Improving the documentation and findability of data services and repositories: A review of (meta)data management approaches

Article *in* Computers & Geosciences · August 2022

DOI: 10.1016/j.cageo.2022.105194

Tomáš Řezník
Masaryk University

Lieven Raes
Digtial Flanders

Andrea Perego
European Parliament

Karel Charvat
WirelessInfo/Plan4all

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Improving the documentation and findability of data services and repositories: A review of (meta)data management approaches

Tomáš Řezník[a,*], Lieven Raes[b], Andrew Stott[c], Bart De Lathouwer[d], Andrea Perego[e,1], Karel Charvát[f], Štěpán Kafka[f]

[a] Masaryk University, Faculty of Science, Department of Geography, Brno, Czech Republic
[b] Digitaal Vlaanderen, Brussels, Belgium
[c] University of Cambridge, Centre for Science and Policy, Cambridge, United Kingdom
[d] Open Geospatial Consortium, Wayland, USA
[e] European Parliament, Luxembourg
[f] Help Service – Remote Sensing Ltd., Benešov, Czech Republic

ARTICLE INFO

Keywords:
Metadata review
Open linked metadata
Open linked data
Geodata
Findability
Search engines

ABSTRACT

This scientific review paper aims at challenging a common point of view on metadata as a necessary evil and something mandatory to the data creating and dataset publishing process. Metadata are instead presented as a crucial element to ensure the findability of data services and repositories. This paper describes a way through four levels of metadata management and publication, from default unstructured data, through schema-based metadata with literal values and/or URIs, towards linked open (meta)data providing explicit linkage between reliable data resources. Such research was conducted within the European Union's project PoliVisu. Special attention is given to the following: (1) guidance on publication aimed at the broad audience of search engine users and (2) the publication of geo (meta)data not only via standard technologies, such as the OGC Catalogue Service for Web and open data portals, but also through leading search engines (that are Schema.org-based).

1. Introduction

The importance of metadata is often underestimated. For data users, metadata are commonly regarded as a necessary evil and something mandatory to the data creation and dataset publishing process. Metadata specialists - a small group of specialists in the IT, geo-communities, scientific, and archive communities - have their particular conceptions of an ideal metadata description. Today, several metadata approaches subsume different levels of effort, complexity, and profit (Dublin Core, 2020; FGDC, 1998; ISO 19115, 2003; ISO 19115, 2014; W3C, 2020b; JoinUp, 2020). Metadata are a crucial element in data-driven decision making because of the essential link between reliable data and the reliability of the decision outcome. They play an essential role during the data collection process regarding dataset characteristics, provenance, and data quality. The concept of metadata, as presented within this review, stems from research conducted in the PoliVisu project (PoliVisu, 2020).

The 1990s and the beginning of the new millennia could be characterized as an era of metadata enlightenment. The importance of metadata was stressed where applicable, and the (misleading[2]) definition of "data about data" was emphasized on almost all relevant occasions, such as scientific conferences, scientific/technical papers, workshops, and commercial (software) presentations, etc. (Weibel et al., 1995; Shien-Chiang et al., 2003; FGDC, 1998; Jensen et al., 2000). This period culminated in the adoption of several legally-binding texts that require metadata creation as well as their maintenance. For instance, in Europe, the INSPIRE directive (European Commission, 2007) and it's (chronologically the first) accompanying Commission Regulation, No 1205/2008 on metadata (European Commission, 2008), are the most evident proofs of the above-described efforts originating from the 1990s. The primary motivation of the INSPIRE Directive is clear: (geo)data can only be shared and re-used if they are findable. The FAIR Guiding Principles for scientific data management and stewardship are being followed across the world (Wilkinson et al., 2016).

* Corresponding author. Masaryk University, Department of Geography, Kotlářská 2, 611 37, Brno, Czech Republic.
E-mail address: tomas.reznik@sci.muni.cz (T. Řezník).

[1] The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Parliament.
[2] Metadata describe various assets except for data; such as Web services, applications, models; field sessions, projects, sensors etc. (ISO 19115, 2003).

https://doi.org/10.1016/j.cageo.2022.105194
Received 2 October 2020; Received in revised form 16 November 2021; Accepted 10 July 2022

The next era, since 2004, is most commonly referred to as the era of the "Semantic Web" (W3C, 2015a). It is characterized as an extension of the World Wide Web (WWW) rather than a new concept. The principle remains the same as in the INSPIRE Directive case, i.e., to re-use and share data across the WWW, with the intention of obtaining more knowledge from the flood of existing data. The Semantic Web Framework is considered as an integrator across different contents, information systems, and communities. It has been acknowledged as a feature of the WWW since 1989 in a paper by the founder of both the WWW and the originator of the term "Semantic Web", Tim Berners-Lee.

Semantic metadata are metadata that describe the "meaning" of data (Melton and Buxton, 2006) in order to promote them for re-use and sharing. Another benefit of semantic metadata is not commonly mentioned: the revoking of the artificial boundary between data and metadata.

The contemporary desire for everything open, interconnected, customizable, shareable, scalable, and fit for use with artificial intelligence applications may create the impression that the Semantic Web and its application are followers of traditional approaches, including metadata. This document is a guide about whether to:

- remain using non-semantic metadata approaches;
- follow semantic metadata approaches: and if so, at which level;
- combine both approaches.

This paper provides a review of (section 3), and guidance at a decision making level (section 4) on whether, and, if so, to what extent, a shift towards semantic metadata approaches should be made by a metadata creator/administrator. It also shows how such modification should be made in order for it to be the most beneficial and cost-efficient at the same time (sections 5 and 6).

2. Methods and related work

This study aims to provide geo-based communities with a comparison and evaluation of advances in metadata development towards the open linked (meta)data paradigm, as well as the motivation to adopt this paradigm. The concept of open linked (meta)data enables the artificial boundary between data and metadata maintained in the geosciences (and other domains) to be annulled. Such an artificial boundary originates from separate conceptualisations of metadata, e.g., in the case of Dublin Core (Dublin Core, 2020), CSDGM (FGDC, 1998), and ISO 19115 (ISO 19115, 2003; ISO 19115, 2014).

The interoperability, flexibility, and openness of metadata have become crucial questions for any location-based information system since the number of assets that need to be described through metadata has multiplied.

We may identify many studies on metadata in geosciences, including linked open (meta)data approaches, as described in detail below. As far as the authors are aware, no study similar to this one has been performed so far. This study provides a synthesizing picture in contrast to state-of-the-art studies, as it comprises the following three major features: (1) a review of metadata management in geosciences, explaining the evolutions/revolutions in metadata approaches & metadata publication systems, (2) guidance on publication aimed at the broad audience of search engines users, and (3) guidance on whether, and, if so, to what extent, a shift towards semantic metadata approaches should be undertaken. Geosciences make up the sixth largest component within the Linked Open Data (LOD) cloud, after life sciences, government, linguistics, publications, and social networking (McCrae, 2020a).

Scientific papers indexed at the Web of Science (Clarivate Analytics, 2020) and applicable to metadata within geosciences were analysed from the following points of view:

- numbers of available papers concerning the ten most common phrases related to the scope of this paper (see Fig. 1 for results);
- analysis of review papers relevant to the scope of open linked metadata; both within and beyond geosciences; (see Fig. 2 for results);
- analysis of papers relevant to the scope of open linked metadata concerning applicability at the following levels (see Fig. 3 for results):
  - inspirations beyond geosciences that may apply to geosciences,
  - applications that are entirely or partly related to geosciences (as a multidisciplinary approach).

In general, the topic of metadata is strongly accented in geosciences. The description of an application accompanied by metadata is the most common one (Di et al., 2009; Ho et al., 2011; Maue et al., 2012; Zhang et al., 2013; Da Silva et al., 2014; Kalantari et al., 2014; Hu et al., 2015; Lafia et al., 2016; Palma et al., 2016; Reznik et al., 2016; McGee et al., 2017; Neumaier et al., 2018). Such papers mostly deal with an explicitly defined scope, like the integration of open linked metadata into cloud platforms, sensor webs, geovisual analytics, Volunteer Geographic Information, e-government, libraries for geo-resources, cartographic models, and Web services. The only geo application of open linked metadata to open data, in general, has so far been presented by Neumaier et al. (2018).

The books by Moellering et al. (2005) and Nogueras-Iso et al. (2005) still contain the most comprehensive overviews of international as well as national metadata initiatives, structures and standards, despite being published fifteen years ago. Since then, we may identify a number of relevant studies dealing, for example, with the following aspects: positioning metadata in the framework of international standardization (Furner, 2020), the evaluation of and outlook on the international standardization of geographic information metadata (Brodeur et al., 2019), and metadata life cycles, spirals, use cases, hierarchies, and their interconnections through links (Habermann, 2018). An analysis of types of metadata and their use by Danko (2012) has a unique position as a similar study has not since been published. The management of metadata collections remains an open challenge, primarily concerning their completeness, topicality, and accuracy, as Giles (2011) described.

Analysis of all the studies examined for this paper confirmed the following. Dublin Core (Dublin Core, 2020) and ISO 19115 (ISO 19115, 2003; ISO 19115, 2014; together with related ISO, 19100 series standards) remain the only two significant players regarding international metadata standards for geosciences. Australia and New Zealand (ANZLIC, 2020), China (Baiquan et al., 2013), Europe (European Commission, 2008.) and the United States of America (FGDC, 1998.) are all examples countries/regions shifting towards ISO 19115 and Dublin Core in geosciences, although national metadata standards are also being used in parallel.

Only five review papers were found concerning open linked metadata in geosciences. Smits and Friis-Christenson (2006) as well as Govedarica et al. (2010) discuss the potential of open linked metadata in general. Di et al. (2009) present an in-depth discussion for sensor metadata. Such work was extended by Tagliolato et al. (2019), who developed semantic profiles for SensorML descriptions. An in-depth discussion of Earth observation data is provided by Harris and Olby (2001). The most comprehensive recent review paper on the topic of semantic geoinformation modelling was written by Kokla and Guilbert (2020). One section of this paper is dedicated to semantic search and knowledge discovery that is based on ontologies. No review paper with a holistic view of metadata in geosciences has yet been found. The most comprehensive summaries have so far been published by W3C/OGC Spatial Data on Web Best Practices (W3C, 2017b). The presented research is in line with these best practices, as it follows and enhances them in terms of open linked metadata.

This review aims to fill relevant knowledge gaps with respect to metadata approaches and publication systems at various levels. The concept of open linked metadata is emphasized as the most novel one, and as the one not described sufficiently in existing papers.

3. Characteristics of metadata approaches & metadata publication systems

This section provides an overview at the management level of different stages of metadata approaches. These metadata (management) approaches are tightly connected to metadata publication systems. The primary intention of the following text is to present an overview that is demarcated by historic milestones. Note that the presented levels indicate different metadata approaches; there is no intention to compare levels as better or worse. The main difference is historical: a level with a lower number was created earlier. Nevertheless, metadata have been produced at all levels now and will presumably also continue to be in the future.

