GloSIS: The Global Soil Information System Web Ontology

Raul Palma [a,*], Bogusz Janiak [a], Luís M. de Sousa [b], Kathi Schleidt [c], Tomáš Řezník [d], Fenny van Egmond [b], Johan Leenaars [b], Dimitrios Moshou [e], Abdul Mouazen [f], Peter Wilson [g], David Medyckyj-Scott [h], Alistair Ritchie [h], Yusuf Yigini [i] and Ronald Vargas [i]

[a] *Poznań Supercomputing and Networking Center - PSNC, Poznań, Poland*
E-mails: rpalma@man.poznan.pl, bjaniak@man.poznan.pl
[b] *ISRIC - World Soil Information, Wageningen, The Netherlands*
E-mails: luis.desousa@isric.org, fenny.vanegmond@isric.org, johan.leenaars@isric.org
[c] *DataCove, Vienna, Austria*
E-mail: kathi@datacove.eu
[d] *Masaryk University, Faculty of Science, Department of Geography, Kotlářská 2, 611 37, Brno, Czech Republic*
E-mail: tomas.reznik@sci.muni.cz
[e] *Aristotle University of Thessaloniki, Thessaloniki, Greece*
E-mail: dmoshou@agro.auth.gr
[f] *Department of Environment, Ghent University, Gent, Belgium*
E-mail: Abdul.Mouazen@UGent.be
[g] *CSIRO - The Commonwealth Scientific and Industrial Research Organisation, Canberra, Australia*
E-mail: peter.wilson@csiro.au
[h] *Manaaki Whenua - Landcare Research, Lincoln, New Zealand*
E-mails: medyckyj-scottd@landcareresearch.co.nz, ritchiea@landcareresearch.co.nz
[i] *FAO - Food and Agriculture Organisation of the United Nations, Rome, Italy*
E-mails: yusuf.yigini@fao.org, ronald.vargas@fao.org

Abstract.

Established in 2012 by members of the Food and Agriculture Organisation (FAO), the Global Soil Partnership (GSP) is a global network of stakeholders promoting sound land and soil management practices towards a sustainable world food system. However, soil survey largely remains a local or regional activity, bound to heterogeneous methods and conventions. Recognising the relevance of global and trans-national policies towards sustainable land management practices, the GSP elected data harmonisation and exchange as one of its key lines of action. Building upon international standards and previous work towards a global soil data ontology, an improved domain model was eventually developed within the GSP [54], the basis for a Global Soil Information System (GloSIS). This work also identified the semantic web as a possible avenue to operationalise the domain model.

This article presents the GloSIS web ontology, an implementation of the GloSIS domain model with the Web Ontology Language (OWL). Thoroughly employing a host of semantic web standards (SOSA, SKOS, GeoSPARQL, QUDT), GloSIS lays out not only a soil data semantic ontology but also an extensive set of ready-to-use code-lists for soil description and physio-chemical analysis. Various examples are provided on the provision and use of GloSIS-compliant linked data, showcasing the contribution of this ontology to the discovery, exploration, integration and access of soil data.

Keywords: Soil, Sustainability, Semantic model, SOSA/SSN, SKOS, GloSIS

1. Introduction and motivation

1.1. The importance of soils and related risks

Human population more than tripled since the end of World War II [35]. This growth has been accompanied by the densification of urban areas, with the share of population living in cities doubling, having surpassed 50% in 2010 [15]. Supporting this population has required unprecedented growth in food production. Nevertheless, dramatic increases in food output per unit area have meant an expansion of global agricultural area by just 30% in the past seven decades [39].

Albeit a success, this transformation and expansion of food production systems has placed unprecedented stress on soils. These are non-renewable natural resources, that if mismanaged can rapidly degrade down to a non-productive state. Soils around the globe are presently impacted by the over-use of fertilisers, chemical contamination, loss of organic matter, salanisation, acidification and outright erosion [30]. These trends pose serious risks not only to food supply, but also to ecosystems, as they provide a myriad of services at the local, landscape and global levels [3, 17, 50].

Addressing these risks often requires an holistic approach, with policies and practices envisioned at a global scale. For instance, the reduction of soil erosion through land rehabilitation and development [8, 53], the protection of food production [16, 46, 47], or the preservation of biodiversity [4, 21, 51] and human livelihood [9]. However, the data necessary to develop such policies is collected, analysed and represented at many different scales, as these remain primarily region or country specific activities. The data harmonisation necessary towards the sustainable use of soils at the global scale remains a challenge [1].

1.2. GSP and its goals

The Global Soil Partnership (GSP) was established in 2012 by members of the Food and Agriculture Organisation of the United Nations (FAO) as a network of stakeholders in the soil domain. Its broad goals are to raise awareness to the importance of soils in attaining a sustainable agriculture and to promote good practices in land and soil management. The GSP involved the majority of the world's national soil information institutions, gathered around the International Network of Soil Information Institutions (INSII).

*Corresponding author. E-mail: rpalma@man.poznan.pl.

The GSP defined five pillars of action structuring its activities:

- Pillar 1 – **Soil management** – promote the sustainable management of soil resources for soil protection, conservation and sustainable productivity.
- Pillar 2 – **Awareness raising** – encourage investment, technical cooperation, policy, education, awareness and extension in soil.
- Pillar 3 – **Research** – promote targeted soil research and development focusing on identified gaps, priorities and synergies with related productive, environmental and social development actions.
- Pillar 4 – **Information and data** – enhance the quantity and quality of soil data and information: data collection (generation), analysis, validation, reporting, monitoring and integration with other disciplines.
- Pillar 5 – **Harmonisation** – targeting methods, measurements and indicator for the sustainable management and protection of soil resources.

The Action Plan for Pillar 5 [1] acknowledges various difficulties with the harmonisation of soil data. In particular, soil data is most often collected and curated by national or regional institutions, focused on their local context, largely abstract from international or global concerns. This lack of heterogeneity severely limits the availability, sharing and use of soil data. The transfer of data, methods and practices, between regions, or from global to local initiatives, is thus prone to hurdles and errors, putting at risk sustainable soil management goals.

Among the key priorities towards harmonisation identified in the Action Plan for Pillar 5 is the development of a soil information exchange infrastructure. This is broadly defined as "[...] a conceptual soil feature information model provid[ing] the framework for harmonisation such that the efficient exchange and collation of globally consistent data and information can occur". Data exchange is put forth both as an essential component of soil data harmonisation and also as a vector to that end, facilitating data integration, analysis and interpretation.

In the Action Plan for Pillar 4 [2] the GSP lays out the guidelines for the development of an authoritative global soil information. This system is envisioned as fulfilling three main functions:

- answer critical questions at the global scale;

- provide the global context for more local decisions;
- supply fundamental soil data to understand Earth-system processes to enable management of the major natural resource issues facing the world.

Draft implementation guidelines are laid out in Action Plan for Pillar 4, pointing to a federated system in which soil institutions provide access to their data through web services, all compliant to a common data exchange specification. The latter is leveraged on the outcome of Pillar 5, concerning the exchange of soil profile observations and descriptions, laboratory and field analytical data, plus derived products such as digital soil maps. Soil data exchange is thus set at the core of GSP, an unavoidable stepping stone to achieve its goals. As set out in the Action Plan for Pillar 5: "Pillar 5 is a basic foundation of Pillar 4, and an enabling mechanism for all GSP pillars providing and using global soil information."

1.3. International Consultancy towards a global soil information exchange

In 2019 the GSP launched a call for an international consultancy to assess the state-of-the-art in soil information exchanges and propose a path towards its operationalisation in line with the goals of Pillar 5. The results of this consultancy are gathered in [54].

During this work a detailed set of requirements was inventoried, sourced from meetings and interviews with various GSP stakeholders. Among these requirements is the will to re-use existing ontologies and exchange mechanisms as much as possible and assess the suitability of each regarding implementation (with Pillar 4 in view). In light of these requirements the international consultancy assessed the following soil ontologies and data models:

- **ANZSoilML** [44] – an ontology developed by Australia and New Zealand directed at a SOAP/XML web services implementation;
- **INSPIRE** [45] – INSPIRE Application Schema for Soil Spatial Data Themer;
- **ISO 28258** [24] – the abstract ontology layed out in the soil quality standard approved by the International Standards Organisation (ISO);
- **OGC Soil IE** [36] – an international soil data exchange experiment conducted by the Open Geospatial Consortium (OGC);
- **WoSIS** [6] – the data model of the World Soil Information Service;
- **SOTER** [37] – data model developed within the Soil and Terrain (SOTER) database programme.

The consultancy identified relevant similarities between ANZSoilML, INSPIRE, ISO 28258 and the OGC Soil IE. All of these models re-use the Observations and Measurements (O&M) domain model [13], an umbrella specification for the observation of natural phenomena. Observations and Measurements was adopted in parallel by the ISO as a standard (ISO 19156) [23]. The relational data models of WoSIS and SOTER do not share the same abstraction, but were nonetheless considered for the sizeable data they collect in an harmonised manner, providing insight on aspects such as the code-list necessary for implementation of a soil data exchange.

The ISO 28258 model was selected as the most suitable starting point to operationalise the sought for exchange mechanism. The model was augmented with container classes encapsulating the Guidelines for Soil Description issued by the FAO [25], an abstraction of the code-lists necessary for the exchange. The resulting model was documented as a UML class diagram.

Regarding implementation, the consultancy concluded on the suitability of both XML and RDF. XML was early on put forth as an implementation vehicle for Observations and Measurements [12], whereas the more recent of publication of the Sensor, Observation, Sample, and Actuator ontology (SOSA) [26], an RDF-based counterpart to O&M, presents a clear path to an implementation on the semantic web.

2. Background and related work

The GloSIS domain model and web ontology follow on the steps of various earlier attempts at a framework for the exchange of soil data and knowledge. This section reviews the most relevant.

2.1. ISO 28258

The international standard "Soil quality — Digital exchange of soil-related data" (ISO number 28253) resulted from a joint effort by the ISO technical committee "Soil quality" and the technical committee "Soil characterisation" of the European Committee for Standardisation (CEN). Recognising a growing need to combine soil data with other data kinds – especially environmental – this standard set out to produce a general framework for the recording and unambiguous exchange of soil data, consistent with other international standards and independent of particular software systems.

This standard was from the onset developed to target an XML based implementation. Its goal was not necessarily to attain a common understanding of the domain, rather to design a digital soil data exchange infrastructure. Therefore the accompanying UML domain model on which the XML exchange schema is rooted was merely a means to an end. Also recognising the relevance of spatial positioning in soil data, the standard adopted the Geography Markup Language (GML) as a geo-spatial extension to the XML encoding.

Even though not necessarily focused on a domain model, ISO 28258 captures a relatively wide range of concepts from soil surveying and physio-chemical analysis. The domain model is a close application of the meta-model proposed by O&M to the soil domain, supporting both analytical and descriptive results. Among the features of interest identified in the model can be highlighted:

- `Site` - representing the surrounding environment of a soil investigation, target of observations such as terrain or land use.
- `Plot` - the location or feature where a soil investigation is conducted, usually leading to a soil profile description and/or to the collection of soil material for physio-chemical analysis. Further specialised into `Surface`, `TrialPit` and `Borehole`.
- `Profile` - an ordered set of soil horizons or layers comprising the soil pedon at a specific spatial location. The object of soil classification.
- `ProfileElement` - an element of a soil profile, characterised by an upper and lower depth. Specialised into `Horizon` – a pedo-genetically homogeneous segment of the soil profile – and – `Layer` an arbitrary and heterogeneous segment of the soil profile.
- `SoilSpecimen` - an homogenised sample of soil material collected at a specific soil depth. Usually meant for physio-chemical analysis.

Meant as an asset for global use, ISO 28258 did not went into further specialisation. It does not propose attribute catalogues, vocabularies or code-lists of any kind, remaining open to the different soil description and classification systems used around the world. Although specifying a class for the traditional concept of "mapping soil unit" used in vector based soil mapping, the standard does not actually support the raster data paradigm. ISO 28258 was conceived as an empty container, to be subject of further specialisation for the actual encoding of soil data (possibly at regional or national scale). However, the standard has so far never been applied in this context it was designed for. The combination of a XML/GML approach (for which off-the-shelf tools remain scant) with the lack of code-lists possibly made the outright adoption of this standard too abstract for soil data providers.

2.2. *ANZSoilML*

The Australian and New Zealand Soil Mark-up Language (ANZSoilML) is rooted on the OzSoilML domain model [43] developed by CSIRO in Australia to support the exchange of soil and landscape data in that country. OzSoilML was possibly the first thorough application of the O&M framework to the soil domain. Soon after CSIRO joined forces with New Zealand's Manaaki Whenua to refine OZSoilML, eventually resulting in a new domain model renamed ANZSoilML [44]. Its domain model targets the soil properties and related landscape features specified by the institutional soil survey handbooks used in Australia and New Zeeland [34, 38].

