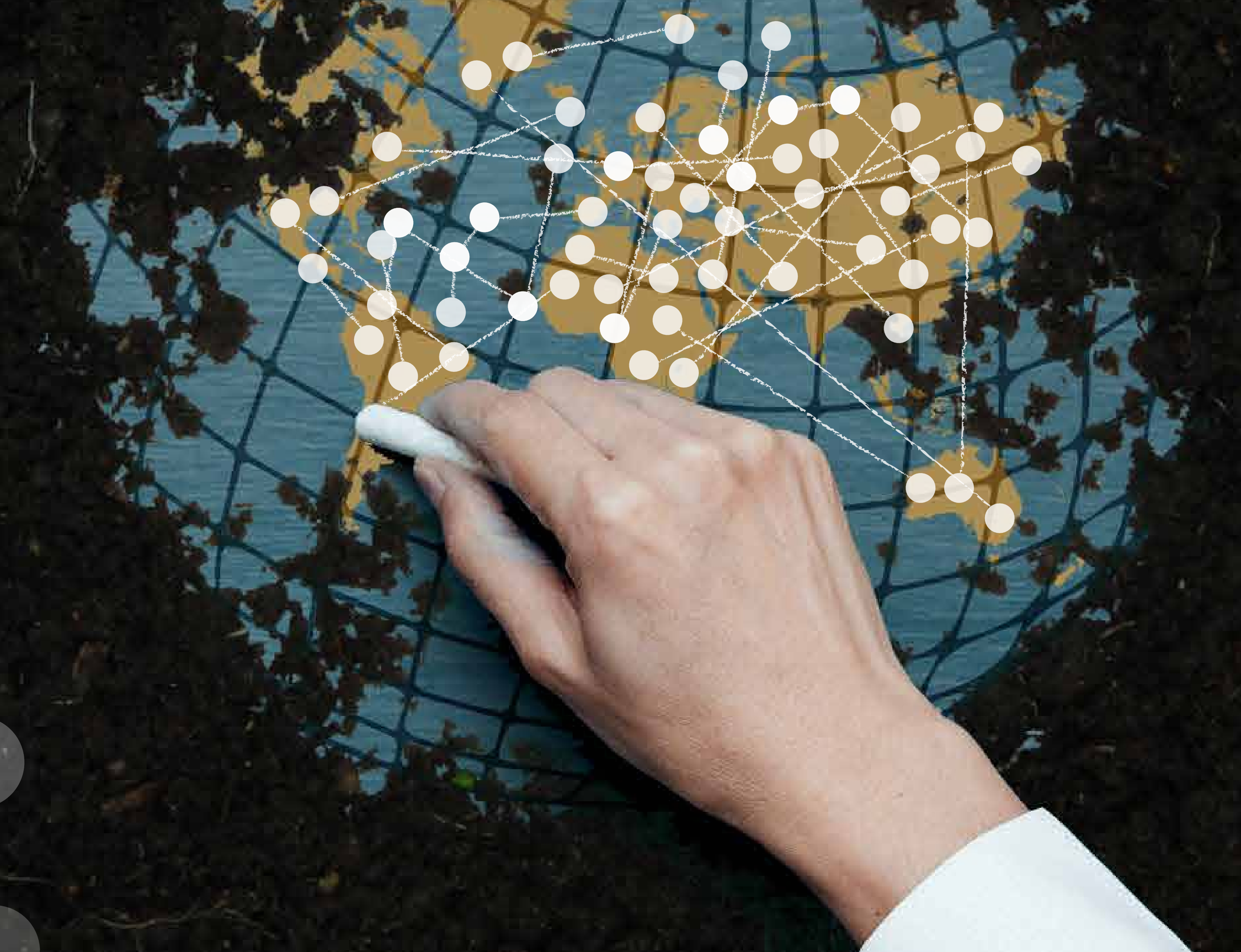
Food and Agriculture Organization of the United Nations

ISRIC World Soil Information

Conceptual design of the Global Soil Information System infrastructure



CSIRO

European Environment Agency

Manaaki Whenua Landcare Research

Global Soil Partnership

Conceptual design of the Global Soil Information System infrastructure

Authors

Luís de Sousa, Bas Kempen, Jorge Mendes de Jesus, Yusuf Yigini, Konstantin Viatkin, David Medyckyj-Scott, Alistair Richie, Peter Wilson, Fenny van Egmond, Rainer Baritz

Contributors

GSP Pillar 4 Working Group

International Network of Soil Information Institutions (INSII)

Published by

the Food and Agriculture Organization of the United Nations

and

International Soil Reference and Information Centre

Required citation:
de Sousa, L., Kempen, B., Mendes de Jesus, J., Yigini, Y., Viatkin, K., Medyckyj-Scott, D. Richie, A., Wilson, P. van Egmond, F. and Baritz, R. 2021. *Conceptual design of the Global Soil Information System infrastructure*. Rome, FAO and Wageningen, Netherlands, ISRIC.

Preface

This document outlines the overall design of the Global Soil Information System (GloSIS) infrastructure at higher level. It explains the architectural and engineering building blocks of the system, presents a number of implementing units, and enumerates some of the technologies on which it may depend. The broad aim is to have an implementation that is lightweight, cheap and easy to deploy by data holders, while at the same time relieving dataproviders from technical details.

Conceptual design of the Global Soil Information System infrastructure was prepared by technical experts from the GSP Soil Data Facility (ISRIC), GSP Secretariat and Pillar 4 WorkingGroup and endorsed by the International network of Soil Information Institutions (INSII) at the fifth INSII Working Session in 2019.

Contents

Figures

Table

1. Background

1.1 The Pillar 4 Implementation Plan

The Global Soil Partnership (GSP) unites stakeholders from policy, research and management to improve theprotection of the limited soil resources, and to promote sustainable soil management. The GSP organises its activities under five pillars of action. For each Pillar, a Plan of Action was drafted, outlining the aims and ambitions of the pillar that was adopted by the GSP Plenary Assembly. Pillar 4 ("Information and data") aims to enhance the quality of soil data and information in order to organise, empower and facilitate soil knowledge,data and impact to address the current environmental and societal challenges. The Pillar 4 Plan of Action addresses this aim through the development of an "enduring and authoritative global system to monitor and forecast the condition of the Earth's soil resources". The plan of action was translated in a Pillar 4Implementation Plan that specifies how the recommendations from the plan of action can be realized. The activities of Pillar 4 are directed towards establishing a Global Soil Information System (GloSIS).

The Pillar 4 Plan of Action identifies three types of soil data products to be distributed through GloSIS:

1. Soil profile and point data,
2. Global polygon coverage,
3. Global grids.

The Pillar 4 Implementation Plan further details these data products (see Chapter 2 'Information System Development'). It proposes a two-tier model for the soil profile and point data. Tier 1 as a 'comprehensive federated database' with minimal data requirement and no stringent requirements on data quality or representativeness. Its objective is to provide access to as many digital profile data as possible. Tier 2 can beregarded a subset of the Tier 1 database, as a 'world reference database' with well-described and analysed soil profiles with thematic subsets of minimal required data, harmonized and quality-assessed. The Global Polygon Coverage is foreseen as an updated and improved version of the FAO/UNESCO Soil Map of the World,building on national soil polygon maps and pragmatic mapping methods such as developed under the EU FP7e-SOTER project. Three types of global gridded products are identified: i) an updated Harmonized World Soil Database, which is a global raster combined with a large soil class and property database ii) fine resolution soil grids version 0, which will be a collection of existing grids (1 km resolution) with no attempt at harmonization, iii) fine resolution soil grids version 1, which will be a collection of consistent, standardized grids of soil properties with global coverage at high resolution (100 – 250 m). These datasets represent the basic soil information: described and analysed soil properties for a certain soil depth and available in differentdigital formats.