The various metadata approaches (see also use cases in Supplementary material 1) presented in this section can also be understood in the following way:

- Level 0 (section 3.1.1) is the **default unstructured metadata** used within many typical IT systems.
- Level 1 (section 3.1.2) represents a revolution that changes the paradigm of metadata management and publication to **schema-based metadata with literal values.**
- Level 2 (section 3.1.3) is an evolution of Level 1, with an emphasis on the usage of unique identifiers in addition to literals whenever applicable, giving **schema-based metadata with unique identifiers.**
- Level 3 (section 3.1.4) represents a further revolution that changes the paradigm of metadata management and publication, as it removes the artificial border between data and metadata. The metadata is itself **linked open (meta)data.**

The terms 'revolution' and 'evolution' are used here in the following ways. 'Revolution' means that the metadata management from the previous level is thrown out and entirely replaced by a new one. In contrast 'evolution' implies that the majority of the features of the prior level are kept.

3.1. *The data management step-up process*

3.1.1. *Level 0: default unstructured metadata*

The most basic level of metadata management relies solely on automatically created metadata in a system. Typical examples in a data repository are file size or date of last modification, or information implicitly associated with a data resource, such as a file name or type of file (format).

Three primary benefits may be defined for Level 0 metadata management as follows:

1. **Easy to deploy.** Automatically/implicitly managed metadata are a part of several IT systems, including database solutions. In many cases, automatic/implicit metadata management is the default one. Even if not, it can be deployed quickly from the administrator's console. Level 0 does not require any specialized metadata software, as metadata remain a natural part of data. As a consequence of such tight bindings, the life cycle of metadata is equal to the life cycle of data.
2. **User-friendliness.** No training is needed, as metadata are so simple that everyone understands metadata elements like file size, file name, date of last modification, access rights etc.
3. **Tight (meta)data bindings.** Metadata are automatically created/updated/deleted together with the data they describe. Life cycles of data and metadata are equal.

The simple approach of **Level 0** also brings **challenges** that are tightly connected to the benefits of easy deployment, user-friendliness and tight (meta)data bindings:

1. **Fixed metadata elements.** Metadata elements are clear to users; however, they are based on the capabilities of the used system. Their customization/extension is usually not supported (or to a minimal extent). It is very complicated, or even not feasible, to add other metadata elements, such as spatial extent, even if there is a user requirement.
2. **Absence of publication.** Metadata are presented only within a system; their export and/or presentation in a different way from the default one is not supported.
3. **Limited findability.** Metadata are presented only at the level of data display. For example, searching for relevant data according to given criteria is not supported, i.e., in a catalogue service based on metadata. Findability capabilities are limited only to metadata automatically created/implicitly associated with a data resource if supported by the application logic of the used system.

3.1.2. *Level 1: schema-based metadata with literal values*

"Traditional", i.e. non-semantic, metadata approaches can be characterised as the management of textual key-value pairs. For instance, we have a key "title" and value "MySampleGeodata". Such a concept is at least as old as relational databases (Codd, 1970) and has remained in use since the 1970s. All the many different metadata standards and definitions vary, among others, in the following aspects:

- **Wrapper**: whether metadata are used as stand-alone data (ISO 19115, 2003; ISO 19115, 2014) or as a fraction of something bigger (Dublin Core, 2020), like a Website, database, Web server response, or a table linked to other data etc.;
- **Complexity and structure**: a metadata record could be represented by a flat list of (by default up to 15) key-value pairs (as in Dublin Core, 2020), through a more complex structure is possible, with dozens of metadata elements and very complex structures (as in GeoDCAT-, 2015.), hundreds of metadata elements, and several hierarchical levels (as in ISO 19115, 2003; ISO 19115, 2014).
- **Exchange format**: This could be null, in the case of metadata managed in a table within an internal database, or it could be CSV (Comma Separated Value); export table structures, such as XLS (Microsoft Excel Spreadsheet) up to XML (eXtensible Markup Language), according to the standards' XML schema definition; and RDF (Resource Description Framework – W3C, 2014a). The Level 1 ("traditional") approaches encode metadata in any of the above-mentioned (and several other) exchange formats, including their combinations. However, not all the formats are "self-describing". Consumers of exchange formats such as CSV and XLS would need to know that they contain metadata in a particular layout (which might mean a further level of metadata!), or the CSV or XLS produced might have to comply with strict standards to be ingestible into established catalogues or search services.
- **Publication**: in general, a broad portfolio of cataloguing solutions is available. However, the given structure determines which publication tools are suitable. For instance, the OGC Catalogue Service for the Web (OGC, 2016) supports by default a structure based on Dublin Core, but, through its application profiles, also structures following ISO 19115/19119, ebRIM (registries) (OGC, 2006), or (North American) CSDGM (FGDC, 1998).

Level 1: schema-based metadata with literal values address the challenges of Level 0 in terms of fixed metadata elements, absence of publication, and limited findability. The following **benefits** document the shift towards Level 1:

1. **Strict, yet flexible, structure.** A structure that follows a given schema is the essential feature of Level 1. Such an approach brings a number of clearly defined metadata elements and their organisation. Metadata elements are commonly defined by their names, textual descriptions, cardinalities (including obligations), data types, and

domains. However, a user may set up his/her metadata profile, as far as it follows the rules given by the schema. A user may change an optional metadata element to conditional/mandatory, or change the data type from a character string to a code list, etc. Under certain conditions, a user may also add his/her new metadata elements that are not included in the schema.

2. **Metadata publication.** The schema is usually defined on both the conceptual level (textual/tabular descriptions accompanied with graphics, most typically expressed as a UML class diagram, e.g. in ISO 19115, 2003; ISO 19115, 2014) and the implementation level. XML Schema (XSD) is a typical technology that is being used to capture encoding rules. A schema definition on the implementation level makes it possible to incorporate (semi)automatic validation tools, like XML schema validation and Schematron validations (Schematron, 2020).
3. **Findability layer.** Searching and findability represent application logic on top of the published metadata records. Catalogue services make it possible to search for and discover relevant resources according to given criteria. Examples in this direction could be: "Show me all the datasets that provide measurements between 2017 and 2019 in Pilsen city by using traffic intensity detectors".
4. **Interoperability.** The ability of systems to work with each other is initiated, as the structure as well as catalogue APIs (Application Programming Interfaces) are clearly defined and originate from an (international) standard. Metadata may be transferred from one system to another; related catalogues may be connected, etc.

Level 1: schema-based metadata with literal values bring **challenges**, similarly to any change of a paradigm. The following challenges were identified:

1. **Difficult to deploy.** Metadata management based on schema-based metadata with literal values usually requires specific metadata tools and staff acquainted with their set up, customizations, and maintenance. The required efforts are equal to the complexity of a given structure/schema. Deployment can be relatively easy for structures/schemas like Dublin Core, with up to 15 metadata elements in a flat structure (Dublin Core, 2020), but very complex for ISO 19115/19119, which includes hundreds of hierarchically organised metadata elements (ISO 19115, 2003; ISO 19115, 2014). Note that efforts with respect to ISO 19100 series metadata can vary considerably, as most of their metadata elements are optional. An ISO 19100 metadata profile can contain dozens or hundreds of metadata elements. Among other issues, ambiguities in Level 0 metadata are dependent on the documentation of a code list granularity. Also, the deployment of catalogue services is specific and requires trained staff.
2. **User-friendliness.** More complex metadata structures/schemas are difficult to understand for both metadata administrators and users. Complex hierarchical structures may decrease the clarity of the information presented in the metadata.
3. **Loose (meta)data bindings.** Metadata require, in Level 1, specific structures, tools and handling, which results in various (and commonly isolated) life cycles of data and metadata.
4. **Ambiguities.** The provided character strings and/or code list values are commonly non-intuitive, as they capture not easily interpretable values such as 'DTM', '007' or 'NoiMesAboGro2mHei.'

3.1.3. *Level 2: schema-based metadata with unique identifiers*

Level 2 is understood, in contrast to the previous two levels, as the first semantic-oriented level. Values are presented as identifiers in comparison to Level 1 to reduce ambiguities. The shift to Level 2 addresses the above-raised points as follows:

- **Wrapper**: remains the same as in Level 1. There are no changes regarding the metadata wrapper (this is the most important change in comparison to Level 0). Websites, databases, Web server responses, tables linked to other data etc. Are still being used in Level 2;
- **Complexity and structure**: where applicable, it could use a triple-based (i.e. subject-predicate-object; W3C, 2014a) or even more complex structure. For example, the relation between the "Coordinate reference system" and "urn:ogc:def:crs:EPSG:4326" (EPSG, 2019) is a typical unique identifier used in geosciences metadata. A user is then sure of what kind of coordinate system is meant. The unique identifier "urn:ogc:def:crs:EPSG:4326" is used explicitly by the Open Geospatial Consortium (OGC) for the two-dimensional expression of WGS84 (World Geodetic System 1984). In Level 2, unique identifiers are re-used as much as possible. New unique identifiers can be created if existing ones are not available and/or not applicable for the given purpose. However, although this ensures uniqueness, it does not necessarily aid the association or combination of datasets unless the same unique identifiers are used for the same entities throughout, say, an organisation.
- **Exchange format**: This remains the same as in Level 1. The given structure can still be exchanged in a table within an internal database, through CSV, export table structures like XLS, XML, or RDF.
- **Publication**: This can remain the same as in Level 1, i.e. a broad portfolio of cataloguing solutions is available in general, while the given structure determines suitable publication tools. The benefits of the findability layer (publication) appear when a cataloguing solution has an application logic that supports interpretations of unique identifiers. A user may then see the types and definitions of a value provided as a unique identifier instead of seeing the unique identifier itself.

We may identify the following differences on top of the modifications mentioned above. The semantic approach based on unique identifiers adds the "types and definitions of data", a term which requires an explanation so as not to be interpreted differently. The key-value pair concept is addressed in the semantic web as follows:

- *Keys* that add the types and definitions of the names of things/relevant concepts (such as noise, which can be defined according to the altitude above the ground, and the methodology, etc. provided); examples of *Keys* are 'keywords', 'spatial extent', 'coordinate reference system', 'identification of an organisation' etc.;
- *Values* that are provided together with their types and definitions (such as units of measure); examples of *Values* are 'gemetKeyword: noise_measurement', 'vocab.gettytgn/7011723' (for Pilsen), 'urn:ogc:def:crs:EPSG:4326' (for WGS84), czechGov:Ministry_of_Transportation'. Values are unique identifiers agreed and followed in a domain/area of use.

Level 2 schema-based metadata with unique identifiers **have identical benefits as** described **in Level 1**, i.e.:

1. **Strict, yet flexible, structure.**
2. **Metadata publication.**
3. **Findability layer.**
4. **Interoperability.**

Furthermore, Level 2 of metadata management brings **one new motivation** for its use:

5. **Clarity.** The unique identifier-based approach aims at explicit designation. It is evident through a unique identifier 'vocab.gettytgn/1014734' that the described "East York" is the one in Canada, Ontario and not the one in the United States, Pennsylvania, or any other "East York" around the world. The unique identifier-based approach is used for locations, keywords, identifications of underlying data resources, and coordinate transformation systems, etc.

The fifth benefit is described more in detail below.

In any context in which data exist, there are values that represent certain concepts; human decisions, including policy-making, are then (evidence-)based on interpretations of such values. Associating specific values with specific concepts is one way of assigning types and definitions to data made up of those values. For example, the following values demonstrate the same concepts of traffic noise measurements: “day” and “night”, “1” and “0”, “Lday” and “Ln”, or “L07-19” and “L20-06”. A metadata registry captures the allowed values for some key as managed by some registration authority. A mechanism by which the names of “things” and the values assigned to them can be managed makes them easier to be found and interpreted in various data sources.

