Development options for a Soil Information Workflow and System

Overview of methods, standards, and tools

F.M. van Egmond | T.I.S. van der Woude | et al.

November 2023

Development options for a Soil Information Workflow and System

Overview of methods, standards, and tools

F.M. van Egmond | T.I.S. van der Woude | et al.

November 2023

ISRIC-World Soil Information
Droevendaalsesteeg 3 (Building 101)
6700 PB Wageningen The Netherlands
E-mail: info@isric.org

The designations employed and the presentation of material in this information product do not imply the expression of any opinion whatsoever on the part of ISRIC concerning the legal status of any country, territory, city or area or of is authorities, or concerning the delimitation of its frontiers or boundaries. Despite the fact that this publication is created with utmost care, the authors(s) and/or publisher(s) and/or ISRIC cannot be held liable for any damage caused by the use of this publication or any content therein in whatever form, whether or not caused by possible errors or faults nor for any consequences thereof.

Additional information on ISRIC can be accessed through www.isric.org

Citation

Van Egmond, F., van der Woude, T., et al., 2023. Development options for a Soil Information Workflow and System, ISRIC – World Soil Information, Wageningen (https://doi.org/10.17027/isric-tmkb-pr58)

This report is only available as an online PDF version.

ISRIC
World Soil Information

Contents

Lists of tables

Lists of figures

Project acknowledgements

This project is funded through the Bill & Melinda Gates Foundation. The foundation has made great progress in the past ten years in exploring new and innovative digital soil mapping technologies and national soil information systems (SIS) in Africa and South Asia. Investments in transferrable technologies and other data assets have allowed for large strides towards improving soil health and farmer livelihoods. Nonetheless, although the value of the assets and technologies are known, the foundation would like to accelerate further impact, introduce greater intentionality and efficiency in the development of data assets, and be more responsive or adaptive to local demand. This project aims to perform a review of SISs, identifying what intervention approaches have worked well, which have not, which solutions work best and where to take innovation to scale. This information can help direct improved SIS intervention design and strengthen impact.

The partner is the Centre for Agriculture and Bioscience International (CABI). The subcontractors are the University of California, Davis (UC Davis) and the International Soil Reference and Information Centre (ISRIC – World Soil Information).
This report is one of the outcomes of the project.

We thank Ulan Turdukulov, Paul van Genuchten, Luis de Sousa, Bas Kempen, Gerard Heuvelink, Niels H. Batjes, Ad van Oostrum, Stephan Mantel, Andre Kooiman, Maria Ruiperez Gonzalez, Laura Poggio, Giulio Genova, Zhanguo Bai and Emily Toner of ISRIC – World Soil Information for their contributions by providing input and guidance on the report.

We thank Martin Parr (CABI), Melissa Allan (CABI), Pascale Bodevin (CABI) and Mariah Coley (CABI-Associate) for their guidance and review of the report.

We thank Ashenafi Abduljelil (CIAT, guest researcher at ISRIC), Yusuf Yigini (FAO-GSP), Marcos Angelini (FAO-GSP) and Isabel Luotto (FAO-GSP) for their suggestions and advice to improve the report.

Abbreviations

AAs	Atomic Absorption Spectroscopy
AES	Advanced Encryption Standard
AfSIS	African Soil Information System
API	Application Programming Interface
CABI	Centre for Agriculture and Bioscience International
CEC	Cation Exchange Capacity
CKAN	Comprehensive Knowledge Archive Network
DCAT	Drug, Chemical & Associated Technologies
DOI	Digital Object Identifier
DSM	Digital Soil Mapping
EJP SOIL	European Joint Program on Soil
FAIR	findability, accessibility, interoperability, and reusability
FAO	Food and Agriculture Organization of the United Nations
GeoTIFF	Geo Tagged Image File Format
ICP Forest	International Co-operative Program on Assessment and Monitoring of Air Pollution Effects on Forests
INSPIRE	Infrastructure for Spatial Information in the European Community
ISFM	Integrated Soil Fertility Management
ISO	International Standards Organization
ISRIC	International Soil Reference and Information Centre (also known as ISRIC – World Soil Information)
JSON-LD	JavaScript Object Notation for Linked Data
LUCAS	Land Use/Land Cover Area Frame Survey
MIR	Mid Infrared
NetCDF	Network Common Data Form
NIR	Near infrared
ODK	Open Data Kit
RDF	Resource Description Framework
SIS	Soil Information System
SOC	soil organic carbon
SQL	Structured Query Language
UC Davis	University of California, Davis
USDA	United States Department of Agriculture
W3C	World Wide Web Consortium
WRB	World Reference Base
XML	Extensible Markup Language
XRD	X-ray diffraction
XRF	X-ray Fluorescence

1. Introduction

1.1 Goal

Access to data can facilitate better informed decision making. A soil information system (SIS) is used to efficiently use, produce, organize, analyze, and serve soil data and information in a country, region or at any other scale. The goal of this document is to offer an aid for designing a SIS for soil data practitioners (users and producers). It provides an overview of the options, choices, results, and boundary conditions, and provides links to more detailed resources to execute the design and implementation.

1.2 Definitions

Soil

Soil is the epidermis of the Earth formed by various soil forming factors: climate, relief, parent material, and organisms acting over time, including humans. **Soils** provide many functions essential to humans, crops, vegetation, and water. These vital components of the soil rely on physical, chemical, and biological soil properties. Commonly measured and used basic soil properties are pH, cation exchange capacity (CEC), organic matter and carbon, bulk density, water retention, texture and particle size fractions, particle size distribution, proportion of coarse fragments, rooting depth, metals, some biological properties. Together they characterize the soil, its functions, quality, and health. This information can be used for various purposes, such as land use planning, crop production, water and nutrient management and climate change mitigation and adaptation. The relevant information on soil can be stored and managed in a SIS.

Soil Information System

A SIS is defined as an integrated information system that facilitates the storage, analysis, management and dissemination of soil data and information. The system often aims to provide users with access to a wide range of soil-related data and information, including soil properties, classifications, maps, and associated environmental data. It may contain multiple data sets, models, and tools in support of improved decision making by end-users. This definition primarily refers to the technological aspects. In this project, a soil information system encompasses the entire soil information workflow.

