Data model for the Soils4Africa project

Luís M. de Sousa   Luís Calisto   Paul van Genuchten
Ulan Turdukulov   Bas Kempen

juni 3, 2024

Abstract

This document provides a description of the data model created at ISRIC for the Soils4Africa project in 2022. This data model is intended to host soil description data collected during field work, measurements conducted in laboratory, as well as data derived from spectral based models. The data model is encoded according to the SQL standard language (plus SQL-MM for spatial entities).

In addition, this document also introduces the generic data model implementing the ISO 28258 standard, from which the Soils4Africa data model derives. Details are given on how the latter was specialised, together with maintenance and development guidelines.

Contents

1 Introduction

1.1 Motivation

In late 2021 a requirement emerged in the Soils4Africa project for a relational database model directed at the storage of soil properties observations. At first, the data model of the World Soil Information Service (WoSIS) was considered.

However, this path presented an insurmountable challenge from the onset, as the WoSIS data model was developed *ad hoc*, unrelated to the state-of-the-art in soil ontology. In addition, this data model does not always follows best practices in SQL and relational database modelling. Among the issues with the WoSIS data model, the following can be highlighted:

- absence of thesauri or other mechanisms for controlled content;
- data type mismatches, e.g. numerical values stored in text fields;
- functional dependencies (marked with the concept of "attribute");
- unclear semantics (in the data model and documentation).

While not the most severe issue, the lack of controlled thesauri renders the WoSIS data model nearly unusable outside the context for which it was developed. The other issues carry risks for maintenance and data accuracy that can easily become too expensive to address.

1.2 The option for an international standard

A second path was thus considered, developing a more general data model following the state-of-the-art in soil ontology. Several models have been proposed as basis for the exchange of soil data that offer a starting point for a data model. No thorough comparison was undertaken between the different options, the domain model proposed in the ISO 28258 standard was adopted outright for a first prototype. ISO 28258 is the only truly international standard directed at soil data exchange, with the alternatives being primarily regional in nature (e.g. INSPIRE, ANZSoilML). Soils4Africa taking place in a different continent, and with the goal of a generic data model in mind, the international model was the default choice. Further details on how these different models compare are offered in Section 2.2.

The ISO 28258 domain model provides a "shell" architecture meant to be specialised for particular contexts. Initial data model prototyping proved that to be the case. Moreover, the parenthood of ISO 28258 to the GloSIS web ontology also facilitated the use of controlled content referenceable on the web.

1.3 Repositories

A deployment of the Soils4Africa data model for the PostGres Database Management System (DBMS) has been maintained with the graphile-migrate migration tool. This PostGres deployment is maintained at a Git repository [1]. The core data model, implementing the ISO 28258 domain model, is maintained at a separate repository [2], so it may be forked for other purposes if necessary. Both of these data models can be deployed to any other SQL-compliant DBMS. Section 8 provides further details on these repositories and how to use them.

[1] https://git.wur.nl/isric/soils4africa/database
[2] https://git.wur.nl/isric/databases/iso28258

1.4 Structure

This document is structured as follows. Section 2 briefly presents the ISO 28258 standard and the domain model it proposes, also comparing with other relevant models in soil ontology. The features of interest included in the data model are introduced next, in Section 3, after which the observations and measurements assets are presented in Section 4. Section 5 details the procedures employed to populate the thesauri and their intended use. Meta-data is addressed in Section 6. The specialisation of the ISO 28258 model to Soils4Africa, with resulting assets, is exposed in Section 7. Some relevant operational aspects of the data model are discussed in Section 8. The document concludes by pointing directions for future work in Section 9.

1.5 Format

This document is encoded with the Markdown language, meant to be compiled with the Pandoc system [3]. It can also be consulted directly at the Git repository where it is maintained [4].

1.6 Glossary

- **Data model**: a logical structure for the storage of a data (usually in a digital system). In most cases a data model implements the informational aspect of a domain model. In this document relations in a data model are represented with lower case e.g., `element`.
- **Domain Model**: an abstraction synthesising the information (and in some cases behaviour) of a specific domain. Often represented with a visual language like UML. A domain model can also be referred as "ontology". In this document classes or concepts in a domain model are represented with capital characters, e.g. `SoilElement`.
- **Feature of Interest**: a class in the Observations and Measurements domain model representing the subject of an observation or measurement. I.e. what is meant to be observed.
- **Observations and Measurements (O&M)**: a standard sanctioned by ISO and the OGC providing a domain model for information captured with human instruments and methods on natural phenomena.
- **ontology**: an information abstraction resulting from the application of Ontology principles to the information/computer science domain. Expressed with lower case "o", an ontology is a domain model, usually not including behavioural aspects.
- **Semantic Web**: network of standard and specifications issue by the OGC for the digital exchange of data over the internet. It includes the Unified Resource Identifier (URI), the Resource Description Framework (RDF),

[3] https://pandoc.org/
[4] https://git.wur.nl/isric/soils4africa/data-model-doc

the Web Ontology Language (OWL), the SPARQL query language and much more.

- **Thesaurus**: a controlled set of terms that may be associated to a specific class property or table column. E.g. the set (*red*, *green*, *blue*) would be a thesaurus for the *colour* property. The concept of thesaurus is close to those of code-lists and vocabulary.

2 The ISO 28258 domain model

2.1 Overview

The international standard "Soil quality - Digital exchange of soil-related data" (ISO number 28258) ("Soil quality – Digital exchange of soil-related data" 2013) is the result of 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 types of data - especially environmental - these committees set out to produce a general framework for the unambiguous recording and exchange of soil data, consistent with other international standards and independent of particular software systems.

The ISO 28258 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 direct application of the meta-model proposed in the Observations and Measurements (O&M) standard ("Geographic information – Observations and measurements" 2011) to the soil domain. It aims to support both analytical and descriptive results.

ISO 28258 identifies the following features of interest:

- `Site` - representing the surrounding environment of a soil investigation, the subject of observations such as terrain or land use.
- `Plot` - the location or spatial 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. `Plot` is 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.

Figure 1 presents a simplified diagram of the ISO 28258 domain model showing the relevant relations between features of interest.

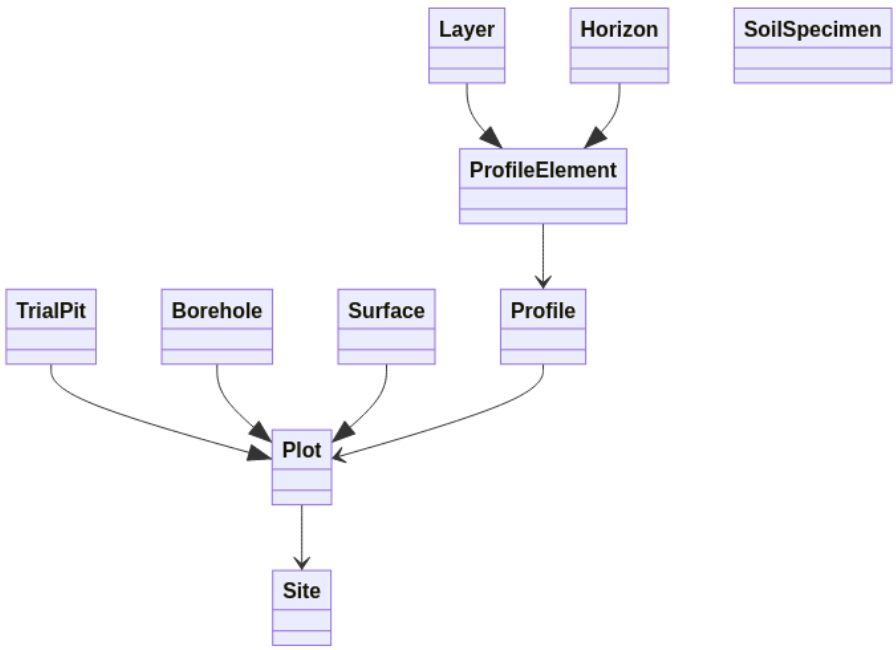

Figure 1: Simplified UML class diagram with ISO 28258 features of interest.

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 Relation with other soil ontologies

The ISO 28258 domain model is semantically rooted in O&M (see Section 4 for details), re-using UML stereotypes from that standard throughout. This architecture automatically aligns ISO 28258 with other standards from ISO and the OGC, such as Sampling Features. Moreover, it also provides alignment with other soil ontologies that follow a similar philosophy.

2.2.1 INSPIRE Soil Theme

Perhaps the most relevant among the domain models related to ISO 28258 is the Soil Theme domain model published by the European Commission in the context of the INSPIRE directive (Soil 2013). The core of these two models, concerning the abstraction of observations and measurements is by and large the same, with the concept of `Observation` expressed as a triple: `Property`, `Procedure` and `Unit`. As for the features of interest there are differences mainly in how the spatial surrounding of a soil investigation is abstracted. The concepts of `Plot` and `Site` are also present in INSPIRE but in a leaner way, without the specialisations found in ISO 28258. INSPIRE further adds the concept of `SoilBody`, a wider spatial area in which various soil investigations are conducted. And the concepts of `RectifiedGridCoverage` and `ReferenceableGridCoverage` provide a backbone for gridded data.