As a result, the semantic approach prefers unique identifiers over text strings to populate a certain value. A value for noise could, for instance, look like “urn:noiseAuthority:measurements:registry:noisePeriod:day” instead of “day”.

The added value for a user is receiving the types and definitions relating to this value, e.g. that a noise measurement conducted during the day means between 7 a.m. and 7 p.m., at a height of 2 m above the ground, while other bias noises were suppressed by further processing. In particular, sensor networks including the IoT (Internet of Things; Gubbi et al., 2013) benefit from semantic metadata through concepts like SOSA (Semantic Sensor Network Ontology; W3C, 2017a), a joint effort of the World Wide Web Consortium and the Open Geospatial Consortium.

Versioning in a unique identifier provides an opportunity to capture the evolutions of a resource. Versioning commonly includes a version number, e.g. “1.3.0”, or a date of revision/update, most commonly in line with ISO 8601-01 (2019), e.g. a date like “2021-11-10” or with a timestamp like “2021-11-10 T12:00:07”. Moreover, EPSG (2019), for example, assumes the latest version when the version number is missing. For instance, the unique identifier “urn:ogc:def:crs:EPSG:4326” contains “::” precisely for such a purpose. The greatest challenge in versioning lies at the implementation level. All implementations using such unique identifiers with resource versioning must follow all the evolutions, a non-trial task.

Linking a value to a corresponding registry, thesaurus, and/or gazetteer can be achieved for any application. When moving to semantic approaches, starting from Level 2, geosciences are being seamlessly integrated into the e-government concept: geo- and non-geo- resources are handled equally and can be linked together.

The challenges identified at Level 2 of metadata management remain the same as in Level 1 with one exception: ambiguities are no longer a challenge, rather a benefit. The following challenges from Level 1 remain valid also for Level 2 metadata management:

1. **Difficult to deploy.**
2. **User-friendliness.**
3. **Loose (meta)data bindings.**

3.1.4. Level 3: linked open (meta)data

The shift to Level 3 of metadata management includes the adoption of referenced URIs. Open data available on the Web are linked to other data through URIs. The shift to Level 3: linked open (meta)data addresses these points as follows:

- **Wrapper**: This is enhanced in comparison to previous levels. In general, anything WWW-related can be re-used: from a Website or a Web server response. However, some kinds of resources remain inaccessible to metadata applications - for instance, even a Semantic Web application cannot address the metadata of an e-mail.
- **Complexity and structure**: any (meta)data structure identified in Levels 1 and 2 can be used as an underlying one (Dublin Core, CSDGM, ISO, 19115, etc.). Level 3 is the most complex as it provides URIs in a dereferenceable way. Unique identifiers in Level 2, like 'gemetKeyword:noise_measurement', are provided as URL links, such as 'https://www.eionet.europa.eu/gemet/en/concept/5646'. A user may click on the link and get to another source of information. Such an approach enables relevant pieces of information to be connected and improves the decision making process on top of the new connections (links). Moreover, Level 3 allows users and applications to identify equivalencies, hierarchies (parents, children, siblings), broader terms (like measuring – GEMET, 2020a) or related terms (like noise analysis – GEMET, 2020b), and homonyms (words/terms which sound alike or are spelt alike but have different meanings) etc. Another point of view lies in the structure of metadata element itself.
- The shift between different levels can be illustrated on the example of a creator metadata element as follows:
  - At Level 0, only the implicitly provided name of the file owner at the current time is available. This also means that a contributor's identification depends on the system used.
  - At Level 1, key-value pairs with literal values are provided. For instance, a key *creator* has a value 'John First'.
  - At Level 2, values are supported by unique identifiers to identify the corresponding resources uniquely. For instance, a key *creator* has a value as the unique identifier 'JohnFirst003'. Such an approach enables us to identify the right John First explicitly.
  - At Level 3, a collection of FOAF (Friend of a Friend; for further details see FOAF (2014) objects is provided. For instance, the (semantic) triple can be expressed as 'dc:creator = “http://myOrganisation.policy/staff/JohnFirst”'. Complexity should be depicted in ontologies, as the object (“value”) http://myOrganisation.policy/staff/JohnFirst can be a predicate (“key”) to other objects (“values”), e.g. linking http://myOrganisation.policy/staff/JohnFirst with information about him on a web page.

Such an example tends to illustrate the fact that statements on semantic metadata approaches are not black or white, as semantic approaches are not 'all or nothing'.

- **Exchange format**: three formats are defined as the default ones: RDF/XML (W3C, 2014c), Turtle (W3C, 2014b), and JSON-LD (W3C, 2020a), as they are all capable of handling links to other information resources and are “self-defining” in the sense of allowing a consuming process to interpret the metadata without any additional information. However, referenceable URIs may also be handled in XML through XLink (W3C, 2010), as in <gmx:Anchor> tags in ISO 19139 (ISO/TS 19139, 2007; ISO/TS 19139-1, 2019) compliant encoding.
- **Publication**: follows the structure(s) used and the exchange format (s). For instance, CKAN (Comprehensive Knowledge Archive Network) with proper extensions (GitHub, 2020) is used as a wide-spread cataloguing solution. Linked open (meta)data also open new publication possibilities for leading search engines. Such an approach may attract more users than before. Moreover, such attraction may appear in terms of user-friendliness, as depicted in Fig. 4. Note that any search engine following Schema.org can be re-used for such publication. The leading search engines mostly require metadata to be embedded in Web pages using HTML + RDFa (W3C, 2015b) or JSON-LD (W3C, 2020a) snippets. If metadata are expressed in RDF, this could facilitate indexing, provided that they are not available only separately, but embedded in HTML pages following SEO (Search Engine Optimisation) techniques. Interoperability is not granted. Moreover, these rules are different for different information resources. As an example, see the differences in recipes (Fig. 4) and datasets (Fig. 5). When metadata are expressed using Dublin Core (Dublin Core, 2020) and DCAT (W3C, 2020b), they can be indexed without the need to convert them to Schema.org. Moreover, the Schema.org terms for datasets and catalogues are modelled on DCAT, so the correspondence is pretty straightforward.

A successful publication in Level 3 also requires the following decisions:

- which leading search engine(s) are desired,
- which information resources are desired.

Search engines mostly omit or simplify the metadata of geo resources (typically by avoiding spatial extent information) due to their specificity. According to Clarke (2012), semantic metadata can be used to increase traffic from search engines, as such metadata can provide search engines with more information about the content being searched.

For this reason, several of the leading search engines have begun working together, via Schema.org (2020), on the development of standards that facilitate a greater level of metadata exposure. As a consequence, the leading search engines provide the user with an answer that best matches his/her search history, i.e. a user profile. To sum up, semantic metadata are one of the ingredients of machine learning algorithms - and this not only within the leading search engines.

There are dozens of kinds of Schema.org-based rich (meta)data content supported by the leading search engines: from 'article' through 'dataset', 'event' or 'recipe' to 'video'. In general, 'dataset'-rich results seem to be the most common way of describing geoscience resources. However, 'dataset'-rich results do not appear on the entry pages of some leading search engines (Brickley et al., 2019). The primary advantage of rich results is lost – a user-friendly visualisation that is common to non-geo metadata as depicted in Fig. 4. Instead, geo metadata are presented without the benefits of rich results, as depicted in Fig. 5.

The benefits and challenges of Level 3, open linked (meta)data, are described in greater detail in section 4 due to their complexity and novelty.

4. The benefits and challenges of open linked (meta)data management and publication

The text in the following sections presents the benefits of, and challenges facing open linked (meta)data management and publication in more detail.

Level 3 is a further shift in comparison to Level 2. The main benefit of Level 2 is clarity in the types and definitions of data. Level 3 uses, contrary to Level 2, dereferenceable URIs that point to other relevant resources. Thesauri, gazetteers and/or registries are used as primary sources. For instance, the Level 2 unique identifier 'vocab.gettytgn/1014734' is modified for Level 3 as follows: "http://vocab.getty.edu/page/tgn/1014734". The linked open (meta)data approach increases user-friendliness, as a user can click on a link and obtain the information directly from its source. Moreover, the application logic of a system is capable of employing improved processes that automatically connect relevant pieces of information.

4.1. Benefits

As noted at the beginning of section 3, Level 3 represents a revolution with respect to both metadata management and publication. For this reason, the benefits, as well as challenges, relating to it will be described in greater depth in comparison with the previous levels.

4.1.1. Openness

Open data is a paradigm that, in the last decade, has been emphasized more and more in two major communities: in research and in public administrations. Regarding the latter, the open data paradigm enables the re-use and creation of added value and evidence-based decision making on top of already published data. Also, for this reason, public administration bodies commonly desire to publish openly as much data as is feasible. Such a situation leads, in some cases, to a heap of data in which it is difficult for users to orient themselves. For example, Fig. 6 shows an overwhelming number of results on available visualisations at the EU Open Data Portal (https://data.europa.eu/euodp/en/visualisation-home), which, however, are not connected in a user-friendly way. Such difficulties from a user experience point of view result in quite a low number of visitors to the EU Open Data Portal in comparison to national geoportals.

4.1.2. Dissemination to the masses

The basic weakness of both open data portals and geoportals is the fact that users need to know that they exist. Finding the right tool for findability may be an even more severe obstacle than searching within the discovered tool. Geoportals are not commonly known to people outside of the geo bubble, while open data portals are not commonly known to people outside of the open data (e-government) bubble.

Level 3 (the linked open (meta)data approach) facilitates **the user-friendly publication of available (meta)data in leading search**

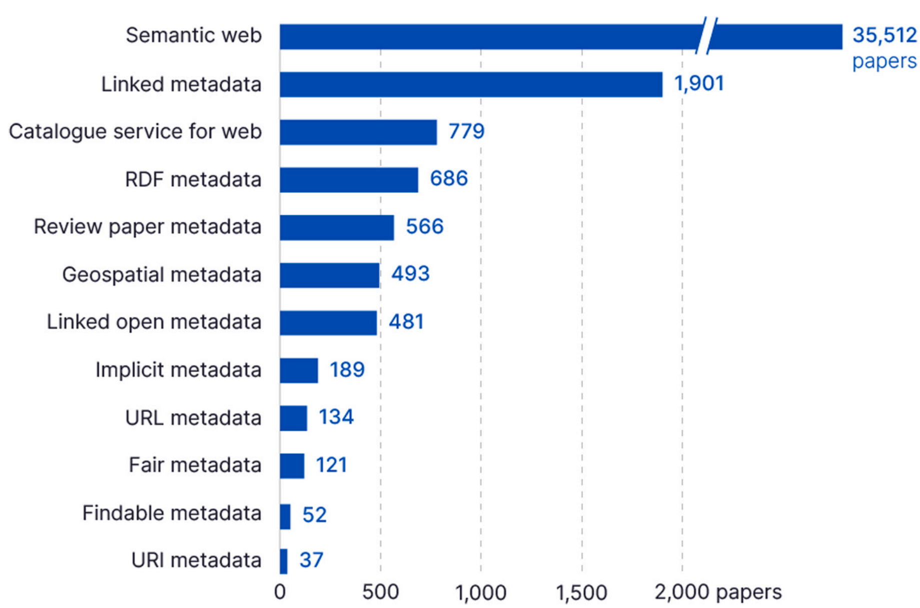

Fig. 1. Numbers of papers available at the Web of Science concerning eleven phrases related to the scope of this paper. Note that the analysis of papers for the term 'semantic web' was not performed in this study as it would have been too demanding for the capacity of our team.

