D6.1 Technical design of the Soil Information System

Project No. 869200

30 November 2021

Deliverable: Soils4Africa_D6.1_v01

Version 1.0

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

Contact details

Director of Coordinating Institute – ISRIC: Rik van den Bosch

Project Coordinator: Mary Steverink-Mosugu

Address: Droevendaalsesteeg 3, 6708 PB Wageningen (Building 101), The Netherlands

Postal: PO Box 353, 6700 AJ Wageningen, The Netherlands

Phone: +31 317 48 7634

Email: mary.steverink-mosugu@isric.org

Project details

Project number	862900
Project acronym	Soils4Africa
Project name	Soil Information System for Africa
Starting date	01/06/2020
Duration In months	48
Call (part) identifier	H2020-SFS-2019-2
Topic	SFS-35-2019-2020 Sustainable Intensification in Africa

Document details

Work Package	6
Deliverable number	D6.1
Version	1
Filename	Soil4Africa_D6.1_Technical design of the SIS_v01
Type of deliverable	Report
Dissemination level	Public
Lead partners	ISRIC
Contributing partners	IITA, ARC, SZIU, RCMRD, JRC
Author	Ulan Turdukulov, Bas Kempen, Jorge S. Mendes de Jesus, Luis Calisto, Paul van Genuchten, Laura Poggio
Contributors	Jeroen Huising, Garry Patterson, Adam Csorba, Olatunbosun Obileye, Allan Oware, Phanuel Ayuka, Arwyn Jones
Due date	30 November 2021
Submission date	30 November 2021

This report only reflects the views of the author(s). The Commission is not liable for any use that may be made of the information contained therein.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

Table of Contents

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

List of Figures

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

Abbreviations

API	Application Programming Interface
ARC	Agricultural Research Council, South Africa
CPU	Central Processing Unit
DB	Database
DMT	Data Management Tool
DSM	Digital Soil Mapping
ETL	Extract-Transform and Load
EU	European Union
HPC	High Performance Computing
FAIR	Findable, Accessible, Interoperable and Reusable principles of data sharing
FNSSA	Food, Nutrition, Security and Sustainable Agriculture
GDPR	General Data Protection Regulation
GIS	Geographic Information System
GLOSIS	Global Soil Information System
GLOSOLAN	Global Soil Laboratory Network
GSP	Global Soil Partnership
ICRAF	International Centre for Research in Agroforesty
IaC	Infrastructure-as-Code
IITA	International Institute of Tropical Agriculture
ISRIC	International Soil Reference and Information Centre
KIMS	Knowledge Information Management System
LADP	Lightweight Directory Access Protocol
LIMS	Laboratory Information Management System
LUCAS	Land Use and Coverage Area Frame Survey (EC)
OGC	Open Geospatial Consortium
RCMRD	Regional Centre for Mapping of Resources for Development
RDBMS	Relational Database Management System
S4A	Soils4Africa
SIS	Soil Information System
TBD	To be decided
TLS	Transport Layer Security protocol
WUR	Wageningen University & Research
WP	Work Package

I. Introduction

A Soil Information System for Africa

The Soils4Africa[1] project (2020-2024) aims to provide an open-access Soil Information System (SIS) hosting a set of key soil quality indicators under agricultural land use across Africa. These indicators are to be based on field data collected from 20,000 sampling sites, according to a sound methodology for repeated soil monitoring. This soil information system will become part of the knowledge and information system of the EU-Africa Partnership on Food and Nutrition Security and Sustainable Agriculture (FNSSA)[2] and will be later hosted by an African organisation with the requisite capacity to manage the system. This system will inform decision making and other activities towards sustainable agricultural intensification in Africa and facilitate future monitoring and evaluation and it will enhance the performance of other land uses (Fatunbi & Abishek, 2020).

This report marks the start of Work Package (WP) 6 in the Soils4Africa (S4A) project. The main objectives of WP6 are: develop the SIS following the requirements specified in WP3, design and build it, and train technical staff of the SIS hosting institute in its operation and management.

The focus of this report will be the first objective of WP 6 - developing technical design of the SIS, including descriptions of:

- Application Programming Interfaces (APIs) to ingest, store and manage field data from Field campaign (WP4)
- APIs to ingest, store and manage analytical data from Soil analysis (WP5)
- data exchange services based on open standards
- hardware requirements and (open source) software resources

Overall, S4A SIS requirements can be summarised as a **Knowledge Information Management System (KIMS)** that follows Findable, Accessible, Interoperable and Reusable (FAIR) data sharing principles. The SIS should be able to become a node in the GLOSIS[3] federated system of soil information systems (under development in the Global Soil Partnership) and in the GLOSOLAN[4] soil spectrometry network. The SIS should allow for the creation of data, metadata, and web services (such as those of the Open Geospatial Consortium - OGC) and support metadata exchange standards. Its architecture should support existing and newly developed OGC standards (also known as OGC APIs) for spatial data exchange with an outlook into the future. Finally, the SIS should be extensible to provide future endpoints to the collected S4A soil data such as additional API and ontology interfaces.

This report marks the completion of the first task in WP6 (Task 6.1: Technical design of the SIS) and is the result of several bilateral and group discussions with project partners taking part in WP4 (Field campaign) and WP5 (Soil analysis), as well as collecting the experiences of the LUCAS survey. Several meetings were held among ISRIC staff who will be implementing the SIS in WP6 to discuss the present design. Findings were translated into a conceptual design that was presented to the project consortium during the S4A annual project meeting, and later into a draft design that was presented and discussed during a WP6-Task 6.1 meeting.

[1] https://www.soils4africa-h2020.eu/the-project

[2] https://knowledge4policy.ec.europa.eu/publication/eu-africa-research-innovation-partnership-food-nutrition-security-sustainable_en

[3] https://www.fao.org/global-soil-partnership/areas-of-work/soil-information-and-data/en/

[4] https://www.fao.org/global-soil-partnership/glosolan/en/

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

SOILS4AFRICA

The report is organised as follows. Chapter 2 looks at the two main data flows into the S4A SIS. It briefly presents state of field data collection workflow and the current state of the laboratory data systems at ARC (project partner responsible for the laboratory analyses). It looks at the requirements for successful data processing and discusses use of APIs and remote data access. Chapter 3 illustrates the overall architecture of the S4A SIS. A detailed description of seven key components of the SIS is given, including their functionality, software, and deployment requirements, including use of containerisation tools, and managing SIS infrastructure through the code. Lastly, in this chapter we mention common issues of authentication and security in the context of the SIS. Finally, chapter 3 concludes the report.

Figure 1 gives a schematic representation of the data flows among the work packages in the Soils4Africa project.