As core feature of interest ANZSoilML defines the abstract `SoilFeature` class, a specialisation of the `LandscapeFeature` abstract class that may relate to an instance of `SpatialEntity` (the latter providing geo-spatial expression). `SoilFeature` specialises into three concrete classes: `SoilSurface`, `SoilHorizon` and `Soil`. The later can be associated with one or more instances of `SoilProfile`. These concrete classes correspond to concepts well familiar to soil surveyors.

In a drive to re-use existing domain models as much as possible, ANZSoilML specifies a set of classes for the description of soil composition that import concepts from GeoSciML [42]. Soil sampling and soil mapping are heavily reliant on O&M, with the `SoilProfile` class bridging to `SF_SpatialSamplingFeature`. GeoSciML is further used to provide meta-data on laboratory analysis, in conjunction with the `OM_Observation` complex from O&M.

In addition, a set of vocabularies were developed for ANZSoilML, providing values for categorical soil properties and laboratory analysis methods. However, these vocabularies were never made mandatory, with the goal of leaving the model open to use with alternative vocabularies. More recently these vocabularies were transformed into RDF resources, in order to be managed with modern Semantic Web technologies.

ANZSoilML was developed with the SOAP/XML web services specified by the OGC in mind. Thus the exchange mechanism is rooted on the concept of *ComplexFeature* and its hierarchical paradigm of data encoding with XML. Since 2019 CSIRO and Manaaki Whenua have been involved with the ELFIE initiative set up by OGC, a series of interoperability experiments promoting the exchange of environmental data on a linked data pattern. Future editions of ANZSoilML should therefore become increasingly aligned with the Semantic Web.

While focused on the particular context of two countries, ANZSoilML remains a pioneer in digital soil data exchange, providing a clear road map to similar initiatives. And its curated vocabularies make it one of the most well developed frameworks of the kind.

2.3. The Soil Theme in INSPIRE

The INSPIRE directive of the European Union came into force in 2007 aiming to create a spatial environmental data infrastructure for the Union. Its broad goals are three fold: (i) facilitate policy-making across borders, (ii) enable the interchange of environmental information among public sector organisations, and (iii) promote public access to spatial information across Europe. INSPIRE was implemented in various stages, broadly corresponding to environmental subdomains, called Themes. EU member states have been legally required to fully implement INSPIRE since 2019.

A detailed data specification for the soil theme was published by the European Commission in 2013 [45], supported by a detailed domain model documented as a UML class diagram. The INSPIRE domain model targets inventories of soil conditions and soil properties with soil monitoring over time in mind, but also soil mapping, primarily derived from soil inventory data. The model is far more developed in its inventory aspect, relying heavily on O&M in the specification of soil properties observations (both numerical and descriptive). While the domain model is documented as UML, there is no enforcing policy from the European Commission regarding implementation. Guidelines have been published by the INSPIRE Maintenance and Implementation Group (MIG) on possible implementation technologies, such as GeoPackage [1].

[1]https://github.com/INSPIRE-MIF/gp-geopackage-encodings

The features of interest identified in this model match familiar concepts in soil surveying: a `SoilProfile` class encapsulates the idea of an ordered collection of `SoilHorizons` or `SoilLayers` (*vide* Figure 2.3). A `SoilProfile` is described at a particular location captured by the `SoilPlot` class. A `SoilSite` class represents the surroundings of the `SoilProfile`, the spatial context in which the profile lays. Concerning mapping, the model specifies the `SoilBody` class, a spatial area associated with one or more profiles, meant for vector based cartography. In addition, the classes `RectifiedGridCoverage` and `ReferenceableGridCoverage` provide the backbone for raster based soil mapping.

An infrastructure has been set in place to register the code-lists of all INSPIRE themes, currently maintained by the Joint Research Centre [2]. The Soil properties identified as relevant by the European Commission are represented in this registry. However, they are in most cases composed solely by broad concepts that must be further redefined by member states. E.g. for the `SoilLayer` and `SoilHorizon` classes generic items are provided named `biologicalParameter` and `physicalParameter` that must be extended with actual biological and physical observable soil parameters.

The European Commission has set up a dedicated platform named INSPIRE Geoportal [3] functioning as a single access point to the INSPIRE-compliant data services provided by the EU member states. The Geoportal regularly harvests meta-data from the member states' discovery services, keeping a log of data availability. Users are thus able to discover, consult or download INSPIRE-compliant datasets without restrictions. However, not all member states make data services available for the soil theme, in many cases only *ad hoc* geo-spatial datasets are provided and in assorted file formats. The Geoportal does not seem to conduct any kind of semantic validation of the datasets and data services provided by member states.

The INSPIRE Soil Theme is possibly the most used soil ontology in the world, legally the working basis for twenty seven sovereign countries and at least as many soil survey institutions. However, it is at the same time the least accomplished regarding implementation of encoding mechanisms and operationalisation.

[2]https://inspire.ec.europa.eu/registry

[3]https://inspire-geoportal.ec.europa.eu/

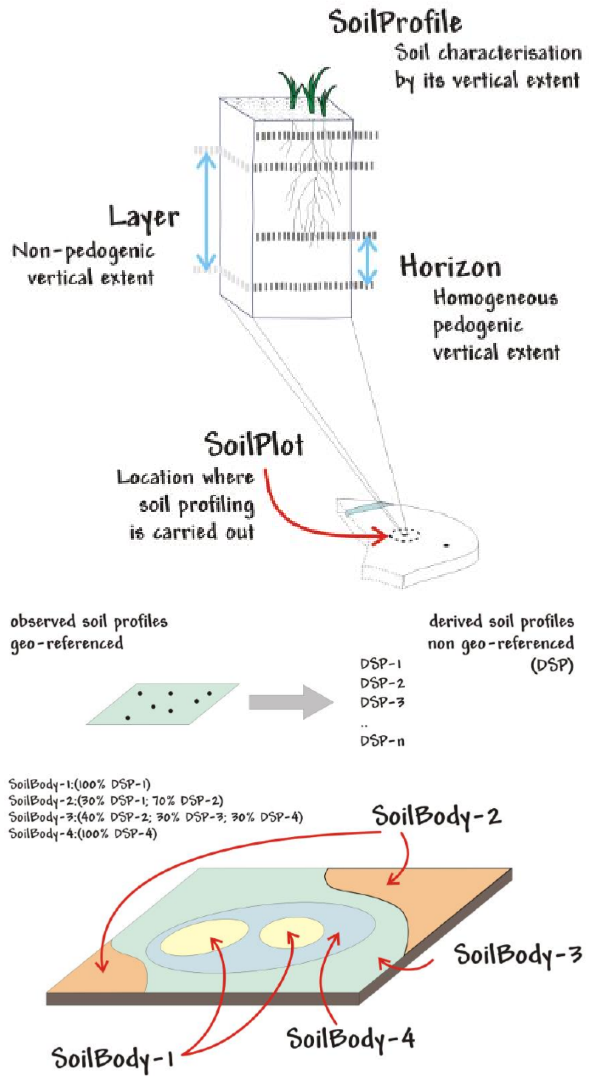


Fig. 1. Visual representation of the main feature of interest in the INSPIRE domain model. **Authorisation must be obtained for republishing**.

2.4. OGC Soil IE

The Working Group on Soil Information Standards (WGSIS) of the International Union of Soil Sciences (IUSS) acknowledged the parallel efforts in Oceania (ANZSoilML), Europe (INSPIRE) and by ISO towards a soil information exchange mechanism. However, in the perspective of the WGSIS these concurrent initiatives were leading to a dispersed landscape in need of consolidation. Under the auspices of the OGC, the WGSIS set out the Soil Interoperability Experiment (SoilIE), aiming to reconcile the existing soil information domain models into a single exchange paradigm. Among its goals, SoilIE attempted to synthesise a simplified domain model, easier to operationalise, and test it with a set of exemplary user cases. The end goal of the experiment was to be a prototype for an international soil information exchange standard to be adopted by the OGC.

The experiment went beyond the definition of a domain model and its translation into an exchange schema. The resulting XSD schema was used as bedrock for a series of web services implementing the user cases. This approached showcased the employment of a soil information schema used as actual exchange mechanism and not as a prescription for data structuring by providers.

Contrary to the "empty shell" approach of ISO 28258, SoilIE went on to define in detail the soil properties subject to exchange. To this end the experiment relied primarily on the FAO Guidelines for Soil Description [25], with additional guidance from the USDA Field Book for Describing and Sampling Soils [41]. As with previous efforts, SoilIE relied heavily on O&M to express the aspect of soil sampling and analysis, but going into considerable more detail, defining controlled content for its various classes. This is one of the merits of SoilIE, the application of O&M to explicitly decouple procedures, properties, units and results, addressing poor, bur pervasive practices in soil science and soil information conflating various of these concepts into massive, and largely unmanageable "properties" lists.

The resulting domain model is sub-divided into four sub-models, each addressing a specific aspect of soil information: (i) soil classification; (ii) soil profile description; (iii) sampling and field/laboratory observations; and (iv) sensor-based monitoring of dynamic soil properties. Left out of the experiment were soil mapping and landscape/land-use characterisation.

Perhaps at the behest of its initial goals, the SoilIE domain model became the most complex and detailed ontology of soil information resulting from an international effort. The number of features of interest and sampling features is considerably larger than in other ontologies, with more intricate relationships. In the Soil Description aspect the main class is `SoilFeature`, an umbrella concept for features of interest. `SoilFeature` is specialised into `Soil` and `Horizon` and is composed by a `Component` class, harbouring various physio-chemical soil properties. In the Soil Sampling and Observations aspect a host of sampling features is specified: `Plot`, `Layer`, `Station`, `Site` and `Sample`. `Soil` and `Horizon` also appear in this sub-model, but solely as features of interest and not directly related to sampling features. The concept of `SoilProfile` is introduced in its own sub-model, and as a different concept to `Soil`. However, it is the latter encapsulating the traditional concept of a vertical set of soil horizons, and containing most descriptive properties. The SoilIE report acknowledges the contentious nature of this distinction of concepts.

The experimental implementation took an hybrid approach. The domain model was encoded as a XML schema (known as SoilIEML) following the principles laid out in ISO 19136 [22], reliant on GML for geospatial features. This XML schema was the base for a series of OGC-compliant web services (Web Feature Service (WFS) in particular). The Simple Knowledge Organisation System (SKOS) was selected as preferred vehicle for controlled content (e.g. code-lists). The integration of the Semantic Web based SKOS with the XML schema proved problematic, with `XLINK` attributes eventually used to refer SKOS based URIs. Bespoke URI resolution services services were set up to de-reference SKOS concepts.

SoilIE reached many of the goals it set out to achieve, particularly in the operationalisation of a soil information exchange mechanism based on O&M, XML/GML and web services. However it is possibly hampered by a complex domain model, not always easy to square with common understanding in soil surveying. Among its many merits, SoilIE also showcased the role of the Semantic Web in information exchange, for instance in the way it provides unique identifiers and locators to controlled content.

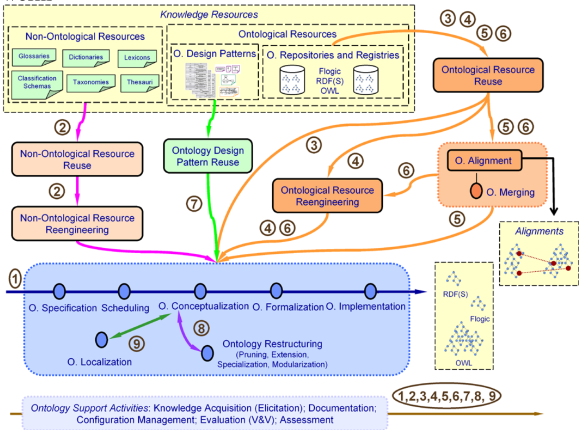
No international standard emerged out of SoilIE and no follow up applications are known. The controlled vocabularies and respective services are no longer online. However, the majority of the participants became involved in the efforts by the GSP towards GloSIS, pouring in their experience and eventually steering towards a full Semantic Web approach.