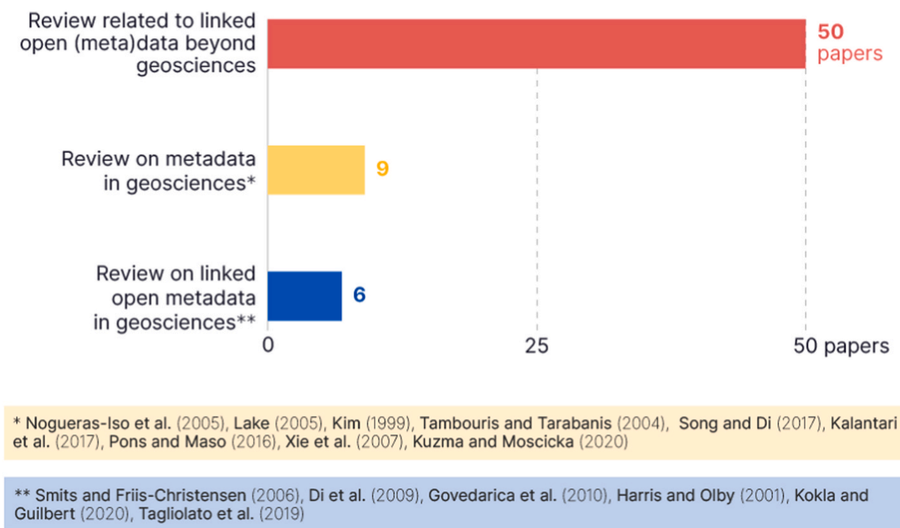

Fig. 2. Analysis of review papers relevant to the scope of open linked metadata at the Web of Science.

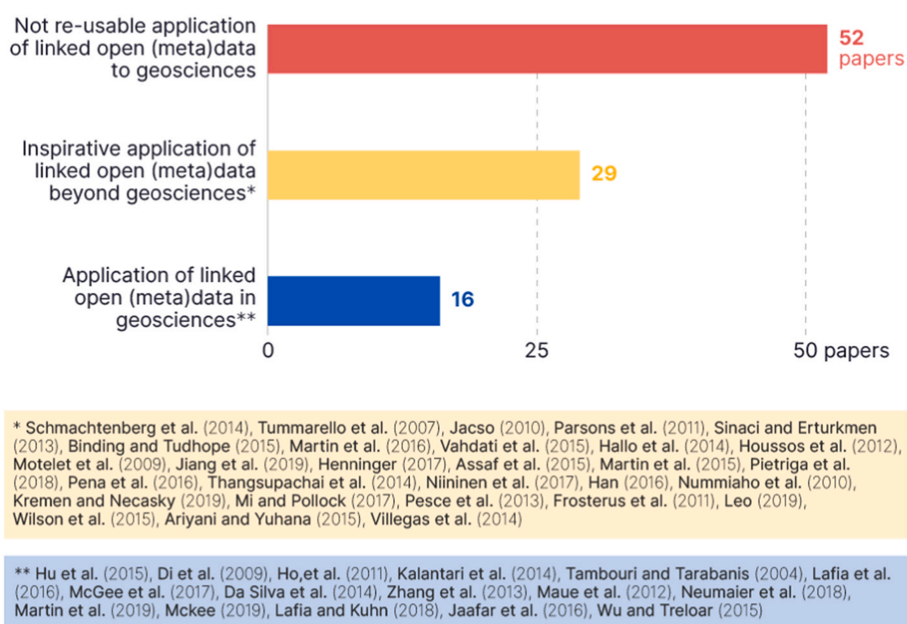

Fig. 3. Analysis of papers relevant to the scope of open linked metadata concerning their applicability in geosciences.

engines via publication through Schema.org (see Fig. 7).

4.1.3. Describing only relevant aspects

An easy-to-use structure, especially when compared with the complexity of standard geo metadata structures, is another feature and benefit of semantic (meta)data. Contemporary "traditional" metadata standards, like CSDGM (The Content Standard for Digital Geospatial Metadata; published by the Federal Geographic Data Committee – FGDC – of the United Nations; FGDC, 1998) or ISO 19115 Geographic information – Metadata (ISO 19115, 2003; ISO 19115, 2014), have very complex structures including hundreds of metadata elements in several hierarchical levels. In these, a metadata creator/administrator is also pushed to document mandatory metadata elements that are 1) not needed and 2) do not describe a resource appropriately according to his/her scope of applications. Such an impulse often results in non-equivalent descriptions of identical or similar concepts. Semantic approaches enable only those descriptions that are relevant according to the scope of the metadata application to be documented. This benefit may easily become a disadvantage as there is no minimal set of describing metadata elements, as in the case of the core metadata elements in ISO 19115 or a legally required set of elements in INSPIRE.

4.1.4. Links within/between information resources

Semantic approaches by default aim at linking open data. The added value lies, among other advantages, in clearly linking relevant pieces of information, such as in an example of visualisations of traffic measurements (see Fig. 8). When supporting semantics, we may visualise the most relevant resources as the primary ones and leave others to be shown as links, if more information is desired. Semantic approaches assist in estimations of the most relevant resources. For instance, a user is searching for a "river"; however, the metadata contains the term "stream". A catalogue service will provide datasets to a user for both terms, i.e. "river" as well as "stream", thanks to an associated thesaurus that has indicated the terms "river" and "stream" as synonyms.

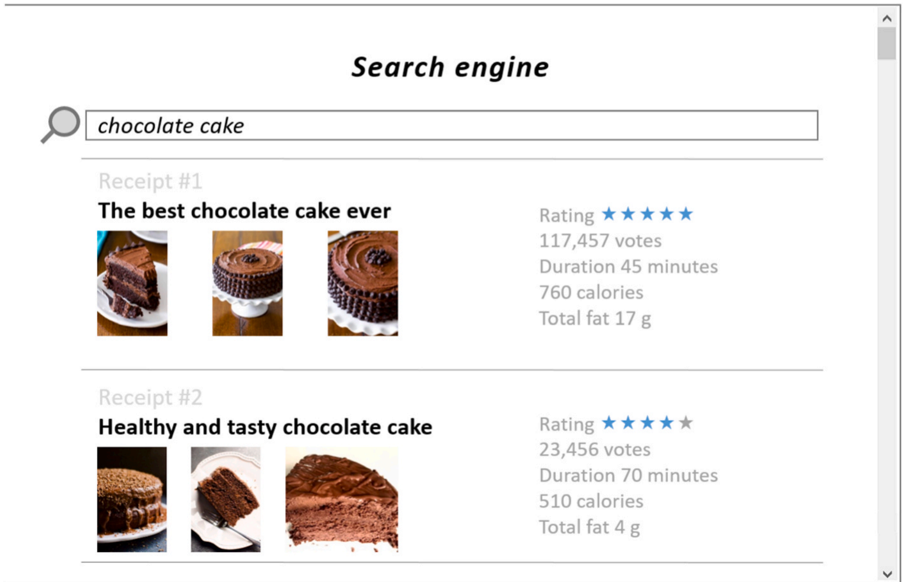

Fig. 4. Demonstration of Schema.org-based rich results as a user-friendly means of linked open (meta)data publication, as appearing in search engines (images adopted from https://sallysbakingaddiction.com/triple-chocolate-layer-cake/ and https://ifoodreal.com/healthy-chocolate-cake/, modified).

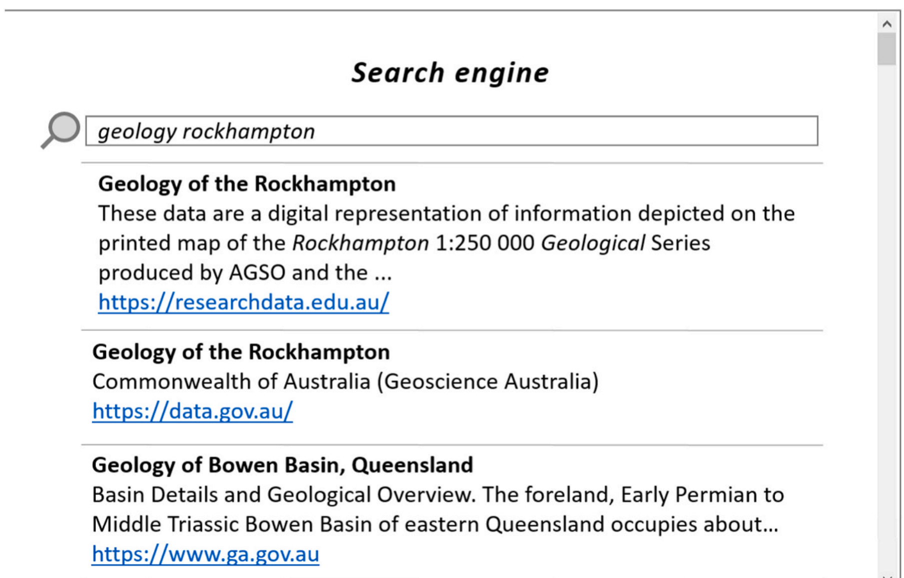

Fig. 5. Demonstration of a Schema.org-based "rich result" of a geo-domain dataset, as appearing in search engines (texts adopted from https://researchdata.edu.au/, https://data.gov.au/ and https://www.ga.gov.au, modified).

4.1.5. Revoke the (artificial) boundary between data and metadata

Level 0 and Level 3 have a common aspect: they both apply the same rules to data as well as metadata. Level 3 revokes an artificial boundary that is the cost of the paradigm used in Levels 1 and 2. Level 3, open linked (meta)data, uses the triple-based construction for data and metadata. Data and metadata follow a common life cycle. Metadata accompany data where desired and at several levels, such as a series of datasets, a single dataset, dataset visualisation, the e-shop offering the dataset, the layers of a dataset, the object type as part of a layer, and object instance, etc. Findability, as well as other processes, may be designed in a new, more complex way.

4.2. Challenges

The following challenges will be, similarly to benefits, described in more depth:

4.2.1. Updating mechanisms

Updates need to be set, the most common in geosciences through ETL (Extract, Transform, Load) mechanisms. The updating mechanisms are mostly automatic; however, a certain amount of manual input is usually needed, which influences the regular costs invested into metadata solutions.