The recent Global Soil Organic Carbon map was the first gridded soil map product produced for GloSIS througha participatory, country-based approach at 1 km spatial resolution following a semi-standardized approach based mapping guidelines (FAO, 2018; Pillar 4 Working 2017). Besides the above-mentioned basic soil data products, the Pillar 4 Implementation Plan outlines the plans for developing a system for monitoring, forecasting and status reporting of the soil resource called 'SoilSTAT', using soil indicators, as rasterized spatialassessments and derived statistics This system may use the GloSIS soil data and other national soil data about indicators as input. A capacity development programme to support soil information institutes with the development of GloSIS data products and services concludes the implementation plan. Figure 1 provides a graphic overview of the various GloSIS components.

Underlying GloSIS is a spatial data infrastructure that connects the various GloSIS components and provides the tools to store, process and share the spatial data hosted in GloSIS. The infrastructure is aimed at the exchange and retrieval of soil information collected by (national) soil information institutes through a web-based platform. This is aimed to be achieved in a decentralised manner, with source institutions largely retaining their data and controlling outside access to their products. The broad vision is that of a federation ofsoil information systems (SIS), and interoperable data sets. Ideally, countries and other data providers have their own soil information systems with data storage and data access. Through GloSIS these systems will 'communicate' (exchange) data in a commonly recognisable way. This federated system of the GSP and its supporting capacity development programme will empower countries and other data providers to develop theirfacilitating a standardised soil data infrastructure and easy querying and exchange of soil data at national, institutional, regional and international scales.

Figure 1. Overview of the GloSIS components based on the Pillar 4 Implementation Plan

There are examples of federated systems (or platforms) for the distribution of scientific data, tool integration and report generation that will be taken as examples of best practices. The technical development of these systems generally followed three mantras: simplicity, transparency and continuous development. Examples ofother community-driven projects that proved successful from which this proposal draws inspiration are amongothers the Global Biodiversity Information Facility, OneGeology, NASA data portal and the Digital Observatoryfor Protected Areas

1.2 Scope

This document provides a proposal for the design of the GloSIS infrastructure at higher level. It considers therequired architectural and engineering building blocks. The architectural building blocks are mostly abstract, setting out structures and formalising knowledge into an information model. The engineering building blocks concern primarily the technologies that realise the structures set out in the architecture. It presents a numberof implementing units and enumerates some of the technologies on which it may depend. The broad aim is tohave an implementation that is lightweight, cheap and easy to deploy by data holders, while at the same timerelieving data providers from technical details.

Note that the soil data that will populate GloSIS and their specifications as well as SoilSTAT will be detailed in other, related documents; this is not part of this document.

Chapter 2 presents the general design of GloSIS, lists its building blocks and outlines how data providers canparticipate in the GloSIS federation. Chapters 3 to 6 describe the building blocks in more detail. Chapter 7 presents strategies to deal with the expected emergence of further data exchange standards for soil data. Chapter 8 discusses the extension of the data exchange standard from point data to other data types such asgrid and polygons. Chapter 9 deals with quality of service and quality of user experience. Finally, the social architecture of GloSIS is addressed in Chapter 10 that highlights the necessity for a framework for system governance.

2. GloSIS design

2.1 Building blocks

GloSIS is envisioned as a federation of soil information systems, which share interoperable soil data sets via web services. Figure 2 gives a simplified, graphical representation of GloSIS. GloSIS connects soil informationusers that are depicted at the right hand side of the figure with soil data providers that are depicted on the left. The **soil information systems** that host and publish the soil data of the data providers are referred to as "**nodes**" in the federation. These nodes could be national (country) systems, regional systems (e.g. the Latin American Soil Information System, SISLAC) or a soil information system of an (inter)national research organization or NGO that wishes to share its soil data. GloSIS will connect users with providers through a single access point: the discovery hub. (Note that each data provider can (continue to) serve its own data through its own portal using its own software and services. The discovery hub is not a replacement for nationalor institutional soil web portals). Exchange of soil data coming from different countries and institutes requiresadoption of common standards for data exchange.

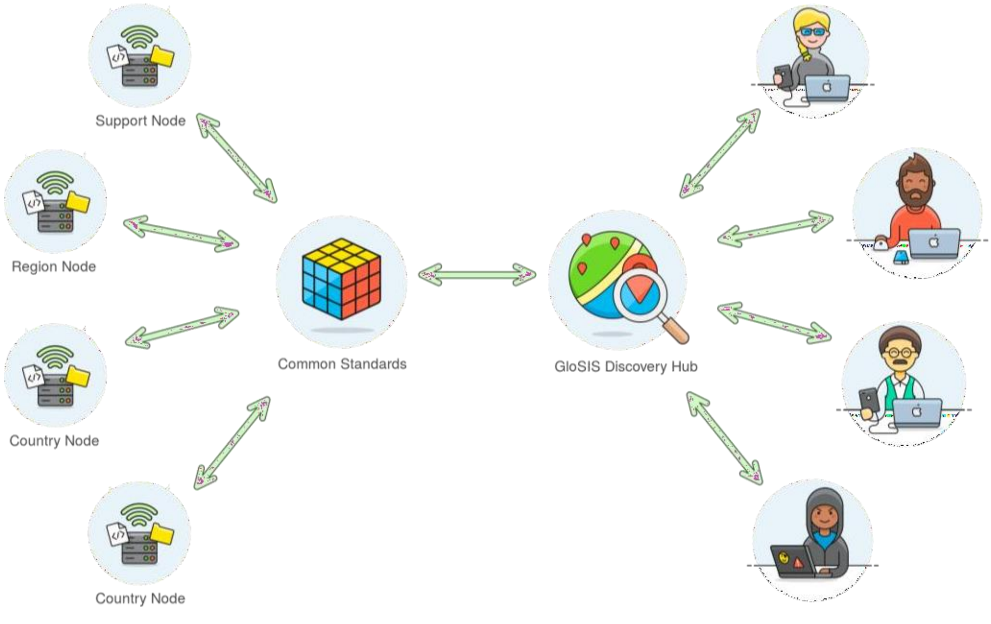


Figure 2. The GloSIS federation that connects soil data users to soil data providers via one single access point

Based on this general design of GloSIS, four main building blocks can be devised. We will provide a brief overview of these components here. More details on each of these are provided in the Chapters 3 to 6 below,laying out the necessary steps needed to bring together different data sources of heterogeneous nature. The components are:

- **Domain model** – Abstract, architectural component that defines how data are organised; it also embodies a common understanding of what soil profile data are. The GloSIS domain model defines thestructure of the data sharing vehicle and determines the structure of the database underlying the reference GloSIS node (see section 2.2).

- **Data Exchange Standards** – Standards for data publication and exchange allow different parties to send and receive soil (point) data through a well-recognised medium. Existing standards such as those issued by the Open Geospatial Consortium can be adopted for GloSIS.

- **GloSIS Node** - A soil information system connected to the internet and able to publish soil data according to the GloSIS data exchange specifications; it can be a server, or cluster of servers, reachablethrough a URL. GloSIS nodes can be existing soil information systems but also newly implemented systems.

  - The **Support Node** is a specialised instance of a GloSIS node hosted and maintained by the GSP. It is intended to host data from institutions that are not able or do not have the ambition to set up their own GloSIS node.

- **Discovery Hub** - A web-based gateway to the GloSIS nodes offering data browsing and discovery functionalities of data within GloSIS. It brings all nodes together, through harvesting of web services, into a single point of access for users. It usually consists of view and catalogue services. Internally, a registry of the available services, allowed terms and their meanings (vocabularies), are stored and maintained.

2.2 Participation levels

There are large differences between data holders in how soil data are stored, managed and disseminated. Some countries maintain full-fledged soil information systems while others have their soil data stored in tableson a local computer or even only on paper.

Implementing a sound soil information system (SIS) requires advanced knowledge and skills, including the implementation and management of relational database systems and the employment of standards for the creation, exchange and publication of geospatial data on the internet. Designing and setting-up a soil information system from scratch could be technically overwhelming for data holders with limited knowledge and skills in setting up (geospatial) data systems.