There is not one single way to design a SIS because the **SIS profile** (see Box 1) of a country tends to vary depending on end-user needs, data availability, and technical skills. A SIS profile is, therefore, essential to design a sustainable system that links to the use cases and provides a viable business model. The steps of designing and building a SIS are described in a Soil Data workflow, see the following section.

Soil Information Workflows

A workflow is defined as a 'sequence of processes through which a piece of work passes from initiation to completion'. The workflow concept can also be applied in the context of soil data: a '*soil information workflow*'. These are several steps that convert soil data into actionable information, see box 2.

Box 1: Criteria for defining a SIS profile

In this document, the criteria for defining a SIS profile are the combination of organizational, legislative, institutional, and financial contexts; known and potential future stakeholders (data users, providers, funders); current data collection and management; available hardware and software infrastructure; and the soil information presence and situation in a country or region.

In this document, we follow the DIKW pyramid in defining data and information. Where data are the observed, measured, assigned, or calculated values that characterize a (soil) property. Information is the interpretation of data, a synthesis, methodology. The foundation of information is data. Users can derive knowledge from the information provided or a knowledge layer can be added to the SIS. This would contain for example reference documents, papers, reports, sustainable soil management practices overviews, etc.

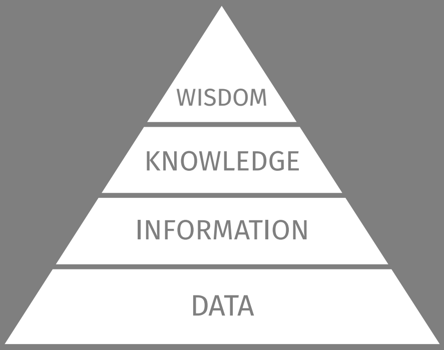


Figure 1: DIKW pyramid[1]

[1] https://en.wikipedia.org/wiki/DIKW_pyramid

Soil information workflows can vary widely depending on the user needs and specific circumstances in which the workflows are set up and need to function. An initial step in this process is to relate the user's needs to specific soil information workflow components. This report considers seven components: 1) needs assessment, 2) data collection, 3) laboratory analysis, 4) soil archiving, 5) data organization, 6) modelling and mapping primary soil data (soil properties/types), 7) applying soil information, and 8) data and information serving, as shown in figure 2.

The optimal choices of a SIS in methods, tools, standards and implementation options for each step of the soil information workflow are determined by the use cases, see Chapter 2 *Soil Information user consideration*, and the SIS profile (Box 1).

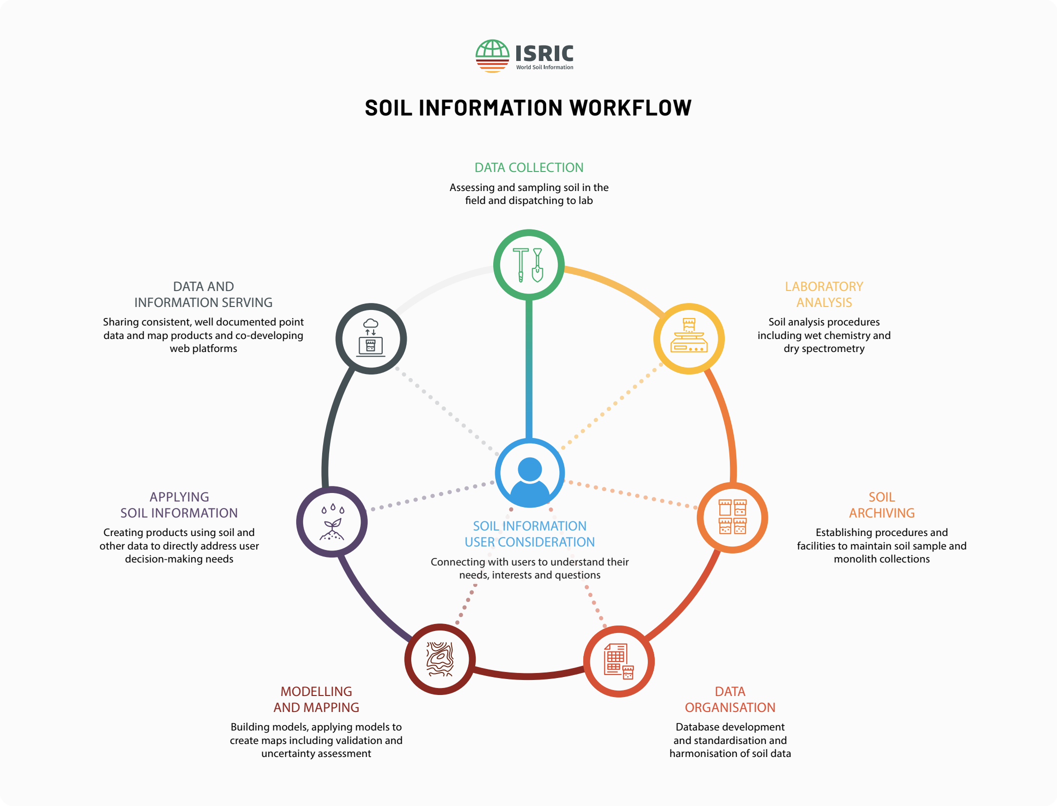
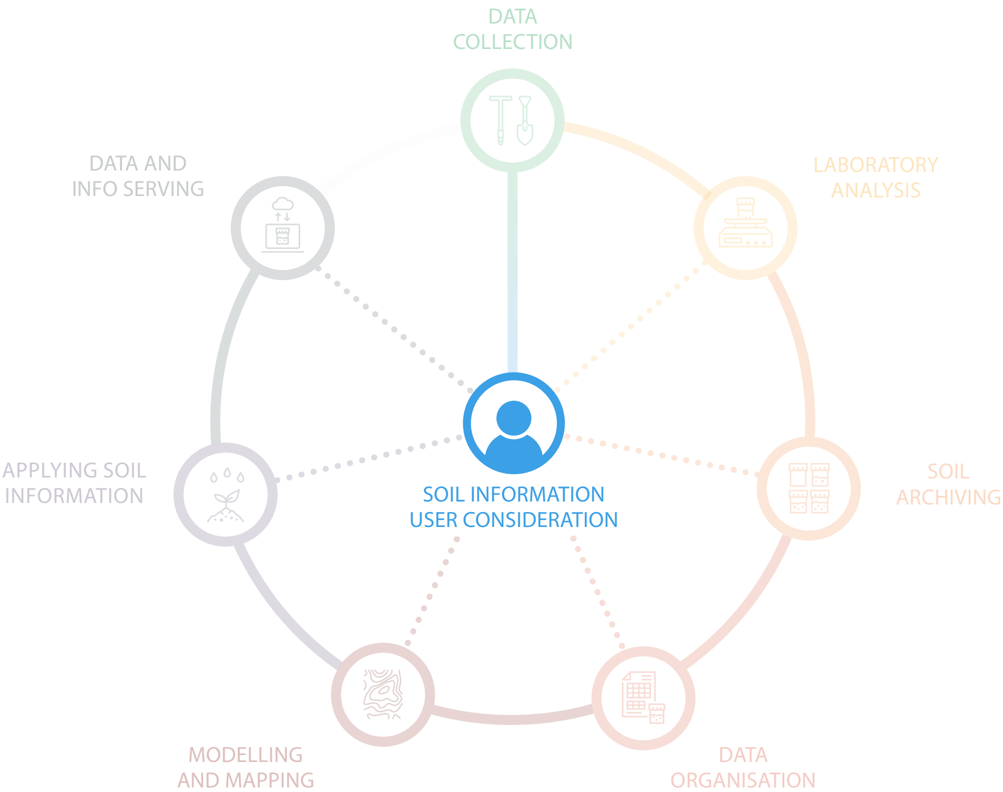