4.2.2. Lack of concepts

The LOD (Linked Open Data) cloud (McCrae, 2020a) provides an excellent basis for (meta)data integration with relevant semantically-rich content. The LOD cloud is poorly balanced when speaking about different scientific domains as well as concepts in a scientific domain. The LOD cloud (Fig. 9) in November 2021 contained

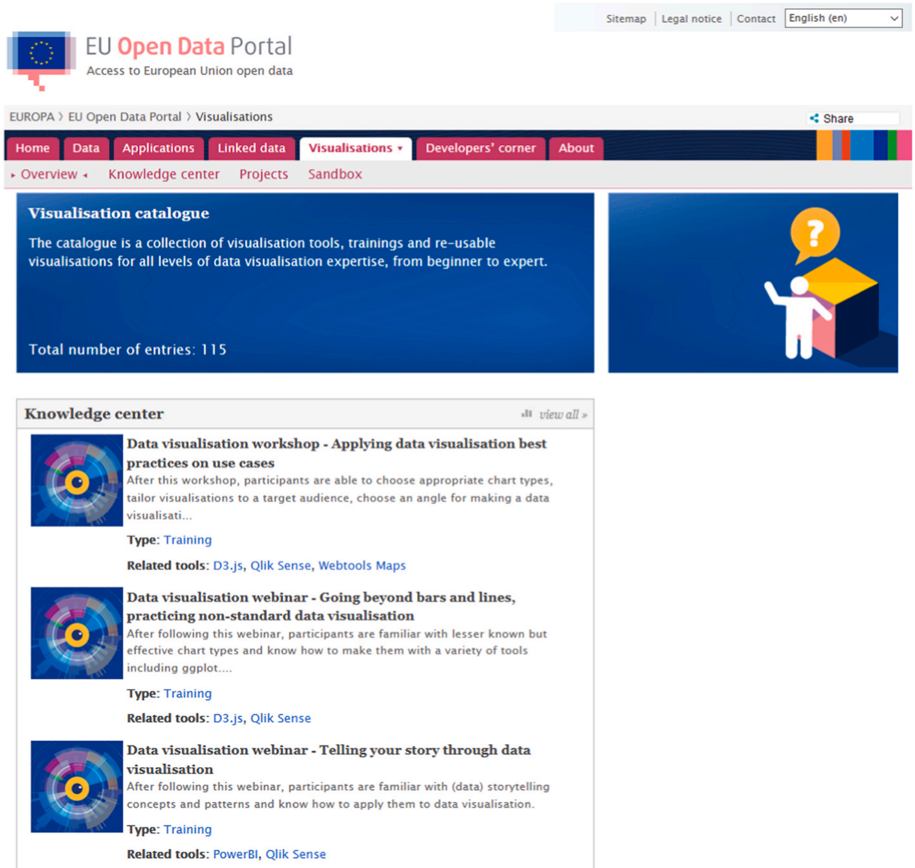

Fig. 6. Results brought to a user when searching for visualisations at the EU Open Data Portal (adopted from: https://data.europa.eu/euodp/en/visualisation-home).

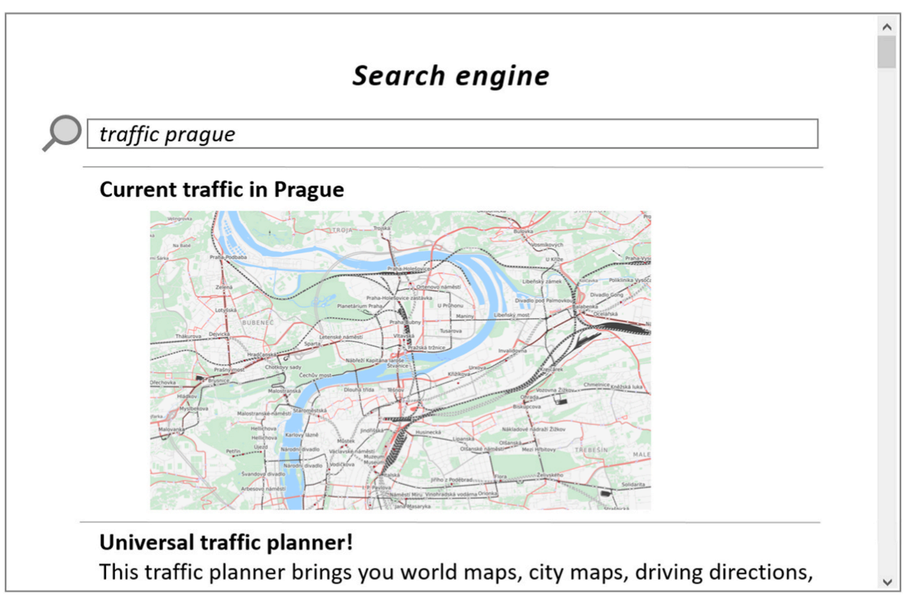

Fig. 7. Demonstration of Schema.org-based metadata: primarily, answers are presented to users directly instead of metadata behind the answers (although metadata were used) (map adopted from https://www.openstreetmap.org/, modified).

1301 datasets with 16,283 links.

Geosciences, identified in the LOD cloud as a separate domain, contributed 44 datasets; see also Fig. 10. It should be noted that geosciences are mentioned in the LOD cloud as 'geography'. For example, in reality, the 'geography' domain in the LOD cloud contains, among others, the Geological Survey of Austria (GBA) – Thesaurus. Therefore, the authors in this paper prefer the term 'geosciences' when speaking about the 'geography' domain in the LOD cloud. The major geoscience databases within the LOD cloud are the DBpedia (a semantic equivalent of Wikipedia), LinkedGeoData (including a semantic version of OpenStreetMap) and GeoNames (as the primary source for geocoding, with 25 million geographical names and 150 million web service requests per

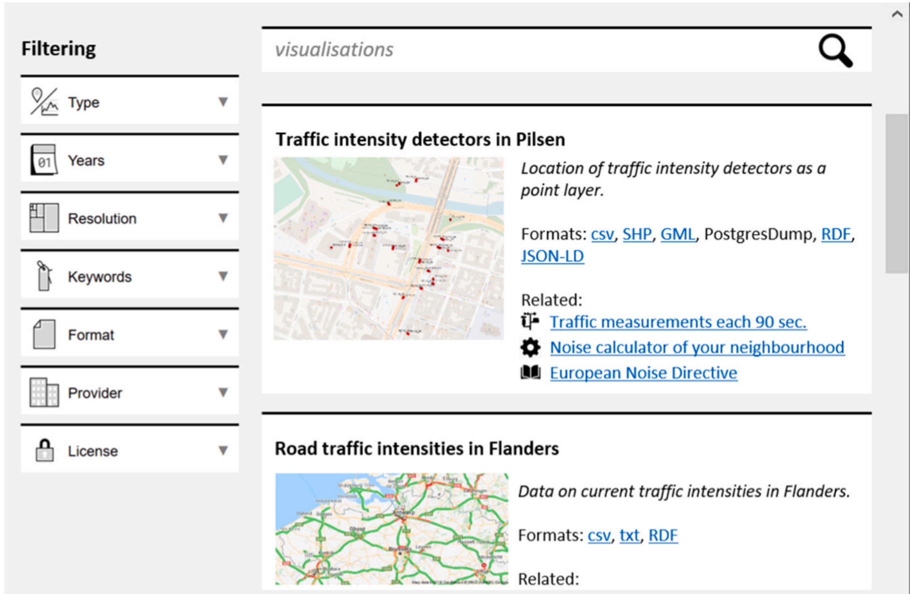

Fig. 8. Visualisation of open data in a structured and linked way: PoliVisu prototype on Traffic intensity detectors in Pilsen (development inspired by https://data.technologiestiftung-berlin.de/en and http://inspire-geoportal.ec.europa.eu).

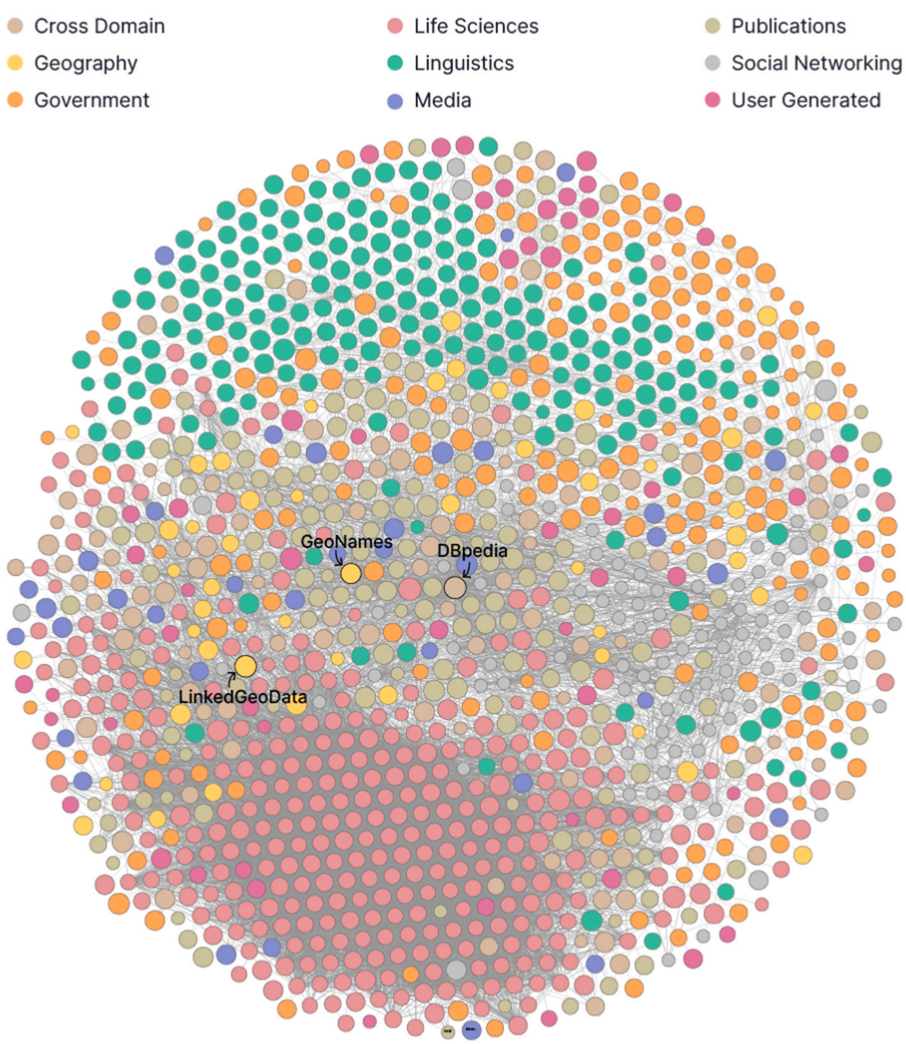

Fig. 9. All the databases within the LOD cloud and their linkages. Adopted from: https://lod-cloud.net.

Fig. 10. The selection of geosciences within the LOD cloud. Adopted from: https://lod-cloud.net.

day). As depicted in Fig. 10, these three databases are the main hubs, as all the remaining semantic databases in geosciences are linked to them.

Regarding their trajectory, geosciences are growing faster than the LOD cloud in general. From another point of view, the geosciences are the sixth most represented within the LOD cloud (after life sciences, government, linguistics, publications and social networking). Another perspective is philosophical: "What is and what is not related to geosciences?" For instance, OECD Linked Data (McCrae, 2020b) contains geolocated information; however, it is not classified within the LOD cloud as "geo". Therefore, the inclusion of only 44 linked geo datasets in the LOD cloud is misleading.

4.2.3. *Weak support in geosciences*

This situation seems to be a result of the rigidness of the geo community. With the exception of a pioneers, we can see a lack of best practice (Hu et al., 2015; Di et al., 2009; Ho et al., 2011; Narock and Fox, 2012; Kalantari et al., 2014; Wilson et al., 2014; Lafia et al., 2016; McGee et al., 2017; Da Silva et al., 2014; Zhang et al., 2013; Maue et al., 2012; Neumaier et al., 2018). The geo community, having reached a peak with respect to the use of non-semantic metadata approaches (i.e. Levels 1 and 2), still, as a whole, hesitates whether or not to adopt semantic metadata approaches. It seems that the identified challenges, together with a lack of application support, prevent the geosciences community from taking significant steps towards using semantic metadata. Findability mechanisms present the most visible obstacle. Geoportals have not evolved to comply with publication techniques (the so-called SEO techniques) used by all Web developers to ensure that a Web site is indexed by search engines. These not only include embedding metadata in Web pages - via HTML + RDFa and/or JSON-LD snippets, but also - and more importantly - the use of basic Web publication best practices (such as having a URL for each page to be indexed). Nevertheless, some catalogue platforms are moving in this direction, such as GeoNetwork (2020). The leading search engines have adopted semantic web principles while geo catalogues mostly remain according to how such catalogues were designed and built ten years ago, even though semantic approaches were successfully tested within a geo catalogue as early as in 2006 (HarmonISA, 2020). A shift to the development of semantic applications will also mean a shift on the part of the geo community to the employment of semantic-based use cases (and vice-versa).

