDI perspective toward IT are identified.

Lastly, user adoption is also considered. Users may include decision makers, scientists or citizens. Our approach is to establish a stronger interface between people and metadata. It serves the objective of collaboratively building metadata, since the lack of metadata is a recurring flaw. Another perspective is that it also serves the objective to engage different disciplines in the exploitation and curation of ontologies.

Different validations of our approach are considered as perspectives. The first is our capacity to engage more and more organizations in this effort. It can be measured through the number of organizations adopting our proposed developments to collect low hanging fruits from SW and EGKG, presented in Section 4.1. We target European adoption but also cross-authority adoption within one country. Another validation is the capacity of our EGKG to be populated from different authorities. The evaluation of how the EGKG can contribute to solve unsolved issues in European SDIs is, in our opinion, partly tackled in the earlier development. For example, supporting user discovery of available datasets based on a feature of interest is a contribution to user-oriented catalogues. More work is needed on similarity measures and navigation methods over the EGKG in order to unlock the full potential of such a conceptual structure.

Author Contributions: Bénédicte Bucher is the lead author of the paper, writing the original draft and revising as well as conceptualization, investigating, funding and methodology. Esa Tiainen contributed in writing, reviewing and editing, investigation an methodology. Thomas Ellett von Brasch contributed in writing, review and editing, and investigation. Paul Janssen contributed in writing original draft on OWA and CWA, reviewing and editing. Dimitris Kotzinos contributed in writing original draft on cloud technologies, reviewing and editing. Marjan Čeh contributed in writing original draft on Geographic space context ontologies, Martijn Rijsdijk contributed in project administration, Erwin Folmer contributed in writing major revisions, reviewing and editing, Marie-Dominique Van Damme contributed in conceptualization and investigation. Mehdi Zrhal contributed in reviewing and editing during major revisions. All authors have read and agreed to the published version of the manuscript.

Funding: This work has partly been funded by EuroSDR and also partly by ERA4CS, an ERA-NET initiated by JPI Climate with co-funding from the European Union (Grant n° 690462) for the URCLIM project.

Acknowledgments: The authors acknowledge the contribution of participants to EuroSDR and Eurogeographics seminars and workshops on INSPIRE or on Linked Data. We deeply acknowledge the work of anonymous reviewers and of Lars Bodum, their constructive critics of a first version of this paper and suggestions for revision.

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

References

1. European Parliament; Council of European Union. *Directive 2007/2/EC Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*; OJL 108, 25.4.2007; European Parliament & Council of European Union: Brussels, Belgium, 2007; pp. 1–14. Available online: http://data.europa.eu/eli/dir/2007/2/oj (accessed on 20 January 2020).
2. The White House. *Office of Management and Budget (2002) Circular No. A-16 Revised*; The White House: Washington, DC, USA, 2002.
3. UN-GGIM: Europe Work Group on Data Integration. *Report on the Territorial Dimension in SDG Indicators: Geospatial Data Analysis and Its Integration with Statistical Data*; v1.4; Europe Work Group on Data Integration, Instituto Nacional de Estatística: Lisboa, Portugal, 2019.
4. Crompvoet, J.; Rajabifard, A.; Bregt, A.; Williamson, I. Assessing the Worldwide developments of National Spatial Data Clearinghouses. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 665–689. [CrossRef]
5. Masser, I. All shapes and sizes: The first generation of national spatial data infrastructures. *Int. J. Geogr. Inf. Sci.* **1999**, *13*, 67–84. [CrossRef]