ISRIC
World Soil Information
SOIL INFORMATION WORKFLOW
DATA COLLECTION
Assessing and sampling soil in the field and dispatching to lab
LABORATORY ANALYSIS
Soil analysis procedures including wet chemistry and dry spectrometry
SOIL ARCHIVING
Establishing procedures and facilities to maintain soil sample and monolith collections
DATA ORGANISATION
Database development and standardisation and harmonisation of soil data
MODELLING AND MAPPING
Building models, applying models to create maps including validation and uncertainty assessment
APPLYING SOIL INFORMATION
Creating products using soil and other data to directly address user decision-making needs
DATA AND INFORMATION SERVING
Sharing consistent, well documented point data and map products and co-developing web platforms
SOIL INFORMATION USER CONSIDERATION
Connecting with users to understand their needs, interests and questions

Figure 2 Soil Information workflow components, implementation is usually clockwise.

1.3 Outline

Each of the eight components in Figure 2 is addressed in a chapter of this report. The report starts with a description of the soil information user considerations.


DATA COLLECTION
LABORATORY ANALYSIS
SOIL ARCHIVING
DATA ORGANISATION
MODELLING AND MAPPING
APPLYING SOIL INFORMATION
DATA AND INFO SERVING
SOIL INFORMATION USER CONSIDERATION

2. Soil Information user consideration

The user considerations on soil information are essential when setting up a SIS. The **user considerations or needs** can contribute to strategic, technical, operational, and institutional requirements for a SIS. If the (end) user considerations are not consulted, the success and the sustainability of a SIS could be jeopardized.

Two examples are given here to illustrate this. At a national level, the Ministry of Agriculture has established that there is a need for a SIS which can be used by farmers. This SIS should be a central place for all relevant soil information that farmers can consult. However, the foremost need of the farmer is to have fertilizer or other land management recommendations, not information on basic soil properties. The farmer may find the SIS too complicated, does not know how to use it, or the content might not suit his/her needs. As a result, the farmer will not utilize it, but extension workers and land use planners will. Another example is that a SIS infrastructure is set-up to share soil data, but the data producers do not trust the system or its funders to share their data safely, thus the SIS is not used. Metadata publication functionalities in a SIS could overcome this issue as data ownership and licenses to use the data are clearly stated.

End user needs are regularly overlooked and often not considered during the conceptualization and design of a SIS. The user needs inform the SIS designers and developers about which data is needed, which maps should be provided and how, what functionalities the SIS should have and for which applications the SIS is needed. It is often assumed too soon that the end user needs are known and, for that reason, do not need to be consulted. When not taking the (end) user needs into account, for instance by actively engaging with users, the SIS could fail to serve the purposes and needs of users. User needs are, therefore, the base of the success and sustainability of a SIS.

It is possible that the user needs change during the project. By assessing the user needs during multiple phases of the project, these updated user needs can be taken along in the project to create the maximum impact.

One method to obtain user needs is by means of a **user needs assessment**. A user needs assessment[2] is a process through which user needs are identified. It can help to understand the current situation and to identify gaps. It is a tool for making decisions about how to serve the potential users of an information system (Watkins et al., 2012).

The user needs assessment of a SIS could follow these steps:

1. Investigate potential success and sustainability of the SIS;
2. Define the use cases;
3. Identify potential users;
4. Collect users' needs;
5. Define SIS requirements to address the needs;
6. SIS adoption by end-users.

This chapter introduces the various steps to assess the user needs during the development of a SIS.

[2] https://edepot.wur.nl/537347

2.1 Investigate potential success and sustainability of the SIS

The potential and definition of success and sustainability of the SIS should be investigated within the hosting organization before the development of a SIS. It is important to identify the capacity gaps (skills, knowledge, data assets and digital tools) that need to be addressed to ensure a successful and sustainable SIS. The user needs assessments can help to identify these gaps by asking the following questions:

Goal:

- What problems or needs does the creation of this SIS aim to solve[3]? For example, soil fertility recommendations or soil water conservation.
- What are the strategic and institutional requirements of the SIS?
- How would the SIS be used to guide actions/outcomes?
- Who is/are the primary audience(s) for the SIS[2]?
- How will you ensure the sustainability of the SIS past the initial design and development project?
- How should the SIS enable users to work with the soil information?
- What are potential future problems or needs to be solved?