4.2.4. *Invested efforts*

Efforts are needed when shifting from one level to another, no matter how great or small the shift. The most expensive are revolutions; i.e. shifts from Level 0 to Level 1 and from Level 2 to Level 3. The geo community has made the shifts to Levels 1 and 2 over the last two decades. However, the will to finance another revolution in terms of metadata management and publication seems to be low. Creation and maintenance costs on the one hand and low benefits, especially when the scope of application is not sufficiently specified enough, on the other hand are the major economic disadvantages. See section 6 for further discussion.

4.2.5. *Formalisation*

Two extreme situations could be identified:

- Limited loose semantics (typical for "linked data") makes implementation easier and more re-useable.

- Strong formalised semantics allow for powerful reasoning but make it hard to combine information from different sources.

The occurrence of both situations simultaneously is unlikely, at least not with formalisms like OWL (Web Ontology Language), which are based on classical logic (Augusto, 2019), or a more precise descriptive logic (Horrocks et al., 2003; Horrocks, 2005). As a result, a successful semantic web application is either powerful, but limited in scope and hard to integrate with other systems, or it is “dumb” but easy to integrate and use. It is therefore necessary to decide which type of success is desired.

4.2.6. Performance

As far as the authors are aware, the most complex semantically interlinked geodatabase has been developed within the FOODIE and DataBio projects (Reznik et al., 2017): more than 700 million triples were maintained in the Virtuoso triple store, with responses within seconds when using the Poznań Supercomputing and Networking Centre (in Poland). Nevertheless, the performance of a semantic-based findability service becomes worse in cases of:

- the presence of low-end hardware (on the part of the server),
- the inputting of complex queries (usually dozens or more conditions in one query) or
- the existence of very strong formalisation (especially the number of linkages and the capability of related stores to respond under stress conditions).

5. Open linked (meta)data: incremental versus an “all in one go” strategy

As noted earlier, open linked (meta)data is not an ‘all or nothing’ concept. Two major approaches are feasible: incremental implementation on the one hand and “all in one go” implementation on the other.

As stated in section 3, Level 3 open linked (meta)data approaches are considered as another revolution in terms of (meta)data management and publication. Open-linked (meta)data are often understood as a completely new paradigm that requires abandoning the existing approach. This section (5) tries to summarize the advantages of incremental implementation as a revolution divided into several smaller steps, on the one hand, and of a complete revolution, i.e., changing everything at once, on the other.

Both revolutionary approaches are valid; both have advantages and disadvantages. The steps identified in Fig. 11 need to be considered:

The strengths and weaknesses of each approach are summarised in Table 8.

6. Discussion

The lack of best practices in semantic approaches in geosciences sometimes seems to have a paralysing effect on their wider adoption (and vice versa). This discussion is intended as a guide to selecting the most appropriate approach for an organisation at a particular point in time, taking account of the type of content, the organisation’s objectives, the extent of openness and connectivity with other data, and the resources and skills available.

6.1. Suitability for a semantic approach

6.1.1. How suitable is the resource for this approach?

As also discussed in section 4, semantic approaches are applicable only for some kinds of resources. For instance, e-mails cannot be (at least so far) linked with/to any other related concept in existing semantic approaches. Video, audio and images can be linked only partially (Isaac and Haslhofer, 2013; Sikos, 2017). The basic geo resources like datasets,

Table 8
Strengths and weaknesses of incremental implementation versus “all in one go” approaches.

	Incremental implementation	All in one go
Strengths	· preservation of existing infrastructure · open linked (meta)data “only” in a publication data store · existing processes remain unchanged	· linkages between all the defined data resources · one (new) infrastructure for data and metadata · metadata and data life cycles are identical
Weaknesses	· linkages only between some data resources · metadata and data life cycles remain different · difficult updating mechanisms · training needed to set up and maintain the combined infrastructure	· costs for changing the infrastructure · training needed to set up and maintain the combined infrastructure · missing best practices · division of publicly unavailable (meta)data · linkage to historical data (before shifting to the linked open (meta) data approach)

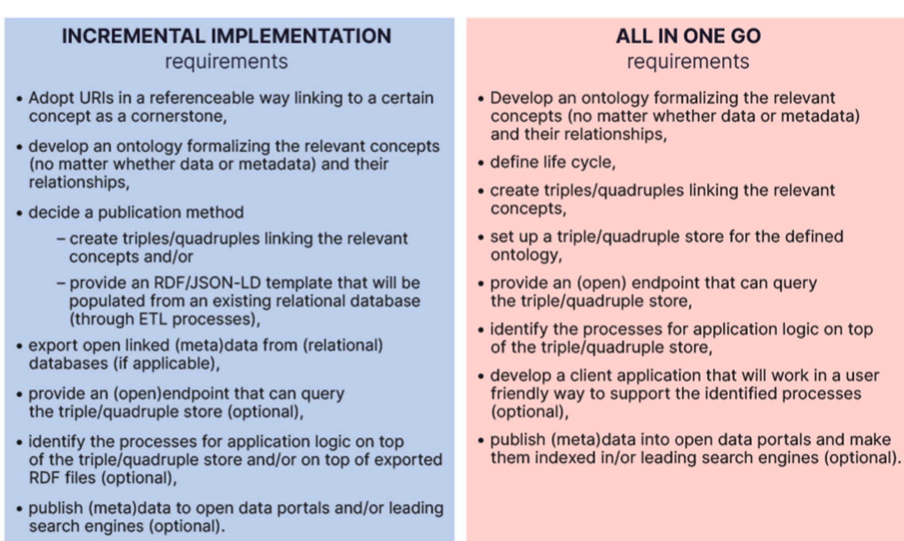

Fig. 11. Steps identified for ‘incremental implementation’ as well as for the “all in one go” approach.

web services, (map) compositions, and/or applications could follow the best practices defined within the OGC document on GeoDCAT-AP (Raes et al., 2019). Even such a portfolio of resources is not commonly described by semantic approaches in geosciences. Linking to relevant concepts is both the basic idea and the greatest benefit of semantic approaches. Higher impact would be achieved in cases when other kinds of resources are supported within implementations.

6.1.2. How much will the (meta)data be shared, now or in the future?

Semantic approaches are based on linking concepts, and one of the main objectives has been to make it easier both to share data in a meaningful way and to use data from other sources, sometimes by combining them with data already held. Therefore, the greater the expectation that data will be shared, the greater the realisable value of a semantic approach.

Many semantic approaches are associated with data openness goals. However, this does not mean that all the metadata need to be publicly available now, or in the future. It would be possible to design a semantic approach with a view to a future sharing of the organisation's data and with the possibility of using data from other sources more easily straight away. It would also be possible to publish some parts of the metadata but not other parts, with a clear and managed division between publicly available metadata and confidential metadata.

6.1.3. How much can existing semantic resources be used?

The more that semantic concepts and ontologies relevant to your data have already been developed, the easier the application of a semantic approach, and the greater the potential benefits of linking with other data. Over the last ten years, significant progress has been made in developing useful standard resources. Some key resources are.

- The LOD cloud: as presented in section 4.
- Thesauri (that may be a part of the LOD cloud)
  - domain-related thesauri like GEMET, AGROVOC, USGS Thesaurus, …
  - gazetteers, i.e. thesauri with geolocated concepts, like:
    - Getty Thesaurus of Geographical Names (TGN),
    - GeoNet Name Server (GNS),
    - The World Gazetteer,
    - GeoNames (the most used one),
  - registries that provide types and definitions typically for the coding of list values, such as the INSPIRE registries (INSPIRE, 2020) or the EPSG registry.
  - other resources relevant to the scope of a (metadata) application.

It is not essential that all necessary concepts have been developed. There are cases in which equivalent concepts do not (yet) exist or are not applicable and/or it is not desirable/could be misleading to link to existing concepts. In such cases, a (meta)data creator/administrator has an opportunity to define an ontology as well as to set up URIs. However, if no concepts are linked, many of the benefits of a semantic approach are lost.

6.2. What benefits would be gained from a semantic approach?

6.2.1. How important is precision and uniqueness in the metadata?

Traditional metadata management provides key-value pairs that already aim at adding types and definitions of data. However, these types and definitions of data may be:

- known only to some community;
- very limited, as the key-value pair concept does not allow more complex explanations.

The benefits of semantic approaches include explicit types and definitions of data for the provided (meta)data, including explicit position: for instance, when "Dublin" is indicated as a place of origin, semantic metadata is able to indicate whether it is "Dublin – the capital of Ireland", "Dublin – a city in Ohio, the United States of America", or some other.

6.2.2. How important is it that public search engines can find your content?

In summary, full-text based search, as provided by the leading search (Web) engines, brings vast amounts of users from various domains. Full-text search engines have several times higher numbers of users than the most visited geoportal on the planet. Such users are eager to discover also geodata. It is important to emphasise that not all users are willing to receive the answer in the form of a map. Nevertheless, the question is often a location-based one, such as "What noise from traffic do I encounter when walking from my home to my work?" (Kraak and Brown, 2000).

Two related levels can be identified:

- If you are capable of delivering a direct answer instead of metadata, do it.
- Only if you cannot directly provide an answer, present the user with metadata on a resource where (s)he can find the searched information.

For instance, a user is searching for "2 + 2"; (s)he immediately receives an answer. The same applies when searching for a location-based answer like "noise map Flanders". A user receives a preview of a collection of maps instead of textual metadata (Schema.org, 2020).

6.2.3. How important is it for your metadata to be integrated with sensor networks?

The integration of (meta)data within and beyond sensor networks brings new perspectives as well as added value (Di et al., 2009; Tagliolato et al., 2019). Complex queries in semantic approaches may become even more complex as they can also address the original (sensor) measurements. The joint Open Geospatial Consortium's and World Wide Web Consortium's recommendation called the Sensor Network Ontology (W3C, 2017a) addresses this step in detail. Such a document also discusses the differences between so-called live and static datasets.

6.3. What is the ability to deliver a semantic approach?

6.3.1. What skills are available?

Human resources skilled in metadata management are also needed for semantic approaches. However, it may not be sufficient to understand a metadata standard, define a metadata profile, develop a template for metadata encoding, publish a Web service, or define validation mechanisms. The following changes in comparison to "traditional" metadata approaches may be identified:

1. Analysis of user requirements (as "traditionally" this step is omitted within geosciences at the metadata level (Ho et al., 2011; Lafia et al., 2016; Maue et al., 2012; Neumaier et al., 2018),
2. Definition of high-level business process workflows including a decision on the re-use of the existing ontology versus developing a new one(s) (Kokla and Guilbert, 2020),
3. Determination of whether the required (meta)data will also be of a sensitive nature and require any special handling,
4. Determination of whether (meta)data will also be findable by leading search engines that follow Schema.org,
5. Decision about semantic-ready storage and its implementation (typically, a triple/quadruple store in comparison to "traditional" relational database),
6. Decision about exchange formats and their implementation,
7. The setting up of publication mechanisms (typically, an open API supporting SPARQL queries),

8. Definition and development of quality assurance and (or) quality control measures (especially to verify whether the underlying, i.e. interlinked, concepts are still reachable).