6. Masser, I.; Crompvoets, J. *Building European Spatial Data Infrastructures*; Esri Press: Redlands, CA, USA, 2015.
7. Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
8. Noy, N. Industry-scale Knowledge Graphs: Lessons and Challenges. *Queue* **2019**, *17*, 48–75. [CrossRef]
9. Goodwin, J.; Dolbear, C.; Hart, G. Geographical linked data: The administrative geography of Great Britain on the semantic web. *Trans. GIS* **2008**, *12*, 19–30. [CrossRef]
10. Scharffe, F.; Atemezing, G.; Troncy, R.; Gandon, F.; Villata, S.; Bucher, B.; Hamdi, F.; Bihanic, L.; Képéklian, G.; Cotton, F.; et al. Enabling Linked Data Publication with the Datalift Platform. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012.
11. Hamdi, F.; Abadie, N.; Bucher, B.; Feliachi, A. GeomRDF: A geodata converter with a Fine-Grained Structured Representation of Geometry in the Web. *arXiv* **2015**, arXiv:1503.04864.
12. Debruyne, C.; Meehan, A.; Clinton, É.; McNerney, L.; Nautiyal, A.; Lavin, P.; O'Sullivan, D. Ireland's Authoritative Geospatial Linked Data. In Proceedings of the 16th International Semantic Web Conference, Vienna, Austria, 21–25 October 2017; Springer: Cham, Switzerland, 2017; pp. 66–74.
13. Ronzhin, S.; Folmer, E.; Maria, P.; Brattinga, M.; Beek, W.; Lemmens, R.; van't Veer, R. Kadaster Knowledge Graph: Beyond the Fifth Star of Open Data. *Information* **2019**, *10*, 310. [CrossRef]
14. Platform Linked Data Netherlands. Available online: http://platformlinkeddata.nl (accessed on 20 January 2020).
15. Use Of INSPIRE Data Workshop. Available online: https://eurogeographics.org/calendar-event/use-of-inspire-data-past-experiences-and-scenarios-for-the-future/ (accessed on 20 January 2020).
16. Bucher, B.; Tiainen, E.; Ellett, T.; Acheson, E.; Laurent, D.; Boissel, S. Data Linking by Indirect Spatial Referencing Systems. In Proceedings of the EuroSDR-Eurogeographics Workshop Report, Paris, France, 5–6 September 2018.
17. Bucher, B.; Laurent, D.; Jansen, P. Preserving Semantics, Tractability and Evolution on a Multi-Scale Geographic Information Infrastructure: Cases for Extending INSPIRE Data Specifications. In Proceedings of the Report of Eurogeographics-EuroSDR workshop on INSPIRE Data Extension, Warsaw, Poland, 27–28 November 2018.
18. The General Finnish Ontology YSO. Available online: https://finto.fi/en/ (accessed on 20 January 2020).
19. Atzemoglou, A.; Roussakis, Y.; Kritikos, K.; Lappas, Y.; Grinias, E.; Kotzinos, D. Transforming Geological and Hydrogeological Data to Linked (Open) Data for Groundwater Management. In Proceedings of the 10th International Hydrogeological Congress, Thessaloniki, Greece, 8–10 October 2014.
20. Atzemoglou, A.; Kotzinos, D.; Grinias, E.; Spanou, N.; Pappas, C. Transforming Geological and Landslide Susceptibility Mapping Data to Linked (Open) Data for Hazard Management. *Bull. Geol. Soc. Greece* **2016**, *50*, 1683–1692. [CrossRef]
21. Frank, A.U. Tiers of ontology and consistency constraints in geographical information systems. *Int. J. Geogr. Inf. Sci.* **2001**, *15*, 667–678. [CrossRef]
22. Kuhn, W. Semantic reference systems. *Int. J. Geogr. Inf. Sci.* **2003**, *17*, 405–409. [CrossRef]
23. Ceh, M.; Podobnikar, T.; Smole, D. Semantic similarity measures within the semantic framework of the universal ontology of geographical space. In *Progress in Spatial Data Handling, Proceedings of the 12th International Symposium on Spatial Data Handling, Vienna, Austria, 12–14 July 2006*; Riedl, A., Kainz, W., Elmes, G., Eds.; Springer: Berlin, Germany, 2006; pp. 417–434.
24. Fonseca, G.; Câmara, A.; Miguel, M. A Framework for Measuring the Interoperability of Geo-Ontologies. *Spat. Cognit. Comput.* **2006**, *6*, 309–331. [CrossRef]
25. Tambassi, T. *The Philosophy of Geo-Ontologies*; Springer: Cham, Switzerland, 2017.
26. Probst, F.; Lutz, C. Giving Meaning to GI Web Service Descriptions. In Proceedings of the 7th AGILE Conference on Geograpic Infirmation Science, Heraklion, Crete, 29 April–1 May 2004.
27. Kuhn, W.; Raubal, M. *Implementing Semantic Reference Systems*; AGILE: Lyon, France, 2003.
28. Heath, T.; Bizer, C. Linked Data: Evolving the Web into a Global Data Space. In *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1st ed.; Morgan & Claypool: Burlington, MA, USA, 2011; pp. 1–136.
29. Berners-Lee, T. Linked Data. Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 20 January 2020).