3. Methodology

3.1. NeOn Methodology

NeOn Methodology (*NEtworking ONtologies*) is one of the state of the art methodologies for building ontologies, which was developed in the course of NeOn Project's[4] work. The main objective of this methodology is to support both the collaborative aspects of ontology development and the reuse and dynamic evolution of ontology networks [20]. The methodology identifies nine scenarios for building ontologies and ontology networks, which can be combined in different ways, but always including Scenario 1:

Fig. 2. NeOn Methodology - scenarios for building Ontology Networks



1. From specification to implementation
2. Reusing and re-engineering non-ontological resources (NORs)
3. Reusing ontological resources
4. Reusing and re-engineering ontological resources
5. Reusing and merging ontological resources
6. Reusing, merging and re-engineering ontological resources
7. Reusing ontology design patterns
8. Restructuring ontological resources
9. Localising ontological resources

The NeOn methodology was adopted and used as reference for building the GloSIS ontology, applying some of the identified scenarios. In particular the following scenarios were used:

– Scenario 1: From specification to implementation - this scenario is made up of the core activities that have to be performed in any ontology development.
– Scenario 2: Reusing and re-engineering non-ontological resources - developers should decide according to the requirements in the ORSD (Ontology Requirements Specification Document) which existing NORs can be reused to build an ontology and then transform the selected NORs

[4]https://neon-project.org/

into ontologies; this will be further described in section 4.2.1.

- Scenario 3: Reusing ontological resources - existing ontological resources are used for building ontology networks - as a whole, only one part or module, or ontology statements; this will be further described in section 4.2.2.
- Scenario 7: Reusing ontology design patterns - ontology developers access ODPs (Ontology Design Patterns) repositories to reduce modeling difficulties, to speed up the modeling process, or to check the adequacy of modeling decisions.

3.2. Iterative-Incremental Model

The iterative-incremental model is based on the continuous improvement and extension of the ontology network resulted from performing multiple iterations with cyclic feedback and adaptation. This model was used for building GloSIS ontology. Firstly, a set of basic requirements was created; from these requirements, a subset was chosen and considered in the development of the ontology network. The partial result was reviewed, the risk of continuing with the next iteration was analysed, and the initial set of requirements was increased and/or modified in the next iteration. The ontology is continuously being reviewed and improved, according to the requirements and needs of the users/developers. The process is further described in section 5.

4. Ontology Specification

4.1. Requirements

The GloSIS data model shall as far as possible support the general requirements listed below; these requirements have been gleaned from the various inputs received as well as the discussions to date. The requirements presented below have been defined in line with the principles of software engineering.

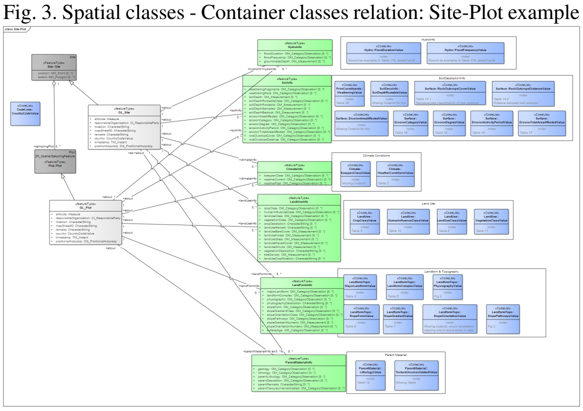
- Re-use existing standardisation efforts to avoid developing a completely new data model.
  - ∗ Re-use ANZSoilML as a basis/integrate relevant concepts.
  - ∗ Re-use ISO 28258 as a basis.
  - ∗ Integrate relevant concepts from the OGC Soil Interoperability Experiment.
  - ∗ Integrate relevant concepts from the SOTER/ISRIC model.
  - ∗ Resulting data model should be simple and easy to use.
- Support the properties pertaining to soil body as defined in the UN FAO Guidelines for soil description in a general way.
  - ∗ Design a generalised mechanism providing data users insight as to what properties are available pertaining to a specific soil body.
    - ∗ Codelists/federation of vocabularies/registries (ontologies) shall be developed for linking the data model with explicit soil body properties.
    - ∗ Include vocabularies/registries (ontologies), but in an abstract form. This means that vocabularies may be added/modified/deleted without changing the domain model itself.
    - ∗ AGROVOC terms should be used as a basis to avoid terms duplications.
  - ∗ The data model shall specify the main "groups" of soil body properties according to the UN FAO Guidelines for soil description.
- The data model shall support the properties inventoried by the GSP in the report "Specifications for the Tier 1 and Tier 2 soil profile databases of the Global Soil Information System (GloSIS)" [5].
- Decision on which concepts (Observed Properties) are considered as attributes and which should be provided as observations (as access to measurement metadata may be required) need to be reached.
- Concept for indicating observed properties available on the soil features should be supported.
- Platform agnostic soil data model, i.e. abstract specification (in the terms of the Open Geospatial Consortium), should be elaborated to provide a common basis for all ongoing and future developments.
- Provide mappings between the newly developed data model and all the existing data models.

4.2. Conceptualisation and Implementation

Linked (Open) Data is one of the most popular methods for publishing data on the Web as it can provide many benefits, including improved accessibility and easier integration that foster data reuse and exploitation. In order to enable the publication of soil datasets as Linked Data, the first step involves a definition of the target ontology that would be used for representing such data in semantic format. Instead of creating such an ontology from scratch, the GloSIS domain model was used as the base from which to derive the ontology.

The source UML model is composed of two main class types, the container classes, which are abstract classes used only for grouping observations (measurements) in a more readable manner, and spatial classes, which are the main GloSIS classes. The spatial classes are connected to the related observations via the connection with the container classes. Each of these two main types of classes was transformed and post-processed to generate the final ontology.

Fig. 3. Spatial classes - Container classes relation: Site-Plot example



Based on the decisions described earlier, ISO 28258:2013 Soil quality – Digital exchange of soil-related data incl. Amd 1 (ISO 28258) was taken as the basis for GloSIS data model development. In order to better understand the steps taken for this task, one must first understand the basic structure of ISO 28258. At the most abstract level, the two core components of ISO 28258 pertain on the one hand to a set of spatial object types describing soil objects as well as artefacts generated by soil sampling, on the other hand various observations or measurements of physiochemical properties on these objects. When extending this model for a specific usage area, one must determine if the information being extended is of a more static type, and thus should be appended to the spatial object, or of a more dynamic nature, or also a value that can be determined via vastly different methodologies, and thus should be provided as an observation on the spatial object.

The initial challenge in creating the GloSIS data model was identifying which spatial object types are required for the provision of the necessary information. Based on the GloSIS data requirements the following spatial data types were identified:

- Site
- Plot
- Surface
- Sample
- Specimen
- Profile
- Horizon
- Layer
- Grid

In a second step, the information requirements to each of these spatial object types was agreed upon with the experts, whereby basis was provided by the FAO Guidelines for Soil Description [25] and the GSP report “Specifications for the Tier 1 and Tier 2 soil profile databases of the Global Soil Information System” [5]. For this purpose, a spreadsheet was created with a row for every possible property, a column for each of the spatial object types. This matrix guided all further modelling work. Based on the understanding of the information requirements to each of these spatial object types, a decision had to be reached on how this information will be linked to the spatial objects. Based on the constraints laid down by ISO 28258, there were two main options available:

1. provide this information as an attribute of a specialised spatial object class;
2. provide this information as an O&M Observation referencing a specialised spatial object class.

While the first option is simpler to implement, the second allows for far more flexibility and precision pertaining to the information content. This is of particular relevance in the GloSIS context, as the model must support a very heterogeneous data provider community; one cannot mandate how data is to be ascertained, instead being grateful that data is available at all. Thus, we believe that through the wide use of the O&M Observation model, we can allow for well-structured provision of both data as we wish it to be, following the agreed methods and procedures, as well as other available data, whereby derivations from the agreed methods and procedures can be properly documented.

As the GloSIS model was created from a UML model, it had to be transformed to ontology, and then aligned with SOSA/SSN and O&M. Based on the acquired knowledge and previous experience (e.g., FOODIE project), a semi-automatic transformation process was carried out with the help of the tool called ShapeChange [5]. ShapeChange is a Java tool that takes application schemas constructed according to ISO 19109 from a UML model and derives implementation representations. Herein, the goal was to carry out a transformation from UML into RDF/OWL. ShapeChange enables the generation of an ontology following the ISO/IS 19150-2 standard, which defines rules for mapping ISO geographic information from UML models to OWL ontologies.

The output ontology generated by ShapeChange provided a good starting point to produce the final GloSIS ontology, but it required substantial post-processing tasks, as described in the following sections. In particular, due to the presence of the container classes and specific requirements (e.g., reuse the ISO 28258 standard as a basis, use of properties pertaining to soil body as defined in the UN FAO Guidelines for soil description like observed property, used methodology, structure, etc.), the transformation using ShapeChange gave only a basic framework.

4.2.1. Reusing and Reengineering Non-Ontological Resources

The GloSIS model was created as a UML model and released as an Enterprise Architect project[6]; therefore, the goal was to carry out a transformation of this model into a semantic format. However, it had to be modified before a successful transformation using ShapeChange could be carried out. In particular, it was necessary to add an ApplicationSchema in the Stereotype of each package and assign the targetNamespace property to the GloSIS namespace value: http://w3id.org/glosis/model. This change was applied to all GloSIS packages, namely: CodeLists, General, Layer-Horizon, Observation, Profile, Site-Plot, and Surface, and thereafter they were saved as XMI 1.0 (XML Metadata Interchange)[7]. The model complexity required publishing each package to a separated XMI 1.0 file.

Another significant change that was required before moving to ShapeChange's transformation was the addition of missing DataTypes information. It was not the model's issue per se but rather a disconnect between the ShapeChange requirement for working with XMI files and how the Enterprise Architect performs export. Hence, missing DataTypes information in the XMI files was added to each package manually, including the name, visibility, and boolean properties associated with each `DataType_ID` used in the package. Some of the most commonly used DataTypes include:

- `OM_CategoryObservation`,
- `OM_Measurement`,
- `OM_TruthObservation`,
- `OM_ComplexObservation`,
- `CharacterString`.

The primary mechanism for providing arguments to ShapeChange is the configuration file. Through this, one can specify how elements from the source model are transformed into the target one. A configuration file itself is an XML (Extensible Markup Language) file, where the root element `<ShapeChangeConfiguration>` wraps five core elements: `<input>`, `<targets>`, `<log>`, `<dialog>` and `<transformers>`. GloSIS implementation re-used the default configuration provided with ShapeChange for testing purposes.[8].

The vanilla configuration file had to be adjusted for GloSIS transformation needs. Namely, each package had its corresponding configuration file through the `<input>` node. Besides a few operational changes made for convenience, the imperative changes had to be made to the `<targets>` node. ShapeChange's documentation brings a boilerplate code for the UML to RDF/OWL transformation[9]. With that said, to customise the conversion desirably, several changes were needed to the referenced snippet:

- Removing inputs="TRF" from `<TargetOwl>` node, as no transformer was used.
- Adjusting the value of URIbase to http://w3id.org/glosis/model.
- Adding source targetParameter, which describes the source model.
- Appending namespaces of additional vocabularies (under `<namespaces>` node) that define target terms used in the customised transformation

[5] https://shapechange.net/
[6] https://sparxsystems.com/products/ea/index.html
[7] https://shapechange.net/app-schemas/xmi/
[8] https://shapechange.net/resources/test/testXMI.xml
[9] https://shapechange.net/targets/ontology/uml-rdfowl-based-isois-19150-2/

rules. These include: `ssn`[10], `sosa`[11], `lcc-cr`[12], `iso19115-1`[13], `om`[14], `foaf`[15].

- Introducing additional mapping rules under the `<RdfTypeMapEntries>` node:
  1. `OM_CategoryObservation` xx `sosa:ObservableProperty`
  2. `OM_Measurement` xx `sosa:Observation`
  3. `CountryCodeValue` xx `lcc-cr:Alpha2Code`
  4. `DQ_PositionalAccuracy` xx `ssn:Property`
  5. `CI_ResponsibleParty` xx `foaf:Agent`
  6. `TM_Instant` xx `xsd:dateTime`
- Introducing three new rules under the `<EncodingRule>` node:
  1. rule-owl-prop-voidable-as-minCardinality , which encodes lower bound of multiplicity with 0 for voidable properties;
  2. rule-owl-prop-multiplicityAsUnqualifiedCardinalityRestriction, which makes the multiplicity of a UML property encoded with unqualified cardinality restriction;
  3. rule-owl-prop-globalScopeByUniquePropertyName, which makes ShapeChange determine if the name of a UML property from the application schema is going to be unique within the ontology into which its OWL property representation has to be placed. If it is unique, the property converts to a globally scoped one.

Once the configuration was completed, the transformation was carried out by invoking the ShapeChange processor in the command line with the customised config file as an input.

```
java -jar ShapeChange-2.9.1.jar
-Dfile.encoding=UTF-8 -c
myInfo/GloSIS-config.xml
```

The crude result of the transformation contained all container classes from the UML model represented as subclasses of `geo:Feature` and their relationship to spatial data types. Alongside the properties in the container classes, also known as container types. All container types were modeled as Object Properties with inchoate and shallow connections to the SOSA/SSN taxonomy.