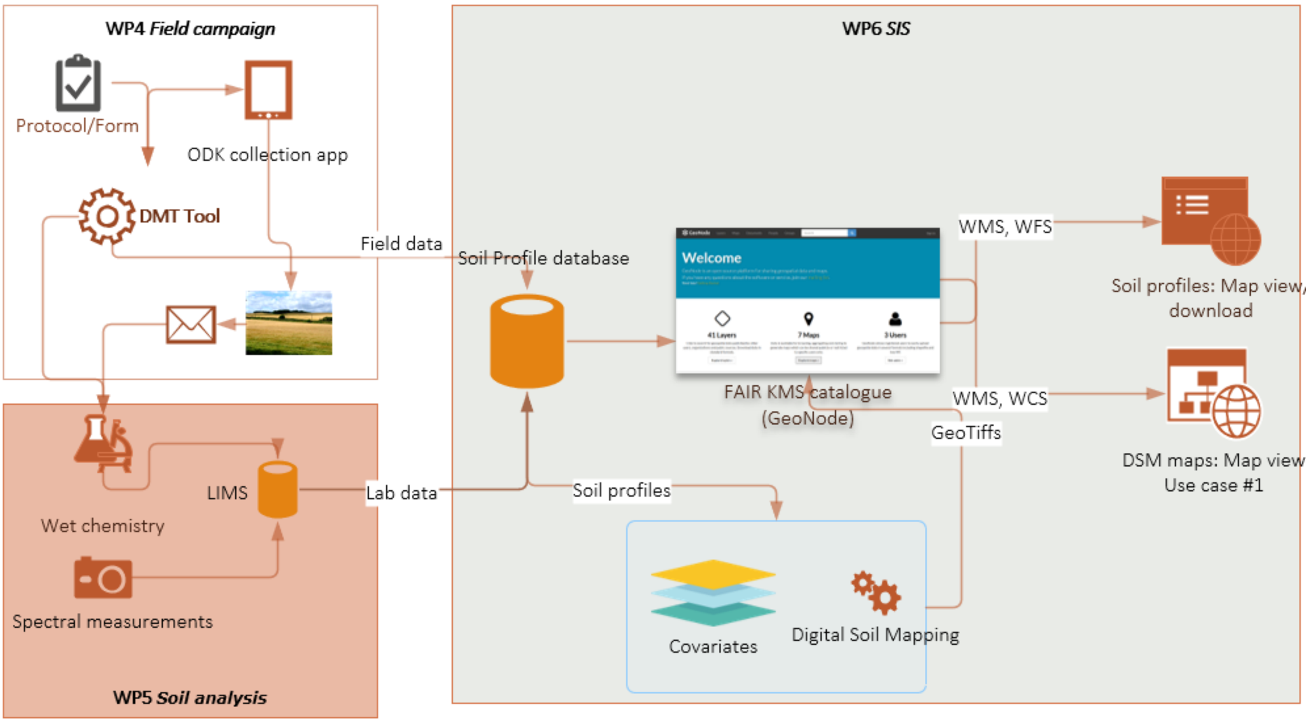


Figure 1 Schematic representation of interdependencies among the work packages in the Soils4Africa project.

II. Data inputs into the SIS

There are two main inputs into the S4A SIS: data coming from the field campaign (WP4) and analytical data (wet chemistry and spectral measurements) coming from the laboratory (WP5). The following sections will look at specifications and requirements for each of these data flows.

Field data

Software architecture of a tool to support field data collection is currently defined by the WP4 leadership and consists of an ODK server, along with an ODK collection app made available from the Google Appstore. Collected data will undergo a quality control procedure that is currently being developed by IITA (WP4 leader); the so-called field Data Management Tool (DMT) based on the ONA platform[5]. Figure 2 represents a simplified workflow of WP4 where the DMT plays a key role of pre-checking assigned forms as well as a post-quality checking step for submitted field data. At the time of writing, the DMT tool was under design and its exact implementation is as yeyetown.

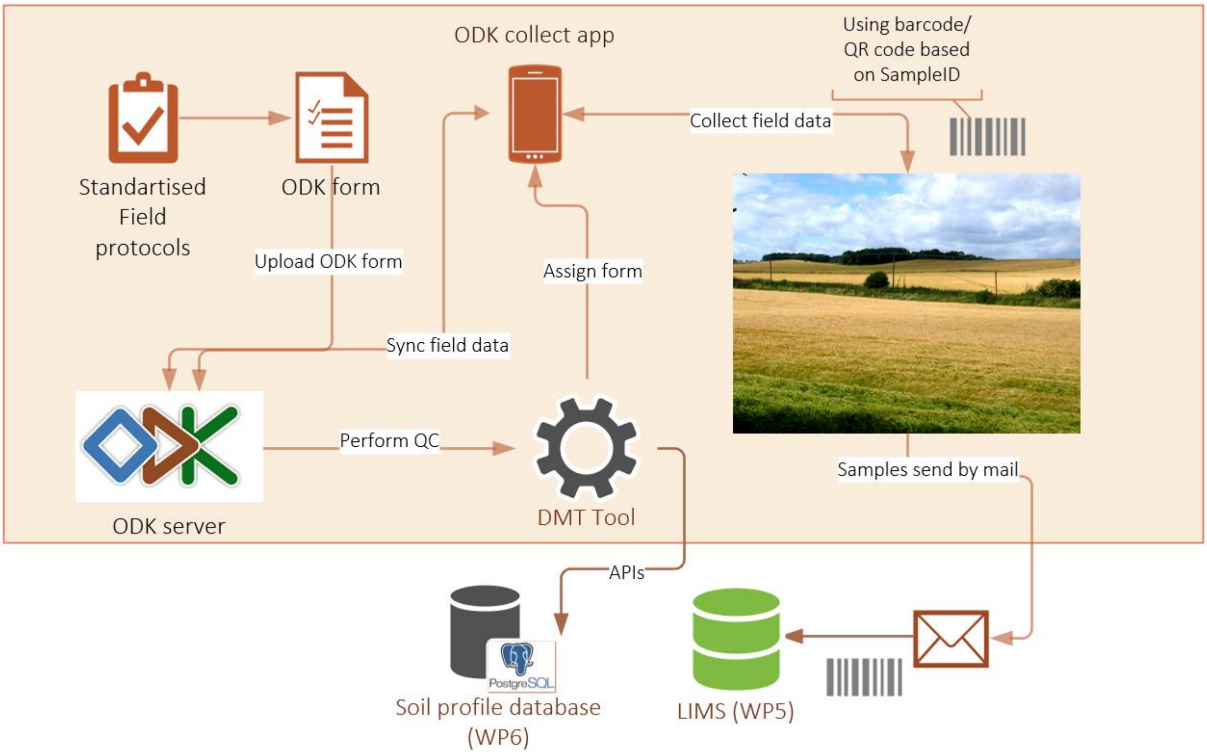


Figure 2 Simplified workflow diagram with workflows related to field data processing (WP4).

From the perspective of the SIS design, there are two important requirements that must be met in data flows from WP4. These were defined earlier in WP3 report (de Sousa, Turdukulov, & Kempen, 2021) and can be reworded as follows:

1. Data entries (values and value codes, so called "code lists") in field protocols and ODK forms should conform to the standards of the FAO defined in the "Guidelines for soil description" (FAO, Land and Water Division, 2006).

[5] https://ona.io/

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

2. Submitted ODK forms should indicate lineage of the data collection process: recording responsibilities in data collection and data verification steps.

The usage of common code lists is especially important in relation to the "reusable" aspect of data in the future FAIR portal - it implies data are to be harmonised and standardised for current and future reuse. To meet the requirement, the system must make use of standardised code lists when providing the data collected during the sampling campaign. Both code lists and APIs are currently under development in WP4, and the exact specifications are yet unknown.

Analytical data

Laboratory measurements of basic soil properties and spectral measurements will be performed by project partner ARC in South Africa. In addition, measurements of pesticide levels in a small subset of samples will be done by Wageningen University. Three different levels of analysis will be carried out, namely:

1. Spectral analysis (mid infra-red or MIR, 4000-400 cm^{-1}) on all 30 000 samples taken at 20,000 sampling locations.
2. Traditional "wet chemistry" analyses (as per LUCAS protocols) on a selected subset of about 20% of the samples for most chemical properties and more for properties that predict less well with spectroscopy (for example P, K) as specified in the Grant Agreement.
3. Among the 20 000 locations, 250 "reference sites" will be selected where soil profiles will be excavated, described, classified, and sampled in more detail, including topsoil and subsoil horizons.

The wet chemistry analysis results will also be used to develop spectral calibration models that correlate wet chemistry measurements with spectral measurements. These models are subsequently used to predict soil properties from the spectral measurements. Figure 3 represents a simplified workflow related to analytical data processing.