2.2.2 OGC SoilIE

In 2016 the OGC hosted an initiative named Soil Interoperability Experiment (SoilIE) ("OGC Soil Data Interoperability Experiment" 2016) with similar goals to those of ISO 28258. Also focused on data exchange, SoilIE would go into far more detail concerning features of interest. 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.

The SoilIE domain model yields familiar concepts such as `Site`, `Plot`, `Soil`, `Layer`, `Horizon` or `Sample`. But these are complemented by many other classes, in what is a far broader set of features of interest, with more intricate relationships. However, to what observations and measures is concerned, the same patterns proposed in the O&M standard are applied in this domain model too.

2.2.3 GloSIS

The GloSIS web ontology is essentially a translation of the ISO 28258 domain model to the Semantic Web, employing the Ontology Web Language (OWL). While semantically it is the same model, GloSIS introduces large sets of ready to use code-lists, including:

- Descriptive properties values (transposed from the FAO Guidelines of Soil Description (Jahn et al. 2006)).
- Physio-chemical properties (for Layer, Horizon, Plot and Profile).
- Procedures associated with pysio-chemical properties (re-used from the Africa Soil Profiles project (Leenaars, Van Oostrum, and Ruiperez Gonzalez 2014)).

By adopting the Semantic Web paradigm, this ontology automatically expresses all its content with Universal Resource Identifiers (URIs), than can easily be rendered derenferencable with a service such as W3ID (Group 2022). The GloSIS web ontology has in this way become one of the most extensive resources on soil ontology on the web.

2.3 Issues identified

During the course of this work various issues were identified with the ISO domain model that required addressing in the adaptation to the Soils4Africa project (details in Section 7). In particular:

1. The `Surface` and `Site` concepts revealed too similar and difficult to distinguish by soil scientists. Although yielding slightly different properties, the domain model is not fully clear. Moreover, a polygon type of spatial feature is expected to have a one-to-one relation with a `Profile`.
2. The `SoilSpecimen` concept is defined with a single depth property, whereas in soil surveying a sample collected in the field is always reported with two depths (upper and lower boundaries). A specimen (or sample) is regarded as a tangible segment of the soil profile whose material is homogenised.
3. `SoilSpecimen` and `Layer` appear also as too similar concepts. Both report to an arbitrary segment or stratum of the soil profile, in most cases unrelated to pedo-genetic horizon boundaries. Moreover, soil properties assessed in laboratories from soil specimens are often reported in reference to a soil layer by data providers. The depth issue noted above blurs the distinction between the two concepts even further.

3 Features of Interest

3.1 General

The ISO 28258 domain model was translated into a relational data model in the most seamless way possible. Whenever practical, classes and their attributes were translated directly into relations and attributes (or tables and columns). Generalisations were dealt with case-by-case, applying the child relations rules, i.e. creating individual entities for the children only if they bear diverse relations with other classes. This section reviews the entities and relationships implementing the features of interest (FoIs) and related assets.

3.2 Project and Site

The first concept to introduce is that of `Project`, a general placeholder providing the context of the data collection activity. The ISO domain model considers this a prerequisite for the proper use or reuse of these data. Within a project one or more soil investigations take place. It can be a soil sampling campaign, a regular soil survey, or some other organised process of soil data collection. The table `project` contains a single field for the project name, exposing the open nature of this concept (Figure 2).

A project can be related to one or more other projects. For instance, if a certain field campaign occurs at regular time intervals, the user might wish to record each as a single project, but related to others undertaken at a different time. Hierarchical relations can also be recorded this way, expressing a certain project as sub-project of another one. The table `project_related` provides the `role` field in which the user may express the nature of the relationship.

A soil investigation takes place within a certain spatial area or extent: the `Site`. A site is not a spatial feature of interest, but provides the link between the spatial features of interest (`Plot`) to the `Project`. It can be expressed either as a location (point) or spatial extent (polygon). The fields `position` and `extent` in the `site` table provide for this information, being that only one may be non empty for each record.

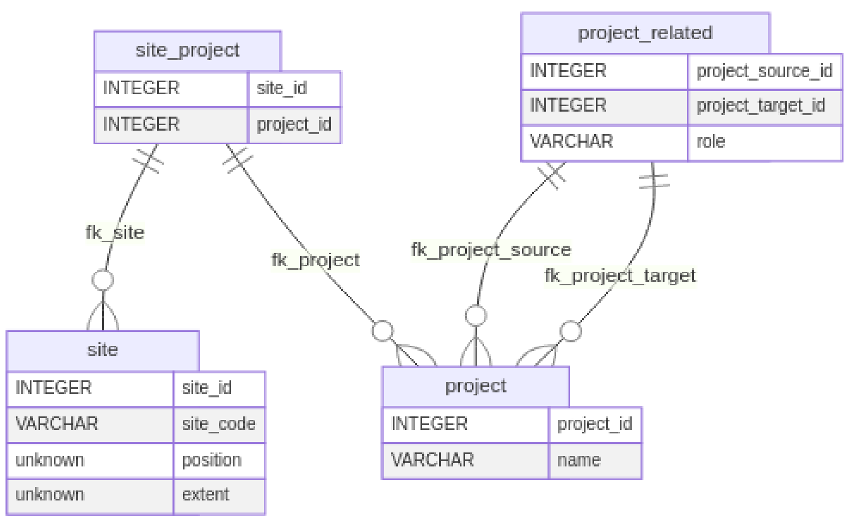

Figure 2: Entity-Relationship diagram for the `project` and `site` entities.

3.3 Spatial Features

All features of interest in the ISO 28258 model relate to a site, directly or indirectly. It is not possible to express information without previously defining the spatial extent within which the soil investigation took place. A single site may host more than one investigation, thus assumed to have an heterogeneous soil composition.

The soil investigation itself is conducted in a precise spatial location - the `Plot` - the first spatial feature of interest. The `Plot` can be of three different kinds: `Surface`, `Pit` or `Borehole`. The `Surface` corresponds spatially to a polygonal feature, a spatial extent within which the character of the soil tends to be homogeneous. `Surface` can express soil investigations recorded with high positional inaccuracy. The model also defines a hierarchical relation between surfaces.

Both the `Borehole` and the `Pit` correspond to point type spatial features, translating locations recorded with good positional accuracy. A borehole represents soil investigations conducted with an auger or similar boring instrument, whereas a pit indicates a soil excavation activity. In practice, both `Borehole` and `Pit` yield the same properties and relations, undistinguishable as data structures.

The `Plot` was thus modelled with two tables: one for the `Surface` and another - named `plot` - for `Borehole` and `Pit`. Both of these have a mandatory relation to a site. The `plot` table has a `Point` type column as spatial feature and `surface` a `Polygon` type column. As Figure 3 shows, the `plot` table contains a different set of columns.

The `plot` and `surface` tables provide the spatial hook on which to record a soil `Profile` (a vertical sequence of soil horizons). The resulting `profile` table is rather simple, yielding a `code` column and two foreign keys, one for `plot` and another for `surface`. A `CHECK` constraint forces one, and only one of these two foreign keys to be used simultaneously.

Note the nature of the relation between `Plot` and `Profile` being one-to-many. Meaning that more than one profile can be associated with the same plot (of whatever kind). This kind of association is not meant to related profiles collected at different points in time to a same plot. In such case the two profiles would be in different projects and would thus relate to different plots.

3.4 Other Features of Interest

There are three further features that do not have spatial expression but still are passible or measurement. Two are the profile elements: `Layer` and `Horizon`, the third being the `SoilSpecimen`.

`Layer` and `Horizon` present a similar nature: a section of the soil profile starting and ending at defined depths. Each instance is recorded in a particular order, re-creating the full soil profile. Semantically, `Layer` and `Horizon` differ in their