To support data holders in setting up a (national) soil information system, the **GloSIS template node** is provided. The template node is a deployable infrastructure ready to host and serve all important componentsof a soil information system (see Annex 1 for a description of the architecture of the template node), in compliance with the data exchange standards and services (see Chapters 3 and 4) that will be used in GloSIS.Thus, by deploying a template node, and populating it with data, a data holder acquires a fully functional soilinformation system that can be easily linked to the GloSIS federation. The template node can be a solution for data holders who wish to set up their own soil information system but have limited technical capacity to doso.

Sharing interoperable soil data will not necessarily rely on the implementation of a national or institutional soilinformation system. Besides supporting capacity building, the global data infrastructure will offer componentsand solutions for those partners which do not have their own information systems well-developed yet throughthe support node. The federation-based architecture described here allows soil data holders to choose betweenthree different levels of participation to GloSIS, hereby acknowledging the differences in technical level, technical skills and resources of data holders as well as differences in ambitions that data holders might havefor setting up and maintaining their own SIS:

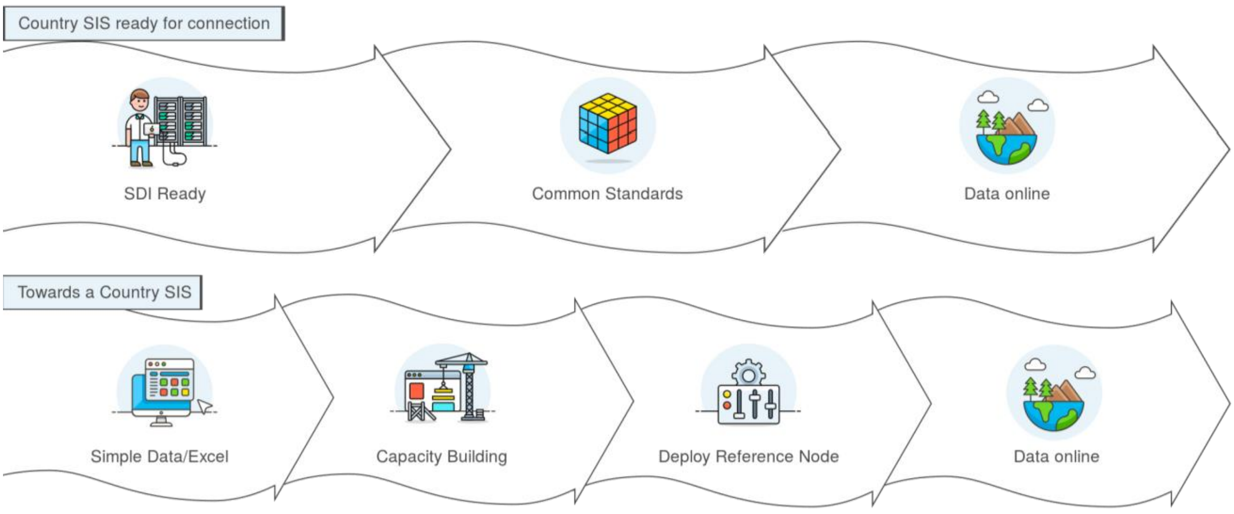
1. **Tailored implementation** – for data holders with an existing SIS. The data holder must implement the GloSIS data exchange specifications (see Chapter 4) in its data services to guarantee full interoperabilitywith the GloSIS federation, but maintains its current SIS data model. Because existing SIS are based on different technologies and structures and store soil data differently, this requires tailor-made solutionsfor mapping the SIS data model to the data exchange model.

2. **Template implementation** – for data holders that wish to setup and populate their own SIS. They implement the GloSIS template node. Data holders opting for this possibility shall primarily be concernedwith the compliance of their data and its loading into the database of the template node. Data exchangeand publication is taken care of by the template node.

3. **Support implementation** – for data holders lacking the resources, knowledge or desire to set up and maintain a SIS. These data holders can submit the soil data they wish to share to the support node. Thesupport node is a GloSIS template node that will be centrally managed. The data will be stored in the database of the support node and published through web services of this node, thus automatically complying with the GloSIS domain model and the data exchange specification.

These three implementations are described in more detail in chapter 5.

Figure 3 shows the three different implementations graphically. The *top* figure shows the 'tailored implementation': data holders that already have and maintain a soil information system implement the GloSISdata exchange specifications in their services, i.e. they bring their data to the common GloSIS standard for data exchange, hereby making it automatically available to general users through the discovery hub. The *middle* figure shows the 'template implementation': data holders with soil data stored in simple databases or plain tables and with the ambition to establish a (national) soil information system will be trained in setting upa template node. Once a template node is deployed the data are served through a national web portal and arediscoverable through the discovery hub if allowed by the data holder. The *bottom* figure shows the 'support implementation': data holders with their soil data stored in simple databases or plain tables standardize theirsoil profile data so that their data can be easily added to the database of the support node. They then send their data to the GSP, who will act as a custodian of the data, who will store the data the support node. The data will be discoverable through the discovery hub if allowed by the data provider.



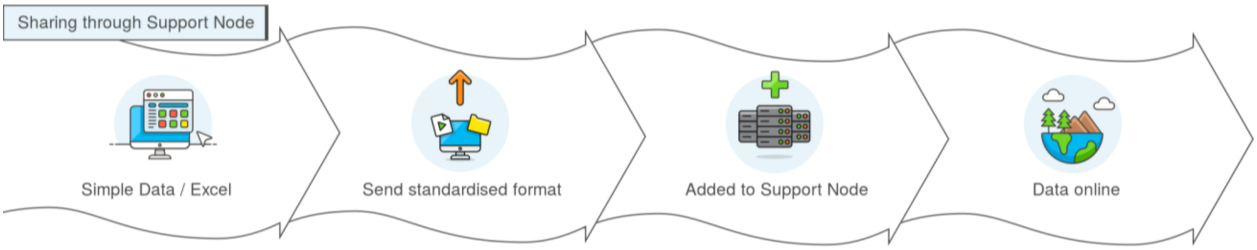


Figure 3. Three different levels of participation to the GloSIS federation

3. Domain model

GloSIS enables users to access distributed soil data. It presents searchable, viewable and retrievable soil dataas defined in the GSP Pillars. This only works if the data providers speak the same 'language', i.e. if providersstructure their soil data in a unified and harmonized way and have a common understanding of logical conceptsof soil information such as the profile or an horizon.

The structure of the data objects served by the GloSIS nodes is defined by a domain model. The domain model is an abstract representation of a thematic domain, in this case soils. It describes how different elementswithin this domain relate to one another and are organised into a common structure. For instance, a soil pedoncan be described by a profile that may consist of a number of horizons that each have a specified number of characteristics. These characteristics can be described and/or measured in a laboratory with *i* methods in *j* units. A domain model therefore organises elements of (soil) data and formalises how these elements relate to one another and to properties of real-world entities such as the pedon. It is the architectural component of GloSIS that guides the engineering components and makes sharing of harmonised data possible through thecreation of a common 'language' among the parties involved (data providers, developers and stakeholders in general). Note that the domain model should not be confused with a *data model*. The domain model can be implemented as data model that determines the logical structure of the contents of a database.

Thus, the domain model sets the rules and requirements for data insertion into the system. A strong and refined domain model enforces data validation by itself. For instance, because a soil horizon must always be part of a soil profile and soil measurements must always be assigned to a soil layer or horizon. For this same reason it also prevents data records from arriving at inconsistent states later on.

In summary, it defines how local soil data (which can be coded, defined and stored in many different ways), are exported into a web service in a unified way. Only by using an agreed domain model an interoperable system can be guaranteed in which data are discoverable and easily retrievable.

An ISO soil domain model exists (ISO 28258), as well as various other national or regional models, such as INSPIRE, SOTER,and ANZSoil. It is an important criterion for GloSIS, that the data shared by soil informationsystems that use already existing models are interoperable with those applying the domain model recommended for GloSIS. GSP Pillar 5 ('Harmonization') will evaluate the various available soil data models and assess their suitability for GloSIS.