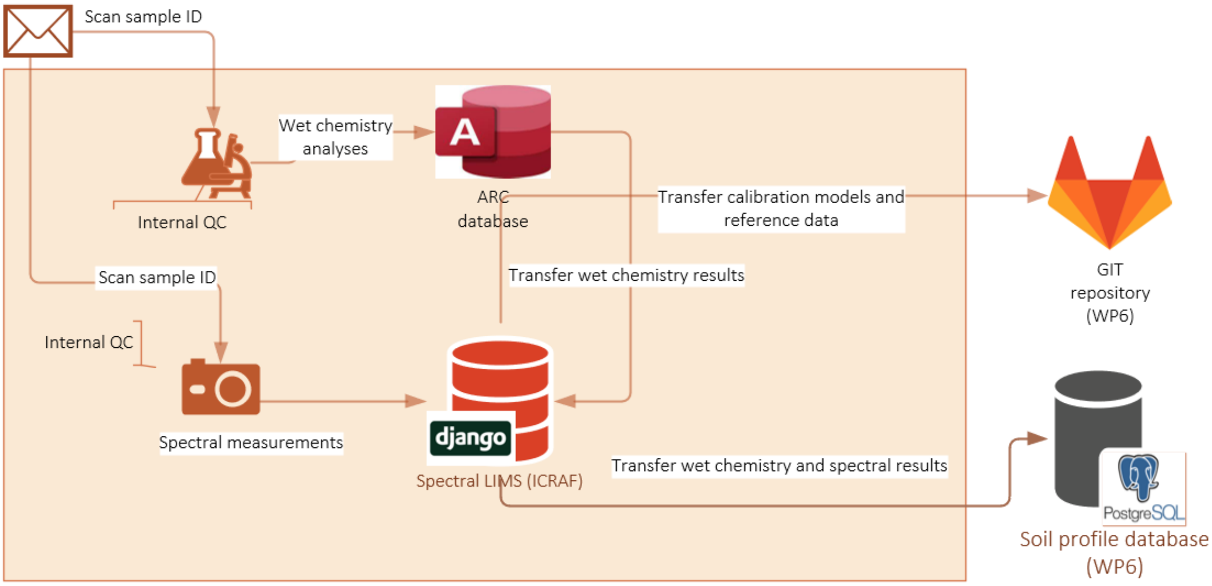


Figure 3 Simplified diagram with workflows related to analytical data processing (WP5).

At the time of writing, ARC is undergoing developments of its Laboratory Information Management System (LIMS). The current MS Access database that stores the results of the wet chemistry analysis will be replaced by a different system. It is not yet clear how the remote data access to the laboratory results will look like. Once ARC has its updated systems infrastructure in place, a

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

SOILS4AFRICA

solution to connect to the laboratory systems will be identified and incorporated in the technical design.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

III. S4A Soil Information System

This chapter describes in detail the architecture of the S4A Soil Information System. The key components of the system are:

1. Soil profile database: a single place to store data from the field observations and laboratory measurements.
2. Digital Soil Mapping (DSM) workflow: component that uses collected soil data along with additional environmental data layers (covariates) to generate gridded soil maps.
3. Git repository for storing codebases and spectral calibration models
4. Persistent volume storage to store covariates and soil maps
5. Dashboarding mechanism based on Apache Superset for creating interactive visualisations used for both internal workflows and on the dedicated use cases websites.
6. KIMS catalogue implemented as a GeoNode web portal for uploading and publishing geospatial data and OGC services.
7. Use cases

In the subsequent sections, each of these components will be described. The chapter will conclude with short sections on authentication and soil information use cases in the context of the SIS. Figure 4 below gives a schematic overview of the proposed SIS components.

With the rise of microservice architecture, Docker as a containerization tool has gained increasing popularity with time. Most of the SIS components (except for the Soil profile database) will be containerised and managed as Infrastructure-as-Code (IaC) process. As the name suggests, IaC is about managing the infrastructure through code which can be versioned through any version control system like Git for reliability for a better continuous delivery practice[6].

The proposed S4A architecture requires communication between components, using publicly exposed networks (i.e., through internet). Therefore, the architecture will implement Transport Layer Security (TLS) protocol in all its communications between SIS component (e.g., HTTPS, or Database access). Penetration testing and external audit will be taken into consideration when implementing the SIS infrastructure.

Common authentication / authorization mechanisms will be assured using Lightweight Directory Access Protocol (LDAP) and/or OAuth2 protocols, in a centralized system storing usernames and passwords, providing a single login point and registration access to all S4A SIS components. User information and registration will follow the European Union's General Data Protection Regulation (GDPR).

[6] https://www.nexastack.com/en/blog/infrastructure-as-code-and-containers

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

SOILS4AFRICA

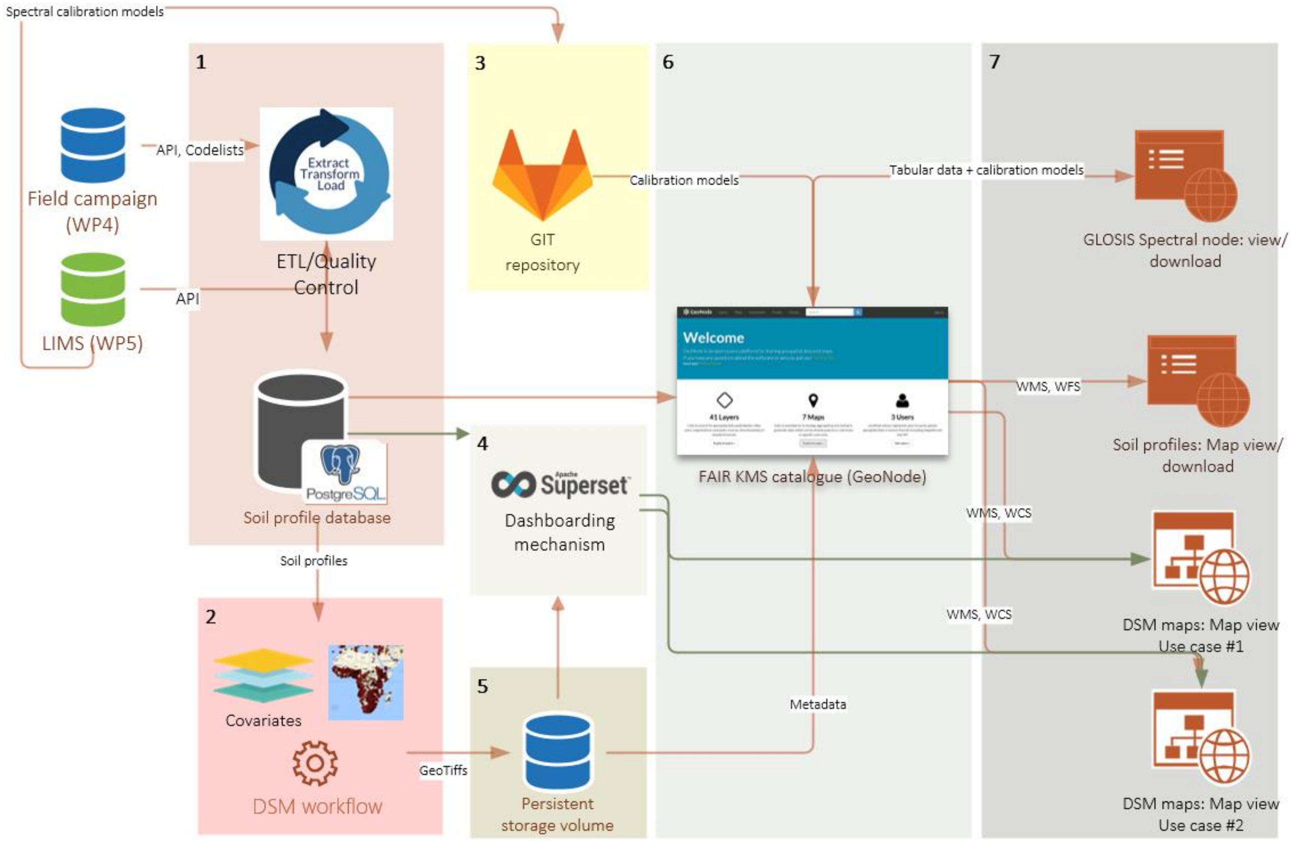


Figure 4 Schematic representation of the proposed S4A architecture.