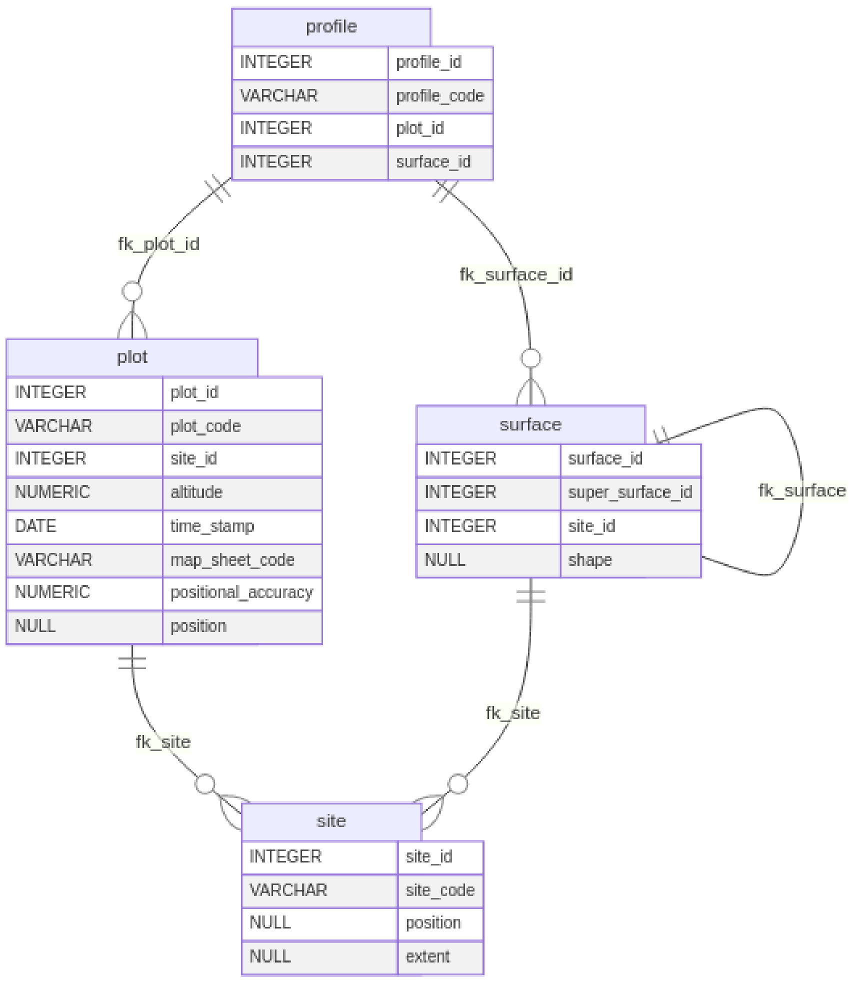

Figure 3: Entity-Relationship diagram for the `profile` and `plot` entities.

composition, the latter is homogeneous, the former heterogeneous. An horizon is identified by field observation, through sharp transitions of colour, composition or texture. Layer depths are usually pre-determined prior to field work. However, this difference does not translate into different properties or relations in the data model. Therefore, both `Layer` and `Horizon` are captured in a single table: `element`. The field `type` provides the means to distinguish between layer or horizon, if needed be. Figure 4 provides an overview of these relations.

The simultaneous inclusion of the properties `upper_depth`, `lower_depth` and `order` rapports important redundancies that are worth noting. The depth fields are recorded as positive integers representing centimetres from the surface, increasing downwards. The data model forces `upper_depth` to be lower than `lower_depth`. However the model cannot prevent overlapping profile elements from being recorded. Likewise, it cannot guarantee consistency between `order` and the depth columns. These redundancies are translated from the domain model "as is" since they portray common practice in soil survey. For these issues to be fully addressed a business rules layer is necessary (e.g. with database stored procedures).

The concept of `SoilSpecimen` in ISO 28258 is derived from the ISO 10381 standard for Soil Sampling (Standardization Organization) 2002). In essence it is a portion of soil matter (implicitly assumed as homogenous) collected at a certain depth, meant to be transported to a storage facility where it may be further prepared and analysed with different methods. It does not appear directly associated to any specific feature in the ISO 28258 domain model, but its parent structure in ISO 10381 indicates possible associations with spatial features that may function as sampling platforms. The WoSIS database includes a similar concept, `sample` that appears associated with a site or plot. The same approach was thus taken in the ISO 28258 data model with a one-to-many association between `plot` and `specimen`. An additional table, `specimen_prep_process`, provides essential attributes to record how a sample is transported and stored.

4 Observations & Measurements

4.1 Main concepts

O&M ("Geographic information – Observations and measurements" 2011) presents a general framework to encode measurements of natural phenomena of any kind. The concept of `Observation` is at its core, in essence a triplet of three other concepts:

- `Property`: an individual characteristic of a feature of interest. E.g. sand fraction in a profile element.
- `Procedure`: an action performed on a feature of interest (or sampling feature) in order to measure a property. E.g. sieving with 2 mm and 0.05 mm grades.

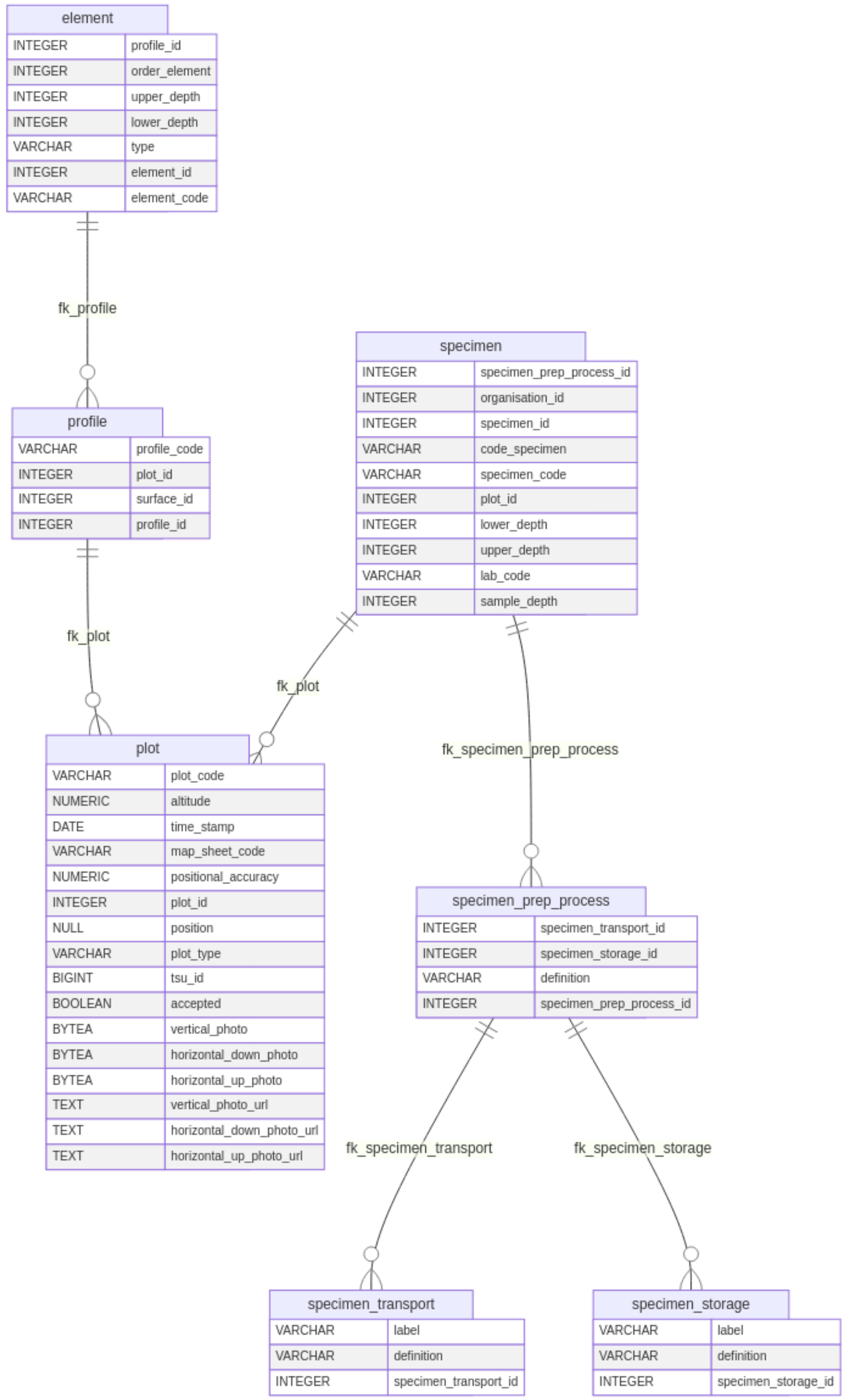

Figure 4: Entity-Relationship diagram for the `element` and `specimen` entities.

- `Unit of measure`: defines the magnitude of a measurement executed with a particular procedure on a property. E.g. per mille.

The `Result` class provides the actual placeholder for measurement results. It refers to a feature of interest and an observation and provides a matching value. The nature of this value is left completely open, it can be of any kind.

4.2 Physio-chemical Observations

Observations of physio-chemical properties of the soil provide a full introduction to the implementation of O&M in a relational database. These observations produce numerical results and refer to relatively well known procedures of wet chemistry and related soil specimen treatment.

The model proposed by O&M can be transposed to a relational structure nearly literally, however, there is an important choice to make regarding the nature of the relevant feature of interest. Since `Result` ought to refer to a feature, it must be translated into as many relational tables as many different features of interest exist. Therefore corresponding result tables are necessary for `surface`, `plot`, `profile`, `element` and `specimen`.

Considering physio-chemical observations, in practice only `Element` and `Specimen` yield observations of this nature, with numerical values as result. For `element` the observation tables include the suffix `_phys_chem` whereas for specimen it is `numerical_specimen`. If further features of interest come to require numerical observations this suffix system must be harmonised.

The `observation` table is in essence a ternary association, with foreign keys to `property`, `procedure` and `results`. In addition the columns `value_min` and `value_max` provide an interval of admissible values to guarantee data consistency (e.g. per mille). Figure 5 provides an overview of the observation assets for the `element` entity.