Listing 1: Container Type

```
glosis:Concentrations.mineralConcSize
a owl:ObjectProperty ; rdfs:domain
glosis:Concentrations ; rdfs:range
sosa:ObservableProperty ;
skos:definition "Result should be of
type
MineralConcSizeValue\nObservedProperty
= MineralConcSize"@en .
```

After the transformation, the spatial object classes were represented as subclasses of `gsp:Feature` and connections between those classes, and container classes were represented as object properties with range and domain.

Listing 2: Spatial Object Class

```
glosis:GL_Plot a  owl:Class ;
rdfs:subClassOf gsp:Feature .
```

Listing 3: Connection

```
glosis:GL_Plot.climateInfo a
owl:ObjectProperty ; rdfs:domain
glosis:GL_Plot ; rdfs:range
glosis:ClimateInfo .
```

4.2.2. *Reusing Ontological Resources*

SOSA/SSN is a lightweight but self-contained core ontology. It has already been used in GloSIS as the base model to represent observations. Nonetheless, various `DataType` elements present in the UML representation required a more complex approach.

The post-processing part required cleaning the ontology at first. Namely, removing container classes alongside the pointers between them and spatial object classes. Secondly, the development of object properties while aligning them to SOSA/SSN considering their data type. The latter was a complex task that is presented with regard to `DataType` elements. `CharacterString` was the simplest of these. All container types that were associated with it were modeled as `owl:DataTypeProperty`, with a range of simple string and literal definition.

[10] Semantic Sensor Network Ontology: http://www.w3.org/ns/ssn/

[11] Sensor, Observation, Sample, and Actuator (SOSA) Ontology: http://www.w3.org/ns/sosa/

[12] Country and Subdivision Representation Ontology: https://www.omg.org/spec/LCC/Countries/CountryRepresentation/

[13] ISO 19115 standard 2018 release for citation and responsible party information: https://def.isotc211.org/ontologies/iso19115/-1/2018/CitationAndResponsiblePartyInformation.rdf

[14] The Ontology of units of Measure (OM) 2.0: https://github.com/HajoRijgersberg/OM/raw/master/om-2.0.rdf

[15] Friend Of a Friend ontology: http://xmlns.com/foaf/0.1/

Listing 4: Container Type - CharacterString

```
glosis_sp:physiographyDescription a
owl:DatatypeProperty ; rdfs:range
xsd:string ; skos:definition
"Description of the local
physiography"@en .
```

There was considerably more variability with post-processing various observation types and measurements. All of them were represented as subclasses of `sosa:Observation`.

Listing 5: Modeling Observations

```
glosis_lh:Fragments a owl:Class ;
rdfs:subClassOf sosa:Observation ;
rdfs:subClassOf [ a owl:Restriction ;
    owl:onProperty
    sosa:hasFeatureOfInterest ;
    owl:allValuesFrom [owl:unionOf
    (glosis_lh:GL_Layer
    glosis_lh:GL_Horizon) ] ] ;
     rdfs:subClassOf [ a owl:Restriction ;
         owl:onProperty sosa:hasResult ;
         owl:allValuesFrom
         glosis_lh:FragmentsValue ] ;
     rdfs:subClassOf [ a owl:Restriction ;
         owl:onProperty
         sosa:observedProperty ;
         owl:someValuesFrom
         glosis_cl:FragmentsPropertyCode ]
         .
```

Moreover, they were restricted by constraining the various `owl properties`. A feature of interest restriction was applied uniformly across all observations, connecting them to the spatial object(s).

Listing 6: Feature of Interest restriction

```
rdfs:subClassOf [ a owl:Restriction ;
owl:onProperty
sosa:hasFeatureOfInterest ;
owl:allValuesFrom [owl:unionOf
(glosis_lh:GL_Layer
glosis_lh:GL_Horizon) ] ] ;
```

The result restriction is represented differently depending on the type. The string is represented with `sosa:hasSimpleResult`.

Listing 7: Simple result restriction

```
rdfs:subClassOf [ a owl:Restriction ;
owl:onProperty sosa:hasSimpleResult ;
owl:allValuesFrom xsd:string ] ;
```

In the case of the result being an auxiliary class containing a code-list, the model would incorporate `sosa:hasResult` instead.

Listing 8: Result restriction

```
rdfs:subClassOf [ a owl:Restriction ;
owl:onProperty sosa:hasResult ;
owl:allValuesFrom
glosis_lh:FragmentsValue ] ;
```

Each code-list is modeled using a class and a concept scheme. The concept scheme is defined as an individual of type `skos:ConceptScheme`, while the class is defined as a subclass of `skos:Concept`. Both elements are pointing to each other via `rdfs:seeAlso property`. Then, each code-list value is modeled as an individual of `type`: the defined class and `skos:Concept`, and in the scheme the associated `ConceptScheme` individual. Furthermore, the class includes an enumeration of all the code-list value individuals as a Collection[16].

Listing 9: Code List

```
glosis_cl:rootsAbundanceValueCode a
skos:ConceptScheme ; skos:prefLabel
"Code list for RootsAbundanceValue -
codelist scheme"@en; rdfs:label "Code
list for RootsAbundanceValue -
codelist scheme"@en; skos:note "This
code list provides the
RootsAbundanceValue."@en;
skos:definition "table 80" ;
rdfs:seeAlso
glosis_cl:RootsAbundanceValueCode .

## The code list Class
glosis_cl:RootsAbundanceValueCode a
owl:Class; rdfs:subClassOf
skos:Concept ; rdfs:label "Code list
for RootsAbundanceValue - codelist
class"@en; rdfs:comment "This code
list provides the
RootsAbundanceValue."@en;
rdfs:seeAlso
glosis_cl:rootsAbundanceValueCode ;
owl:oneOf (
glosis_cl:rootsAbundanceValueCode-N
glosis_cl:rootsAbundanceValueCode-V
glosis_cl:rootsAbundanceValueCode-F
glosis_cl:rootsAbundanceValueCode-C
glosis_cl:rootsAbundanceValueCode-M )
.
```

[16] https://www.w3.org/TR/rdf-schema/#ch_collectionvocab

```
## One individual value
glosis_cl:rootsAbundanceValueCode-N a
skos:Concept,
glosis_cl:RootsAbundanceValueCode;
skos:topConceptOf
glosis_cl:rootsAbundanceValueCode;
skos:prefLabel "None"@en ;
skos:notation "N" ; skos:definition
"< 2 mm (number)0;> 2 mm (number)0" ;
skos:inScheme
glosis_cl:rootsAbundanceValueCode .
```

In order to facilitate the reuse, extension, and maintenance, code-lists were modeled in a separated module.

If the result is a numerical value, the model uses `sosa:hasResult` restriction, similar to the code-list approach. The auxiliary class that we link to the observation represents a numeric value type (integer, float, boolean). The class itself is defined as a subclass of `quadt:QuantityValue`, and it is restricted by constraining the properties `quadt:numericValue` and `qudt:unit` to a particular numeric type (e.g., `xsd:integer`) and unit of measurement (e.g., percent), respectively.

Listing 10: Numeric Value

```
glosis_sp:LandUseGrassValue a
owl:Class ; rdfs:subClassOf
qudt:QuantityValue ; rdfs:subClassOf
[ a owl:Restriction ; owl:onProperty
qudt:numericValue ; owl:allValuesFrom
xsd:integer ] ; rdfs:subClassOf [ a
owl:Restriction ; owl:onProperty
qudt:unit ; owl:hasValue
unit:PERCENT] .
```

Finally, the last restriction is linking the observation with the observed property, defined as an instance of `sosa:ObservableProperty`.

Listing 11: Observed Property

```
glosis_sp:parentLithologyProperty a
sosa:ObservableProperty ;
rdfs:label
"parentLithologyProperty"@en;
rdfs:isDefinedBy "GfSD Table 12"@en.
```

There are few cases where `sosa:observedProperty` links the observation with a code-list.

Listing 12: Code List for ObservableProperty

```
glosis_cl:PhysioChemicalPropertyCode
a owl:Class; rdfs:subClassOf
skos:Concept,
sosa:ObservableProperty ;
```

In those cases the code-list for the observed property is created based on the same approach to the one presented for the result. The only difference is that the class representing the corresponding code-list is defined as a sublass of `sosa:ObservableProperty` instead of `skos:Concept`.

ShapeChange's transformation resulted in spatial objects being represented only as subclasses of geosparql `Feature`[17] (See Listing 2). One of the post-processing goals was to enrich these classes and remove redundant connections between spatial objects and container classes (See Listing 3). To achieve it the spatial object classes were then aligned with the ISO 28258 standard. As there is no ontology available for such a standard, an additional module for modeling the relevant parts of this standard, was created manually. All properties directly associated with the spatial objects were captured as data type or object type properties and restricted with range and cardinality.

Listing 13: Spatial Object Class aligned with iso28258

```
glosis_sp:GL_Plot a owl:Class ;
rdfs:subClassOf iso28258:Plot ;
rdfs:subClassOf [ a
owl:Restriction ;
owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty glosis_sp:location
] ; rdfs:subClassOf [ a
owl:Restriction ; owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty glosis_sp:remarks
] ; rdfs:subClassOf [ a
owl:Restriction ; owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty
glosis_sp:responsibleOrganization
] ; rdfs:subClassOf [ a
owl:Restriction ;
owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty
glosis_sp:positionalAccuracy ] ;
rdfs:subClassOf [ a
owl:Restriction ; owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty glosis_sp:altitude
] ; rdfs:subClassOf [ a
```

[17] http://www.opengis.net/ont/geosparql

```
owl:Restriction ; owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty
glosis_sp:timestamp ] ;
rdfs:subClassOf [ a
owl:Restriction ; owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty
glosis_sp:mapSheetID ] ;
rdfs:subClassOf [ a
owl:Restriction ; owl:cardinality
"1"^^xsd:nonNegativeInteger ;
owl:onProperty glosis_sp:country
] .
```

4.2.3. Introduction of Procedure code-lists

A long standing issue in the semantics of soil science is the conflation of soil property and laboratory analysis concepts. *Ad hoc* soil datasets often commingle in a single item the soil property, the laboratory process used to assess it, and on occasion even the units of measure. The OGC SoilIE [36] identified this as a major hindrance to the correct exchange of soil information. Some of the soil properties inventoried in the GloSIS domain model yielded this problem.

In order to address this and further exemplify the rich use of the resulting GloSIS web ontology, a thorough inventory of physio-chemical analysis processes was gathered. The primary source of this inventory was the output of the Africa Soil Profiles Database [31], with further insight gathered from the WoSIS database and procedures manual [7]. A further spreadsheet was developed with this information, adding also bibliographic references and existing on-line resources detailing each laboratory process.

A small transformation was created to produce a new module inthe GloSIS web ontology from this spreadsheet, following on the framework applied with the ShapeChange transformation and making use of the SOSA/SSN and SKOS ontologies. Each laboratory process is expressed both as an instance of `sosa:Procedure` and of `skos:Concept`. The SKOS ontology is employed not only to formalise the description of the procedure, but also to build a hierarchical structure between less or more detailed laboratory methods (applying the `skos:broader` and `skos:narrower` predicates). In its turn, the SOSA/SSN ontology provided the means to relate procedures with soil properties, through the enrichment of `sosa:Observation` individuals (as shown in Listing 5). Listing 14 provides and example with a classical laboratory process to assess total Nitrogen content in the soil.

Listing 14: Procedure instance for the Kjedahl process of Nitrogen content assessment.

```
glosis_proc:nitrogenTotalProcedure-TotalN_kjeldahl
    a skos:Concept,
    glosis_proc:NitrogenTotalProcedure;
  skos:topConceptOf
      glosis_proc:nitrogenTotalProcedure;
  skos:prefLabel "TotalN_kjeldahl"@en ;
  skos:notation "TotalN_kjeldahl" ;
  skos:definition "Method of Kjeldahl
      (digestion)" ;
  skos:scopeNote
      <https://en.wikipedia.org/wiki/Kjeldahl_method>
      ;
  skos:scopeNote "Kjeldahl, J. (1883) 'Neue
      Methode zur Bestimmung des
      Stickstoffs in organischen Korpern'
      (New method for the determination of
      nitrogen in organic substances),
      Zeitschrift fur analytische Chemie,
      22 (1) : 366-383." ;
  skos:inScheme
      glosis_proc:nitrogenTotalProcedure .
```

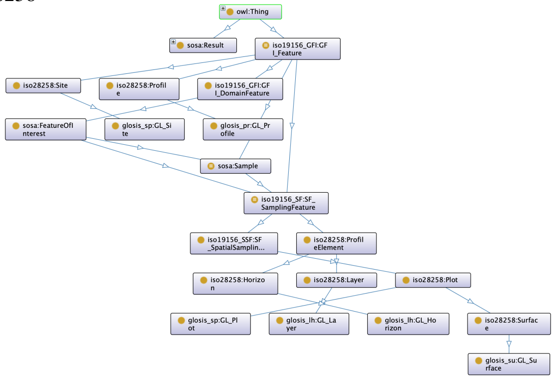
4.3. Ontology Overview

Considering readability and having in mind the best software development practices (e.g., “Do not Repeat Yourself“), the ontology was implemented following a modular approach as a networked ontology, facilitating its reusability, extensibility, and maintainability. For instance, all code-lists were implemented within the “code-list” module, and observations referenced across multiple modules were moved into a separate module called the “common module”. Additionally, as mentioned above, one of the most crucial aspects of post-processing was to align all the spatial objects with the ISO 28258 standard. That task was far from being straightforward since there is no existing ontology for this standard that could be used as a reference. Therefore, the "iso28258” module was created to introduce ISO features that were indispensable for connecting GloSIS ontology with an ISO 28258 standard. For this task, it was necessary to rely on the documentation of the standard. Additionally, this module includes alignment between elements in different ISO standards and other ontologies relevant to GloSIS. Some of these alignments include the definition of the following classes to be equivalent:

- `gsp:Feature` and `iso19156_GFI:GFI_Feature`;
- `sosa:Sample` and `iso19156_SF:SF_SamplingFeature`;
- `sosa:Observation` and `iso19156_OB:OM_Observation`.