1. Soil profile database

Summary

PURPOSE	Match and store soil profile observations and lab results
INCOMING CONNECTIONS	Via API from WP4 (Field campaign) and WP5 (LIMS)
OUTGOING CONNECTION	To DSM workflow and to KIMS catalogue
SOFTWARE	PostgreSQL (version 12 or higher) with PostGIS extension, Python for ETL workflow
DEPLOYMENT	Hosted database
PROCESSING NEEDED	ETL workflow for quality assessment, design of S4A data model and population of DB, developing retrieval endpoints (APIs)

This component of infrastructure is responsible for quality control and storing soil profile observations and field data as well as wet chemistry and spectral measurements coming from the laboratory. Matching field observations and laboratory data will be based on common sample identifiers and further quality-assessed through Extract-Transform and Load (ETL) workflow. Field and lab data will be received following API specifications from both WP4 (field campaign) and WP5 (lab analyses).

Extract-Transform and Load (ETL) workflow

The ETL workflow is the procedure of importing data into the standardized S4A data model. It will focus on the quality of correspondence of soil profile observations collected in the field campaign to the wet chemistry and spectral measurements. It will perform plausibility checks of soil property values (and units of measurement) and, where possible, harmonise using consistency procedures.

Measures for geographical accuracy (i.e., location) of the point data, as well as a first approximation for the uncertainty associated with the operationally defined analytical methods, will be presented for possible consideration in the subsequent digital soil mapping. The ETL workflow will be implemented as a Python script with additional SQL queries.

Data structures

For developing a data model for the S4A database we will look at existing data models for storing soil data such as the WOSIS[7] and GLOSIS data models. These could form a starting point for a data model for S4A. Especially the GLOSIS data model would be an obvious choice since it will be globally supported domain model. Although this model is still under development within the Global Soil Partnership, the structure of the data model could be used and adapted for S4A . Figure 5 shows graphical representation of related, WOSIS data model. The S4A data model aligned with GLOSIS data model would look similarly in the future.

[7] https://www.isric.org/explore/wosis

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

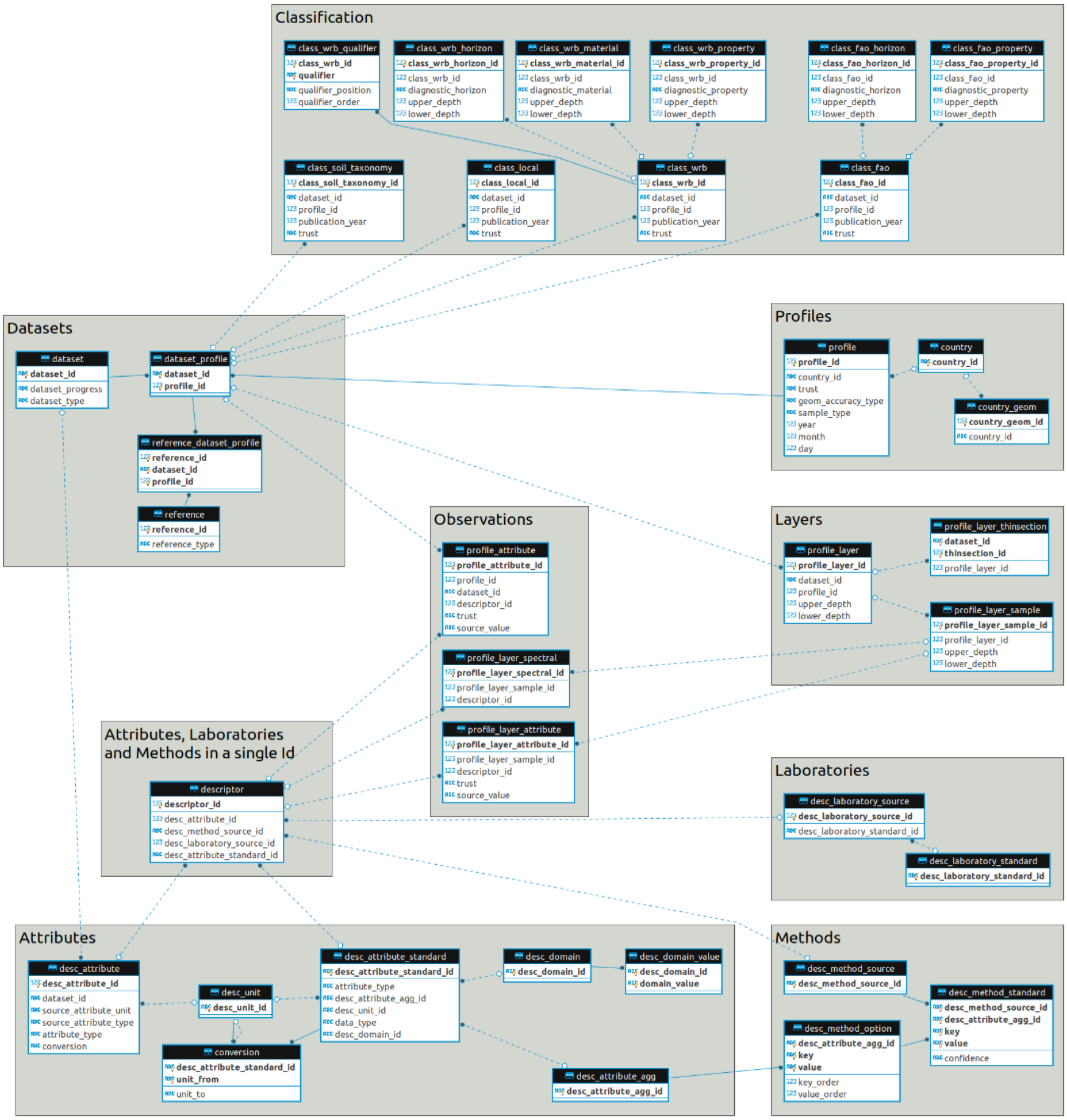


Figure 5 A simplified representation of the WOSIS data model.

A dataset in this model is defined as a series of observations having similar lineage. In practice, it can be a range of laboratory reports within a certain timeframe, on samples from a certain region, sampled by a certain group of surveyors, and/or modelled using a certain model calibration. Lineage of a dataset is captured in a metadata record describing that dataset.

Software

We propose to use PostgreSQL (version 12 or higher) for storing soil datasets. PostgreSQL is a free and open-source Relational Database Management System (RDBMS) emphasizing extensibility and SQL compliance. PostgreSQL is designed to handle a range of workloads, from single machines to data warehouses or Web services with many concurrent users. It has automatically updatable views, materialized views, triggers, foreign keys, and stored procedures. (The PostgreSQL Global Development Group, 2021).

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the EU Horizon 2020 Research and Innovation Programme under Grant Agreement.

With PostGIS extension, PostgreSQL supports a wide variety of geographic information system (GIS) data types. Support for geospatial data types is especially important in S4A project since soil data are inherently spatial. GIS data type support are essential for ETL and in other parts of DSM workflows.

Deployment

Soil profile database will initially be hosted at the premises of Wageningen University & Research (WUR). If needed, the database can easily be migrated to any other hosted environment or dedicated database server in the future at the SIS hosting institute in Africa.

Having a single storage for all S4A soil datasets has several advantages besides consistency. In the future, various endpoints can be developed to this database. For instance, APIs to spectral estimation services can be developed from the S4A soil profile database that allow queries of samples based on their spectral responses, wet chemistry results or site characteristics. In addition, a semantic web service can be developed to serve as an interface to exchange soil data stored in the S4A database with other soil data providers. Both a spectral estimation service and a semantic soil ontology are currently in the active development in GLOSIS and GLOSSOLAN and EJP[8] initiatives.