The tables for `property`, `procedure` and `results` are thesauri, not meant for direct modification by users. They provide the controlled content in this segment of the data model. In all three tables the natural columns are a human readable `label` and an `uri` to an entry in an on-line controlled vocabulary. This configuration is specifically designed to align the data model with the GloSIS web ontology (details in Section 5), but it also facilitates referencing other on-line sources of controlled content. The `procedure` table includes an additional foreign key to itself, for the encoding of hierarchical procedures, with different levels of detail.

Closing this segment of the data model is the `result` table. It refers both to an `observation` and a feature of interest (`element` for `result_phys_chem` and `specimen` for `result_numeric_specimen`). A numeric column named `value` hosts the actual measurement. To each numerical result table a trigger is associated that on insertion or update verifies the value against the `value_min` and `value_max` columns in the associated observation. In case it is outside the admissible interval, an exception is raised and the value remains unchanged.

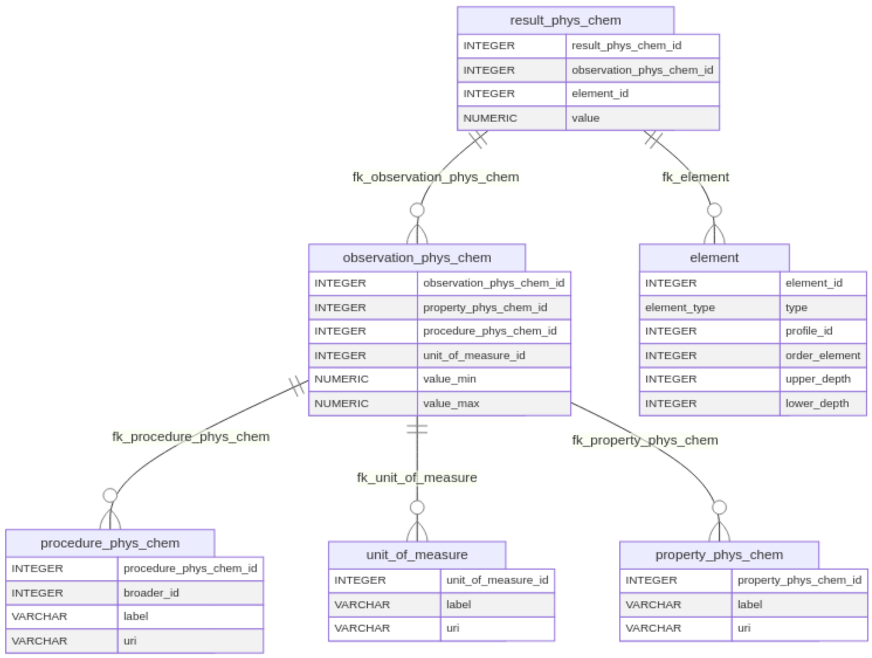

Figure 5: Entity-Relationship diagram for numerical observations on *element*.

4.3 Descriptive Observations

Various of the soil properties assessed during field work provide qualitative or descriptive results from direct observation, dispensing laboratory analysis. Examples are water drainage or soil classification. Results from these observations are thus textual in nature, ideally in reference to controlled content sources.

Compared to numerical observations, the main difference in the translation of the O&M pattern to a relational model for descriptive observations is in the nature of the result, it becomes a reference (foreign key) to a thesaurus. A thesaurus provides a list of controlled terms or items, that as a rule may not be modified by system users.

All features of interest can be subject to descriptive observations. Therefore the O&M pattern is replicated to each one, with a suffix to the table making the distinction. An exception was opened to the procedures thesaurus. In this case a procedure is a publication (field manual, soil description guidelines, etc) of which few are expected and should apply equally to different features of interest.

Figure 6 presents the tables for the descriptive observations related to `Surface`, thus yielding the suffix `_surface`. Similar tables exist for `plot`, `profile` and `element`.

A `property_desc` table provides the thesaurus for properties, just like in the numerical observations case. `thesaurus_desc` provides the additional controlled lists of descriptive terms to associate with properties. `procedure_desc` hosts then the thesaurus of publications. The `observation_desc` table relates these previous concepts in a ternary relationship as before.

The `result_desc` table is simply composed by two references, one to the feature of interest and another to the observation. The latter is actually a composed foreign key with the primary keys of the property and the respective item in the descriptive thesaurus. A joint unique constraint is applied on the foreign keys to the feature of interest and the property, guaranteeing that only one result is recorded for each property of each feature of interest.

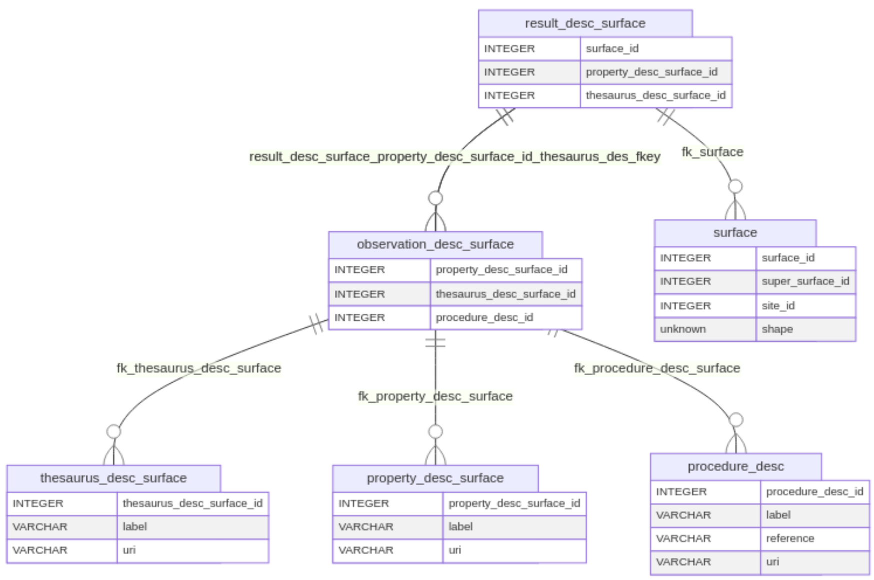

Figure 6: Entity-Relationship diagram for descriptive observations associated with the Surface feature of interest.

5 Thesauri

5.1 GloSIS code-lists

One of the advancements GloSIS brings to soil ontology is the introduction of comprehensive and structured code-lists. Making the best use of the Semantic Web, these code-lists are structured according the Simple Knowledge Ontology System (SKOS) (Miles and Bechhofer 2009) with each item fully de-referenceable. SKOS also makes these code-lists easily extendable.

GloSIS currently comprises code-lists for three cases:

- **Descriptive property values**: as digitised from the FAO Guidelines for Soil Description during the consultancy work for Pillar 5 of GSP (Řezník and Schleidt 2020). Some 830 codes are currently found in these code-lists.
- **Physio-chemical properties**: in accordance to the Tier1/Tier2 inventory gathered within the Pillar 4 of GSP [5], comprising 80 individual items.
- **Physio-chemical procedures**: adapted from the large catalogue assembled within the Africa Soil Profiles project (AfSP) (Leenaars, Van Oostrum, and Ruiperez Gonzalez 2014), currently totalling over 200 individuals.

Beyond these code-lists, the GloSIS ontology also embodies a large collection of descriptive soil properties. However, these are not modelled as code-lists, but rather as direct instances of the `Property` class from the SOSA ontology. Over 160 such properties currently exist in the ontology, across all features of interest.

These collections of controlled content provide a wealth of information form which to start a soil information system or an *ad hoc* database as the one described in this document. The GloSIS ontology is currently hosted at the W3ID service set-up by the W3C [6], guaranteeing an important level of resilience. Moreover, GloSIS remains in active development with accompanying tools that will facilitate the involvement of soil scientists.

SoilIE attempted a similar approach to controlled content, also producing code-lists with de-referenceble items. However, these are no longer on-line so it is not possible to compare their implementation.

5.2 SPARQL transformations

A series of transformations were created to obtain relational database records from the GloSIS ontology. They are coded as SPARQL queries and stored in the `sparql` folder of the code repository [7]. Each of these queries obtain as output a set of SQL `INSERT` instructions that for each code-list item create a counterpart record in the corresponding thesaurus in the database.

[5] "Specifications for the Tier 1 and Tier 2 soil profile databases of the Global Soil Information System", unpublished draft report.
[6] https://w3id.org/glosis/model
[7] https://git.wur.nl/isric/databases/iso28258/-/tree/master/sparql

The query in Listing 1 populates the thesaurus for descriptive properties associated with the `GL_Profile` class in GloSIS (`Profile` in ISO 28258). First, it identifies the relevant observations, those whose feature of interest is the `Gl_Profile` class (with the `sosa:hasFeatureOfInterest` predicate). Secondly, it identifies the associated properties with the `sosa:observedProperty` element. It then identifies the associated result and corresponding values code-list (`sosa:hasResult`). Using the `BIND` function the resulting SQL `INSERT` instruction is produced as a string. Similar SPARQL queries were developed for all other features of interest.