Management and descriptions of domain models, concepts (ontologies) and property types used, and analyticalmethods will be done through a set of authoritative web pages (a registry) that once published will remain atthe published location for the life of GloSIS.

4. Data exchange standards

4.1 OGC Standards

To be able to exchange soil data and information contained in the soil databases of the GloSIS nodes with users through the discovery hub (see Figure 2), a 'vehicle' is required that is able to send 'data packages' across the GloSIS federation and then publish the soil data contained in these packages in a GIS of a user forinstance.

The Open Geospatial Consortium (OGC[1]) has issued ISO standards for publication and exchange of geo-spatial data that will be adopted by GloSIS for this purpose. These technical standards describe specifications for describing data to be exchanged, how requests for data are specified and how data are encoded. The standardsare well proven for implementation in many thousands of geospatial systems world-wide. Commonly used standards include ISO 19128 – web feature services (WFS) for data publication, ISO 19142 - web map services(WMS) for map publication, Web Coverage Service Interface Standard (WCS) for coverage (raster) publication and ISO 19109:2015 which defines rules for creating and documenting application schemas built from domainmodels that described how data are organised for sharing. GloSIS will implement the WFS standard for exchange and publication of soil point data.

The WFS standard provides a protocol that defines how spatial features, such as georeferenced soil profile data, are transmitted in an Internet environment between a service provider and a client and the type of payload (the 'data package') being transmitted. The payload can for instance be the result of a user request for specific soil characteristics from the database (a query) or metadata describing the data. This flexibility is important as users can only requests subsets of data that they require and also in alternative formats.

The next section provides a more technical description of two versions of OGC WFS standard, one encoded with GML and one encoded with JSON, that are both candidates for use by GloSIS. Finally, we note here thatdata exchange within GloSIS is governed by the endorsed GSP data policy, of which a summary is provided in Annex 2.

4.2 Web Feature Service (WFS)

4.2.1 WFS 2.0

The Geographic Markup Language (GML) is an XML based grammar defined by the OGC to represent geo-spatial data in polygon data format. GML provides a standard format to store and share geo-spatial data, butit extends far beyond. For the particular case of GloSIS, GML provides a facility called Application Schema forpolygon data such as soil profile data, by which domain-specific data types may be defined to complement thegeo-spatial data. In practice, an application schema allows a GML document to contain not only the geo-spatialinformation of a particular feature, but also structured domain specific data. For example, with the correct application schema, the GML file of a certain soil profile can contain also its various horizons, the properties ofthose horizons and the methods used to measure these, all in a structured way. Examples of implemented application schemas are provided by INSPIRE and ISO 28258.

Version 2.0 of the Web Feature Service (WFS) standardises the publication of geo-spatial data with GML, thusencompassing application schema. This standard extends the concept of a spatial object with that of complexfeature. A complex feature is a spatial object that beyond its own attributes, may include attributes from

[1] The Open Geospatial Consortium (OGC), an international voluntary consensus standards organization, originated in 1994. In the OGC, more than 500 commercial, governmental, non-profit and research organizations worldwide collaborate in a consensus process encouraging development and implementation of open standards for geospatial content and services, sensor web and Internet of Things, GIS data processing and data sharing.

related objects (typically the end points of one-to-may relationships in a relational database). For instance, aquery for a particular soil profile may obtain all related horizons with a complex feature (beyond the profile attributes). The application schema in WFS 2.0 defines the structure of these complex features, i.e. it determines that a profile is composed by a set of horizons, and therefore when it is queried, all its horizons must be included in the response within a complex feature. In a WFS 2.0 compliant programme like GeoServer,the user is able to map directly the tables and columns of a relational database to the structure of an applicationschema. WFS 2.0 is therefore a clear option for the interchange of domain specific objects related to geo- spatial features such as soil data.

Many tools and clients support the geospatial standards such as a WFS resource bearing a specific applicationschema, making it possible to access data held in third party data nodes, download the data and use the soil data in tools like QGIS Desktop, R and Python. For instance, the French Geological Survey (BRGM) provides aplug-in for QGIS that translates between GML application schema files, thus allowing for data to be imported into databases like PostGIS or SQLite and vice-versa.

4.2.2 WFS 3.0

In recent years the OGC has fostered work towards a new version of the WFS standard (WFS 3.0). Technical development is by now largely complete; in late 2018 the OGC ran the necessary public consultation on the new version proposal, with final endorsement expected during 2019. One of the main changes introduced by this new version is the detachment from XML: data exchanged with WFS 3.0 can also be encoded with JSON and HTML. JSON stands for "JavaScript Object Notation", it is a simple data interchange format that began asa data notation for the world wide web and in the linked data communities. However, it has proven convenientand powerful enough to be used in many other contexts. Its simplicity and ease of read by humans has madeit more popular than XML among computer scientists.

By detaching itself from XML, WFS 3.0 also distances itself from GML and the Application Schema facility; in this new version, no specific domain specific object encoding is endorsed. However, in what respects JSON, asimilar facility exists. JSON Schema is meta-syntax for JSON, originally conceived as a formalism to validate the content of JSON documents. Just as the GML Application Schema, JSON schema allows the definition of how a document is structured: which objects it admits, the content of those objects and the relationship between them. It too can therefore be used to interchange complex object structures associated with geo- spatial features.

WFS 3.0 with JSON Schema and JSON-LD (JSON linked data) is a novel alternative to the established (and aging) WFS 2.0 plus GML Application Schema. Both are worthy alternatives to implement the GloSIS domain model and realise the overall GloSIS vision. Tables 1 and 2 list the advantages and disadvantages of both versions of the WFS standard. Despite the fact that WFS 3.0 is not yet a fully matured standard and that implementation tools are lacking, the advantages of WFS 3.0 over WFS 2.0 are such that data exchange in GloSIS will be based on WFS 3.0, but it is important to note that WFS 3.0 and WFS 2.0 are not mutually exclusive. A WFS 3.0 interface can be mapped to a WFS 2.0 implementation, which facilitates compatibility for soil information systems with WFS 2.0 implementations.

Table 1. Advantages and disadvantages of the WFS 2.0 standard

WFS 2.0 - Advantages	WFS 2.0 – Disadvantages
Implemented by multiple client software	XML is heavy and resource consuming. It might be hard to transfer MBs of XML files when with a bad internet connection and/or in remote locations.
Usage of XML and implementation of complex data relationships	Complicated development implanting multiple XML standards (Observations & Measurements, GML).
Possibility of app-schema usage: meaning SoilML-like structures	SoilML development and application schema are technically "very thick", requiring special tools (e.g. Enterprise Architect) and high levels of technical know-how. This also makes maintenance harder.
Possibility to download soil data using GML to QGIS	Web services responding with XML are not easy to use in website development, neither in programming environments requesting data (R, Python, etc.).
Replication of GloSIS database into the user's computer	Application schema community is avoiding this sort of developments at the moment.

Table 2. Advantages and disadvantages of the WFS 3.0 standard

WFS 3.0 - Advantages	WFS 3.0 – Disadvantages
Lightweight standard based on JSON with possibility of implementing linked data with JSON-LD, easy to use and non-demanding on resources for data transfer an management.	Still under appreciation by OCG - should become a standard in 2019.
Based on REST, the same technology used to develop APIs and websites. Facilitates the development of viewer applications to GloSIS, and its integration with computing languages like R or Julia.	Requires *ad-hoc* implementation by the P4WG; no readily available tools for implementation.
By being based on REST services, WFS 3.0 facilitates data search and indexation by Google.	Off-the-shelf client software not yet available.

More convenient for website development and mobile app development since data are organised using URIs and using as default JSON as payload.
JSON schema allows the definition of domain-specific document structures. This is similar to the GML Application Schema, but considerably lighter and easy to use.
Linked data is a novel facility that can be used with JSON to simplify data discovery.
Development dispenses heavy-weight software.
Internet oriented (WFS 2.0 is data oriented).