2. DSM workflow

Summary

PURPOSE	To generate soil properties maps for use cases based on soil profile database and covariates
INCOMING CONNECTIONS	From DB (soil profile data), from 3rd party data sources (covariates)
OUTGOING CONNECTION	To KIMS catalogue
SOFTWARE	R, Python, GDAL, GRASS, QGIS
DEPLOYMENT	On HPC centre in WUR premises
PROCESSING NEEDED	DSM workflow, metadata creation

Digital Soil Mapping is the process of producing gridded maps of soil properties for a user-defined area of interest with (geo)statistical models that combine soil sample data and environmental data layers (or "covariates") that represent the soil forming factors (typically related to climate, land cover, terrain morphology and vegetation dynamics). Maps are produced at a specific, user-defined spatial resolution and are typically stored in in GeoTiff format. Resolution will depend on the user needs and available input observations. In S4A, the input data for the DSM models include the soil data collected during the project (possibly supplemented with additional soil data from other sources) and environmental covariates with continental or global coverage. Covariates will be obtained from existing remote sensing catalogues (Landsat, Sentinel, MODIS) and will be derived using appropriate satellite imagery processing services (such as Google Earth Engine, SentinelHub). Nationally or locally available covariates can be integrated. In addition to generating maps of basic soil properties and soil quality indicators, the workflow and software environment will also be used for generating maps for use cases.

Software environment

The workflow uses statistical and machine learning methods for digital soil mapping, relying exclusively on open-source tools (R, Python, QGIS, GRASS). An example of a suitable DSM

[8] https://ejpsoil.eu/

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

workflow is the workflow developed to produce the latest version of SoilGrids (Poggio, et al., 2021) This workflow could act as a starting point for developing a DSM workflow for the S4A project.

Deployment

The DSM workflow will be implemented using Linux containers[9]. A container includes all software with a specific version and the code that is needed to run the workflow. The use of containers will allow easy deployment and migration of the workflow to a different infrastructure. There will be two main implementations of the DSM workflow foreseen in the project: a) for developing (high-resolution) maps at continental level, and b) for developing soil maps at national or sub-national level.

a. *Continental level*: The continental implementation will require High Performance Computing (HPC) facilities. An example is Anunna[10], an HPC cluster managed by Wageningen University and Research (WUR). Anunna provides many CPUs with adequate amount of memory per CPU. The Slurm Workload Manager is used to manage the cluster workload. Singularity[11] is used for access and running of containers. Having a containerized workflow ensures efficient transfer of the workflow to the SIS hosting institute. Academic, national, or commercial cloud infrastructures could be used as alternatives as the container technology ensure the portability of the workflow.

b. *National/User level*: a simplified workflow could be deployed for developing soil information products on a limited geographic extent. The core will be in common with the continental implementation, but the workflow will be simplified to potentially run-on single workstations as not to depend on access to an HPC environment. Again, containers will be used to ensure easier transfer of applications and reproducibility.

3. Git repository

Summary

PURPOSE	To store spectral calibration models and various codebases for DSM workflow, soil sampling design, tutorials
INCOMING CONNECTIONS	Direct upload from WP4 (spectral calibration models), direct upload from various workflows (DSM, sampling)
OUTGOING CONNECTION	To KIMS catalogue
SOFTWARE	Any Git (i.e., GitLab Community Edition 14 or higher)
DEPLOYMENT	Any Git (GitLab is hosted in WUR premises)
PROCESSING NEEDED	None, but can trigger workflows if needed

The current practice to share spectral calibration models is to prepare one table (for instance a `csv' table) with both the spectra and the wet chemistry calibration data, linked through a sample identifier. For instance, ICRAF develops its models in the R language and bundles these models with the datasets it the model uses (de Sousa, Turdukulov, & Kempen, 2021). However, no code marking mechanisms are used presently (e.g., tags, branches) and models are just stamped with a date to identify a particular state of the repository (de Sousa, Turdukulov & Kempen, 2021).

To anticipate future use of version control mechanism on these models, we propose to use Git. Git is software for tracking changes in any set of files, usually used for coordinating work among

[9] https://www.docker.com/resources/what-container

[10] https://www.wur.nl/en/show/High-Performance-Computing-Cluster-HPC-Anunna.htm

[11] https://singularity.hpcng.org/

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

programmers collaboratively developing source code during software development. Its goals include speed, data integrity, and support for distributed, non-linear workflows and "it is by far, the most widely used modern version control system in the world" (Atlasian, 2021).

A Git repository can also be used for controlling other scripts and codebases developed in the S4A project: soil sampling workflow, digital soil mapping codebase, ETL workflows. For all of these, a version control system will be needed since the code changes and modifications are expected among project partners and national institutes working on the same codebase in non-liner workflows.

Git can also be used as an issue tracker and code workflow manager. It has ability to create wiki pages that allows proper communication and documentation of various software components that will be deployed in SIS – a necessity, since SIS will be later hosted by an African organisation and such transition will require proper documentation. In addition to the code, Git can store files, thus effectively Git can be used as a common holder of tutorials and training data needed for capacity building objective in the S4A project (another specific objective of WP6 – task 6.5).

One last useful feature of Git is that it can be enabled with a trigger mechanism to generate certain workflows based on the recent code merges (i.e., ingest data, quality control workflow, soil mapping, etc). This feature is known as Continuous Integration/Continuous Development (CI/CD) method for automating the code development and can be useful in future extension of SIS. CI/CD can be used, for instance, to convert a calibration model and associated reference data into a GeoPackage (GPKG) format. GPKG is an open, non-proprietary, platform-independent, and standards-based data format for geographic information system implemented as an SQLite database container that later can be uploaded into KIMS portal and displayed in the map interface of the GeoNode instance. A GeoPackage can contain data, metadata, lineage information and visualisation guidance.

4. Dashboarding mechanism

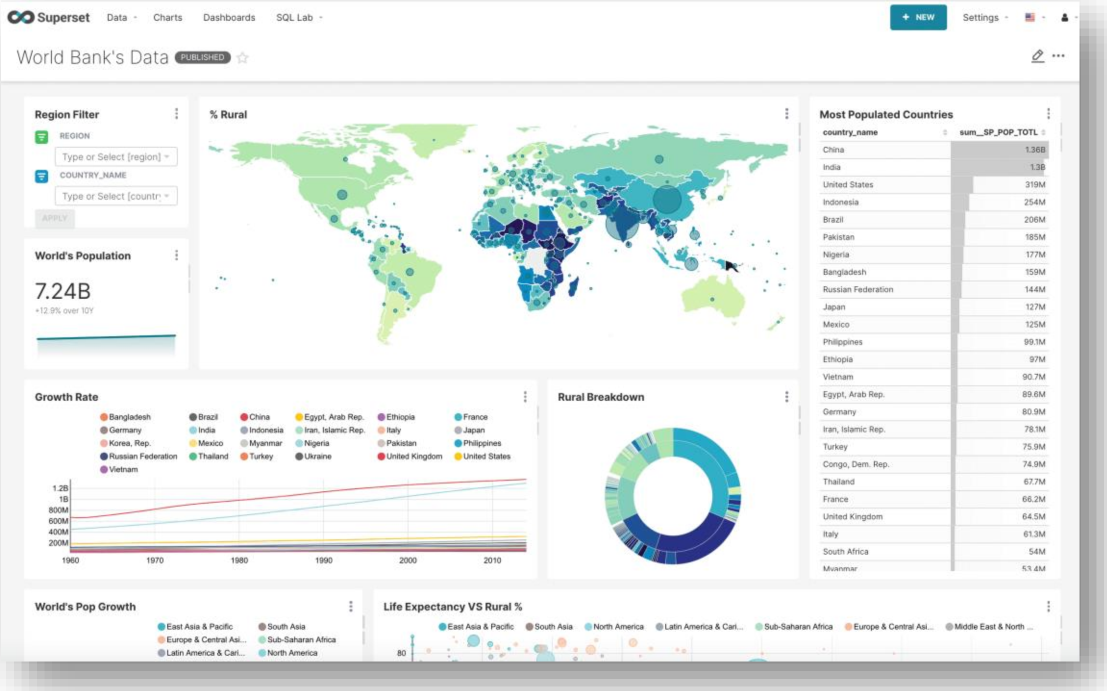
Summary

PURPOSE	Dashboarding mechanism for ETL workflow, for use cases
INCOMING CONNECTIONS	Database connections, APIs
OUTGOING CONNECTION	To ETL, to use cases
SOFTWARE	Apache Superset (version 1.3 or higher)
DEPLOYMENT	On Kubernetes