Listing 1 SPARQL query transforming descriptive observations for the 'GL_Profile' class into SQL 'INSERT' instructions.

```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:   <http://www.w3.org/2002/07/owl#>
PREFIX sosa:  <http://www.w3.org/ns/sosa/>
PREFIX skos:  <http://www.w3.org/2004/02/skos/core#>
PREFIX glosis_su:  <http://w3id.org/glosis/model/v1.0.0/profile#>

SELECT ?query
WHERE {
    ?obs rdfs:subClassOf sosa:Observation .
    ?obs rdfs:subClassOf ?rest_f .
    ?obs rdfs:subClassOf ?rest_p .
    ?obs rdfs:subClassOf ?rest_r .
    ?rest_f owl:onProperty sosa:hasFeatureOfInterest .
    ?rest_f owl:allValuesFrom glosis_su:GL_Profile .
    ?rest_p owl:onProperty sosa:observedProperty .
    ?rest_p owl:hasValue ?prop .
    ?rest_r owl:onProperty sosa:hasResult .
    ?rest_r owl:someValuesFrom ?code_list .
    ?value a ?code_list .
    ?value skos:prefLabel ?l .
    BIND (CONCAT("INSERT INTO core.observation_desc_profile (property_desc_profile_id, thesa
                   ?prop,
                   "\'), (SELECT thesaurus_desc_profile_id FROM core.thesaurus_desc_profile
                   ?value,
                   "\'));") AS ?query)
}
```

Similar queries are used to obtain the `INSERT` instructions for the code-list values themselves, that populate the thesauri. In Listing 2 is again the case for the Profile feature of interest. The matching triples in the `WHERE` clause are essentially the same as in Listing 1.

Obtaining the physio-chemical properties for the profile element is a more straightforward operation (Listing 3). The parent observation class

Listing 2 SPARQL query transforming code-lists of descriptive observations values for the 'GL_Profile' class into SQL 'INSERT' instructions.

```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:   <http://www.w3.org/2002/07/owl#>
PREFIX sosa:  <http://www.w3.org/ns/sosa/>
PREFIX skos:  <http://www.w3.org/2004/02/skos/core#>
PREFIX glosis_pr:  <http://w3id.org/glosis/model/v1.0.0/profile#>

SELECT ?query
WHERE {
    ?obs rdfs:subClassOf sosa:Observation .
    ?obs rdfs:subClassOf ?rest_f .
    ?obs rdfs:subClassOf ?rest_r .
    ?rest_f owl:onProperty sosa:hasFeatureOfInterest .
    ?rest_f owl:allValuesFrom glosis_pr:GL_Profile .
    ?rest_r owl:onProperty sosa:hasResult .
    ?rest_r owl:someValuesFrom ?code_list .
    ?value a ?code_list .
    ?value skos:prefLabel ?l .
    BIND (CONCAT('INSERT INTO core.thesaurus_desc_profile (label, uri) VALUES (\'', ?l,
          '\', \'', ?value, '\');') AS ?query)
}
```

`glosis_lh:PhysioChemical` is used to identify all the properties linked by this kind of observation, then retrieving the human readable strings to include in the thesauri.

The procedures thesaurus is the simplest to obtain, since all relevant items are instances of the SOSA class `Procedure` (Listing 4). However, in this case the can code-list is hierarchical and thus the SKOS predicate `broader` must be taken into account. Using the `OPTIONAL` function, the query generates an inner SQL `SELECT` query to identify the parent of each procedure, in case it exists.

6 Meta-data

6.1 Requirements

O&M does not consider meta-data directly, an aspect that is beyond the scope of the ontology. However, ISRIC identified the need to register individuals responsible for certain laboratory measurements and field observations. This information can be critical to keep track of laboratory work. Also regarding field work, this information can be crucial to trace irregularities and obtain clarification from the institutions involved.

vCard, the meta-data ontology specified by the W3C (Iannella and McKinney

Listing 3 SPARQL query transforming physio-chemical properties into SQL 'INSERT' instructions.

```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:   <http://www.w3.org/2002/07/owl#>
PREFIX sosa:  <http://www.w3.org/ns/sosa/>
PREFIX skos:  <http://www.w3.org/2004/02/skos/core#>
PREFIX glosis_lh:  <http://w3id.org/glosis/model/v1.0.0/layerhorizon#>

SELECT DISTINCT ?query
WHERE {
    ?obs rdfs:subClassOf  glosis_lh:PhysioChemical .
    ?obs rdfs:subClassOf ?rest_p .
    ?rest_p owl:onProperty sosa:observedProperty .
    ?rest_p owl:hasValue ?prop .
    ?prop skos:prefLabel ?label
    BIND (CONCAT('INSERT INTO core.property_phys_chem (label, uri) VALUES (\'', ?label,
          '\', \'', ?prop, '\');') AS ?query)
}
```

2014) and the ISO 19115 standard for geo-spatial meta-data ("geographic information — metadata" 2014) were initially considered as semantic sources. The Dublin Core vocabulary (Baker 2005) was also taken into consideration. vCard came to be the main source as it matched closer the organisational information commonly related to soil surveys.

6.2 Data model

6.2.1 Overview

The current data model is primarily based on the vCard classes `Organisation`, `Individual` and `Address` with respective tables for each (Figure 7). An organisation is typically a set of individuals that work together towards a same goal (or set of goals), but its meaning can be taken more broadly depending on the context. In most countries organisations are legal entities. Semantically the relevance is in distinguishing the collective from an individual. Organisations have a name and can be contacted through an e-mail address or a telephone number. They also may have a URL locating them in the digital space. Organisations can be set up in a hierarchical fashion, through a parent-child relationship. Organisations can also be sub-divided in various units (`organisation_unit` table).

The individual is a person, usually subject to a set of legal rights and obligations. The individual has a name and honorific title and can be contacted by e-mail or telephone. An individual can also have a URL to a web page of interest. Individuals can be part of an indefinite number of organisations (through the `organisation_individual` table). They may instead relate directly to a particular unit inside an organisation.

Listing 4 SPARQL query transforming physio-chemical analysis procedures into SQL ‘INSERT‘ instructions.

```
PREFIX owl:   <http://www.w3.org/2002/07/owl#>
PREFIX sosa:  <http://www.w3.org/ns/sosa/>
PREFIX skos:  <http://www.w3.org/2004/02/skos/core#>
PREFIX glosis_proc: <http://w3id.org/glosis/model/v1.0.0/procedure#>

SELECT ?query
WHERE {
    ?parent a sosa:Procedure .
    ?uri a ?parent .
    ?uri skos:prefLabel ?label .
    OPTIONAL {?uri skos:broader ?broader} .
    BIND (CONCAT('INSERT INTO core.procedure_phys_chem (uri, label, broader_id) VALUES (\''
         '\', \'', ?label, '\', (SELECT procedure_phys_chem_id FROM core.procedure_phys_che
         ?broader ,'\')  );') AS ?query)
}
```

The `address` table provides structure for physical postal addresses of individuals and organisations. Currently the relations to this table are one-to-many, allowing for various individuals to have the same address. This is a somewhat permissive structure, following the open nature of vCard, that may be made more restrictive if necessary.

6.2.2 Encapsulation

The entities created for the meta-data model are stored in their own database schema, named `metadata`. This limits the number of tables in the main schema (named `core`) facilitating human interaction with the database and its documentation. Foreign keys from the `core` schema to the `metadata` schema provide the appropriate relations between the two.

Not all database management systems (DBMS) implement the concept of schema laid out in the SQL standard. Some include the concept but do not provided standard interaction. Therefore this option for encapsulating the meta-data entities in their own schema limits the range of DBMS that can host this database.

6.3 Relations with ISO 28258 entities

With a meta-data data model established, relations with the ISO 28258 entities were devised, reflecting expectations on the field work conducted within Soils4Africa and similar projects. These relations identify the organisations and individuals responsible for data collection and/or asset storage. A simple matrix of relations was developed to aid discussion with domain experts (Table 1).