Funding:

- What budget is available to develop the SIS?
- What budget is available to sustain the SIS past the initial project?
- What additional funding possibilities are there to develop the SIS?
- Is funding present for maintenance, upkeep, product evolution and outreach2?
- Does the SIS require a business model? If yes, what should this model look like?
- How should the SIS enable users to work with the soil information and what is the added value of the system for them, are there additional revenues foreseen for the system?
- Are there any concurring projects/activities that may ensure the sustainability of the SIS?
- What national reporting obligations and commitments are in place?

Development:

- What partnerships participate in the development and implementation of the SIS2?
- Which roles do these partnerships have in terms of their involvement with funding, tool development, data collection, needs assessment?
- Are the contracts of the partnerships in place?
- Is the branding of the SIS agreed upon by the project partners?

Human Capacity:

- What is the current capacity in-country for collecting (or analyzing / archiving / organizing / modelling / serving) data on soil?
- Are there existing laboratories or institutes already collecting these data?
- Is skilled staff available or is training of staff needed?
- Has the staff time to work on SIS development?

[3] Extracted from Survey Instrument final created by UC Davis dd October 2022.

Legal:

- Are data licensing and data sharing policies in place?
- Are additional national policies or legislation needed?
- What is the institutional setting in the country?
- How will governance around the SIS and its data be organized?
- What social, legal, financial, institutional, or public support or limitations are there around the SIS[2]?

Data:

- Is a data catalogue/data ecosystem map available?
- Is there a list of soil datasets and their owners, present and available?
- Do the datasets have similar or known data licenses?
- Is the metadata of the datasets described?
- How are soil samples collected and analyzed: standard soil description guidelines, digital data collection tools/apps, laboratory methods, spectroscopy, other[3]?

Technical capacity[4]:

- What file systems/formats are used for storing spatial data?
- How is data currently stored? In an online repository or on local computers?
- Does the organization publish maps on the internet?
- How are current websites / web applications maintained and hosted?
- Is the equipment present that is needed to host the SIS?
- Are there back-up facilities for the data and system?

The purpose of answering these initial questions is to assess the context of institutions in funding, legal, development, capacity to host, maintain and use a SIS. In other words, to build an initial SIS profile. Once this is well identified and described the choice can be made to create a functional SIS. Once this decision is taken the design and subsequent implementation of the SIS can start along the steps in the soil data workflow.

2.2 Define the use cases

The next step is to define the use cases that you would like to address in the SIS. **Use cases** are defined to provide a thematic focus for the Soil Information System. It describes the key applications for the SIS, the main stakeholders, the key issues that are addressed and the main information that is used. Often a SIS can address multiple use cases, making its implementation potentially more cost-effective. And in time, additional use cases can be added to the initial ones, allowing the SIS to contribute to knowledge and decision-making across a wider set of soil management domains or scenarios.

An example of a use case could be promoting **Integrated Soil Fertility Management (ISFM)**. ISFM helps to improve effectiveness and efficiency of agronomic practices, including fertilizer recommendations and organic matter management, and thereby boost crop production and farm income. Land users and their intermediaries can use the SIS to obtain advice on the

[4] A more detailed question list on technical capacity related to data registry, IT infrastructure and resources is provided in Annex I.

soil fertility status and improvement of soil fertility using spatial nutrient gap analysis based on yield response data. The user needs assessment is hereby important as it describes which information should be in the SIS.

A second example of a use case is Soil Water Conservation. **Soil & water conservation (SWC)** focus on engaging stakeholders to more sustainable land use and land management practices. Catchment managers, authorities, extension staff and farmer organizations can use the system to obtain information on land use and land management practices and their suitability with regards to soil and water conservation (for example erosion) and climate change adaptation. The user needs assessment is hereby important as it defines which information should be provided by the SIS. The two use cases are summarized in the table below.

More examples can be found in chapter 8 Applying soil information.

*Table 1 summary of example use cases**

Key applications for the SIS	Main stakeholders	Key issues that are addressed	Main information that is used
Promoting Integrated Soil Fertility Management	Land users; intermediaries; farmers; farmer organizations	Fertilizer recommendations; organic matter management; crop production; farm income	Soil nutrients (and gaps); yield response; crop data; climate data; soil water availability; cost/benefit information on the measure
Soil and water conservation practices	Catchment managers; authorities; extension staff; farmer organizations.	information on land use and land management practices	Rainfall intensity data; soil texture data; slope gradient; land use and land cover data; cost/benefit information on the practice.

*The examples give a short overview of applications of a SIS. There are many more aspects to consider per use case such as cost/benefit ratio, policies, etc. This is researched in WP1: A review of soil information systems and their history.

2.3 Identify potential users

The following step is the identification of potential users, their role and importance for the projects. This step helps to:

i) Provide focus for the users of data in the need assessment;
ii) Define and cluster users with whom to conduct the workshops and subsequent project activities;
iii) Define, optimize, and target the organization and content of the stakeholder workshops, the nature of the questions to be asked and discussion to be held with users;
iv) Ensure the needs assessment fits in the institutional landscape and relates to past and on-going initiatives and development thinking.

This step is in support of the collection of the users' need. The identification of potential users can be performed as a desk study and by soliciting the network of key stakeholders in a country.

2.4 Collect users' needs

Once the potential users are identified, the user needs can be collected. There are various tools to assess the needs of a soil information user. *A guide to Assessing Needs* (Watkins et al., 2012) gives an elaborate guideline on user needs assessment with detailed description of the methods. In addition, CGIAR created an User Research Toolkite which lists user (needs) research methods including suggested time, required expertise, materials and participants. Four examples are described below.