4.2.3 Updates to data exchange standards

Data exchange standards, such as WFS, are in constant evolution, if for anything else, at least to follow on thedevelopments of the WWW and its core protocols (e.g. HTTP, SOAP, JSON). The evolution, and even succession, of standards should therefore be taken as inevitable. How future standards impact GloSIS largelydepends on the direction taken by the GSP consortium. A new data exchange standard may not be deemed relevant enough for implementation, thus in such case no changes to the GloSIS implementation are made. Ifon the other hand a new standard is considered worth of implementation, then the components responsible for data publication and data search must be updated. Other components, focused on data management or mapping, should not be subject to change when a new data exchange standard is adopted.

It is important to realise that irrespective of the direction it takes, GloSIS will require constant technical oversight to deal with circumstances like this (see Chapter 10). Informational technologies are still developingat neck breaking speeds, it is easy for a system to become outdated or obsolete.

5. GloSIS Nodes

National, regional or institutional soil information systems that are part of the GloSIS federation are referredto as 'GloSIS nodes'. A soil information system is 'being part of GloSIS' when it provides an interoperable web service to share and publish its data (i.e. applying the GloSIS data exchange standards to export local soil data to a web service). In chapter 2, three different implementations to participate in the GloSIS federation were outlined and the GloSIS template node was introduced as an easy solution to deploy a soil information system for data holders with limited technical capabilities. Here these implementations are described in more detail.

5.1 Tailored implementation

This implementation applies to data providers that already have a soil information system established and wishto share (part of) their data through GloSIS.

The paradigm of data exchange using OGC standards has the benefit of minimizing the effort required for an existing soil information system to join the GloSIS federation compared to other options. There will be data providers with sophisticated SIS that possibly wish to make their data available through GloSIS. By adheringto existing standards, GloSIS guarantees that joining the federation does *not* imply the re-programming or re-configuration of these SDIs, i.e. data providers do not have to run a parallel GloSIS-compliant system besides their own SIS. The approach to be taken by such data providers largely depends on whether the technologies they use natively support the standards elected to network the federation (e.g. WFS).

In the case where the data provider already runs standards enabled technology, inclusion to the GloSIS federation can be reasonably simple. All that is required is the set-up or extension of data services feeding from the existing data stores (e.g. database tables and views) to comply with the standard(s) used by GloSIS. At an abstract level, this corresponds to a translation of the data provider's data model to the GloSIS domain model.

For data providers that do not possess standards enabled technologies a few more steps are required. An approach is to add an OGC-compliant feature server, like GeoServer, to the existing SDI stack. With such server in place it is then a matter of linking to the appropriate data stores and implement the standard(s) usedby GloSIS. A further alternative is to deploy the template node and then synchronize the reference database with the data provider's data store(s). Many off-the-shelf database management system technologies providesynchronization mechanisms natively, as is the case with PostgreSQL and MariaDB. There are also reliable third-party tools specialised in database synchronisation, e.g. SymmetricDS.

5.2 Template implementation

The 'template implementation' is an implementation designed for soil data providers that do not have set up their own soil information system but have the ambition to do so.

The reference implementation serves as cheap and fast vehicle to set up a SIS that can serve as a GloSIS node. The reference implementation contains tools for the upload and management of data in a database structure respecting the GloSIS domain model, and then automatically publishing its contents complying withthe GloSIS data exchange specification. These tools are combined in a software bundle that a data provider can install on a server. The software and technologies included in the bundle ensure that the installed SIS is compliant with GloSIS standards. This means that after installing the reference implementation, a data provider is 'GloSIS ready'. The GloSIS template node could also be used as a stand-alone national or institutional soil information system.

The architecture of the reference implementation, including software solutions, is described in more detail in Annex 1 of this document.

5.3 Support implementation

To complete this federative structure a support node is necessary to host data not served elsewhere. The support node is referred to as the "centralized data repository" in the Pillar 4 Implementation Plan. Technically,the support node is a GloSIS template node deployed to, and managed by, the GSP. Its goal is to host soil profile data from institutions lacking either the desire, capacity or the resources to set up their own GloSIS node. A protocol and template shall by established determining how such institutions may submit their data tothe GSP, in order to be loaded into the support node and thus automatically joining the GloSIS federation.

6. Discovery Hub

The infrastructure building block bringing together this proposal for a GloSIS implementation is the DiscoveryHub. It performs two essential functions: (i) a search engine of soil data, and (ii) a registrar of all the nodes compliant with the GloSIS data exchange specifications. The discovery hub is a web-based access point to the GloSIS node federation, the go-to place for anyone seeking soil (profile) data. It must offer simple ways of data search across the whole federation, visualization of data and the mechanisms for regular nodes to register themselves and also a single point where search engines can index data. An existing examples of suchdiscovery hub is for instance Pangaea.

6.1 Soil data search engine

The data browsing, or data discovery aspect of the discovery hub will provide users with ways to search the federation by geo-spatial location, by meta-data e.g. provider, date or by data fields (possibly using Elastic Search). Such user queries may therefore trigger requests to all nodes of the federation or just a few or only one. This can imply some technical challenges. However, the reliance on well-established standards such ISO19115/ISO19119 Application Profile for the OGC Catalogue Service for the Web (CSW) 2.0 can help simplifying this functionality.

6.2 Node registrar

The node registrar is the other component of the Discovery hub that enables the search functionality, bringingtogether the different data providers. It is, in essence, a catalogue of nodes that are either a reference GloSISi mplementation deployment or are otherwise able to publish soil profile data according to the GloSIS domain model and application schema.

Any soil data provider should be able to propose a node to join the GloSIS federation, e.g., by filling a specific form with contact details and services addresses; the proposed node then enters a registration queue. The registration is complete only after a series of automatic verifications are performed, guaranteeing that the node indeed complies with GloSIS and adheres to the data exchange specifications. For this verification, the registrar issues probing requests that on the one hand gather information about the node, and on the other validate its GloSIS compliance. These requests could include: a list of available datasets, a list of profiles forwhich a certain variable has been observed, list of horizons in a certain profile, etc.

The Discovery Hub must also guarantee that each survey, soil profile, horizon and observation can be correctlyidentified in the federation, including its respective origin. This is achieved using universally unique identifiers(UUIDs) or Digital Object Identifiers (DOI) as keys to such objects. This way it is guaranteed that two differentobjects stored at different nodes are not assigned the same identifier. GloSIS will rely on UUIDs.

A further step in the node registration is the verification and validation of meta-data services. As outlined in Chapter 5, a GloSIS node must be able to provide a CSW supporting the base data WFS. The registrar must thus validate this service and, if necessary, regularly harvest its contents to expedite the federated search.

This process of verification/validation must be fully automated, taking place without any human intervention.At the end, the node is discarded from the registration queue, being either accepted into the federation or rejected. In the latter case a validation report may be issued to the proponent of the node to facilitate the correction of any problems. Registered nodes must be regularly submitted to this verification process, to guarantee their continuous compliance with the GloSIS schema. The frequency and detail of this regular verification is to be defined by the consortium.

6.3 Universally Unique Identifiers (UUID) and Universal Resource Identifier (URI) policy

Once the federation is in place the pool of data sets and data objects available for search is likely to expand rapidly. Correctly identifying the origin and meta-data of a particular object can thus become challenge. The identification of objects (e.g. a soil profile) across different institutions is an additional problem. For instance,if two different institutions use a simple numerical sequence to identify its profiles and horizons, most likely they will have overlapping identifiers. In computer science this issue is referred as a collision of records.

Universally Unique Identifiers (UUIDs) offer a mechanism to deal with the situations above, providing a well- known identifier format that is valid across different systems. The UUID concept is standardised by the Open Software Foundation (OSF) and is been included in various other international standards. In its latest form (version 4) a UUID is a collections of 36 digits organised in five groups, e.g.: 6edd7d2a-04d4-4dee-a609- 338a21c12e66. Since UUIDs are so long and generated randomly, it is virtually impossible for two separate systems to generate the same UUID.

Unique identifiers are a key component of a functional data federation. With each object must be mandatorilyidentified by a mechanisms like the UUID, guarantees one the one hand that no record collisions may occur, and that each object in the federation can swiftly and seamlessly located.