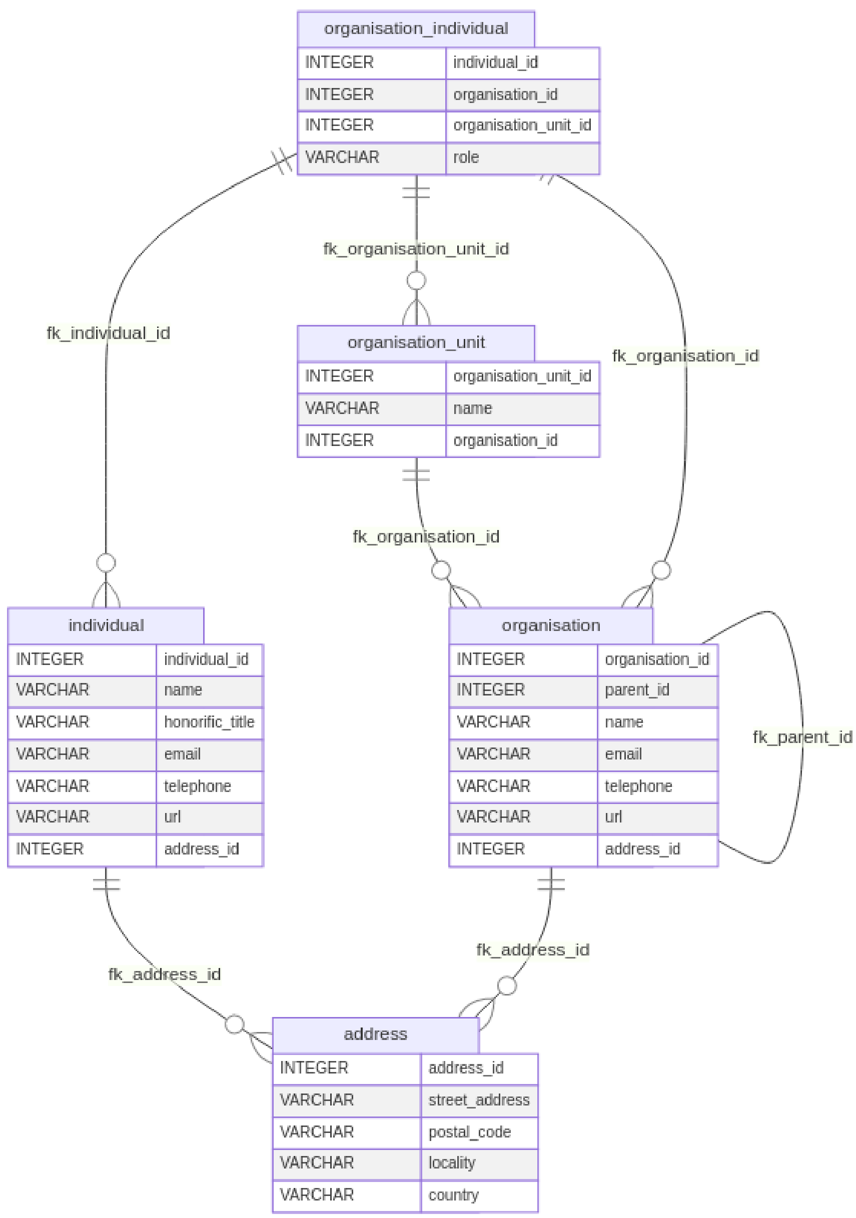

Figure 7: Entity-Relationship diagram for the meta-data tables.

Table 1: Relationships between meta-data and ISO 28258 entities.

	Individual	Organisation
Project	None	One or more
Plot	One or more	None
Surface	One or more	None
Profile	None	None
Element	None	None
Specimen	None	One
Results (Physio-Chemical)	One	None

A project may involve one or more organisations responsible for conducting a survey, or any other kind of soil investigation, including one or more plots. Each plot is surveyed by one or more individuals that in principle must be present at location. This resulted in the tables `project_organisation`, `plot_individual` and `surface_individual`. The individuals associated with a plot are also responsible for all descriptive results gathered on the profiles and profile elements surveyed within the plot.

Neither elements nor profiles refer to individuals or organisations. These entities are part of the investigation conducted on the plot and therefore associated to the respective individual. However, the specimens surveyors collect are sent to a storage facility hosted by an organisation that must be identified. It is also necessary to track individuals responsible for physio-chemical measurements conducted in laboratories. Hence a direct reference from this kind of result to the `individual` table.

7 Specialisations for Soils4Africa

7.1 General

The general data model described in the previous sections was specialised for the Soils4Africa project. This project concerns a wide field campaign across the African continent with soil investigations conducted in 20 000 different plots. These include *in situ* soil descriptions as well as soil specimen collection for later analysis.

Some aspects of the ISO 28258 data model had to be revised to meet the requirements for this project. In particular, the data model had to be expanded to support the activities of soil spectral analysis and modelling, not directly considered in ISO 28258.

7.2 Simplifications

As referred in Section 2 the distinction between the concepts of `Site` and `Surface` proved challenging. While their diverse semantics can be conveyed, the scant set of properties each concept bears in the ISO 28258 domain model does not immediately disclose their purposes.

Moreover, in the Soils4Africa project all plots surveyed correspond to spatial features of the type point (i.e., boreholes or pits). Therefore only the tables `site` and `plot` were retained from the ISO 28258 data model, dropping the `surface` table. The `site` table was translated into the geographic sampling units devised to plan field work, whereas `plot` records the geographic location of soil investigations, to which all relevant above ground properties refer.

7.3 Sampling units

Field work was planned according to a hierarchical spatial structure named sampling units. This structure starts with a squared grid with cells of 2 000 metres in side. Each cell in this top grid is called primary sampling unit (PSU). Each PSU is sub-divided in 400 squares of 1 ha each, named secondary sampling units (SSUs). The SSUs themselves are further sub-divided into 400 squares of 25 m2 each, called Tertiary Sampling Units (TSUs). Field work takes place per TSU, with four plots (boreholes) dug in each.

Specific attributes for each of these sampling units most be recorded in the database, collected in three different tables: (i) `psu`, (ii) `ssu` and (iii) `tsu`. Foreign keys from `tsu` to `ssu` and from `ssu` to `psu` make the hierarchy (Figure 8). Each table records a code in a string field and a geometry. These three tables replace the original `site` table, together corresponding to the same concept.

7.4 Spectral measurements

Soil specimens collected during field work will be target of spectral measurements during the Soils4Africa project. The ISO 28258 domain model does not consider particular measurements conducted with a spectrometer, resulting in a spectrum, i.e. an array of intensity values. However, this kind of measurement still conforms well with the general O&M framework (described in Section 4). Moreover, the SOSA ontology (Janowicz et al. 2019), the counterpart of O&M for the Semantic Web, includes the concept of Sensor, in which a spectrometer fits.

In the Soils4Africa data model, spectral observations thus start with the introduction of the `sensor` entity, harbouring the minimal description of the hardware used. `procedure_spectral` provides the details on the modes of operation of each sensor, specifying the range of the electromagnetic spectrum observed and the respective spectral resolution. `observation_spectral` brings both together, further allowing for a unit of measure, plus minimum and maximum admissible values. The actual result is stored as a JSON-B object (Hallam-Baker 2022) in the `result_spectral` entity. As Figure 9 shows, the result of a spectral

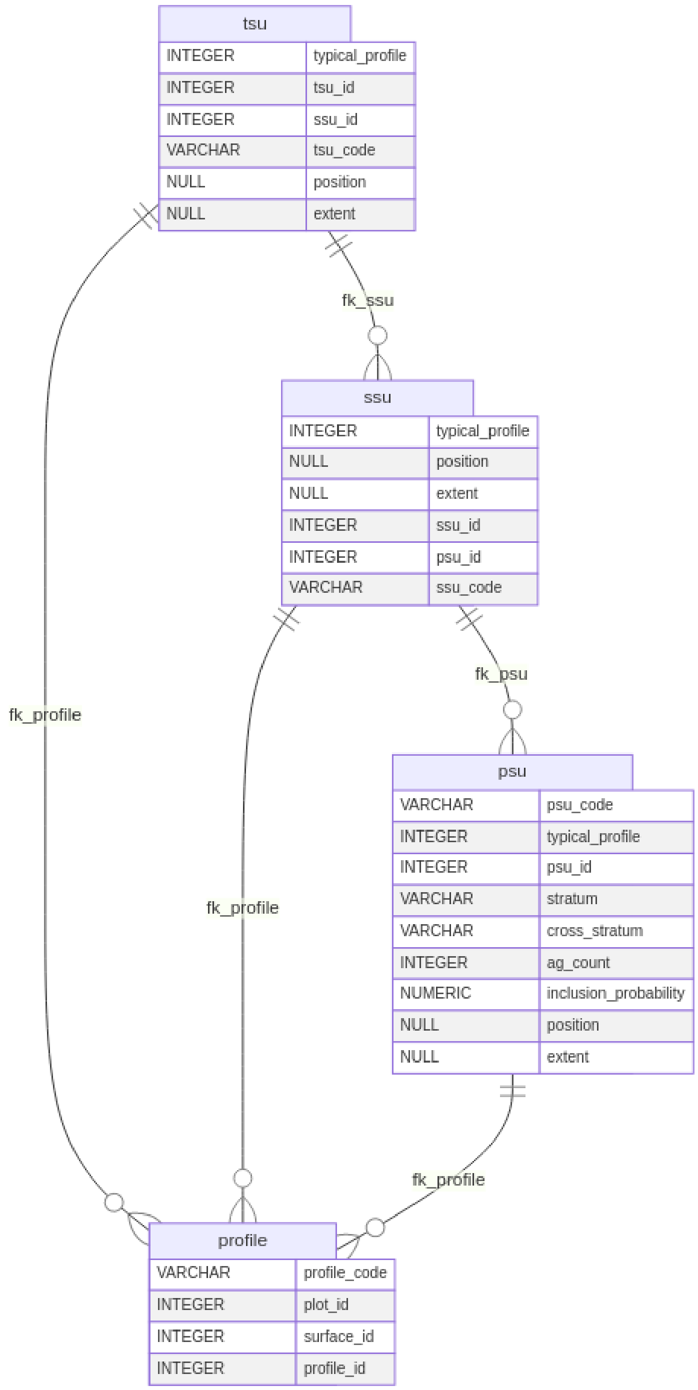

Figure 8: Entity-Relationship diagram for sampling units.

measurement refers to an individual record in the `specimen` table, identifying the of soil material on which the spectrometer was applied.