6.3.2. To what extent would you need to define your own concepts concerning semantic web definitions?

Semantic approaches could be used even when there are no relevant publicly available concepts for integration. However, in such a case, the efforts invested into semantic approaches will be considerably higher, as the definition of one's own concepts is not a trivial task (Corcho et al., 2003). This means that a registry/thesaurus/OWL ontology or any similar entity needs to be created within your organisation. Additional requirements arise both in terms of skills and financial resources.

6.3.3. How much funding is available?

In general, semantic approaches are initially costlier to implement in comparison to "traditional" metadata approaches; however, semantic approaches may give greater benefits, allowing other costs to be avoided over the data lifecycle. The highest cost is likely to be human resources with the necessary semantic design and implementation skills - either within the organisation (including the costs of developing those skills if necessary) or by contracting temporary specialist skills.

6.3.4. Suitability of existing IT infrastructure and services?

The implementation of semantic (meta)data may involve building interfaces and ensuring compatibility with the existing IT corporate infrastructure. If organisation policies permit, it may be possible - and, in the short term, desirable - to use cloud services to host the semantic element in order to avoid costs and delays in implementing semantic software in the corporate IT infrastructure.

6.4. Reaching a decision on the most suitable approach

> The decision on which approach to adopt is not just technical. Still, it should consider issues of business strategy, manageability, costs of scarce technical and non-technical resources and the benefits case for the organisation. It will need to look not just at the short-term and pragmatic issues but also at the longer-term benefits of a semantic approach. It will often be worth preparing a Business Case for the proposed approach so that there is a clear and sustainable basis for the formal decision.

The choice among the options will depend on the circumstances of individual organisations, but broadly four approaches are possible.

- Where there are clear and substantial benefits of a semantic approach, particularly in making the organisation's data more publicly available and findable, and there is a substantial body of existing semantic data to which to link, organisations should seriously consider committing the necessary financial and human resources to a fully semantic approach.
- Where the immediate benefits are less clear, but there is nevertheless a vast amount of data involved, the advantages of precise and unique metadata may nevertheless justify a fully semantic approach.
- Where the organisation is not yet sure about the strength of the benefits case, it might choose in the short-term to adopt the semi-semantic approach of using unique semantic identifiers in its (meta)data. This would allow some of the benefits of unique data to be realised and may also be a suitable way of developing the organisation's skills for fuller semantic approaches at a later stage. This could be particularly attractive where the organisation is in a position to develop ontologies that other stakeholders could use.
- Where the organisation's data is not suitable for a semantic approach or is unlikely to be shared with others, then traditional meta-data approaches may be adequate. However, to enable linking of data within the organisation itself, there may still be an advantage in using URIs to describe key entities uniquely and commonly across different departments of the organisation.

7. Conclusions

The presented paper provides comprehensive reference material for adopting the principles of linked open (meta)data. Three major advantages can be defined for adopting such semantic-based (meta)data approaches:

- clear types and definitions of data used for decision making,
- flexibility of metadata descriptions,
- sharing relevant information with the broad audience of Schema.org-based search engines users.

The first point aims at better decisions, as data in semantic approaches are linked to explicit evidence on the procedures used, units of measures, and the quality of the data (measurements) etc. Benefits of the first point can appear with even meagre investments, as linking data can be handled through URIs in existing (geo) solutions. A (dereferenceable) URI provides clearer information on types and definitions as well as further relevant information in comparison to the "traditional" maintenance of (free text) character strings in metadata.

The second point aims at the flexibility of metadata descriptions. Metadata creators/administrators are no longer pushed to provide mandatory metadata elements according to some complex (international) standard. Instead, only relevant (meta)data are linked together according to the needs of an application. This also means that metadata may be easily provided also in cases not supported by an (international) metadata standard. For instance, ISO 19115 does not support a description of the data structure in the metadata. In contrast, semantic approaches enable metadata to be linked to feature types, attribute types, and/or any similar structure of the described dataset. Semantic approaches enable the removal of the artificial boundary between data and metadata that has been present within geosciences for more than three decades.

Sharing relevant information with the broad audience of Schema.org-based search engines users allows, if desired, to step outside geo and open data bubbles. It also addresses a common challenge that has, so far, been only partially addressed within geo/open data communities. That is, discovering a proper tool for finding the data seems to be in several cases even more complicated than discovering the data in that a tool. Such a step requires knowledge as well as investments in time and effort. Metadata can be advertised in an attractive user-friendly way in leading search engines.

Four levels of metadata management and publication were identified and described:

- 'Level 0: default unstructured metadata', which is generated automatically/implicitly within many typical IT systems.
- 'Level 1: schema-based metadata with literal values', which is a 'revolution' where the semantics of these values is ambiguous (clear usually only within a domain and/or interest group).
- 'Level 2: schema-based metadata with unique identifiers', which is an evolution of Level 1 with an emphasis on the use of unique identifiers that are added to literal values whenever applicable, these providing explicit identification of reliable data resources.
- 'Level 3: linked open (meta)data', which provides explicit linkage (and types plus definitions) between reliable data resources. Moreover, Level 3 facilitates the indexing of (meta)data by leading search engines.

The decision on which level to adopt is not just technical. Still, it will need to consider issues of business strategy, manageability, opportunity costs of scarce technical and non-technical resources, and the benefits case for the organisation. It will need to look not just at short-term and pragmatic issues but also at the longer-term benefits of a semantic approach. It will often be worth preparing a Business Case for the proposed approach so that there is a clear and sustainable basis for the formal decision. In any case, two approaches are feasible: incremental implementations as a revolution divided into several smaller steps on the one hand, and complete revolution, i.e. 'all in one go' implementation on the other hand.

The benefits of semantic approaches increase with the number of involved stakeholders. To date, the Linked Open Data (LOD) cloud contains 1301 datasets with 16,283 links. Geosciences are the sixth most-represented component within the LOD cloud (after life sciences, government, linguistics, publications, and social networking) with 44 datasets. However, the real number of linked geo datasets seems to be higher. For instance, OECD Linked Data contains geolocated information, which is classified as "government" and not assigned to geosciences.

Future work will focus on alignment of the developed guidelines with existing best practices for geo (meta)data. Such work remains a part of the activity of the Metadata Working Group of the Open Geospatial Consortium (OGC). Revisions and amendments to the OGC GeoDCAT-AP with respect to the outcomes of this paper are one of the anticipated results.

Funding sources

This research was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 769608 titled Policy Development based on Advanced Geospatial Data Analytics and Visualisation (PoliVisu) and by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 818346 called "Sino-EU Soil Observatory for intelligent Land Use Management" (SIEUSOIL).

Author contributions

Tomáš Řezník conducted the study, wrote the text, prepared figures&tables and revised the text. Lieven Raes conducted the study, wrote the text and revised the text. Andrew Stott conducted the study, wrote the text and revised the text. Bart De Lathouwer wrote and revised the text. Andrea Perego wrote and revised the text. Karel Charvát revised the text. Štěpán Kafka conducted the study.

Computer code availability

No specific software/script is related to the presented work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2022.105194.

References

Clarivate Analytics, 2020. Web of Science. https://apps.webofknowledge.com.
ANZLIC, 2020. Metadata – ANZLIC Spatial Resource Discovery and Access Toolkit 2009. https://www.anzlic.gov.au/resourcesarchived-foundation-spatial-data-framework/metadata.
Augusto, L.M., 2019. Formal Logic: Classical Problems and Proofs. College Publications, London.
Baiquan, X., Shiqiang, Y., Qianju, W., Jian, L., Xiaoping, W., Keyong, D., 2013. Geospatial data infrastructure: the development of metadata for geo-information in China. In: 35th International Symposium on Remote Sensing of Environment (ISRSE35), Beijing, China, vol. 17. https://doi.org/10.1088/1755-1315/17/1/012259. IOP Conference Series: Earth and Environmental Science.
Brickley, D., Burgess, M., Noy, N., 2019. Google Dataset Search: building a search engine for datasets in an open Web ecosystem. In: The World Wide Web Conference. ACM, New York, NY, USA, pp. 1365–1375. URL: https://dl.acm.org/doi/10.1145/3308558.3313685.
Brodeur, J., Coetzee, S., Danko, D., Garcia, S., Hjelmager, J., 2019. Geographic information metadata—an outlook from the international standardization perspective. ISPRS Int. J. Geo-Inf. 8 (6), 280. https://doi.org/10.3390/ijgi8060280.
Clarke, M., 2012. In: Campbell, R., Pentz, E., Borthwick, I. (Eds.), The Digital Revolution. Academic and Professional Publishing. Chandos Publishing, pp. 79–98. https://doi.org/10.1016/B978-1-84334-669-2.50004-4.
Codd, E.F., 1970. A relational model of data for large shared data banks. Commun. ACM 13 (6), 377–387. https://doi.org/10.1145/362384.362685.
Corcho, O., Fernández-López, M., Gómez-Pérez, A., 2003. Methodologies, tools and languages for building ontologies. Where is their meeting point? Data Knowl. Eng. 46 (1), 41–64. https://doi.org/10.1016/S0169-023X(02)00195-7.
Da Silva, J.R., Castro, J.A., Ribeiro, C., Honrado, J., Lomba, A., Goncalves, J., 2014. Beyond INSPIRE: an ontology for biodiversity metadata records. In: Meersman, R., et al. (Eds.), On the Move to Meaningful Internet Systems: OTM 2014 Workshops. OTM 2014, Lecture Notes in Computer Science, vol. 8842. Springer, Berlin, Heidelberg, pp. 597–607. https://doi.org/10.1007/978-3-662-45550-0_61.
Danko, D.M., 2012. Geospatial metadata. In: Kresse, W., Danko, D.M. (Eds.), Springer Handbook of Geographic Information (S. 191–244). Springer. https://doi.org/10.1007/978-3-540-72680-7_12.
Di, L., Moe, K.L., Yu, G.N., 2009. Metadata requirements analysis for the emerging Sensor Web. Int. J. Digit. Earth 2, 3–17. https://doi.org/10.1080/17538940902866195.
Dublin Core, 2020. Dublin Core Metadata Initiative. https://dublincore.org/specifications/dublin-core/.
EPSG, 2019. https://epsg.io/4326.gml, 4326.
European Commission, 2007. 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). http://data.europa.eu/eli/dir/2007/2/oj.
European Commission, 2008. PDF, EN. Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:326:0012:0030.
FGDC, 1998. Content Standard for Digital Geospatial Metadata (CSDGM). https://www.fgdc.gov/standards/projects/metadata/base-metadata/v2_0698.pdf.
Foaf, 2014. FOAF Vocabulary Specification 0, vol. 99. http://xmlns.com/foaf/spec/.
Furner, J., 2020. Definitions of "metadata": a brief Survey of international standards. J. Assoc. Inf. Sci. Technol. 71 (6), E33–E42. https://doi.org/10.1002/asi.24295.
GEMET, 2020a. Measuring. https://www.eionet.europa.eu/gemet/en/concept/5119.
GEMET, 2020b. Noise Analysis. https://www.eionet.europa.eu/gemet/en/concept/5629.
GeoDCAT-Ap, 2015. A Geospatial Extension for the DCAT Application Profile for Data Portals in Europe. https://joinup.ec.europa.eu/solution/geodcat-application-profile-data-portals-europe/distribution/geodcat-ap-101-pdf.
GeoNetwork, 2020. GeoNetwork. https://geonetwork-opensource.org.
Giles, J.R.A., 2011. Geoscience Metadata—No Pain, No Gain. https://doi.org/10.1130/2011.2482(03.
GitHub, 2020. CKAN. https://github.com/ckan/ckanext-dcat.
Govedarica, M., Boskovic, D., Petrovacki, D., Ninkov, T., Ristic, A., 2010. Metadata catalogues in spatial information systems. Geod. List. 4, 313–334. URL: https://hrcak.srce.hr/64687.
Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of Things (IoT): a vision, architectural elements, and future directions. Future Generat. Comput. Syst. 29 (7), 1645–1660. https://doi.org/10.1016/j.future.2013.01.010.
Habermann, T., 2018. Metadata life cycles, use cases and hierarchies. Geosciences 8 (5), 179. https://doi.org/10.3390/geosciences8050179.
HarmonISA, 2020. Harmonisation of land-use data. http://harmonisa.uni-klu.ac.at.
Harris, R., Olby, N., 2001. Earth observation data archiving in the USA and Europe. Space Pol. 17, 35–48. https://doi.org/10.1016/S0265-9646(00)00052-7.
Ho, Q.V., Lundblad, P., Astrom, T., Jern, M., 2011. A web-enabled visualization toolkit for geovisual analytics. In: Wong, P.C., et al. (Eds.), Proceedings of SPIE, the International Society for Optical Engineering: SPIE: Electronic Imaging Science and Technology. Visualization and Data Analysis, San Francisco USA. https://doi.org/10.1117/12.872250, 78680R.
Horrocks, I., 2005. OWL: a description logic based Ontology Language. In: van Beek, P. (Ed.), Principles and Practice of Constraint Programming - CP 2005. CP 2005, Lecture Notes in Computer Science, vol. 3709. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11564751_2.
Horrocks, I., Patel-Schneider, P.F., van Harmelen, F., 2003. From SHIQ and RDF to OWL: the making of a web Ontology Language. J. Web Semant. 1, 7–26. https://doi.org/10.1016/j.websem.2003.07.001.
Hu, Y.J., Janowicz, K., Prasad, S., Gao, S., 2015. Metadata topic harmonization and semantic search for linked-data-driven geoportals: a case study using ArcGIS online. Trans. GIS 19, 398–416. https://doi.org/10.1111/tgis.12151.
INSPIRE, 2020. Registry. https://inspire.ec.europa.eu/registry.
Isaac, A., Haslhofer, B., 2013. European linked open data - data. europeana.eu. Semantic Web 4 (3), 291–297. https://doi.org/10.3233/SW-120092.
ISO 19115, 2003. ISO 19115:2003 geographic information — metadata. https://www.iso.org/standard/26020.html.
ISO 19115, 2014. ISO 19115-1:2014 geographic information — metadata — Part 1: fundamentals. https://www.iso.org/standard/53798.html.