Stakeholder workshops

A stakeholder workshop is a meeting of different users to identify the users' roles and the challenges and opportunities of a SIS, to specify information needs and information users, to identify capacity requirements for SIS use, to involve these stakeholders and to identify policies and initiatives related to the use cases. An example of a stakeholder workshop layout is given in Annex II.

Key informants' interviews

Interviews are an informative method to learn about user needs. Interviews can be used to gather specific information from key informants. However, they are labor-intensive[5]. An example of an interview protocol is provided in *A guide to Assessing Needs, pg. 106* (Watkins et al., 2012).

The questions can vary from questions on the specific use case, data, and functionality of the SIS. Furthermore, different questions can be asked to data users and to data suppliers. Examples of questions relevant to SIS development are added in Annex III.

Focus group discussions

A focus group discussion[6] gathers people from similar backgrounds to discuss a specific topic. Through this method, multiple people are interviewed simultaneously, and they can hear each other's views. However, not everybody might contribute equally or feel comfortable doing so[4]. An example of a Focus Group Discussion protocol is provided in *A guide to Assessing Needs, pg. 95* (Watkins et al., 2012).

Surveys

A survey or questionnaire is a method to collect information from many people. Surveys can be useful for need assessments as they are easy to develop and easily distributed (Watkins, 2012). Using surveys can improve the representativeness of a user need assessment. An example of a Focus Group Discussion protocol is provided in *A guide to Assessing Needs, pg. 116* (Watkins et al., 2012).

The objectives of the user assessment, and the decisions to be based upon the information collected, should guide the number of interviews to be conducted. This can be approached through using the concept of data saturation, which is widely used in qualitative research to assess whether a number of interviews or focus group discussions is sufficient to address the researcher's objectives (Watkins et al., 2012). Saturation can, in general, be evaluated by

[5] https://edepot.wur.nl/537347
[6] 1485497050-Focus Group Discussion_0.pdf (herd.org.np)

considering how much added information is gleaned from each successive interview; with greater numbers of interviews conducted among members of a given SIS or soil data community, the yield of added information from each new interviewee will decrease (Hennink and Kaiser, 2022). Depending upon your objectives, it may be useful to pursue this strategy subjectively or quite systematically, and specific methodological frameworks exist to guide this process (e.g., Guest et al., 2020).

2.5 Define SIS requirements to address the needs

The collected needs should be translated into **SIS requirements**. By analyzing the information, you will obtain insights into priorities of users, and which gaps the SIS could address. A successful needs assessment understands the needs, but also identifies the functional and technical requirements of the SIS to answer to the identified needs. These requirements propagate in all steps of the soil information workflow. The needs and requirements could be different from your initial idea. Therefore, it is essential to keep an open mind and flexibility, to ensure you are creating a SIS for the users.

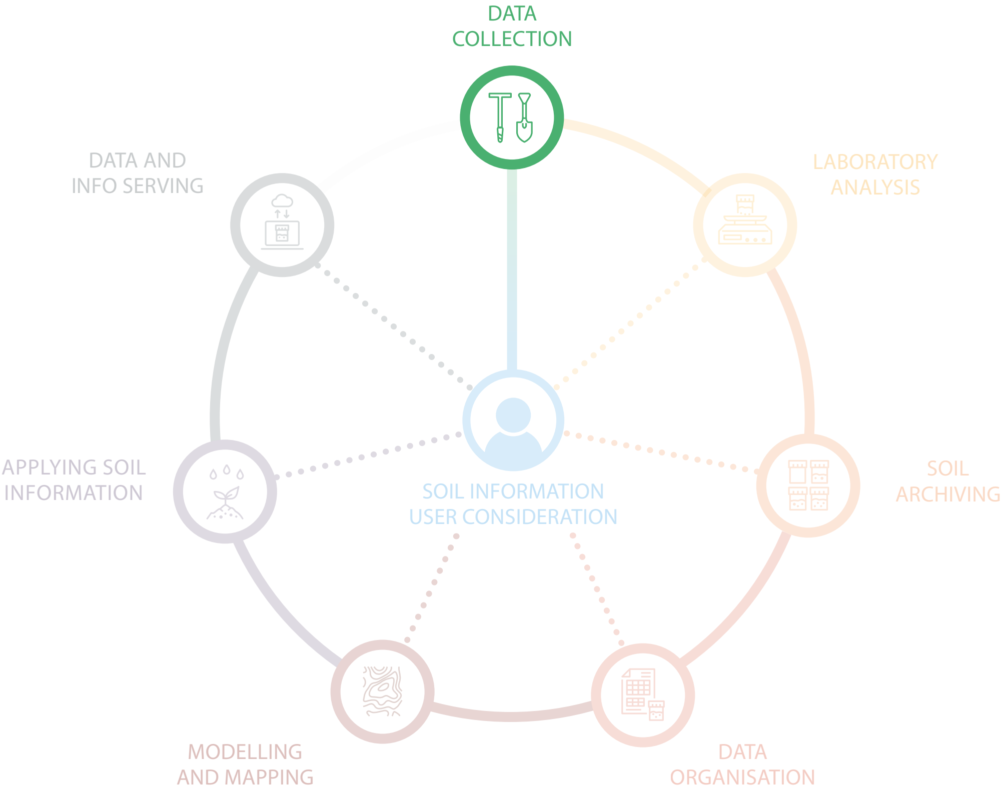
Once the SIS requirements are defined based on the needs, the input is essential for every step of the Soil Information Workflow, as shown in figure 2. As mentioned before, the user needs inform the SIS which data is needed, which maps should be provided, what functionalities the SIS should have and for which applications the SIS is needed.

2.6 SIS adoption and sustainability

The user needs assessment is one part of the engagement with the end-user throughout the whole process. The other parts include:

1. Assessing institutional capacity needs;
2. Applying a test phase with actual users of the system;
3. Organizing adoption workshops using participatory approaches, where users learn to use the system;
4. Ensuring SIS sustainability with capacity building in all phases of the soil information workflow.

By also implementing the other parts, all end-users (data and information producers, as well as information users) can adopt and use the SIS. Capacity building of the end user is extremely important for the sustainability of the SIS. This will be discussed in more depth in other deliverables of this project.