The GloSIS classes are connected to the "iso28258" module and other ISO classes through inheritance as depicted on the diagram below:

Fig. 4. GloSIS Ontology - connection between spatial types and ISO 28258



There are a few important notes that complement the depicted diagram. First, `iso19156_GFI:GFI_Feature` is an equivalent of `gsp:Feature`. Secondly, `sosa:FeatureOfInterest` inherits from `iso19156_GFI:GFI_DomainFeature`. Finally, the alignment between sosa/ssn ontology and ISO 19156: `sosa:Sample` is equivalent to `iso19156_SF:SF_SamplingFeature`, and `sosa:Observation` corresponds to `iso19156_OB:OM_Observation`. Those alignments are explicitly stated in the ISO module of ontology.

4.3.1. *Ontology modules*

The current version of the whole ontology consists of 12 modules. The modular approach allows for the introduction of new extensions and modules whenever it is needed. Contents of the ontology (release v1.0.1):

- **glosis_main**: master module that imports all the components making data model simpler to use;
- **iso28258**: contains all ISO 28258 elements necessary to represent GloSIS, along with the mappings between ISO ontologies, SOSA/SSN, and GeoSPARQL;
- **glosis_layer_horizon**: contains all classes and properties to describe the domain of soil with a certain vertical extension, which is a layer (developed through non-pedogenic processes, displaying an unconformity to possibly over- or underlying adjacent domains) or a horizon (more or less parallel to the surface and is homogeneous for most morphological and analytical characteristics, developed in a parent material through pedogenic processes or made up of in-situ sedimented organic residues of up-growing plants (peat));
- **glosis_siteplot**: contains the classes and properties to describe soil sites (a defined area which is subject to a soil quality investigation) and soil plots (an elementary area where individual observations are made and/or samples are taken);
- **glosis_profile**: contains the classes and properties to describe a soil profile, which is a describable representation of the soil that is characterised by a vertical succession of horizons or at least one or several parent materials layers. Soil profile is an ordered set of soil horizons and/or layers;
- **glosis_surface**: contains the classes and properties to describe soil surfaces (a subtype of a plot with surface shape. Surfaces may be located within other surfaces);
- **glosis_observation**: contains the spatial class to describe the observation process, which is a subtype of `OM_Process`, and it is used to generate the result of the observation;
- **glosis_procedure**: contains the code-lists identifying laboratory processes employed to assess physio-chemical soil properties;
- **glosis_common**: contains all classes and properties that are used among multiple modules;
- **glosis_cl**: contains all the code-lists;
- **glosis_ext_property**: module containing extension to the initially derived ontology;
- **glosis_unit**: module that introduces additional units of measurement that are absent from the qutd ontology.

4.3.2. *Use of Permanent Identifiers*

In line with best practices, the GloSIS ontology has been implemented and released using persistent and resolvable identifiers, allowing access to the ontology on the Web via its URI and ensuring the sustainability of the ontology over time. In particular, the w3id service for persistent identifiers has been used. The base URI of the GloSIS ontology is `https://w3id.org/glosis/model`. This URI redirects to the GLOSIS main module, and it is the only one needed to load the full ontology in an application or ontology editor. Similarly, each individual module is accessible via permanent URIs in the form: `https://w3id.org/glosis/model/`*`module_name`*.

4.3.3. Documentation

The various modules of the GloSIS ontology are documented with a series of HTML pages generated automatically by the Wizard for Documenting Ontologies (WIDOCO) [18]. Written in Java, this software is able to inspect a web ontology and generate human-friendly documentation for all its classes, data types and data properties, in a well organised structure. The output documents apply internal HTML links to facilitate navigation among the different sections. It also integrates with WebVOWL [33] for automatic diagram generation.

WIDOCO is also able to extract some meta-data from the ontology, in order to document its authorship, provenance and licensing. However, it is not able to fully process predicates from the multiple meta-data ontologies in use today ?Doublin Core, VCard, Schema.org, etc). Instead WIDOCO makes available a configuration file in which meta-data can be declared to then be included at generation time. This configuration file contains important meta-data such as authors, contributors and their respective affiliations. Considering the number and varied nature of modules in the GloSIS ontology, it was deemed impractical to maintain a WIDOCO configuration file for each. Such practice would lead to redundancy with the meta-data triples already included in the ontology modules themselves.

A small programme was developed to address the issue above. It inspects the meta-data triples declared in a ontology module, and then produces a specific configuration file for WIDOCO. This programme is included in the GloSIS repository [18], it is able to identify various predicates from the Doublin Core Terms ontology, plus `schema:affiliation` and `foaf:name`. Documenting GloSIS thus becomes a two-step process: first generate the meta-data configuration for WIDOCO and then generate the final HTML documents with WIDOCO itself.

This HTML documentation is also accessible through the W3ID dereferencing mechanism. Making use of content negotiation mappings, the user is presented with the HTML documentation when accessing GloSIS resources directly with a web browser. Otherwise, application access to GloSIS returns the ontology RDF documents.

[18]https://github.com/rapw3k/glosis/tree/master/doc

5. Maintenance

5.1. Versioning

GloSIS uses semantic versioning[19] to denote code changes. This means that the numbers have meanings. The goal of that is to communicate to the user what can be expected from the changes that were made. The general convention looks as follows:

MAJOR.MINOR.MICRO

Incrementing the **MICRO** number means that some bugs were fixed but there are no additional concepts and the existing code should still work without changes.
Incrementing the **MINOR** number means that there are some new concepts introduced, or perhaps there was an extension of an existing one.
Finally, incrementing the **MAJOR** means that the project was updated with significant changes, perhaps a new module was introduced, or there were other major changes in class relationships.

Besides versioning, GloSIS also has releases. Each release presents updated code that is usable and tested. The GloSIS repository does have a simple utility python tool to update the version together with version IRI for each module altogether.

5.2. Scripts for transformation between CSV and OWL

One of the many challenges that the implementation of the GloSIS ontology is facing is the cooperation between developers and domain specialists. The main difficulty that may prevent soil scientists from contributing to the project is the lack of RDF language knowledge. A transformer tool was developed to help with the following:

- Enabling contributions from entities that are not familiar with RDF language;
- Reproducibility;
- Maintainability - comparing changes while maintaining the ordering in the modules.

The tool is capable of performing transformations in two directions:

[19]https://semver.org/

1. from RDF document into CSV file;
2. from the CSV file into an RDF document (specifically a turtle file).

The former requires referencing the RDF file that will be exported to the CSV. The latter requires a specific SPARQL query that will allow translation from tabular data into RDF. The tool currently supports the transformation of two essential modules: code-lists and procedures. Those two are most likely to be the subject of domain experts' contributions as they both consist of enumerated lists that provide concept details. The transformer tool is a Python script that can be executed from the command line. It re-uses the following libraries:

- **rdflib**[20] a python package for working with RDFs. It consists of parsers, serialisers, graph interface a nd allows running SPARQL queries;
- **pytarql**[21] a python implementation of the TARQL tool that allows converting CSV files into RDF;
- **pandas**[22] a python easy-to-use data structures and data analysis tool that helps immensely with processing CSV files;
- **otsrdflib**[23] an extension to the rdflib package. It allows specifying and maintaining order for the Turtle serialisation.

5.2.1. RDF to CSV

The RDF into CSV transformer can automatically recognise between two supported modules: code-list and procedures. It uses the rdflib to load the module (TURTLE file) into a graph. First, it iterates through it to capture all Classes/Procedures and their corresponding instances with the help of regular expressions. Then it collects details from associated triples from each of them. Finally, all acquired pieces of information are arranged into the table and saved as a CSV file using Pandas. The CSV file has a fixed number of columns that are sufficient and compatible with the backward transformation.

```
$ python
transform_to_csv.py [path
to rdf file]
```

5.2.2. CSV to RDF

CSV into RDF transformer tool starts with generating initial RDF representation from the CSV file using the pytarql against provided SPARQL query. The transformer is equipped with two pre-prepared SPARQL files, one for each of the two modules. Unlike the previous transformation, this one requires some amount of post-processing. First of all, the `owl:oneOf` predicate that connects a Class or Procedure to the list of instances should point to the Collection[24]. Building a Collection directly through pytarql did not seem feasible. Therefore some post-processing is required. The rdflib library has a convenient way of introducing Collection to the graph[25]. The first post-processing step utilises the aforementioned functionality. The second one uses a template to append the module header to its content. It will adjust the header's `owl:versionInfo` and `owl:versionIRI` to the value provided through the tool initialisation command. Finally, the post-processing will end with ordering classes to maintain the order inside the Turtle. The ordering is fixed in the following manner:

1. `owl:Ontology` (header)
2. `skos:ConceptScheme`
3. `owl:Class` (code-list module) or `sosa:Procedure` (procedures module)
4. `skos:Concept`

```
$ python
transform_to_rdf.py [path
to input csv] [path to
SPARQL query file]
[output filename]
[version]
```

5.3. Scripts for documentation maintenance

6. Applications of the ontology

This section showcases the use of the GloSIS ontology to represent and query some exemplary soil datasets. First, this sections shows the applicability of the ontology by using it to publish widely known open datasets from Europe and beyond as Linked Data, which are publicly available via the FOODIE endpoint[26]. The generation and publication of the linked

[20] https://rdflib.readthedocs.io/
[21] https://github.com/semanticarts/pytarql
[22] https://pandas.pydata.org/docs/)
[23] https://github.com/scriptotek/otsrdflib
[24] https://www.w3.org/TR/turtle/#collections
[25] https://rdflib.readthedocs.io/en/stable/apidocs/rdflib.html?highlight=Collection#rdflib
[26] https://www.foodie-cloud.org/sparql

Table 1

Namespaces

rdf	<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
g_sp	<http://w3id.org/glosis/model/siteplot#>
g_pr	<http://w3id.org/glosis/model/profile#>
g_lh	<http://w3id.org/glosis/model/layerhorizon#>
g_cl	<http://w3id.org/glosis/model/codelists#>
g_pd	<http://w3id.org/glosis/model/procedure#>
sosa	<http://www.w3.org/ns/sosa/>
qudt	<http://qudt.org/schema/qudt/>
unit	<http://qudt.org/vocab/unit/>
xsd	<http://www.w3.org/2001/XMLSchema#>
rdfs	<http://www.w3.org/2000/01/rdf-schema#>
ogcgs	<http://www.opengis.net/ont/geosparql#>
gn	<http://www.geonames.org/ontology#>
nuts	<http://nuts.geovocab.org/id/>
iso28258	<http://w3id.org/glosis/model/iso28258/2013#>
cap-parcel	<http://lpis.ec.europa.eu/registry/applicationschema/cap-iacs-parcel#>

datasets was carried out using a Linked Data Pipelines tool, developed in the context of different projects (e.g., SIEUSOIL, DEMETER, OPEN IACS), which enables the fetching, preparation, transformation, integration, and publication of linked data in a triple-store[27]. In short, the tool requires a mapping configuration file that specifies how the elements in the source dataset should be transformed to elements in the target ontology (in this case GloSIS). For further information about the tool please refer to its repository in GitHub. Next, this section presents some examples for data retrieval using SPARQL queries over data generated and stored based on the GloSIS ontology. These queries show not only how to retrieve data fromt the original sources, but also how to exploit the linked data. Finally this section introduces a semantic REST API that is built on top of the GloSIS ontology and facilitates the data exploration. This API allows for different applications to consume easily linked data, without the need to know SPARQL, RDF and other semantic technologies.

The following sections use in the code listings the namespaces presented in table 1.