30. Ronzhin, S.; Folmer, E.; Lemmens, R. Technological Aspects of (Linked) Open Data. In *Open Data Exposed*; van Loenen, B., Vancauwenberghe, G., Crompvoets, J., Eds.; T.M.C. Asser Press: The Hague, The Netherlands, 2018; pp. 173–193. [CrossRef]
31. Osterwalder, A.; Pigneur, Y. *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
32. Archer, P.; Dekkers, M.; Goedertier, S.; Loutas, N. *Study on Business Models for Linked Open Government Data*; ISA Programme by PwC EU Services. European Union: Brussels, Belgium. Available online: https://ec.europa.eu/isa2/sites/isa/files/study-on-business-models-open-government_en.pdf (accessed on 20 January 2020).
33. Kobilarov, G.; Scott, T.; Raimond, Y.; Sizemore, C.; Smethurst, M.; Bizer, C.; Lee, R. Media Meets Semantic Web—How the BBC Uses DBpedia and Linked Data to Make Connections. In Proceedings of the European Semantic Web Conference, Heraklion, Greece, 31 May–4 June 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 723–737. Available online: https://link.springer.com/content/pdf/10.1007/978-3-642-02121-3_53.pdf (accessed on 17 September 2018).
34. Ehrlinger, L.; Wöß, W. Towards a Definition of Knowledge Graphs. In Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems, SEMANTICS 2016, Leipzig, Germany, 12–15 September 2016.
35. Wilcke, X.; Bloem, P.; De Boer, V. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Sci.* **2017**, *1*, 39–57. [CrossRef]
36. Paulheim, H. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semant. Web* **2017**, *8*, 489–508. [CrossRef]
37. Ballatore, A.; Bertolotto, M.; Wilson, D. A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS Int. J. Geo Inf.* **2015**, *4*, 471–492. [CrossRef]
38. Brickley, D.; Burgess, M.; Noy, N. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In Proceedings of the Web Conference, San Francisco, CA, USA, 13–17 May 2019; ACM: New York, NY, USA, 2019; pp. 1365–1375.
39. Keet, C.M. Open World Assumption. In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013.
40. Shapes Constraint Language (SHACL). Holger Knublauch; Dimitris Kontokostas. W3C. 20 July 2017. W3C Recommendation. Available online: https://www.w3.org/TR/shacl/ (accessed on 20 January 2020).
41. Brink van den, L.; Janssen, P.; Quak, W.; Stoter, J. Linking spatial data: Automated conversion of geo-information models and GML data to RDF. *Int. J. Spat. Data Infrastruct. Res.* **2014**, *9*, 59–85.
42. Smole, D.; Ceh, M.; Podobnikar, T. Evaluation of inductive logic programming for information extraction from natural language texts to support spatial data recommendation services. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1809–1827. [CrossRef]
43. Tiainen, E. Georef—Linked Data Deployment for Spatial Data. Finnish Initiative, FIG Working Week, Helsinki, Oral Presentation. Available online: http://www.fig.net/resources/proceedings/fig_proceedings/fig2017/ppt/ts02e/TS02E_tiainen_8781_ppt.pdf (accessed on 20 January 2020).
44. Kritikos, K.; Roussakis, Y.; Kotzinos, D. A Cloud-Based, Geospatial Linked Data Management System. In *Transactions on Large-Scale Data- and Knowledge-Centered Systems XX, Special Issue on Advanced Techniques for Big Data Management*; Hameurlain, A., Küng, J., Wagner, R., Sakr, S., Wang, L., Zomaya, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume XX, pp. 59–89.
45. Semic Interoperability Community. The Linked Data Showcase (LDS) Pilot: The Value of Interlinking Data. Available online: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/linked-data-showcase-lds-pilot-value-interlinking-data (accessed on 20 January 2020).
46. Ronzhin, S.; Folmer, E.; Mellum, R.; Ellett von Brasch, T.; Martin, E.; Lopez Romero, E.; Kytö, S.; Hietanen, E.; Latvala, P. Next Generation of Spatial Data Infrastructure: Lessons from Linked Data Implementations across Europe. Available online: https://openels.eu/wp-content/uploads/2019/04/V2_Next_Generation_SDI_Lessons-from-LD-implementations-across-Europe_1.pdf (accessed on 20 January 2020).
47. Bereta, K.; Stamoulis, G. Representation, Querying and Visualisation of Linked Geospatial Data. Available online: http://www.lirmm.fr/rod/slidesRoD04102018/RoD2018-tutorial.pdf (accessed on 20 January 2020).

48. Bucher, B.; Van Damme, M.-D. URCLIM Deliverable D2.1-1, URCLIM Infolab. Project Report, Paris, France. 2018. Available online: https://eurogeographics.org/wp-content/uploads/2018/07/08_BBucher.pdf (accessed on 20 January 2020).
49. Masson, V.W.; Heldens, E.; Bocher, M.; Bonhomme, B.; Bucher, C.; Burmeister, C.; de Munck, T.; Esch, J.; Hidalgo, F.; Kanani-Sühring, Y.-T. City-descriptive input data for urban climate models: Model requirements, data sources and challenges. *Urban Clim.* **2019**, *31*. [CrossRef]