The UUID can be combined with the Universal Resource Identifier (URI) to unequivocally locate any data resource or object (e.g. soil profile) in the federation. Proper URI implementation also facilitates search engineintegration and the creation of linked data between different data objects. For example, a soil profile stored inISRIC and another at FAO could be locate through URIs as follows (the last 36 digits are the UUID):

glosis.org/isric/soil_profile/6edd7d2a-04d4-4dee-a609-

338a21c12e66 glosis.org/fao/soil_profile/fe719ac1-d181-47f3-

8e80-0117ca5eeb03

UUID and URI creation should follow the W3C guidelines.

7. Adoption of future standards

As laid out, this vision is able to fulfil the goals of GSP Pillar 4 by relying solely on existing standards. However,the implementation hereby proposed will accept the results of Pillar 5 and shall integrate the data exchange, such as for instance new version of the SoilML exchange schema (currently under development) being put forward by that pillar.

The strategy to accommodate a new standard (Pillar 5) depends on its form: whether it can be integrated withWFS or not. If it can, then the procedure is in all similar to the one described in Chapter 4, with the implementation of a domain model as an application schema, for instance as a ReST service. In that case a domain model transformation might be required, since the domain model of the new standard is unlikely to perfectly align with the one used in the initial development of GloSIS. But because such a transformation is not technically challenging, it can, for example, be implemented with database views. A new standard of this kind might eventually entirely supersede the original domain model and application schema, if stakeholders soagree.

The integration described above might not be possible if the new standard forces a data transfer protocol aliento WFS (the latter is based on SOAP and ReST) or a markedly different service API. In such case the adoptionof a new standard will require the development of additional software, compliant with the specified data transferprotocol. The effort required largely depends on the specifications of the standard itself. However, it is important to stress that even in such a case the general structure of the GloSIS implementation proposed here should not change in consequence of a new data exchange standard.

8. Extension to other data types

This document primarily concerns the exchange and publication of soil point data in the GloSIS federation. GloSIS however, is not limited to point data only, it also considers other data types (see Figure 1). This chapter briefly discusses extension of data exchange to other data types. We emphasize here that this concernsdata exchange only, GloSIS template node implementation will have functionality to serve raster or polygon datasets (see Annex 1).

8.1 Raster data

Compared to point data, soil data rasters (or grids) are less challenging from an architectural point of view, but pose different hurdles in terms of implementation. Raster maps are very simple elements in information architecture, since each map reports to a single variable (e.g. nitrogen) or enumerate set (e.g. soil class). However, a number of conventions must be enforced to guarantee seamless interchange. These concern aspects like units, data format (integer, Boolean, etc.) or the representation of depth. Like WFS, the Web Coverage Service (WCS) standard also provides means to specify application schemas (or profiles) that couldfacilitate this task.

Until there is an agreed upon standard for exchange of raster data, rasters could be collected centrally, for instance in the support node, and then served from there. An approach somewhat similar to this was recentlyadopted to publish and share the GSOC map.

8.2 Polygon data

Polygonal datasets require a technical implementation that is similar to that drafted above for point data (soilprofiles); the challenges rest with the domain model. The concepts around soil profiles are relatively commonsense to soil scientists, e.g., a profile is divided into horizons, observations from each horizon are analysed according to a particular method in order to characterise soil properties. In contrast, polygon data are used toconvey a myriad of different characteristics, ranging from soil classification, to terrain properties, land use andmanagement, etc. Moreover, these data may proceed from dramatically different acquisition and processing methods. It is therefore not straightforward to reach a synthetic domain model for soil polygons, and trying toencompass all these types of data can easily become an unsuccessful effort. Polygons thus require a sound analysis at the architectural level to devise in a pragmatic way the data that should be stored and exchangedby such spatial features.

An alternative, and more pragmatic, approach to polygon data is to regard polygonal maps simply as irregulartopologies. In such case the problem is reduced to that of rasters, with variables and enumerates becoming single value attributes of each polygon feature.

9. Quality of service and quality of user experience

GloSIS is expected to become a global system used by many users from around the world. These users will have expectations about the usability, availability and performance of GloSIS. They will trust the originality and quality of data presented, and follow the specific data access policies. In the following, the quality of service (QoS) and quality of user experience (QoE) are outlined as a reminder of its importance. The implications of establishing a QoS and QoE framework for the implementation will not be described here. Thiswill be considered during a later stage in the development of GloSIS.

QoS is concerned with the technical reliability and performance of a network service or application. It includesconcepts like the availability of the service (e.g. 24/7/365), scalability in terms of number of concurrent usersand number of data requests (also known as capacity), absolute performance (i.e. speed). The QoS will affectthe performance of the GloSIS network for end users and any third-party applications that use GloSIS. QoSis typically measured using metrics like error rates, throughput, availability and delay or data request responsetimes.

A related aspect is quality of user experience (QoE). This is concerned with the users' experience of using theGloSIS service (users being not only end users but also those users who use GloSIS to supply soil data andthose responsible for the operation of GloSIS). QoE also focuses on the usability of the data that is deliveredvia GloSIS to end users or other applications that use GloSIS, for instance using web statistics. It covers notonly the data but other qualitative aspects such as the provision of metadata, statements regarding accuracy and completeness of the data, proper and descriptive naming of properties and methods, a clear description of data lineage and how the data can be used (data policy and data use rights).

If QoS or QoE are ignored, not properly considered or not sufficiently resourced it will affect the uptake and use of GloSIS. If GloSIS or data published via GloSIS is difficult to use (i.e. QoE very low) or GloSIS or nodes in GloSIS are slow or unstable (i.e. QoS very low) then the result will be that few users will use the system.

QoS framework and evaluation criteria (metrics) which GloSIS nodes will have to meet to be part of the GloSIS ecosystem, need to be defined. A stress testing framework will be defined (metrics), implemented andused to benchmark performance and security of nodes that wish to be added to the GloSIS ecosystem (see section 6.2). The QoS framework will also be used to monitor in (preferably) real-time the performance of GloSIS, for example, the availability of GloSIS nodes and their average request response times. How this framework will fit in the GloSIS architecture will be considered at a later stage in the development of GloSIS.

QoE will need to be assured by GSP Pillars 4 and 5 defining the data, metadata and related standards and theusability criteria that will be implemented in GloSIS.

10. System management

Alongside the technical architecture, a complementary **social architecture** must be put in place for GloSIS.A social architecture addresses the social aspects of the development, implementation, operation and maintenance of GloSIS. It has a focus on system governance, facilitating participation of stakeholders and the management of the agreements that together define how GloSIS stakeholders will collaborate to achievetheir goals.

Given the large number of GSP stakeholders and the complexity of the relationships, addressing the social architecture is a significant collective challenge. The majority of resources (i.e. data, technology, people) from which GloSIS will evolve are owned and managed by numerous agencies based in different countries, so normal, centralised management practices, which imply direct control over resources, cannot be used. Thus, effective management of GloSIS is going to be critical to steer the GSP initiative in the right direction, communicate the collective goals and reconcile the different influences on those involved[2].

Thus to ensure long-term sustainability of GloSIS, the system will require a body that is responsible for its daily management to discuss and decide on proposed rules, policies and mandates, adoption of technical standards, guidelines and tools to develop etc. This body will be embedded in the current GSP governance structure. The global and regional Pillar 4 and 5 Working Groups as well as the INSII network are already in place. It must be investigated what the daily management and administration of GloSIS would require, if there is a need to establish additional working groups or committees and the roles of the P4 and P5 WGs andINSII herein.

Frequent communication is critical to the success of GloSIS. This will ensure shared ownership, buy-in and build trust between participants. A communication plan should be developed and implemented in a way that best meets the needs of different stakeholders. Communication mechanisms that ensure rapid feedback duringimplementation will help implementers improve their nodes and enhance GloSIS's strategic decision making.

[2] These influences include a participant's political and policy contexts, goals, drivers, capabilities, resources and aspirations.