6.1. LUCAS 2015 Topsoil dataset

The LUCAS Programme is an area frame statistical survey organised and managed by Eurostat (the Statistical Office of the EU) to monitor changes in land use and land cover, over time across the EU [29]. Since 2006, Eurostat has carried out LUCAS surveys every three years. The surveys are based on the visual assessment of environmental and structural elements of the landscape in georeferenced control points. The points belong to the intersections of a 2 x 2 km regular grid covering the territory of the EU. This results in around 1 million georeferenced points. In every survey, a sub-sample of these points is selected for the collection of field-based information.

In 2015, the LUCAS survey was carried out in all EU-28 Member States. In total, 27 069 locations were selected for sampling. Samples were eventually collected from 23,902 locations, of which 22,631 were in the EU. Soil samples were collected from a depth of 20*cm* following a common sampling procedure. After the removal of samples that could not be identified, the LUCAS 2015 Soil dataset has 21,859 unique records with soil and agro-environmental data.

The dataset includes the identification code `Point_ID` of the samples and data of physical and chemical properties for each sample. These properties include: Coarse fragments, clay, silt, sand, pH in CaCl2 and in H2O, Electrical Conductivity, Organic carbon, Carbonates, Phosphorus, total nitrogen, and extractable potassium. Additionally, each sample includes the elevation at which the soil sample was taken, land cover class, land use class, and NUTS codes (levels 0,1,2,3) for the country and location where the sample was taken. The full LUCAS topsoil 2015 dataset was transformed into Linked Data and is available in FOODIE endpoint, under the graph: http://w3id.org/glosis/open/LUCAS/topsoildata/.

The following listings present one sample of the dataset represented according to GloSIS ontology. Listing 15 presents the Site instance and its geolocation, representing the location of the sample.

Listing 15: LUCAS site data point #26761786

```
<#site_26761786> a g_sp:GL_Site ;
   rdfs:label "LUCAS #26761786" ;
   ogcgs:hasGeometry <#site_geo_26761786> ;
   gn:parentADM1 nuts:PT1 ;
   gn:parentADM3 nuts:PT150 ;
   gn:parentCountry nuts:PT ;
   gn:parentADM2 nuts:PT15 ;
   iso28258:Site.typicalProfile
       <#profile_26761786> .
<#site_geo_26761786> a ogcgs:Geometry ;
   ogcgs:asWKT "POINT(-8.621613437
       37.336764358)"
```

[27] https://git.man.poznan.pl/stash/projects/DEM/repos/pipelines/browse

Listing 16 presents the Profile and Profile Element (Layer) instance associated to the site.

Listing 16: LUCAS profile data point #26761786

```
<#profile_26761786> a g_pr:GL_Profile ;
   rdfs:label "Profile for #26761786" ;
   iso28258:Profile.element
       <#layer_26761786> .
<#layer_26761786> a g_lh:GL_Layer ;
   rdfs:label "Layer for #26761786" .
```

Listing 17 presents an observation instance associated to the site.

Listing 17: LUCAS site observations #26761786

```
<#lu_26761786> a g_sp:LandUseClass ;
   rdfs:label "Land use for #26761786" ;
   sosa:hasFeatureOfInterest
       <#site_26761786> ;
   sosa:hasResult <#luvalue_U111> ;
   sosa:observedProperty
       g_sp:landUseClassProperty .
<#lc_26761786> a sosa:Observation ;
   rdfs:label "Land cover for #26761786" ;
   sosa:hasFeatureOfInterest
       <#site_26761786> ;
   sosa:hasResult <#lcvalue_375> ;
   sosa:observedProperty
       cap-parcel:landCover .
```

Listing 18 presents two of the observations instances associated to the layer.

Listing 18: LUCAS site observations #26761786

```
<#phCaCl2_26761786> a g_lh:PH ;
   rdfs:label "pH in CaCl2 for #26761786" ;
   sosa:hasFeatureOfInterest
       <#layer_26761786> ;
   sosa:hasResult <#phCaCl2_value_26761786> ;
   sosa:observedProperty
       g_cl:physioChemicalPropertyCode-pH ;
   sosa:usedProcedure
       g_pd:pHProcedure-pHCaCl2 .
<#phCaCl2_value_26761786> a g_lh:PHValue ;
   rdfs:label "pH in CaCl2 value for
       #26761786" ;
   qudt:numericValue "4.30"^^xsd:float ;
   qudt:unit unit:PH .
<#ec_26761786> a g_lh:ElectricalConductivity
    ;
   rdfs:label "EC for #26761786" ;
   sosa:hasFeatureOfInterest
       <#layer_26761786> ;
   sosa:hasResult <#ec_value_26761786> ;
   sosa:observedProperty
       g_lh:electricalConductivityProperty .
<#ec_value_26761786> a
    g_lh:ElectricalConductivityValue ;
   rdfs:label "EC value for #26761786" ;
   qudt:numericValue "4.38"^^xsd:float ;
   qudt:unit unit:MilliS-PER-M .
```

6.2. SRDB

The Global soil respiration database (SRDB) is a compilation of field-measured soil respiration (RS, the soil-to-atmosphere CO2 flux) observations. Originally created over a decade ago, its latest version (V5) [28] has restructured and updated the global RS database, including new fields to include ancillary information (e.g., RS measurement time, collar insertion depth, collar area). The updated SRDB-V5 aims to be a data framework for the scientific community to share seasonal to annual field RS measurements, and it provides opportunities for the biogeochemistry community to better understand the spatial and temporal variability in RS, its components, and the overall carbon cycle. The database is publicly available with a detailed documentation footnotehttps://github.com/bpbond/srdb.

Each record in the database includes fields regarding the record metadata, site data, measurement data, annual and seasonal RS fluxes, and ancillary pools and fluxes. For this transformation, we used only a subset of the site data fields, including Latitude, Longitude, Elevation, Soil bulk density, Sand ratio value, Silt ratio value, and Clay ratio value. The SRDB subset was transformed into Linked Data and is available in FOODIE endpoint, under the graph: http://w3id.org/glosis/open/srdb/.

The following listings present one sample record of the SRDB dataset represented according to GloSIS ontology. Listing 19 presents the Site instance and its geolocation, representing the location of the sample.

Listing 19: SRDB site for study #12211

```
<#site_12211_CN-SN-N180> a g_sp:GL_Site ;
   rdfs:label "Study #12211, site id:
       CN-SN-N180" ;
   ogcgs:hasGeometry
       <#site_geo_12211_CN-SN-N180> ;
   g_sp:altitude "1220" ;
   iso28258:Site.typicalProfile
       <#p_12211_CN-SN-N180> .
<#site_geo_12211_CN-SN-N180> a
    ogcgs:Geometry ;
   ogcgs:asWKT "POINT (107.67 35.22)"
```

Listing 20 presents the Profile and Profile Element (Layer) instance associated to the site.

Listing 20: SRDB profile for study #12211

```
<#p_12211_CN-SN-N180> a g_pr:GL_Profile ;
   rdfs:label "Profile for study #12211
       id:CN-SN-N180" ;
   iso28258:Profile.element
       <#l_12211_CN-SN-N180> .
<#l_12211_CN-SN-N180> a g_lh:GL_Layer ;
   rdfs:label "Layer for study #12211
       id:CN-SN-N180" .
```

Listing 21 presents few observation instances associated to the soil layer.

Listing 21: SRDB observations for study #12211

```
<#bd_12211_CN-SN-N180> a
    g_lh:bulkDensityWholeSoil ;
   rdfs:label "Bulk Density for study #12211
       id:CN-SN-N180" ;
   sosa:hasFeatureOfInterest
       <#l_12211_CN-SN-N180> ;
   sosa:hasResult <#bdv_12211_CN-SN-N180> ;
   sosa:observedProperty
       g_lh:bulkDensityWholeSoilProperty .
<#bdv_12211_CN-SN-N180> a
    g_lh:bulkDensityWholeSoilValue ;
   rdfs:label "BD value for study #12211
       id:CN-SN-N180" ;
   qudt:numericValue "1.3"^^xsd:float ;
   qudt:unit unit:GM-PER-CentiM3 .
<#si_12211_CN-SN-N180> a
    g_lh:ElectricalConductivity ;
   rdfs:label "Silt for study #12211
       id:CN-SN-N180" ;
   sosa:hasFeatureOfInterest
       <#l_12211_CN-SN-N180> ;
   sosa:hasResult <#siv_12211_CN-SN-N180> ;
   sosa:observedProperty
       g_cl:physioChemicalPropertyCode-Textsilt
       .
<#siv_12211_CN-SN-N180> a
    g_lh:SiltFractionTextureValue ;
   rdfs:label "Silt value study #12211
       id:CN-SN-N180" ;
   qudt:numericValue "70"^^xsd:float ;
   qudt:unit unit:PERCENT .
```

6.3. *The WoSIS RDF service*

The World Soil Information Service (WoSIS) is the result of a decade effort towards an harmonised soil observation dataset at the global scale [7]. WoSIS has its core a relational database containing information on more than 200 000 geo-referenced soil profiles, originating from 180 countries different countries. The number of individual soil horizons characterised in this database borders on 900 000, for which almost 6 million individual observation results are recorded. Source datasets are subject to a process of rigurous quality control and harmonisation in order to be added, resulting in a globally consistent dataset, directed at digital soil mapping and environmental application at large scales.

A pilot was conducted to set up a GloSIS-compliant RDF service with WoSIS as data source. This pilot considered in first place ontological alignment. The WoSIS data model follows a substantially different pattern to those found in soil ontologies (*vide* Section 2). For instance, WoSIS does not sport an entity ontologically similar to the `GL_Plot` class, whereas its `profile` entity, a handle for the geo-location of a soil investigation, is closer to `GL_Site` than `GL_Profile`. The WoSIS data model is also foreign to the O&M pattern, including an `attribute` entity that can correspond both to the `Property` and `Procedure` classes in SOSA/SSN. These ontological differences required an *ad hoc* alignment, mapping individual WoSIS attributes to specific GloSIS properties, observations and procedures.

These mappings were encoded in the external schema of the WoSIS relational database as a set of views. These views also perform a transformation to RDF, producing triples expressed in the Turtle language. Listing 22 provides a snipet of one of these views, creating instances of the `GL_Profile` class. The database primary keys are used to compose a URI for each instance, the PostGIS function `ST_AsText` is used to obtain the WKT literal matching the GeoSPARQL `hasGeometry` data property. Listing 23 shows a sample output of this view, including the Turtle URI abbreviations. Similar views were created to produce RDF for soil layers, soil properties, observations, procedures and results.

Listing 22: A view transforming WoSIS profiles into GloSIS compliant RDF.

```
CREATE VIEW rdf.profile AS
SELECT 'wosis_prf:' || p.profile_id || ' a
    glosis_pr:GL_Profile, geo:Point ;' ||
    CHR(10) ||
     '    dcterms:isPartOf wosis_ds:' ||
         d.dataset_id || ' ;' || CHR(10) ||
     '    geo:hasGeometry "' ||
         public.ST_AsText(geom) ||
```

```
        '"^^geo:AsWKT .' || CHR(10) ||
        CHR(10) AS rdf,
    p.profile_id,
    d.dataset_id
FROM wosis.profile p
LEFT JOIN wosis.dataset_profile d
  ON p.profile_id = d.profile_id
LEFT JOIN wosis.dataset s
  ON d.dataset_id = s.dataset_id;
```

Listing 23: Sample output of the database view in Listing 22.

```
@prefix geo:
    <http://www.opengis.net/ont/geosparql#> .
@prefix dcterms: <http://purl.org/dc/terms/>
    .
@prefix glosis_pr:
    <http://w3id.org/glosis/model/profile#> .
@prefix wosis_ds:
    <http://wosis.isric.org/dataset#> .
@prefix wosis_prf:
    <http://wosis.isric.org/profile#> .

wosis_prf:65321 a glosis_pr:GL_Profile,
    geo:Point ;
  dcterms:isPartOf wosis_ds:CU-SOTER ;
  geo:hasGeometry "POINT(-80.25
      22.81999969482422)"^^geo:AsWKT .

wosis_prf:71979 a glosis_pr:GL_Profile,
    geo:Point ;
  dcterms:isPartOf wosis_ds:CU-SOTER ;
  geo:hasGeometry "POINT(-83.83
      22.25)"^^geo:AsWKT .

wosis_prf:71983 a glosis_pr:GL_Profile,
    geo:Point ;
  dcterms:isPartOf wosis_ds:CU-SOTER ;
  geo:hasGeometry "POINT(-81.5
      22.75)"^^geo:AsWKT .
```

Meta-data was added with predicates from Doublin Core, VCard and DCat web ontologies.

A set of triples produced by these RDF transformation views were deployed to the Virtuoso triple store, accessible through a SPARQL endpoint [28] and the Virtuoso Faceted Browser [29]. This pilot RDF service showcases the transformation of a traditional soil observation dataset into a GloSIS-compliant knowledge graph. It exemplifies the geo-location of soil profiles with GeoSPARQL, their composition with soil horizons and respective characterisation with observations of physio-chemical properties.