A dashboarding mechanism is needed in various workflows (e.g., ETL) but will be extensively used in supporting use cases by developing interactive web visualisations. We propose to setup Apache Superset. It is a modern, enterprise-ready business intelligence web application built on open-source tools. Superset is cloud-native, it was designed to scale out to large, distributed environments and can run analytic workloads against most popular database technologies (Superset, 2021). It will be deployed initially in the Kubernetes platform at WUR.

Error! Reference source not found. shows an implementation of Apache Superset in World's Bank Dashboard taken from gallery of Superset examples (http://superset.apache.org/gallery).

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.



Most Populated Countries

country_name	sum__SP_POP_TOTL
China	1.36B
India	1.3B
United States	319M
Indonesia	254M
Brazil	206M
Pakistan	185M
Nigeria	177M
Bangladesh	159M
Russian Federation	144M
Japan	127M
Mexico	125M
Philippines	99.1M
Ethiopia	97M
Vietnam	90.7M
Egypt, Arab Rep.	89.6M
Germany	80.9M
Iran, Islamic Rep.	78.1M
Turkey	75.9M
Congo, Dem. Rep.	74.9M
Thailand	67.7M
France	66.2M
United Kingdom	64.5M
Italy	61.3M
South Africa	54M
Myanmar	53.4M

Figure 6 An example of Apache Superset implementation: World Bank's Dashboard

(http://superset.apache.org/gallery).

5. Storage volumes

Summary

PURPOSE	Persistent data volumes for backup and storing data
INCOMING CONNECTIONS	DSM workflow
OUTGOING CONNECTION	To KIMS catalogue, to ETL workflow
SOFTWARE	TBD (depending on deployment)
DEPLOYMENT	On cloud platform (Google, AWC, Azure), on Kubernetes

Soil properties maps, in cloud optimized GeoTiff format, covariates and other ancillary environmental data necessary to run the DSM workflow will be stored in dedicated data volumes. Such volumes will then be available for use by other components of the infrastructure like GeoNode or the ETL workflow. Storage volumes will facilitate backup operations and make easy migration of the SIS to future hosting organization.

6. KIMS catalogue

Summary

PURPOSE	KIMS data catalogue for all S4A data, GSP/GLOSIS node, OGC service generator
INCOMING CONNECTIONS	From Soil profile DB, from DSM workflow
OUTGOING CONNECTION	To use cases
SOFTWARE	GeoNode (with GeoServer backend)
DEPLOYMENT	On Kubernetes

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

The Soils4Africa project will make all data available as open data and shared according to FAIR principles through user-friendly web services allowing data to be reviewed and downloaded. GeoNode fully facilitates FAIR data exchange and was therefore selected as the SIS' information management system. It forms the core of the S4A SIS architecture.

GeoNode

GeoNode is a web-based geospatial content management system, a platform for the management and publication of geospatial data. Data management tools built into GeoNode allow for integrated creation of data, metadata, and OGC services for uploaded spatial data. Each dataset in the system can be shared publicly or restricted to allow access to only specific users.

GeoNode is also designed to be a flexible platform that software developers can extend, modify or integrate against to meet requirements in their own applications. It brings together mature and stable open-source software projects under a consistent and easy-to-use interface allowing non-specialized users to share data and create interactive maps (GeoNode, 2021).

Error! Reference source not found.7 below shows snapshot of the UN World Food Programme Geographic data repository based on GeoNode with the functionality including *Layers*, *Maps*, *Service*, *Users* and *Datasets*.

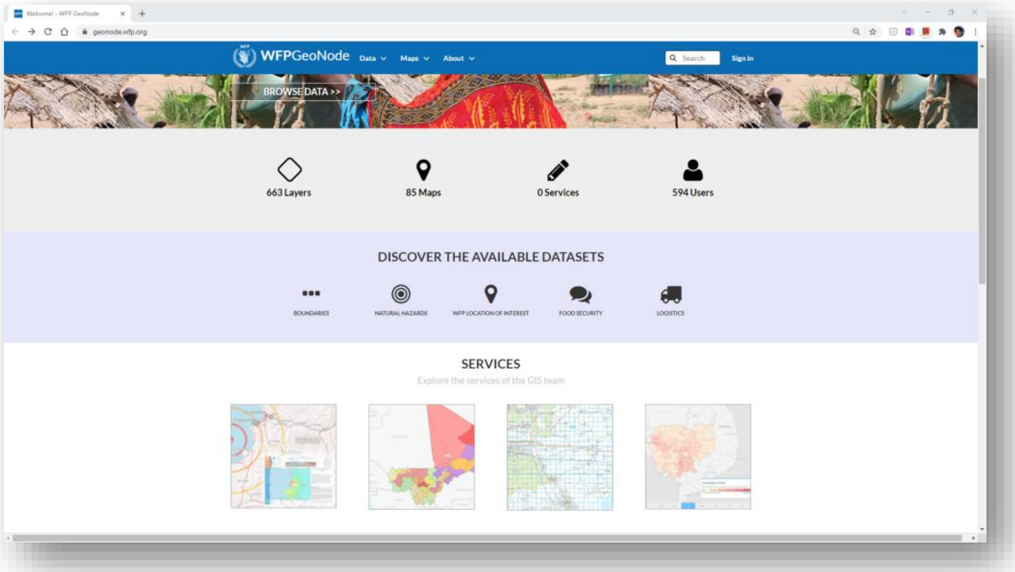


Figure 7 UN World Food Programme Geographic Data Repository based on GeoNode

(https://geonode.wfp.org/).

In essence, GeoNode consists of two other open-source components: pyCSW[12] and GeoServer[13] connected internally through APIs.

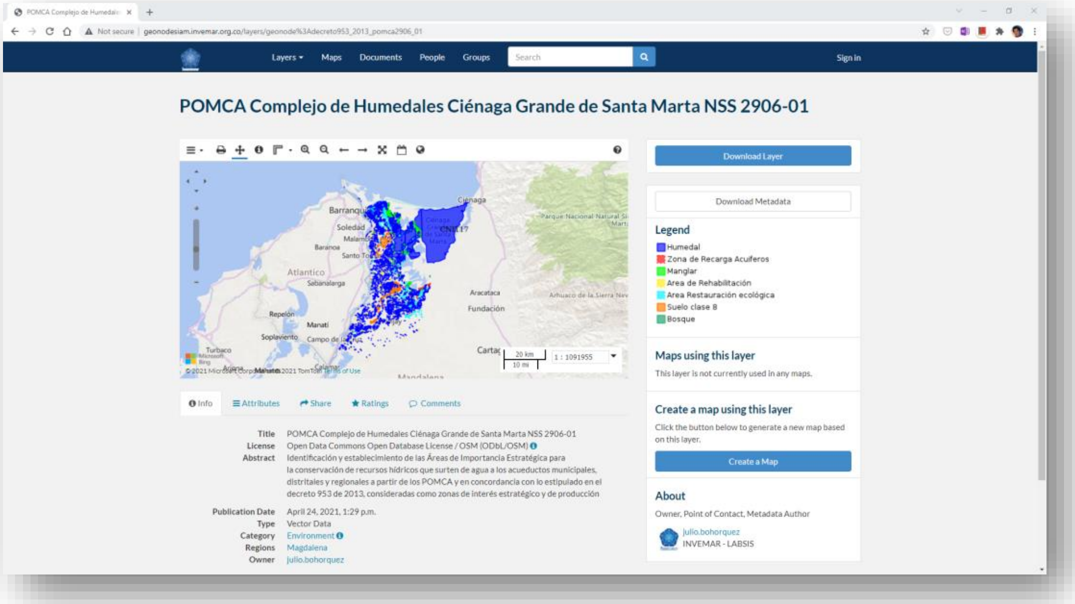
PyCSW

Once geospatial data are uploaded into GeoNode, they can be described using a metadata assistant. GeoNode's built-in assistant offers various license models and citation forms as lists and

[12] https://pycsw.org/

[13] http://geoserver.org/

simplifies and guides through the registration process. Figure 8 represents an instance of uploaded geospatial dataset with map visualisation, associated legend and filled in metadata records from Geographic Online Resources Catalogue.