References

FAO. 2018. *Soil Organic Carbon Mapping Cookbook* (2nd Edition). Rome.
(also available at: http://www.fao.org/3/i8895en/I8895EN.pdf)

Pillar 4 Working Group. 2017. *GSP Guidelines for sharing national data/information to compile a Global Soil Organic Carbon (GSOC) map*. Rome.
(also available at: http://www.fao.org/3/a-bp164e.pdf)

Annex 1. Architecture of the GloSIS template node

This annex describes the functionalities expected from the GloSIS template node and provides some suggestions on the supporting technologies. The template node has two main functionalities: 1) load and management of soil point data, 2) publication of data through web-services. Note that the GloSIS template node can be implemented by a soil data holder as a stand-alone soil information system. It does not necessarilyimply that this SIS is connected to the GloSIS federation and that the soil data are shared. However, implementing a GloSIS template node ensures that the SIS is GloSIS compliant and can be easily linked to GloSIS if the data holder wishes to do so.

Data load and management

The domain model can be straightforwardly transformed into a data model implemented by a relational database management system (DBMS). This system makes the core of the GloSIS template node. Technological options for DBMS abound, however, important constraints must be considered:

1. the DBMS must be spatially enabled (allow for spatial data to be stored), expressively supporting thegeographical nature of soil information;

2. the DBMS must provide strong data structure enforcement, going beyond classical entity-relationship constraints. Technologies such as Oracle/Spatial or PostgreSQL/PostGIS are good examples of DBMSthat meet these criteria.

To use a template node, data providers will have to load their soil data stored in tabular format (e.g. MS Excel)into the database. Data will first be processed and checked for consistency before submission to the database,with further validation using native database constrains. Skilled database administrators can directly access the database using tools like pgAdmin for PostgreSQL and SQL Developer for Oracle, even allowing the insertionof data in bulk using direct SQL queries, thought this will be the exception rather than the rule

Data providers are likely to keep their data in different and dispersed databases, perhaps not even in databases, but rather in loose files. In such contexts data import must be approached differently, also keepingin mind that unstructured data stores are usually consequence of knowledge gaps. To this end, the template node shall make use of Extraction, Transformation and Load (ETL) tools. In a nutshell, such graphical tool maps fields (columns) from existing data stores to fields in the GloSIS data model, allowing for automated data load and specific transformations to conform the original data with the GloSIS domain model. GeoKettleis an example of such tools, with a focus on geographical data (using GeoKettle).

Beyond data load, the template node must also provide tools to manage the database content that are easierto approach than those commonly used by database administrators. One way of achieving this is through a management application offering functionalities to manage users, browse and update data (what aretechnically known as CRUD operations: Create, Read, Update and Delete). This application can take the shapeof web forms, a familiar format that has the convenience of dispensing a client programme. This managementapplication is to be used internally at the data provider institution, and not meant for external access.

Rapid Application Development (RAD) tools suit well the development of such a management application. These provide mechanisms to swiftly arrive at a full featured graphical application from a data model (or database). While the landscape of RAD technologies is far and wide, in the GloSIS implementation those withsounder data modelling constructs are preferable, as is this the case with Django.

A further alternative for data management is a relational database enabled GIS programme. QGIS, for instance,offers users the possibility to develop custom forms to create and edit attributes of spatial features. While easier to set up, such a programme will never be as powerful as an application produced with a RAD tool. In any event, being easy to use, these tools are certainly a good complement for data management tasks.

Data Publication

Basing the implementation of GloSIS on the WFS standard greatly determines the kind of data publishing technologies to include in the template node. GeoServer is an example of an OGC compliant implementation of the WFS standard that provides a plug-in for the development of GML application schemas, for complex datawith WFS 2.0. This plug-in expands an existing web-service to output the geo-spatial data respecting the samestructure of an underlying the domain model. While still in its infancy, WFS 3 has largely been a bottom-up process, meaning that software implementations are already available, as is the case with pyGeoAPI.

Security and accessibility of the data stored in a template node is another important aspect to consider. The data provider may not wish to publish all the data stored in the node, or restrict its access to particular usersor stakeholders. These issues concern solely the data and meta-data services, since none of the other services(DBMS, ETL, management application) is intended for external access. Out-of-the-box software already provideways to control and limit access to data services. For instance, GeoServer provides different levels of access control, be it to services or to specific data sets. This is achieved through a system of roles and access rules, that can be managed by an administrator user (dispensing programming or low level configurations).

Meta-data must also be considered. The domain model shall include all the information needed to produce a standard compliant record for each soil profile. Therefore, meta-data will be included by default in the application schema to be developed. However, it may still be useful to include in GloSIS a meta-data service, for instance, compliant with the Catalogue Services for Web (CSW) standard, also issued by the OGC. This service must publish meta-data records automatically generated or retrieved from the DBMS. Technologies abound for this purpose, including GeoNetwork or pyCSW. Ideally, every data object in the GloSIS federationwould have a corresponding metadata record, accessible through a standard service.

A further technology useful in the template node is a web mapping application. Such application provides a simple and approachable view of the data in the form of a map accessible through a web browser. This mappedview is useful for a rapid inspection of the data, or to provide an overview to external partners and stakeholders.Like with the previous cases, this application shall function "out-of-the-box", dispensing any intervention by the hosting institution; it will use a pre-configured Web Mapping Service (WMS). Various technological optionsare available to set in place a general-purpose web mapping application that dispense programming; a non- exhaustive list:

- Mapbender
- QGis Web Client
- MapGuide
- Geomajas

Figure 3 summarizes the technologies involved in a GloSIS template node. Central is the DBMS that contains the soil point data (profiles, sampled layers with associated descriptive and analytical data). An ETL tool loadsthe data stored in the data repository of the (e.g. a series of MS Excel tables) in the DBMS. The DBMS is supported by a Management Application to handle the data. The data are served via three services: 1) a featureservice for serving soil data across the web; a catalogue service for serving metadata; a mapping service to serve mapped views of the data through a web mapping application. Each service is based on existing open standards of the OGC.

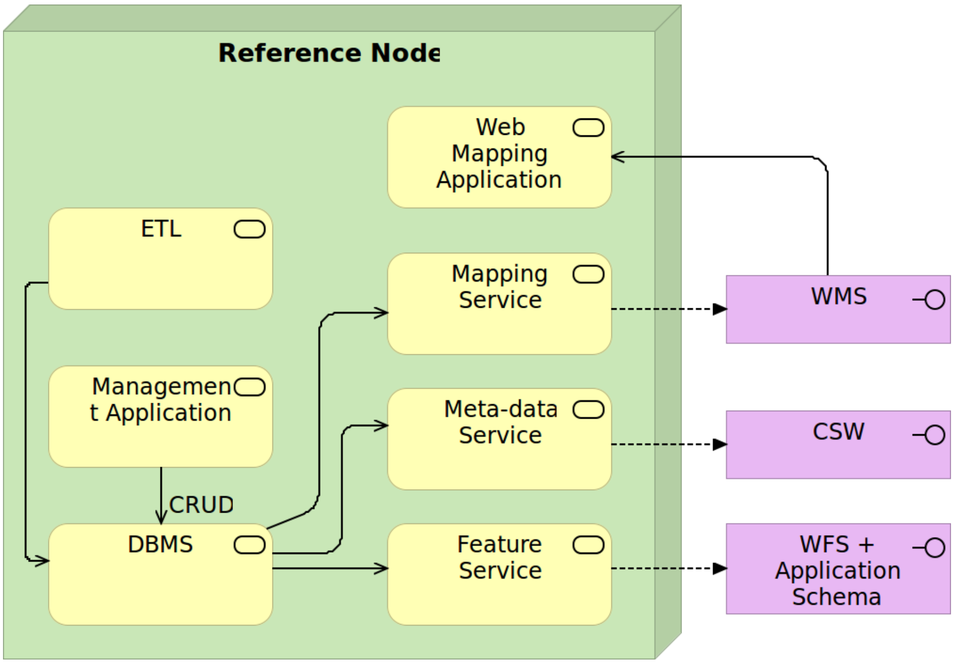


Figure 4. Technologies composing the GloSIS template node

Deployment

With the architecture and engineering aspects of a GloSIS template node defined, a further step is required to guarantee that institutions wishing to deploy the system are able to do so as easily as possible. To that enda deployable infrastructure must be made available: a GloSIS node reference implementation. This infrastructure must include a selection of technologies that fill in the functionalities outlined above:

- DBMS
- ETL tool(s)
- Management application
- Feature Service
- Mapping Service
- Meta-data service
- Web mapping application

Instead of simply defining a set of technologies, their versions and interconnections, the GloSIS reference implementation should be a ready-to-run block, relieving adopting institutions from much of the infrastructureadministration work. Modern container technologies such as Docker present an effective means to create suchcompact block. A container is a lightweight virtual system into which different tools may be installed and connected, thus producing a single bundle, potentially offering different features. A container can then be seamlessly deployed to a host system or a cloud service, with all configurations and tools.