[28] https://virtuoso.isric.org/sparql/

[29] https://virtuoso.isric.org/fct/

6.4. Data discovery and access

This section presents two different approaches to discover and access data represented according to GLOSIS ontology (as from the examples presented in the previous sections). First, the section introduces a set of exemplary SPARQL/GeoSPARQL queries that provide guidance on the interaction with a triple store serving GloSIS-compliant linked data. Then, the section presents an example REST API that allows simplified programmatic access to such data, abstracting all the details on how data is represented, or how to interact with semantic data via SPARQL queries.

A key advantage of producing and publishing GloSIS-compliant linked data is the possibility to access soil-related data from different sources in an integrated manner, as well as to discover and establish links between them, and with other relevant open datasets available in the Linked Open Data (LOD) cloud, e.g., FADN, NUTS, AGROVOC, etc.

6.4.1. SPARQL queries

The following queries use the namespaces listed in Listing 24

Listing 24: Definition of namespaces used in the example SPARQL queries

```
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX qudt: <http://qudt.org/schema/qudt/>
PREFIX glosis_lh: <http://w3id.org/glosis/
    model/layerhorizon#>
PREFIX geo: <http://www.opengis.net/ont/
    geosparql#>
PREFIX geof: <http://www.opengis.net/def/
    function/geosparql/>
PREFIX iso: <http://w3id.org/glosis/model/
    iso28258/2013#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-
    schema#>
PREFIX glosis_proc: <http://w3id.org/glosis/
    model/procedure#>
PREFIX ramon: <http://rdfdata.eionet.europa.
    eu/ramon/ontology/>
PREFIX skos: <http://www.w3.org/2004/02/skos/
    core#>
PREFIX glosis_sp: <http://w3id.org/glosis/
    model/siteplot#>
```

Listing 25 provides a query seeking the average value for the total nitrogen soil property in the top soil of a certain spatial area. Starting from the `glosis_lh:NitrogenTotal` observation, the query identifies all related results, layers, soil profiles and respective geometries. `FILTER` clauses are then used to restrain the selection to soil layers above 30 cm depth that are part of profiles within a geodesic bounding box. Finally, the `AVG` operator is employed to obtain the average nitrogen value.

Listing 25: SPARQL query retrieving average top-soil nitrogen total within a spatial area.

```
SELECT AVG(?value)
WHERE {
   ?obs a glosis_lh:NitrogenTotal ;
      sosa:hasResult ?res ;
      sosa:hasFeatureOfInterest ?lay .
   ?res qudt:numericValue ?value .
   ?lay iso:ProfileElement.lowerDepth ?depth
      ;
      iso:ProfileElement.elementOfProfile ?
         prf .
   ?prf iso:Profile.profileSite ?sit .
   ?sit geo:asWKT ?geom .
   FILTER (xsd:integer(?depth) <= xsd:integer
       ("30"))
   FILTER (geof:sfIntersects(?geom, "POLYGON
       ((-79 19, -79 25, -85 25, -85 19, -79
       19))"^^geo:wktLiteral))
}
```

The query in Listing 26 exemplifies the benefits of linked data, and the rich axiomatisation of GLOSIS ontology. The query retrieves the average value for PH soil property, measured using a specific procedure (e.g., in the top soil of a certain NUTS region, namely Poland. Similar to previous query, it starts by retrieving the values of PH observations (`glosis_lh:PH`), but it retrieves only those measured using specific procedure, namely in a soil/water solution (`glosis_proc:pHProcedure-pHH2O`). Then, the query retrieves the site location where the observations were measured, and filters the result to include only those taken in Poland. The last part requires to retrieve first, in a subquery, the geometry of Poland from the NUTS dataset.

Listing 26: SPARQL query retrieving average top-soil pH, measured in a soil/water solution, within Poland

```
SELECT (AVG(?value) as ?avg_pH_H20)
WHERE {
 ?obs a glosis_lh:PH ;
   sosa:hasResult ?res ;
   sosa:hasFeatureOfInterest ?lay ;
   sosa:usedProcedure glosis_proc:pHProcedure
       -pHH2O .
 ?res qudt:numericValue ?value .
 ?site iso:Site.typicalProfile/iso:Profile.
     element ?lay ;
   geo:hasGeometry/geo:asWKT ?geom .
 { SELECT ?g_nuts
   WHERE {
     ?n a ramon:NUTSRegion .
     ?n rdfs:label ?l .
     ?n geo:asWKT ?g_nuts .
     ?l bif:contains "PL" .
    } limit 1 }
 FILTER (geof:sfIntersects(?geom, ?g_nuts))
}
```

The query in Listing 27 exemplifies the benefits of code-lists and semantic inference. The query retrieves the total number of survey points (from LUCAS) over land use with specific type/supertype (e.g., PRIMARY SECTOR) that have nitrogen total higher than certain threshold (e.g, 2). The query leverages the taxonomic relationships in the code-list for land use (used in LUCAS) to retrieve observations with land use type in any level under the one specified by the user.

Listing 27: SPARQL query retrieving total number of survey points over land use with type/supertype (PRIMARY SECTOR), which have nitrogenTotal higher than 2)

```
SELECT count (distinct ?site) as ?
    total_survey_points
WHERE {
 ?obs a glosis_lh:NitrogenTotal ;
  sosa:hasResult ?res ;
  sosa:hasFeatureOfInterest ?lay .
 ?res qudt:numericValue ?value .
 FILTER (?value > 2)
 ?site iso:Site.typicalProfile/iso:Profile.
     element ?lay .
 ?lu sosa:hasFeatureOfInterest ?site ;
  sosa:observedProperty glosis_lh:
      landUseClassProperty ;
  sosa:hasResult ?lu_res .
 ?lu_res rdf:type|skos:broader* ?lu_code .
 ?lu_code skos:prefLabel "PRIMARY SECTOR"
}
```

Finally, the query in Listing 28 exemplifies even further the the benefits of linked data, and particularly how GLOSIS ontology provides the basis to enable an integrated access to multiple soil data sources. The federated query retrieves NitrogenTotal observations, which have value over the specified threshold, from two different endpoints (FOODIE and ISRIC), and return them in an integrated result set.

Listing 28: Federated SPARQL query retrieving soil observations with nitrogen total higher than 2 from FOODIE and ISRIC data sources

```
SELECT ?obs ?lay ?value
WHERE {{
 SELECT ?obs ?lay ?value
 WHERE {
  ?obs a glosis_lh:NitrogenTotal ;
     sosa:hasResult ?res ;
     sosa:hasFeatureOfInterest ?lay .
  ?res qudt:numericValue ?value .
  FILTER (?value > 2)
 }}
 UNION {
 SELECT ?obs ?lay ?value
 WHERE {
  SERVICE <https://virtuoso.isric.org/sparql
      /> {
   ?obs a glosis_lh:NitrogenTotal ;
      sosa:hasResult ?res ;
      sosa:hasFeatureOfInterest ?lay .
   ?res qudt:numericValue ?value .
   FILTER (?value > 2)
 }}}
} ORDER BY DESC (?value)
```

6.4.2. *Semantic REST API*

Although, the native language to access the RDF data generated based on the model is SPARQL, in order to facilitate the access and consumption of data by potential services/applications, a REST API is created. The REST API returns simple JSON data, which is one of the most popular formats used by Web services to produce/consume data. The API is implemented using GRLC[30] that translates SPARQL queries stored in a Git repository[31] to a REST API on the fly.

Hence, using as starting point the SPARQL from previous section, we created the following API methods:

- `/avg_nitro_for_geo` - implements the exact query in Listing 25, thus, it allows to retrieve the average NitrogenTotal value in a specific geospatial region. The input parameter is the geospatial region of interest, expressed in Well-Known Text (WKT) OGC standard format.
- `/avg_physioChemical_property_for_NUTS` - implements partially the query in Listing 26 , generalising it to retrieve the average value for a specified physioChemical property, in a specified NUTS region code. The input parameters are the NUTS code (e.g., PL, PL41, LT, NO), and the physioChemical property, which can be selected from the predefined list of possible types coming from GLOSIS ontology.
- `/avg_physioChemical_property_for_geo` - same as the previous endpoint, but instead of having as input a NUTS region code, it expects the geospatial region of interest, expressed in WKT format.
- `/avg_physioChemical_property_procedure_for_NUTS` - implements fully the query in Listing 26, generalising it to retrieve the average value for a specified physioChemical property, measured using a specified procedure, in a specified NUTS region code. The input parameters are the NUTS code, the physioChemical property, which can be selected from the predefined list of possible types coming from GLOSIS ontology, and the procedure used for the measurement. This procedure also comes from GLOSIS ontology, and the available options can be retrieved using the `physioChemical_procedures` method.
- `/federated_soil_observations_for_property` - implements the query in Listing 28, generalising it to retrieve the observations, for a specified physioChemical property that have value over a specified threshold (e.g., 2) from multiple data sources (foodie and isric). The input parameters are the threshold number, and the physioChemical property, which can be selected from the predefined list of possible types coming from GLOSIS ontology.
- `/physioChemical_procedures` - allows to retrieve the procedures availabe in GLOSIS ontology for a specified physioChemical property. The input is the physioChemical property, which can be selected from the predefined list of possible types coming from GLOSIS ontology.
- `/total_survey_points_lu_prop_value` - implements the query in Listing 27, generalising it to retrieve the total number of survey points, for a specified physioChemical property with value over a specified threshold (e.g. 2), measured in a land use of specified

[30] http://grlc.io
[31] https://grlc.io/api-git/glosis-ld/api

type (e.g., AGRICULTURE, FORESTRY, 'PRIMARY SECTOR', etc.).

7. Future Work

7.1. Ontological extensions

As it stands, the ontology currently spans soil data exchange in the same breadth as previous initiatives. Focus rests primarily with soil investigations conducted on the field, including the collection of physical samples later to be analysed with wet chemistry methods in a laboratory. There are though advancements in the domain that beg for consideration in a soil data ontology.

Modern instruments allow the collection of high resolution reflectance spectra from soil samples, an activity known as soil proximal sensing. From these spectra estimates of physio-chemical properties can be obtained by statistical models, with relatively high accuracy [52]. Soil spectroscopy instruments are also becoming increasingly relevant in field work, by avoiding expensive activities of sample transport and laboratory analysis [11]. The SOSA ontology already contains assets (such as the `Instrument` class) providing a base framework to extend the GloSIS ontology to proximal sensing. But further investigation is necessary on how best to encode reflectance spectra in a Semantic Web paradigm and reference statistical models.

Another field under active research is the estimation and inventory of measurement uncertainty. Such information is traditionally absent from soil data sources, even though uncertainties stemming from field work and laboratory procedures are known to be relevant [32]. In downstream activities relying heavily on soil data, such as digital soil mapping, and further into decision support, measurement uncertainty is capital in conveying an accurate characterisation and fidelity of resulting products. Since neither O&M nor SOSA consider measurement uncertainty, this remains an open field of research.

Finally a note on soil classification systems. The GloSIS ontology proposes a completely liberal approach, providing simple text data properties without supporting controlled content. The user can therefore use any classification system and even combine various systems. While there are merits to this approach, an alternative pattern with controlled content can be argued for. The World Resource Base of soil resources (WRB) would be the obvious choice for such content, as the only soil classification/description system developed for the world as a whole. However, the WRB system poses its own set of challenges. On overage, it is updated every 5 years, without backwards compatibility. Therefore a soil classified as Vertisol in the 2015 edition might be in a different class in the 2014 edition, yet another still in the 2007 edition and so forth. The INSPIRE Soil Theme opted for the 2007 edition of the WRB (currently legally binding), essentially deterring classification with later versions. In order for a system such as the WRB to be adopted as controlled content, a different evolution paradigm is necessary, taking into account the requirements of digital data exchange. Engagement with the WRB work group of the International Union of Soil Scientists (IUSS) towards this end is indispensable.

7.2. Operational improvements

A future goal is to use the transformer tool as a component in Continuous Integration (CI) and Continuous Delivery (CD). That would allow to automatically re-generate and deploy a new version of the ontology each time a change to the code-lists or procedures is recorded in the supporting spreadsheets. This future improvement can also include automation of other modules, which would allow making changes to the whole ontology content by contributors not familiar with RDF languages.

Also facilitating the use of the ontology is the set up of an on-line browsing service. This can be particularly worthwhile for the use of code-lists, that are somewhat extensive. Since code-lists are encoded with SKOS, some obvious options open in this regard. SKOSMOS [48] is a web application for the publication of controlled vocabularies based on SKOS providing powerful navigation functionalities. An alternative is the ONKI web service [49], a large platform that allows free upload of SKOS-based vocabularies. ONKI automatically provides APIs and web widgets for the resources uploaded.