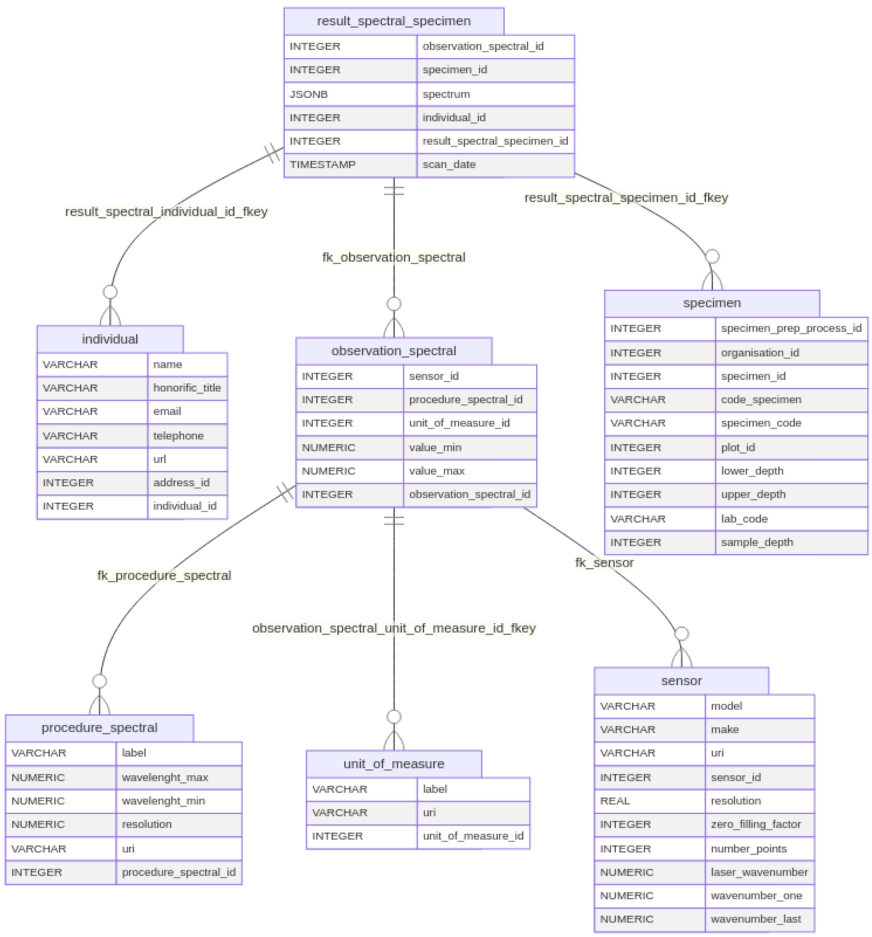

Figure 9: Entity-Relationship diagram for the spectral measurements tables.

7.5 Derived observations

From spectral observations, derived results can be obtained by computer models that predict soil properties from one or more spectra. The `observation_derived` entity thus refers to a specific model and a physio-chemical observation (the target of the model). Each model is identified by the URI of a concrete tag or token at a code forge. The `result_derived` entity represents actual derived results, that beyond referring to a derived observation must also refer to the spectral result on which the model was applied.

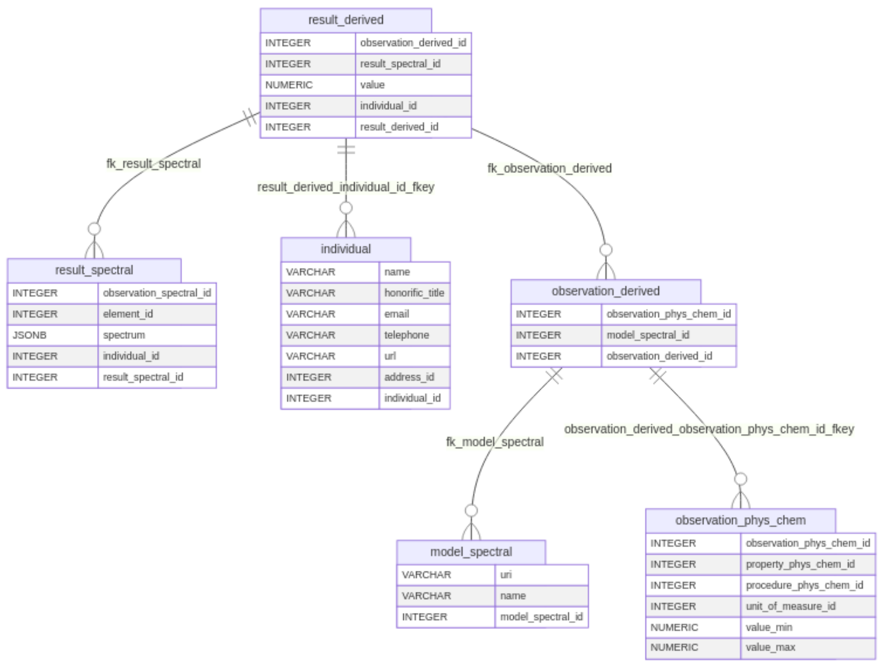

Figure 10: Entity-Relationship diagram for the derived observations tables.

7.6 Multi-lingual thesauri

An additional requirement was later introduced for the inclusion of multi-lingual labels for soil properties and soil description values. The database should start by accommodating English, French and Arabic, and later expand to other languages. A large set of new tables was introduced to implement this requirement. A new table named `language` inventories supported languages, with a many-to-many relationship created between this table and each of the existing thesauri. The diagram in Figure 11 shows this relationship for the descriptive results associated with the Element feature of interest. Note how the label column moved from the `thesaurus_desc_element` table, allowing for a different label per language. Figure 11 also presents the same pattern for the `property_desc_element` table. All tables with the suffix `_label` serve this same purpose for the remaining thesauri.

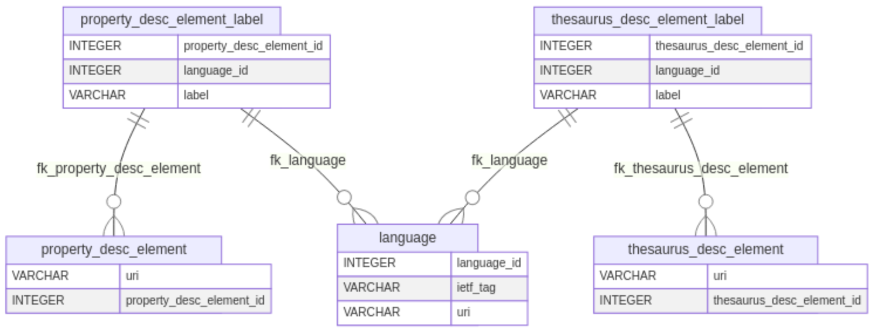

Figure 11: Entity-Relationship diagram with multi-lingual table for the Element descriptive results thesaurus and properties.

7.7 Plot hierarchy

As field work went underway, surveyors started reporting specific attributes for plots from which no soil material was collected. These are plots that were catalogued for field work and eventually visited by surveyors, but for some reason actual sampling or boring was not possible. The extra attributes present information on these reasons. Plots visited but not sampled are known internally in the project as "invalid".

Invalid plots thus have their own attributes and may not be referenced as features of interest from result tables (as no analysis or description are preformed). No profiles may be assigned to the plot either. This forces a distinction between invalid and valid plots with specific tables (`plot_valid` and `plot_invalid`). And since both types of plots retain similar information, such as location or code, a parent table was kept in the model (`plot`). Figure 12 presents this new relationship. A set of thesauri was added to the model to host the controlled content relative to the invalid plot attributes. These are the tables prefixed with `thesaurus_plot_rejection_`, also supporting multi-lingual content.

8 Operational aspects

8.1 Migrations

Both databases covered in this manuscript are developed with the migrations framework from the Graphile project [8]. Migrations provide a versioning mechanism for the incremental development of a database. The tool facilitates rolling back and forth between different points in the development history, by applying or suppressing the SQL instructions that create the database structure.

[8] https://github.com/graphile/migrate

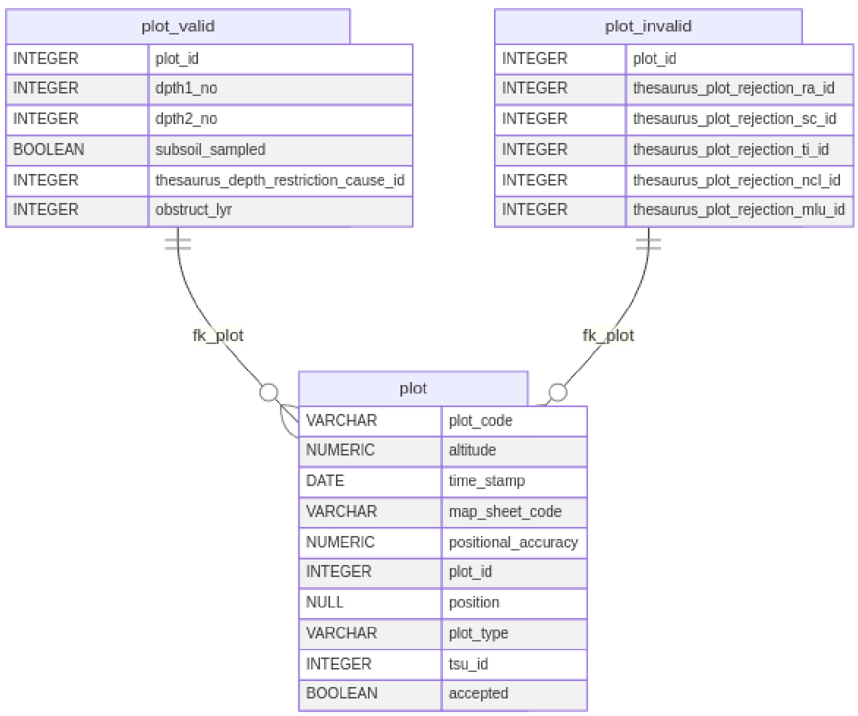

Figure 12: Entity-Relationship diagram for the plot hierarchy.