DATA
COLLECTION
LABORATORY
ANALYSIS
SOIL
ARCHIVING
DATA
ORGANISATION
MODELLING
AND MAPPING
APPLYING SOIL
INFORMATION
DATA AND
INFO SERVING
SOIL INFORMATION
USER CONSIDERATION

3. Soil data collection

Once the use case(s) and the aim of the Soil Information Systems (SIS) are defined by all stakeholders, including data providers, funders and users, the development of the SIS starts with the collection of soil data. This can entail getting access to existing data, collecting new data in the field, or a combination of both. Collection of existing or legacy soil data is addressed in chapter 6 'Data organization' and is briefly described in this chapter Deciding on new data neededas well in the context of defining (new) data needs. The new soil data collection process can be subdivided in two stages:

1. **Design of a field campaign:** from defining exact data needs, the methods, standards, and protocols to use, to planning the logistics of the field campaign including anticipating changes in the design.
2. **Execution of field work:** collecting data and samples up to shipping them to a lab and uploading the field data on the designated servers for data organization and including documenting any changes made during execution compared to the design. Important is that the digital data collection tools integrate with server based/central database of the SIS to ensure online data uploading.

This chapter introduces various standards and tools available to facilitate the collection of new soil or soil-related data that can be considered during the design of a field campaign.

3.1 Overall design steps for a field campaign

The design of a field campaign consists of different steps. Several of these are outlined in more detail in the subchapters below. The first sequential steps are:

- Collect, compile, and review **existing data** or prior information (see chapter 6 on Data relevant for designing the field campaign. These data, general knowledge, experience, and information provide information on the characteristics and variation of soils and landscapes in the study area, which can be used for stratification of the area for sampling or limit the amount of new data to be collected. Existing or legacy data can also be used to guide the field sampling;
- Define **additional data needs** (chapter 3.2), taking into account type (point, map, statistics for (parts of) the area of interest), extent of the area of interest, depth of sampling, scale or spatial resolution (detail or granularity of the to be provided soil information), temporal resolution (when and how often), target parameter (e.g. soil property, class, function);
- Make a choice between **design-based and model-based statistical inference**. The first is typically used for estimation of spatial and temporal aggregates (e.g., spatial, or temporal means or totals), the second for prediction at points in space and/or time (i.e., mapping);
- Define **constraints**: the allocated budget and/or minimum required accuracy measure of the result, fieldwork, fieldwork method constraints, accessibility, transport, and laboratory capacity;
- Decide on **sample support**: which is the area, volume, or time period over which an observation is made. For instance, a soil sample might be a single volume of soil taken at a sampling location and depth or a composite soil sample consisting of multiple single

soil samples taken within a site. Likewise, sensor measurements, such as soil nitrous-oxide emission, can be measured over short and long-time intervals and over small and large areas, which are considered different supports for the measurement;

- Decide on **observation methods** (direct soil observations, sampling and/or sensing) (chapter 3.3), platforms (direct, proximal, UAV, remote) , **registration method** (paper, mobile app) and **sample labeling and tracking method** (bar code-based labeling and tracking system);
- Decide which **field protocols and standards** to use. Draft or adapt protocols and procedures if an existing standard cannot be used. Consider any privacy regulations (e.g., in Europe GDPR[7]) or voluntary protection of privacy sensitive data (such as names, addresses, in several countries coordinates (Fantappiè et al., 2021));
- Decide on a **soil sampling/observation scheme** (chapter 3.4.1) and select the sampling locations (different aims require different approaches). In case of design-based statistical inference: choice of sample size and sampling design type, such as stratified random sampling or cluster random sampling. In case of model-based statistical inference: choice of sample size, sampling pattern type and associated optimization algorithm, such as a nested hierarchical sampling to estimate spatial structure, regular grid sampling, or conditioned Latin hypercube sampling. Examples are outlined in Brus (2019) and Soil Survey Manual (2017);
- Decide on the **timing and feasibility** of the campaign - depends on road accessibility of the sampling area, crop season, weather, drivability, hard soils, and the security situation of the area);
- Consider a **time** aspect (if the sampling campaign should be repeated over time) depending on the soil properties or characteristics that are mapped, different frequencies of sampling make sense. For nutrients in agricultural land the expectation is that the levels change rapidly and monitoring every year or couple of years can be useful. For texture in non-eroding forested areas, changes are not expected within time ranges of 10 years or even more, meaning a larger time in between sampling is advisable. This is typically considered in soil monitoring system design.;
- **Budget** the campaign (drafting a realistic budget, accounting for risks), and if needed make changes in the above steps.

Once this is done these steps can be done consecutively:

- Obtain **access permissions** (often private land needs to be sampled which requires permission by the landowners, standard forms and letters issued by governmental institution can be used, rules vary per country);
- Define **surveyor assignments**, e.g., which area and how many samples per person, and perform workload management (depends on the task, the number of trained surveyors available and the time available and the maximum duration of the sample period);
- Establish **dataflows** (who collects which data and how, when, and where are the collected data stored and transferred to a central repository, requirement of regular backups, how is quality control organized, how will the field and lab data be ingested in the SIS?);

[7] https://gdpr-info.eu/

- Design an approach for **labelling and registration** of soil samples (each soil sample must be labelled consistently with a unique and recognizable code, preferably aligned with lab code on non-wearable material);
- Establish a **list of equipment** and materials needed;
- Organize **transport** of surveyors and soil samples (this may require permits[8]).

3.2 Deciding on new data needed

The decision that additional data is needed in a SIS depends on the aims and use cases of that SIS, which may shift over time. Once the current aims and use cases are defined, the existing and available/accessible soil data can be evaluated against their fitness-for-intended use. This is easier when the soil information is wellorganized and FAIR (see Chapter 6 on Data .