7.3. Human Factors and Education

The GloSIS ontology is one further step in a long lineage of soil ontologies. While it presents clear advances in content and format (not the least by embracing the Semantic Web) by themselves these do not guarantee its complete success. Previous efforts did not always manage to fully engage the soil data provision community, and those that did so were invariably legally enforced. It is therefore capital to keep human factors of ontology use in consideration.

The CI/CD mechanism described above is one step in that direction, by facilitating the dialogue between computer scientists and soil scientists (likely unfamiliar with the innards of the Semantic Web). Providing a simple file format mirroring the actual ontology can be critical to engage and involve domain experts.

To further facilitate engagement with the wider community of soil scientists and soil data provision institutions the establishment of an "Ontology Steering Committee" (OSC) can be decisive. This body could mirror the governance paradigm employed in Open Source projects [19, 40], an assembly of computer scientists and soil scientists collectively guiding ontology development. The actual structure and rules of such body is beyond the scope of this manuscript, however, other concepts from the Open Source community, such as "Request For Change" [10], can provide the necessary templates. Towards this end, engagement with organisations such as the soil standards working group of the IUSS, or the Soil Ontology and Informatics Cluster of ESIP [32] can be paramount

[14] points to ontology as one of the remaining gaps in data science research and education. Its absence is understood to compromise most stages of the research process, starting with data collection and on to the rigour of outcome. However, ontologies and the semantic web in general have already been applied in the educational context to a large swathe of domains [27]. The introduction of soil ontology to soil science and soil data curriculae appear therefore as a natural development. With its extensive code-lists and standards based lineage, GloSIS is a strong candidate for practical application in education. Such development would not only render the use of ontologies commonplace, but also train a new generation of soil scientists themselves capable of evolving ontology in their domain.

References

[1] (2017a). Plan of action for pillar five of the global soil partnership. Technical report, GSP - Global Soil Parnership.

[2] (2017b). Plan of action for pillar four of the global soil partnership. Technical report, GSP - Global Soil Parnership.

[3] Banwart, S., Black, H., Cai, Z., Gicheru, P., Joosten, H., Victoria, R., Milne, E., Noellemeyer, E., Pascual, U., Nziguheba, G., Vargas, R., Bationo, A., Buschiazzo, D., de Brogniez, D., Melillo, J., Richter, D., Termansen, M., van Noordwijk, M., Goverse, T., Ballabio, C., Bhattacharyya, T., Goldhaber, M., Nikolaidis, N., Zhao, Y., Funk, R., Duffy, C., Pan, G., la Scala, N., Gottschalk, P., Batjes, N., Six, J., van Wesemael, B., Stocking, M., Bampa, F., Bernoux, M., Feller, C., Lemanceau, P., and Montanarella, L. (2014). Benefits of soil carbon: report on the outcomes of an international scientific committee on problems of the environment rapid assessment workshop. *Carbon Management*, 5(2):185–192.

[4] Barnes, M. (2015). Aichi targets: Protect biodiversity, not just area. *Nature*, 526(7572):195–195.

[5] Batjes, N., Kempen, B., and van Egmond, F. (2019). Tier 1 and Tier 2 data in the context of the federated Global Soil Information System (GLOSIS). Technical Report 2019/01, ISRIC - World Soil Information.

[6] Batjes, N. H., Ribeiro, E., and Van Oostrum, A. (2020a). Standardised soil profile data to support global mapping and modelling (wosis snapshot 2019). *Earth System Science Data*, 12(1):299–320.

[7] Batjes, N. H., Ribeiro, E., and Van Oostrum, A. (2020b). Standardised soil profile data to support global mapping and modelling (wosis snapshot 2019). *Earth System Science Data*, 12(1):299–320.

[8] Borrelli, P., Robinson, D. A., Fleischer, L. R., Lugato, E., Ballabio, C., Alewell, C., Meusburger, K., Modugno, S., Schütt, B., Ferro, V., Bagarello, V., Oost, K. V., Montanarella, L., and Panagos, P. (2017). An assessment of the global impact of 21st century land use change on soil erosion. *Nature Communications*, 8(1):2013.

[9] Bouma, J. (2015). Engaging soil science in transdisciplinary research facing "wicked" problems in the information society. *Soil Sci. Soc. Am. J.*, 79(2):454–458.

[10] Canfora, G. and Cerulo, L. (2005). Impact analysis by mining software and change request repositories. In *11th IEEE International Software Metrics Symposium (METRICS'05)*, pages 9–pp. IEEE.

[11] Chang, C.-W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. *Soil Science Society of America Journal*, 65(2):480–490.

[12] Cox, S. (2011a). Observations and measurements-xml implementation. version 2.0. Technical report.

[13] Cox, S. (2011b). OGC Abstract Specification Geographic information — Observations and measurements. Technical report, Open Geospatial Consortium.

[14] Daniel, B. K. (2019). Big data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, 50(1):101–113.

[15] Desa, U. (2018). World urbanization prospects 2018. *United Nations Department for Economic and Social Affiars*.

[16] FAO, IFAD, UNICEF, WFP, and WHO (2018). The state of food security and nutrition in the world 2018. building climate resilience for food security and nutrition. Report, FAO.

[17] FAO and ITPS (2015). Status of the world's soil resources (swsr) - main report. Report, Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils.

[18] Garijo, D. (2017). Widoco: a wizard for documenting ontologies. In *International Semantic Web Conference*, pages 94–102. Springer.

[19] German, D. M. (2003). The gnome project: a case study of open source, global software development. *Software Process: Improvement and Practice*, 8(4):201–215.

[32] https://www.esipfed.org/get-involved/collaborate/soil

[20] Gomez-Perez, A. and Suárez-Figueroa, M. C. (2009). Neon methodology for building ontology networks: a scenario-based methodology.

[21] IPBES (2019). Global assessment report on biodiversity and ecosystem services of the intergovernmental science- policy platform on biodiversity and ecosystem services. e. s. brondizio, j. settele, s. díaz, and h. t. ngo (editors). Report, IPBES.

[22] ISO 19136:2007 (2007). Geographic information — Geography Markup Language (GML). Standard, International Organization for Standardization, Geneva, CH.

[23] ISO 19156:2011 (2011). Geographic information – Observations and measurements. Standard, International Organization for Standardization, Geneva, CH.

[24] ISO 28258:2013 (2013). Soil quality – Digital exchange of soil-related data. Standard, International Organization for Standardization, Geneva, CH.

[25] Jahn, R., Blume, H., Asio, V., Spaargaren, O., and Schad, P. (2006). *Guidelines for soil description*. FAO.

[26] Janowicz, K., Haller, A., Cox, S. J., Le Phuoc, D., and Lefrançois, M. (2019). Sosa: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 56:1–10.

[27] Jensen, J. (2019). A systematic literature review of the use of semantic web technologies in formal education. *British Journal of Educational Technology*, 50(2):505–517.

[28] Jian, J., Vargas, R., Anderson-Teixeira, K., Stell, E., Herrmann, V., Horn, M., Kholod, N., Manzon, J., Marchesi, R., Paredes, D., et al. (2021). A restructured and updated global soil respiration database (srdb-v5). *Earth System Science Data*, 13(2):255–267.

[29] Jones, A., Fernandez-Ugalde, O., and Scarpa, S. (2020). Lucas 2015 topsoil survey. presentation of dataset and results, eur 30332 en, publications office of the european union.

[30] Kopittke, P. M., Menzies, N. W., Wang, P., McKenna, B. A., and Lombi, E. (2019). Soil and the intensification of agriculture for global food security. *Environment international*, 132:105078.

[31] Leenaars, J., Van Oostrum, A., and Ruiperez, M. (2014). Africa Soil Profiles Database - Version 1.2. Technical report, ISRIC - World Soil Information.

[32] Libohova, Z., Seybold, C., Adhikari, K., Wills, S., Beaudette, D., Peaslee, S., Lindbo, D., and Owens, P. (2019). The anatomy of uncertainty for soil ph measurements and predictions: Implications for modellers and practitioners. *European journal of soil science*, 70(1):185–199.

[33] Lohmann, S., Link, V., Marbach, E., and Negru, S. (2014). Webvowl: Web-based visualization of ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 154–158. Springer.

[34] Milne, J., Clayden, B., Singleton, P., and Wilson, A. (1995). *New Zealand Soil Description Handbook*. Manaaki Whenua Digital Library, revised edition.

[35] Nations, U. (2019). World population prospects 2019. *Vol (ST/ESA/SE. A/424) Department of Economic and Social Affairs: Population Division*.

[36] OGC 16-088r1 (2016). OGC Soil Data Interoperability Experiment. Engineering report, Open Geospatial Consortium.

[37] Oldeman, L. and Van Engelen, V. (1993). A world soils and terrain digital database (soter) – an improved assessment of land resources. *Geoderma*, 60(1-4):309–325.

[38] on Soil, N. C., (Australia), T., and Publishing, C. (2009). *Australian soil and land survey field handbook*. Number 1. CSIRO PUBLISHING, third edition.

[39] Ramankutty, N., Mehrabi, Z., Waha, K., Jarvis, L., Kremen, C., Herrero, M., and Rieseberg, L. H. (2018). Trends in global agricultural land use: implications for environmental health and food security. *Annual review of plant biology*, 69:789–815.

[40] Riehle, D. (2011). Controlling and steering open source projects. *Computer*, 44(07):93–96.

[41] Schoeneberger, P. J., Wysocki, D. A., and Benham, E. C. (2012). *Field book for describing and sampling soils*. Government Printing Office.

[42] Sen, M. and Duffy, T. (2005). Geosciml: development of a generic geoscience markup language. *Computers & geosciences*, 31(9):1095–1103.

[43] Simons, B., Wilson, P., and Jacquier, D. (2012). Ozsoilml:an australian standard for exchange of soil data. In *Joint SSA and NZSSS Soil Science Conference*.

[44] Simons, B., Wilson, P., Ritchie, A., and Cox, S. (2013). ANZSoilML: an Australian-New Zealand standard for exchange of soil data. In *EGU General Assembly Conference Abstracts*, pages EGU2013–6802.

[45] Soil, I. T. W. G. (2013). D2.8.iii.3 inspire data specification on soil – draft guidelines. Standard, European Commission Joint Research Centre.

[46] Soussana, J.-F., Lutfalla, S., Ehrhardt, F., Rosenstock, T., Lamanna, C., Havlík, P., Richards, M., Wollenberg, E., Chotte, J.-L., Torquebiau, E., Ciais, P., Smith, P., and Lal, R. (2017). Matching policy and science: Rationale for the '4 per 1000 - soils for food security and climate' initiative. *Soil and Tillage Research*.

[47] Springmann, M., Clark, M., Mason-D'Croz, D., Wiebe, K., Bodirsky, B. L., Lassaletta, L., de Vries, W., Vermeulen, S. J., Herrero, M., Carlson, K. M., Jonell, M., Troell, M., DeClerck, F., Gordon, L. J., Zurayk, R., Scarborough, P., Rayner, M., Loken, B., Fanzo, J., Godfray, H. C. J., Tilman, D., Rockström, J., and Willett, W. (2018). Options for keeping the food system within environmental limits. *Nature*.

[48] Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J., Baker, T., Caracciolo, C., and Retterath, A. (2015). Publishing skos vocabularies with skosmos. *Manuscript submitted for review*.

[49] Tuominen, J., Frosterus, M., Viljanen, K., and Hyvönen, E. (2009). Onki skos server for publishing and utilizing skos vocabularies and ontologies as services. In *European Semantic Web Conference*, pages 768–780. Springer.

[50] UNEP (2012). *The benefits of soil carbon - managing soils for multiple, economic, societal and environmental benefits*, pages 19–33. United Nations Environmental Programme, Nairobi.

[51] van der Esch, S., Brink, B. t., Stehfest, E., Bakkenes, M., Sewell, A., Bouwman, A., Meijer, J., Westhoek, H., and van den Berg, M. (2017). Exploring future changes in land use and land condition and the impacts on food, water, climate change and biodiversity: Scenarios for the unccd global land outlook. Report, UNCCD.

[52] Viscarra Rossel, R., Adamchuk, V., Sudduth, K., McKenzie, N., and Lobsey, C. (2011). Chapter five - proximal soil sensing: An effective approach for soil measurements in space and time. In Sparks, D. L., editor, *Advances in Agronomy*, volume 113 of *Advances in Agronomy*, pages 243–291. Academic Press.

[53] WOCAT (2007). *Where the land is greener: Case studies and analysis of soil and water conservation initiatives worldwide*. CTA, UNEP, FAO and CDE, Berne.

[54] Řezník, T. and Schleidt, K. (2020). Data Model Development for the Global Soil Information System (GloSIS). Technical report, GSP - Global Soil Parnership.