Listing 5 provides a simple example on how to deploy the Soils4Africa database with `graphile-migrate`. First the repository is cloned from the code forge, then the `.env` file must be edited to point environmental variables to the correct database cluster. Finally the environmental variables are loaded to the session and the migrations are run.

Listing 5 Simple instruction set to deploy a new instance of the Soils4Africa database.

```
git clone git@git.wur.nl:isric/soils4africa/database.git

cd database

vim .env

source .env

yarn graphile-migrate migrate
```

Further documentation on the use of `graphile-migrate` is beyond the scope of this manuscript. The `README` file for both the Soils4Africa database, as well as for the generic `iso28258` database include basic instructions.

8.2 Incremental dependency

The Soils4Africa database is not a fork of the generic `iso28258` database. It is rather a set of migrations that specialise the latter for the context of the project. In practice when a new instance of the Soils4Africa database is deployed, the complete `iso28258` database is created. The migrations then incrementally apply the specialisations described in Section 7.

This dependency is defined in the migrations configuration file `.gmrc`, specifically in the `"afterReset"` section. It instructs the tool to run the SQL scripts in the `migrations/setup` folder during setup. One of these scripts is named `iso28258_v1.1.sql`, with the prefix `v1.1` referring to the particular version of the `iso28258` database repository. This number evolves as new versions of the database are released.

8.3 Integrating modifications to the ISO 28258 data model

Any modifications to the entities described in Section 3 through Section 6 that are not specific to the Soils4Africa project should be conducted in the `iso28258` repository. For instance, the removal of the `surface` table is an action specific to Soils4Africa, and thus was applied in that repository. However a general modification, e.g. changing a column type, must be applied in the `iso28258` repository. The latter can also evolve simply with the correction of bugs or the introduction of general requirements, as is the case with the `metadata` schema.

In such cases the modification will result in a new `iso28258` setup script. The steps to produce one are the following:

1. Apply the necessary modifications with a new migration in the `iso28258` repository.
2. Create a new tag in the `iso28258` repository, marking a new release.
3. Obtain a backup from the `iso28258` repository with the `pg_dump` tool, marking the resulting file with the version (e.g. `_v1.1`). This dump must include only the `core` and `metadata` schemas, ignore ownership and add data as `INSERT` instructions, as Listing 6 exemplifies.
4. Replace the setup script in the `migrations/setup` folder of the Soils4Africa repository.
5. Update the `.gmrc` file in the Soils4Africa repository to load the new script (`"afterReset"` section).
6. Fully re-run migrations in the Soils4Africa database (`reset` parameter).

Listing 6 Dumping relevant schemas from the iso28258 database as set-up for a derived database.

```
pg_dump iso28258 --inserts --no-owner -n core -n metadata > iso28258_v1.1.sql
```

9 Future work

9.1 Uncertainties

The estimation and recording of measurement uncertainties has been a topic of active research at ISRIC. However, this aspect has so far been left outside soil ontology initiatives. The absence of uncertainty elements in O&M itself is a contributing factor to this state of affairs.

Measurement uncertainty is likely associated with the procedures employed. Particularly those used in the laboratory, but so too in field observations. Beyond those, spectral models also carry intrinsic estimate uncertainties that can be relevant to store.

In the data model, uncertainties are expected to be primarilly associated with entities implementing Procedure-type classes of O&M. However, other aspects may also warrant this kind of information. Positioning uncertainty is an example. Results themselves may be subject to uncertainties too, for instance to convey particular conditions to field work. This area of soil ontology likely requires further refinement in requirements before it can be introduced to the data model.

9.2 JSON-B

The recording of spectral results as a database blob in the JSON-B format is an elegant approach to a kind of data for which the relational model was not initially concieved (targetted at alfa-numerical data). However it poses important challanges that need to be acknowledged.

Whereas descriptive results are controlled through the structural integrity mechanisms associated with thesauri, and numeric results by the admissable intervals declared in the respective observation, database blobs are largely uncontrolled. It is not trivial to guarantee that two spectra referring to the same observation contain the exact same number of discrete spectral measurements. Much less that in fact both match the same segment of the electro-magnetic spectrum.

Controlling JSON-B records likely requires a business rules layer. This layer could be part of the data model as stored procedures. The 2017 edition of the SQL ISO standard added support for JSON (note, not JSON-B), but with a limited set of functionality ("Information technology — Database languages — SQL Technical Reports — Part 6: SQL support for JavaScript Object Notation (JSON)" 2017). Therefore the creation of a business rules layer in the data model entails the selection of a specific technology. Postgres provides an interesting range of functionality on JSON-B but it is not clear it suffices.

While the further specialisation of the Soils4Africa data model into a vendor locked implementation is possibly a defendable course of action, a more conservative approach might be advisable regarding the generic ISO 28258 data model.

9.3 Maintaining the parental link to ISO 28258

Once the Soils4Africa data model starts being used effectively, storing actual results records, the integration procedure with the parent data model, described in Section 8.3, becomes less evident. Some modifications may be straightforward to apply in the same fashion, e.g. adding a new, nullable column to an existing table. But others may imply modifications to existing records, that do not fit within the existing procedure.

Two approaches are possible to this issue. First is to devise an additional strategy to apply changes from the parent model without loss or corruption data. This likely requires an additional backup and restore mechanism. The other approach is to simply let the two data models diverge. While not optimal, this latter approach is by far the cheapest, and if the parent model is not expected to evolve much further, it might not have much consequence.

Bibliography

Baker, Thomas. 2005. "Maintaining Dublin Core as a Semantic Web Vocabulary." In *From Integrated Publication and Information Systems to Information and*

Knowledge Environments, 61–68. Springer.

"Geographic information – Observations and measurements." 2011. Standard. Vol. 2011. Geneva, CH: International Organization for Standardization.

"geographic information — metadata." 2014. Standard. Vol. 2014. geneva, ch: international organization for standardization.

Group, W3C Permanent Identifier Community. 2022. "Permanent Identifiers for the Web." 2022. https://w3id.org/.

Hallam-Baker, P. 2022. "Binary Encodings for JavaScript Object Notation: JSON-b, JSON-c, JSON-d." *Draft-Hallambakerjsonbcd-13 (Work in Progress)*. https://www.ietf.org/archive/id/draft-hallambaker-jsonbcd-23.html.

Iannella, Renato, and James McKinney. 2014. "vCard Ontology-for Describing People and Organizations." *W3C Group Note NOTE-Vcard-Rdf-20140522*.

"Information technology — Database languages — SQL Technical Reports — Part 6: SQL support for JavaScript Object Notation (JSON)." 2017. Standard. Vol. 2017. Geneva, CH: International Organization for Standardization.

Jahn, R, HP Blume, VB Asio, O Spaargaren, and P Schad. 2006. *Guidelines for Soil Description*. FAO.

Janowicz, Krzysztof, Armin Haller, Simon JD Cox, Danh Le Phuoc, and Maxime Lefrançois. 2019. "SOSA: A Lightweight Ontology for Sensors, Observations, Samples, and Actuators." *Journal of Web Semantics* 56: 1–10.

Leenaars, Johan, Ad Van Oostrum, and Maria Ruiperez Gonzalez. 2014. "Africa Soil Profiles Database - Version 1.2." ISRIC - World Soil Information; https://www.isric.org/documents/document-type/isric-report-201401-africa-soil-profiles-database-version-12.

Miles, Alistair, and Sean Bechhofer. 2009. "SKOS Simple Knowledge Organization System Reference." *W3C Recommendation*.

"OGC Soil Data Interoperability Experiment." 2016. Engineering Report. Vol. 2016. Open Geospatial Consortium.

Řezník, T., and K. Schleidt. 2020. "Data Model Development for the Global Soil Information System (GloSIS)." GSP - Global Soil Parnership.

Soil, INSPIRE Thematic Working Group. 2013. "D2.8.III.3 INSPIRE Data Specification on Soil – Draft Guidelines." Standard. European Commission Joint Research Centre.

"Soil quality – Digital exchange of soil-related data." 2013. Standard. Vol. 2013. Geneva, CH: International Organization for Standardization.

Standardization Organization), ISO (International. 2002. "Soil Quality—Sampling—Part 1: Guidance on the Design of Sampling Programmes. ISO 10381-1: 2002." ISO Geneva.