Existing or legacy data

Existing data typically comes in many shapes and sizes. This ranges from well-organized data in libraries, portals and repositories (5.2.3 Libraries, Reference institutes and Museums) and soil information systems, to stored on an institutional, company or personal server or hard drive, with or without metadata or adhering to a data model or in a (less-) common format. It an also come in the form of printed reports and maps, digitized into pdf or other formats. Typical steps to disclose this information and incorporate it into a SIS for future use are to; scan or digitize any analog formats, to georeferenced maps and/or polygons, annotate reports, maps, data with metadata, transposing any digitized point data to a tabular or other format, to transform digital data to a common and standardized format and data model if applicable, to define a license and upload it to a persistent repository or SIS. It is advisable to have a prioritization process and standardized workflow in place when curating a large legacy data set into a SIS.

Fitness-for-intended-use

Only when data and information are insufficient, outdated, or not freely shared, then additional data collection is needed. One exception to this is monitoring data, this is typically a repeated sampling or other observation of the same property and space at different moments in time. A useful approach to evaluate the suitability of existing data and designing a soil data collection scheme is outlined in *Sampling for Natural Resource Monitoring* (De Gruijter et al. 2006), the most important components of which are:

1. Detailed description of the **objective** of the scheme or information requirement:
   - Target universe and domain of interest: respectively the outer boundaries of the target area and period, and the sub-areas and periods within the target universe for which information is required.
   - Target variables: variables for which information is desired. These can be qualitative, such as soil type and suitability class, or quantitative, such as soil pH and clay content.

[8] SIMPLE - Soil Import Legislation is a tool for global soil import rules available at https://www.fao.org/global-soil-partnership/glosolan/en/

- Target parameter: the type of statistic that is desired given the target variable and domain of interest, such as the spatial or temporal mean or median, or the fraction of points in the domain of interest where the target variable is above a critical threshold.
- Target quantity: the combination of a domain, target variable and target parameter, such as the median (parameter) nitrate concentration (target variable) in the subsoil of all agricultural soils in the Netherlands in the year 2020 (domain).

2. Accuracy measure: the quantity used to express the statistical quality of the soil survey or monitoring result. Examples for quantitative data include the width of a 90% confidence interval in estimation, the mean prediction error variance in soil property mapping, or the probability of correct classification in soil type mapping. See also the *Soil sampling quality assurance user's guide* of the US Environmental Protection Agency. For qualitative data this can be, for example, the percentage correctly classified.

3. Accuracy requirement: a threshold on the accuracy measure.

Examples of the application of this method are described in Knotters and van Egmond (2018). After evaluating the fitness-for-intended-use as described above of existing soil data and information, additional data collection needs and options can be identified and evaluated by the value of information method (De Bruin et al., 2001). This method quantifies or qualifies the value of additional information to improve the answer to the question at hand in a use case. It is a variation on a cost-benefit analysis. The scheme of De Gruijter et al. (2006) can be adopted for this purpose if the analysis shows that additional data are required.

3.3 Soil observation methods

Once the data or information needs are defined, the most suitable measurement and observation methods can be determined. These can consist of easy tests and observations in the field that do not require much training, observations, and descriptions by trained soil surveyors, taking physical soil samples to ship to a laboratory. But they can also consist of in situ or proximal sensing with static, handheld, or towable/mountable sensors, UAVs or drones with sensors and cameras, airplanes, and satellites with sensors to measure various parts of the electromagnetic spectrum. From UAV to satellite is usually referred to as remote sensing or Earth Observation. Especially the domain of interest, the target variable, type of result and accuracy requirement determine the choice of method (see chapter 3.2 Deciding on new data needed).

Selection criteria
When selecting **an observation method** or several methods, the approach by de Gruijter et al. (2006) may also be followed, but now not for evaluating data needs but for defining the requirements for observation techniques. A practical approach is to:

- Firstly, define the **target variable(s)** (e.g., clay content or soil classes).
- Secondly the **domain of interest** (plot, field, farm, watershed, province, country, etc.), later on also referred to as area of interest.
- Thirdly, consider the desired **accuracy measure and requirement** depending on the use case. A common line of thinking is that new data should always be as accurate as pos-

sible. However, this usually increases the costs of data acquisition and is therefore not always the best choice. For example, determining a change in soil organic carbon content requires many accurate measurements, whereas defining management zones in a field requires a reliable pattern, but often a proxy measurement (of for example electrical conductivity or gamma-ray total counts) is sufficient.

- A fourth consideration is the **availability and applicability** of soil data acquisition methods of interest. Some sensors, services, or lab methods may not be available in all countries or regions. Some techniques are applied by driving over the area (field or bigger) which is not possible when crops are on the field or in the forest. Visible and near infrared satellite applications are hindered by clouds.
- A fifth consideration is **the costs of data acquisition** and the estimated benefit of the data for the use cases/number of uses. This can be a monetary (higher yields) benefit, risk mitigation (yield loss reduction, water, and food security, maintaining or improving biodiversity), a general benefit for ecosystem services to a society or otherwise.

Based on these considerations several options will remain. A selection of the methods is detailed in paragraph 3.2, and in Annex IV: Common lab, proximal and remote (soil) sensing methods. It is worthwhile to evaluate what the co-benefits of the methods are, such as providing data for multiple soil properties at the same time beyond only the target variables, an educational benefit for land users in an increased soil awareness, flexibility in the execution to increase or reduce the effort based on initial findings and changing circumstances, dependency on on-the-ground data collection and desirability of this given the security and accessibility of the area.

Table 2 Observation methods

Observation methods	How	Results
Simple field observations	GSP Soil Doctors[9] guides, Cornell pH toolkit[10], (national) Visual Soil Assessments[11][12]	Range of mostly descriptive properties of soil, described in classes that provide an indication of soil composition and health.
Soil profile descriptions	Description of soil morphology and other features (e.g., fauna, roots) using established protocols, such as the NRCS handbook for describing and sampling soils (Schoeneberger et al., 2012), the NRCS Soil Survey Manual[13] the FAO guidelines for soil description[14] or the WRB (4th ed) field guide[15].	Complete identification and description of a soil profile, its layers, and a broad range of soil properties per layer. It provides the basis for many futures uses and interpretations.