POMCA Complejo de Humedales Ciénaga Grande de Santa Marta NSS 2906-01

Title: POMCA Complejo de Humedales Ciénaga Grande de Santa Marta NSS 2906-01
License: Open Data Commons Open Database License / OSM (ODbL/OSM)
Abstract: Identificación y establecimiento de las Áreas de Importancia Estratégica para la conservación de recursos hídricos que surten de agua a los acueductos municipales, distritales y regionales a partir de los POMCA y en concordancia con lo estipulado en el decreto 953 de 2013, consideradas como zonas de interés estratégico y de producción
Publication Date: April 24, 2021, 1:29 p.m.
Type: Vector Data
Category: Environment
Regions: Magdalena
Owner: julio.bohorquez

Legend
🟦 Humedal
🟥 Zona de Recarga Acuiferos
🟩 Manglar
🟨 Area de Rehabilitación
🟦 Area Restauración ecológica
🟧 Suelo clase 8
🟩 Bosque

Maps using this layer
This layer is not currently used in any maps.

Create a map using this layer
Click the button below to generate a new map based on this layer.

About
Owner, Point of Contact, Metadata Author
julio.bohorquez
INVEMAR - LABSIS

Figure 8 Snapshot of Wetland datasets of Geographic Online Resources Catalogue

(http://geonodesiam.invemar.org.co/).

pyCSW is the catalogue component of GeoNode. It offers standardised access to metadata records within the catalogue and harvesting records from remote sources through the following specifications:

- Catalogue Service for the Web - the existing OGC CSW standard.
- OGC API – Records. New OGC API family specification that offers the capability to create, modify, and query metadata on the Web.
- The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability.

GeoNode also has capabilities to transform records between ISO19139, DCAT and Dublin Core metadata standards. Therefore, the SIS, based on GeoNode's pyCSW component, can be linked and synchronised with other metadata data catalogues and data repositories such as GeoNetwork, DCAT and Dataverse. This allows the S4A SIS to be integrated into GLOSIS and be an OGC compliant node in GLOSIS – the federation of soil information systems. In addition, GeoNode web pages have schema.org annotations that facilitate the searchability of their content and integration into search engine results (i.e., Google, Bing search).

GeoServer:

GeoServer is an open-source server for sharing published data in GeoNode as geospatial data. GeoServer implemented as a backend component of GeoNode implements industry standard OGC protocols such as:

- Web Map Service (WMS, graphical representation of map layers)
- Web Feature Service / OGC API - Features gives direct API access to vector data
- Web Coverage Service / OGC API - Coverages gives direct API access to grid data

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

SOILS4AFRICA

- Web Processing Service / OGC API - Processes provides standardised access to processing API's

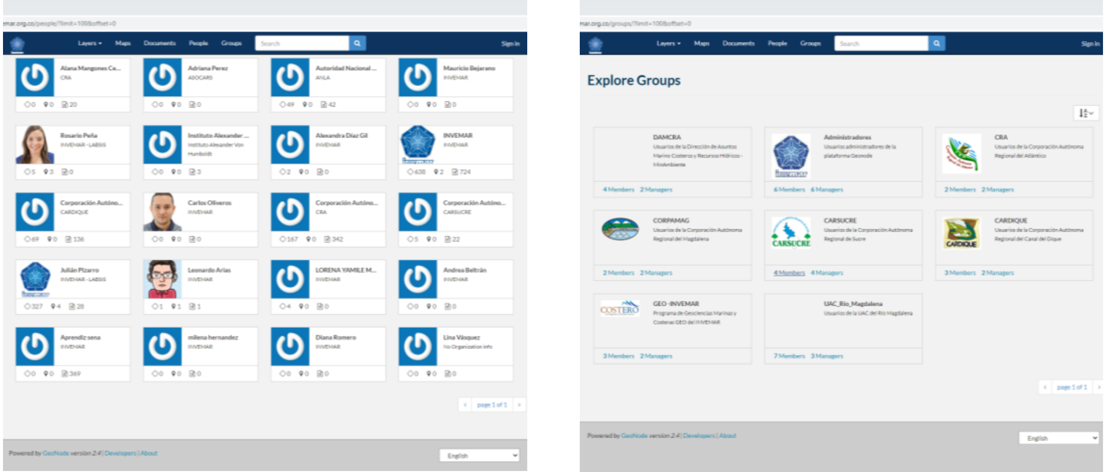
Note, the OGC API family of standards are a new set of OGC standards being developed to make it easy for anyone to provide geospatial data to the web. These standards build upon the legacy of the OGC Web Service standards (WMS, WFS, WCS, WPS, etc.), but define resource-centric APIs that take advantage of modern web development practices. These standards are being constructed as "building blocks" that can be used to assemble novel APIs for web access to geospatial content. (OGC, 2021). Therefore, it is important to ensure future proof design of the SIS by enabling support for recently developed OGC API standards, as these implement OpenAPI Specification allowing them to be used by third party websites, mobile phone applications and data science platforms in the future.

System administration

GeoNode can facilitate the use, management, and quality control of the data it contains. Social features, like such as user profiles, and commenting and rating systems, allow for the development of communities around each platform. For that, GeoNode provides a dashboard user interface for administrators, in which they can:

- Register new users and assign roles
- Register new datasets and maps
- View the usage statistics of the system
- Evaluate error logs of the various components

Figure 9 below shows an example of user (left) and group management (right) dashboard build in GeoNode.



mar.org.co/people/?limit=100&offset=0

Alana Mangones Ce... CRA ○0 ⚲0 20
Adriana Perez ASOCARS ○0 ⚲0 0
Autoridad Nacional ... ANLA ○49 ⚲0 42
Mauricio Bejarano INVEMAR ○0 ⚲0 0
Rosario Peña INVEMAR - LABSIS ○5 ⚲3 0
Instituto Alexander ... Instituto Alexander Von Humboldt ○0 ⚲0 3
Alexandra Díaz Gil INVEMAR ○2 ⚲0 0
INVEMAR INVEMAR ○638 ⚲2 724
Corporación Autóno... CARDIQUE ○69 ⚲0 136
Carlos Oliveros INVEMAR ○0 ⚲0 0
Corporación Autóno... CRA ○167 ⚲0 342
Corporación Autóno... CARSUCRE ○5 ⚲0 22
Julián Pizarro INVEMAR - LABSIS ○327 ⚲4 28
Leonardo Arias INVEMAR ○1 ⚲1 1
LORENA YAMILE M... INVEMAR ○4 ⚲0 0
Andrea Beltrán INVEMAR ○0 ⚲0 0
Aprendiz sena INVEMAR ○0 ⚲0 369
milena hernandez INVEMAR ○0 ⚲0 0
Diana Romero INVEMAR ○0 ⚲0 0
Lina Vásquez No Organization info ○0 ⚲0 0