ISO 8601-01, 2019. ISO 8601-01:2019 date and time — representations for information interchange – Part 1: basic rules. https://www.iso.org/iso-8601-date-and-time-format.html.
ISO/TS 19139, 2007. ISO/TS 19139:2007 Geographic information — metadata — XML schema implementation. https://www.iso.org/standard/32557.html.
ISO/TS 19139-1, 2019. ISO/TS 19139-1:2019 Geographic information — XML schema implementation — Part 1: encoding rules. https://www.iso.org/standard/67253.html.
Jensen, J., Saalfeld, A., Broome, F., Cowen, D., Price, K., Ramsey, D., Lapine, L., 2000. Spatial data acquisition and integration. http://dusk.geo.orst.edu/ucgis/web/research_white/data.html.
JoinUp, 2020. GeoDCAT-AP 1.0.1 PDF. https://joinup.ec.europa.eu/solution/geodcat-application-profile-data-portals-europe.
Kalantari, M., Rajabifard, A., Olfat, H., Williamson, I., 2014. Geospatial metadata 2.0-an approach for volunteered geographic information. Comput. Environ. Urban Syst. 48, 35–48. https://doi.org/10.1016/j.compenvurbsys.2014.06.005.
Kokla, M., Guilbert, E., 2020. A review of geospatial semantic information modeling and elicitation approaches. ISPRS Int. J. Geo-Inf. 9, 146. https://doi.org/10.3390/ijgi9030146.
Kraak, J.-M., Brown, A., 2000. Web Cartography. CRC Press, p. 213.
Lafia, S., Jablonski, J., Kuhn, W., Cooley, S., Medrano, F.A., 2016. Spatial discovery and the research library. Trans. GIS 20, 399–412. https://doi.org/10.1111/tgis.12235.
Maue, P., Michels, H., Roth, M., 2012. Injecting semantic annotations into (geospatial) web service descriptions. Semantic Web 3, 385–395. https://doi.org/10.3233/SW-2012-0061.
McCrae, J.P., 2020a. The Linked Open Data Cloud. https://lod-cloud.net.
McCrae, J.P., 2020b. Organisation for Economic Co-operation and Development (OECD) Linked Data". https://lod-cloud.net/dataset/oecd-linked-data.
McGee, M., Durante, K., Weimer, K.H., 2017. Applications and projections toward a linked data model for describing cartographic resources. J. Map Geogr. Libr. 13, 133–144. https://doi.org/10.1080/15420353.2017.1308291.
Melton, J., Buxton, S., 2006. Metadata – an overview. In: Melton, J., Buxton, S. (Eds.), Queyring XML. Morgan Kaufmann, Burlington, USA, pp. 67–84. https://doi.org/10.1016/B978-155860711-8/50005-8.
Moellering, H., Aalders, H.J.G.L., Crane, A., 2005. World spatial metadata standards. Elsevier, amsterdam, Netherlands, 710 pp. https://doi.org/10.1016/B978-0-08-043949-5.X5000-1.
Narock, T., Fox, P., 2012. From science to e-Science to Semantic e-Science: a Heliophysics case study. Comput. Geosci. 46, 248–254. https://doi.org/10.1016/j.cageo.2011.11.018.
Neumaier, S., Savenkov, V., Polleres, A., 2018. Geo-semantic labelling of open data. Procedia Comput. Sci. 137, 9–20. https://doi.org/10.1016/j.procs.2018.09.002.
Nogueras-Iso, J., Zarazaga-Soria, F.J., Bejar, R., Alvarez, P.J., Muro-Medrano, P.R., 2005. OGC Catalog services: a key element for the development of Spatial Data Infrastructures. Comput. Geosci. 31, 199–209. https://doi.org/10.1016/j.cageo.2004.05.015.
OGC, 2006. OpenGIS® catalogue services — ebRIM (ISO/TS 15000-3) profile of CSW. http://portal.opengeospatial.org/files/?artifact_id=12604&version=1&format=pdf.
OGC, 2016. Catalogue Services 3.0 - General Model. http://www.opengis.net/doc/IS/cat/3.0.
Palma, R., Reznik, T., Esbri, M., Charvat, K., Mazurek, C., 2016. An INSPIRE-based vocabulary for the publication of agricultural linked data. In: Tamma, V., Dragoni, M., Gonçalves, R., Ławrynowicz, A. (Eds.), Ontology Engineering. OWLED 2015, Lecture Notes in Computer Science, vol. 9557. Springer, Cham, pp. 124–133. https://doi.org/10.1007/978-3-319-33245-1_13.
PoliVisu, 2020. PoliVisu EU Project - Policy & Data Results Hub. https://www.polivisu.eu/.
Raes, L., VanDenbroucke, D., Reznik, T., 2019. GeoDCAT-AP. www.opengis.net/doc/dp/GeoDCAT-AP.
Reznik, T., Chudy, R., Micietova, E., 2016. Normalized evaluation of the performance, capacity and availability of catalogue services: a pilot study based on INfrastruture for SPatial InfoRmation in Europe (INSPIRE). Int. J. Digit. Earth 9, 325–341. https://doi.org/10.1080/17538947.2015.1019581.
Reznik, T., Charvat, K., Palma, R., Kozuch, D., Cerba, O., Jedlicka, K., Berzins, R., Bergheim, R., 2017. Integration of open Land use, smart point of interest and open transport map using RDF. In: G20 Summit in Berlin, Meeting of Chief Agricultural Scientists. https://doi.org/10.13140/RG.2.2.28965.40165.
Schemaorg, 2020. Schema.org. https://schema.org.
Schematron, 2020. Schematron. http://schematron.com/.
Shien-Chiang, Y., Kun-Yung, L., Ruey-Shun, C., 2003. Metadata management system: design and implementation. Electron. Libr. 21, 154–164. https://doi.org/10.1108/02640470310470525.
Sikos, L.F., 2017. RDF-powered semantic video annotation tools with concept mapping to Linked Data for next-generation video indexing: a comprehensive review. Multimed. Tool. Appl. 76 (12), 14437–14460. https://doi.org/10.1007/s11042-016-3705-7.
Smits, P.C., Friis-Christensen, A., 2006. Resource discovery in a European spatial data infrastructure. IEEE Trans. Knowl. Data Eng. 19, 85–95. https://doi.org/10.1109/TKDE.2007.250587.
Tagliolato, P., Fugazza, C., Oggioni, A., Carrara, P., 2019. Semantic profiles for easing SensorML description: review and proposal. ISPRS Int. J. Geo-Inf. 8, 340. https://doi.org/10.3390/ijgi8080340.
W3C, 2010. XML Linking Language (XLink) Version 1.1. https://www.w3.org/TR/xlink11/.
W3C, 2014a. RDF. https://www.w3.org/RDF/.
W3C, 2014b. RDF 1.1 Turtle. https://www.w3.org/TR/turtle/.
W3C, 2014c. RDF 1.1 XML Syntax. https://www.w3.org/TR/rdf-syntax-grammar/.
W3C, 2015a. Semantic Web. https://www.w3.org/standards/semanticweb/.
W3C, 2015b. HTML+RDFa 1.1 - second edition. Support for RDFa in HTML4 and HTML5. https://www.w3.org/TR/html-rdfa/.
W3C, 2017a. Semantic Sensor Network Ontology. https://www.w3.org/TR/vocab-ssn/.
W3C, 2017b. Spatial Data on the Web Best Practices. https://www.w3.org/TR/sdw-bp/.
W3C, 2020a. JSON-LD 1.1. https://www.w3.org/TR/json-ld11/.
W3C, 2020b. Data Catalog Vocabulary (DCAT) - Version 2. https://www.w3.org/TR/vocab-dcat-2/.
Weibel, S., Godby, J., Miller, E., 1995. OCLC/NCSAmetadata workshop report. URL: www.oclc.org/oclc/research/conferences/Metadata/dublin_core_report.htm.
Wilkinson, M., Dumontier, M., Aalbersberg, I., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.
Wilson, A., Cox, M., Elsborg, D., Lindholm, D., Traver, T., 2014. A semantically enabled metadata repository for scientific data. Earth Science Informatics 8, 649–661. https://doi.org/10.1007/s12145-014-0175-1.
Zhang, M.D., Yuan, J., Gong, J.Y., Yue, P., 2013. An interlinking approach for linked geospatial data. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. XL-7/W2, 283–287. https://doi.org/10.5194/isprsarchives-XL-7-W2-283-2013.