For data providers wishing to publish large data sets, a single container might be a limited architecture. It is also possible to slice the bundle into various containers against which resources may be balanced. However, the single bundle concept provides an easy means of deployment and an approachable implementation to dataproviders with limited knowledge or limited resources.

A containerised approach also facilitates software updates, particularly those regarding security. Any healthy software is in constant development, and therefore the need to deploy a new version of any of the technologies outlined in this Chapter must be considered. This is easily achievable if the reference implementation is a bundle of containers (one per technology/application). In such case all that is required is the replacement of the respective container. If the technologies are all bundled in a single container then a new bundle must be deployed altogether. However, as long as data and configuration files are kept outside the containers (e.g. inthe host system or in a storage volume) replacing software should always be a seamless operation. Still reporting to the example of Docker, the deployment of new software versions can be greatly facilitated if a repository of container configuration files, known as Dockerfiles, is kept up-to-date. From such configuration files the data provider can easily trigger the deployment of an updated container.

Annex 2. Data policy and publication

The accessibility and exchange of data in GloSIS will be governed by the endorsed GSP data policy. The GSPData Policy was developed by the GSP Secretariat and adopted by the GSP plenary assembly in order to promote soil data sharing for data products identified through Pillar 4 and considering harmonization and interoperability requirements according to Pillar 5. The data policy is applicable to all members of the GSP andall user groups including end users, developers and contributors that share soil data through GloSIS.

GSP data policy aims to ensure that:

- every existing ownership right to shared soil data is respected;
- the specific level of access and the conditions for data sharing are clearly specified;
- the ownership of each dataset and web service is properly acknowledged and well-referenced (lineage);
- the data owners are protected from any liability arising from the use of their original and/or derived data.

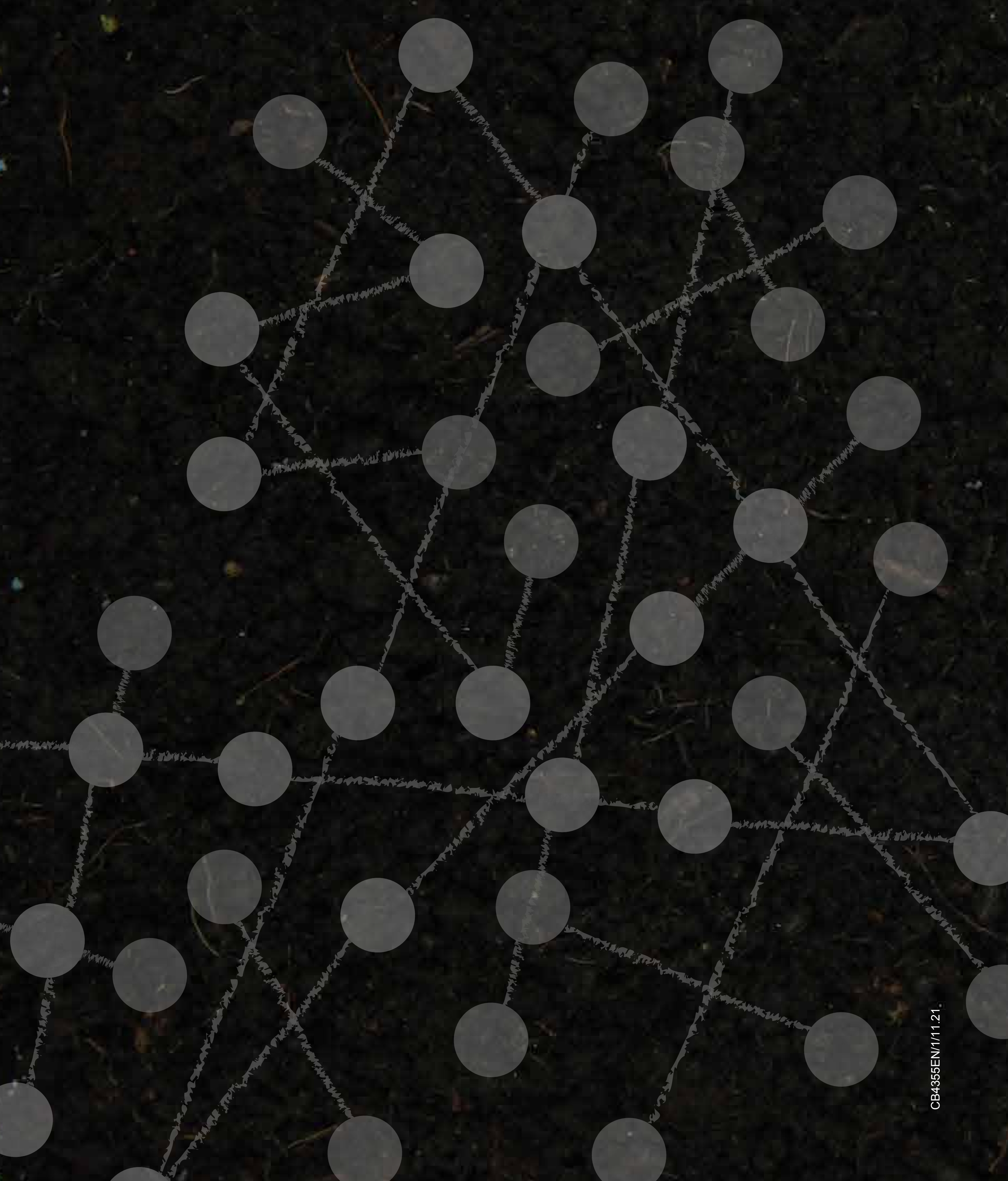
The license and therefore accessibility level for different uses of any dataset entered into GloSIS is therefore defined by the data owner and ensured in the infrastructure by the exchange mechanisms. Some datasets may perhaps be earmarked for use within countries themselves or only for research purposes while other datamay be fully open, this is the preferred choice.

Data shared through web services will be provided and exchanged according to the standards as described inchapter 4 and disclosed on a continuous basis through the web interface of the GloSIS discovery hub and theweb interfaces of the data holder's soil information system if the data holder has one available. User authentication is not foreseen in the setup of GloSIS.

GloSIS aims to follow the FAIR principles for scientific data management and stewardship while respecting the GSP data policy. Therefore access will at all times be controlled by the data provider, so the data provider determines which data will be shared. For instance, the provider may choose to serve only a subset of their data openly and another part for national or institutional use only, instead of sharing their complete data holding openly. The data provider will also determine how data will be shared. For instance, the provider may wish to share only the metadata but not the data itself. Sharing metadata at least ensures thedata are findable. Data access could then be further arranged between the provider and user.

GLOBAL SOIL PARTNERSHIP

The Global Soil Partnership (GSP) is a globally recognized mechanism established in 2012. Our mission is to position soils in the Global Agenda through collective action. Our key objectives are to promote Sustainable Soil Management (SSM) and improve soil governance to guarantee healthy and productive soils, and support the provision of essential ecosystem services towards food security and improved nutrition, climate change adaptation and mitigation, and sustainable development.



CB4355EN/1/11.21