page 1 of 1

Powered by GeoNode version 2.4 | Developers | About
English

mar.org.co/groups/?limit=100&offset=0

Explore Groups

DAMCRA Usuarios de la Dirección de Asuntos Marino Costeros y Recursos Hídricos - MinAmbiente 4 Members 2 Managers
Administradores Usuarios administradores de la plataforma Geonode 6 Members 6 Managers
CRA Usuarios de la Corporación Autónoma Regional del Atlántico 2 Members 2 Managers
CORPAMAG Usuarios de la Corporación Autónoma Regional del Magdalena 2 Members 2 Managers
CARSUCRE Usuarios de la Corporación Autónoma Regional de Sucre 4 Members 4 Managers
CARDIQUE Usuarios de la Corporación Autónoma Regional del Canal del Dique 3 Members 2 Managers
GEO -INVEMAR Programa de Geociencias Marinas y Costeras GEO del INVEMAR 3 Members 2 Managers
UAC_Rio_Magdalena Usuarios de la UAC del Rio Magdalena 7 Members 3 Managers

page 1 of 1

Powered by GeoNode version 2.4 | Developers | About
English

Figure 9 Snapshot of user (left) and group (right) management page of GeoNode of Geographic Online Resources Catalogue (http://geonodesiam.invemar.org.co/)

This functionality of GeoNode allows making the SIS a central community place where S4A and other users can gather around the SIS platform. Third party data providers (i.e., Copernicus – EU's Earth observation programme, National Soil Institutes and Regional partners) will be able to register and publish datasets with an easy-to-use template for filling the metadata and for creating map visualisations. Additionally, it can be used for project partners to upload smaller datasets into the SIS (for instance, datasets with pesticide measurements). More advanced GIS professionals can use the QGIS Bridge tool to publish data to the catalogue from within QGIS - a well-known open-source GIS Desktop tool.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

SOILS4AFRICA

Deployment

GeoNode will initially be deployed in the Kubernetes platform that is hosted at WUR. This task is facilitated by the fact that GeoNode community maintains the official Docker image[14]. If needed, customisations can be added to the existing Docker image since open source GeoNode is designed to be a flexible platform that software developers can extend, modify or integrate against to meet requirements in their own applications.

7. Use cases

Main use case defined in the S4A is the KIMS catalogue based on FAIR principles described above. Envisioned content of the catalogue is:

NUMBER OF BASIC SOIL OBSERVATIONS	~30,000
NUMBER OF REFERENCE SOIL PROFILES	~250
COVARIATES USED FOR SOIL MAPPING	TBD
REMOTE SENSING DATA	TBD
REFERENCE DATASETS (TRANSPORT NETWORK, ADMINISTRATIVE UNITS, LANDCOVER, LANDUSE)	TBD
CALIBRATION FILES SPECTRAL SOIL	TBD
SOIL PREDICTION MODELS (DSM)	TBD

Besides the main use case, deliverable 2.1 (Fatunbi & Abishek, 2020) defined a number of other use cases with key informants that include:

(i) soil information use in integrated landscape management.
(ii) soil data use for a sustainable intensification program in African farming systems.
(iii) soil data use in agricultural extension and advisory services.
(iv) use of soil information in public land resource conservation; and
(v) use of soil information by fertilizer producers and suppliers.

These use case examples characterise practical interventions requiring the development of relevant soil quality indicators, which determine the soil parameters to be included in the SIS. Although, deliverable 2.1 has not defined the exact functionality required for each of these use case, we believe that the proposed architecture (GeoNode with OGC service in combination with Apache Superset) is able to address the future interactivity needs. However, further usability research is needed to establish the exact interfaces and interactive functionalities required to support decision making in each of these use cases.

[14] https://github.com/GeoNode/geonode-project

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

IV. Conclusions

This report presents technical design of the SIS following the description of use cases in WP2 and the SIS technical requirements that were defined in WP3. Although WP2 has not addressed the exact functionality required for each of the above use case and further usability research is needed to establish those, we believe that the proposed architecture is able to address those interactivity needs. Further, this report can effectively be used as a blueprint for building the SIS. The presented design addresses some of the key requirements of the S4A SIS:

- All data published in the catalogue will contain metadata and allowing for easy search and query of published datasets and exchange of metadata records. Therefore, the KIMS, based on GeoNode described in section 6 of the SIS architecture, meets the FAIR principles and can be easily integrated as a node into GLOSIS federated system of soil information and in GLOSOLAN network.
- Soil calibration models along with reference data will be uploaded to the Git and eventually published in the GeoNode portal, thus addressing the key requirements of SIS being a node in GLOSOLAN soil spectral data exchange network.
- OGC Web services and OGC API services published through the catalogue will be used by GIS/IT professionals and will be the basis for generating maps and interactive visualisations for use cases.
- Various endpoints can be developed to S4A Soil profile database such as: a) APIs to spectral estimation services and/or b) an ontology interface to exchange soil data stored in S4A database.

When designing the architecture, special consideration was given to the use of open-source tools (no-vendor locking) and open standards. For instance, existing OGC Web services and recently developed OGC API standards, metadata exchange standards, extensibility options to add API endpoints and ontology web interfaces. Use of open standards provides possibilities for third party developers to use the SIS services developed here in other projects, thus increasing the exposure and relevance of this project.

Most of the proposed components will be deployed using microservices architecture and Infrastructure as Code approach. This ensures easy deployment, portability, and scalability of the future the SIS. Thus, the presented design ensures that S4A SIS architecture is will extensively cover all the required workflows in S4A from data ingestion, to mapping and publishing of soil data in support of sustainable agricultural intensification in Africa and provide the basis for a longer-term soil monitoring programme for the continent.

This design document will guide the development of the SIS in WP6. During all the development stages, an agile development approach will be followed for the implementation of the SIS. This means that the SIS will evolve gradually through iterations. Each iteration will result in a working product (‘rolling releases’) that can be demonstrated to the project partners and stakeholders in the project. A great advantage of having an agile implementation approach is that, when necessary, it allows the design and implementation to be adjusted to take account of stakeholder requirements that might change during the project’s life cycle. In this context, this design document is seen as a living document that will be adapted during the SIS development stage.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

V. References

Atlasian, a. (2021, October 31). *What is Git*. Retrieved from https://www.atlassian.com/git/tutorials/what-is-git

de Sousa, L. M., Turdukulov, U., & Kempen, B. (2021). *Deliverable 3.5: User requirements for the IT infrastructure.* Wageningen: Soils4Africa project. European Union’s Horizon 2020 research and innovation programme grant agreement No 869200.

FAO, Land and Water Division. (2006). Guidelines for soil description. Rome, Italy. Retrieved from https://www.fao.org/soils-2015/resources/fao-publications/news-detail/en/c/262368/

Fatunbi, O. A., & Abishek, A. (2020). *Deliverable 2.1: A set of use cases plus supporting soil quality indicators.* Wageningen, the Netherlands: Soils4Africa project. European Union’s Horizon 2020 research and innovation programme grant agreement No 869200.

GeoNode. (2021, October 31). *GeoNode -Open Source Geospatial Content Management System*. Retrieved from https://geonode.org/

OGC. (2021, Novemeber 6). *OGC API Context*. Retrieved from https://ogcapi.ogc.org/

Poggio, L., de Sousa, L., Batjes, N., Heuvelink, G., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL, 7*, 217-240. doi:doi: 10.5194/soil-7-217-2021

Superset. (2021, November 8). *What is Apache Superset?* Retrieved from https://superset.apache.org/docs/intro#what-is-apache-superset

The PostgreSQL Global Development Group. (2021, Novemeber 1). *PostgreSQL: The World's Most Advanced Open Source Relational Database*. Retrieved from https://www.postgresql.org/

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement.

SOILS4AFRICA