[9] www.fao.org/global-soil-partnership/pillars-action/2-awareness-raising/soil-doctor/en/#c853850
[10] https://cnal.cals.cornell.edu/ph-kits/
[11] https://doi.org/10.1016/j.still.2017.11.012
[12] https://www.isqaper-is.eu/soil-quality/visual-soil-assessment
[13] https://www.nrcs.usda.gov/resources/guides-and-instructions/soil-survey-manual
[14] https://www.fao.org/3/a0541e/a0541e.pdf
[15] https://www3.ls.tum.de/boku/wrb-working-group/, https://www.isric.org/explore/wrb

Table 2 Observation methods - continued

Observation methods	How	Results
Field quantitative estimations	Assessment of features, such as (percentage) mottling, rooting, coarse fragments, clay-, sand-, silt content, acidity, infiltration capacity.	These features can be used in keys for assessing soil types (classification), for soil assessment, pedotransfer function application and for soil (e.g., hydraulic) modelling.
Field sensors point measurements*	With handheld or on-the-go proximal sensors conducting point measurements on the surface of a soil profile.	Options and results: see Annex IV: Common lab, proximal and remote (soil) sensing methods
Field soil sampling for lab analysis	Sampling of a layer (standardized depth range) or horizon (soil profile description) with a soil auger or spade for disturbed (physical and chemical analysis) or undisturbed (soil hydrophysical soil property analysis) in the lab. (NRCS handbook: Schoeneberger et al., 2012)	Disturbed or undisturbed soil samples to be analyzed in the lab (see 4.1 Methods for laboratory analysis)
Field soil sampling for biological analysis	Disturbed soil sampling with a soil auger or spade. Samples are stored cooled as soon as possible (Lane et al., 2022)[16].	Disturbed soil samples to be analyzed in the lab on soil biological properties (see Methods for laboratory analysis
Remote sensing methods	Remote measurement of soil or soil related properties from a UAV, airplane, or satellite	Direct soil property estimation, or often vegetation or elevation related pattern information as sensor to soil patterns underneath

*For most if not all of the proximal and remote soil sensing methods calibration data are needed to derive the final result. These are often soil point observations or lab results or soil profile descriptions that are modelled (e.g., using linear regression, PLSR, machine learning) with the measured sensor values to derive the (core)relation between sensor measurement and soil parameter or target variable.

3.4 Soil sampling

3.4.1 Soil sampling design

When soil samples need to be collected, an important choice is how to select sampling locations from the geographic area and/or time period of interest for which inferences or estimations of target variables (soil properties or soil quality indicators) need to be made. The area/period of interest is sometimes referred to as *target universe* or *population* in sampling theory. The selection of sampling locations is defined by a **sampling design**. The output of a sampling design is a list of (geo-referenced) sampling locations that can be assigned to surveyors.

[16] https://doi.org/10.1016/j.soilbio.2022.108858

Sampling designs may vary depending on the purpose for which the samples are collected. Thus, designing a sampling scheme starts with clearly defining the aim of the soil survey (see Section 3.2 and De Gruijter et al., 2006). More recently, Brus (2022) distinguished three broad aims:

1. Estimation of target parameters (such as means, totals, fractions) of selected target variables for the target universe;
2. Estimation of parameters of selected target variables for several, separate *domains of interest* or *subpopulations* (for instance different land use classes, agro-ecological zones, or farming systems within the target universe);
3. Mapping a target variable across a geographic area and/or time period.

Sampling schemes for 'mapping' (aim 3) can also be used for training other types of prediction models such as calibration models for soil spectroscopy.

There are two main sampling options to choose from when designing a sampling scheme: probability sampling or non-probability sampling, also referred to as 'purposive sampling'. The choice depends on the aim of the soil survey.

Probability sampling is preferred when the aim of the survey is to estimate statistical parameters for the target universe or domains within this universe (aims 1 and 2). Examples include estimation of the total soil organic carbon (SOC) stocks in a country, the average SOC stocks per land cover type in a country or the magnitude of change in soil pH over time after a soil management intervention in an agricultural landscape. Probability sampling designs randomly select sampling units with a (pseudo)random number generator. Probability sampling has the advantages that it provides unbiased estimates of population parameters of (soil) properties of interest and the associated uncertainty, and that it does not rely on models and hence model assumptions that could be questioned. Hence it is an optimal choice when the aim is soil monitoring. The main disadvantage is operational, as it puts stringent requirements on fieldwork. Surveyors must visit the preselected sampling locations and are not allowed to select locations themselves that might be more convenient to visit. When a sampling location cannot be sampled it must be replaced by a new sampling location from a pre-defined back-up list. This means that the surveyor might need to travel a substantial distance to the new location.

Purposive sampling is preferred when the aim of the soil survey is (digital soil) mapping, or calibration of prediction models that can include spectral or other sensing models or (mechanistic) process models that rely on soil data (aim 3). Purposive sampling designs typically optimize sampling locations in geographic or feature space in such a way that it gives the most accurate map (or more general, calibrated prediction model) possible given the sample size. Furthermore, surveyors can be more flexible with moving sampling locations in case selected locations cannot be sampled. The main disadvantage is that models (and thus model assumptions) are required to estimate parameters of target variables (including change detection over time) of which the unbiasedness and validity can be questioned.

De Gruijter et al. (2006) wrote a standard handbook that contains a wealth of information on designing sampling schemes for survey and monitoring. Brus (2022) provides a comprehensive and thorough overview of spatial sampling[17] designs based on probability and purposive sampling, with examples of the statistical software R. A condensed summary on sampling approaches for mapping and monitoring is provided in an EJP SOIL[18] report (Teuling et al. (2021). An inventory on European soil monitoring systems that shows a cross section of options for this has been compiled by EJP SOIL as well (Bispo et al., 2021).

Field campaigns considerations
For field campaigns, some considerations can contribute to an efficient field campaign. It is advisable that surveyors have a digital or paper printed protocol and the sampling layout with them at all times during the fieldwork.

Another important aspect is the bagging and labelling for quality assurance. Labelling and bagging should be done consistently to ensure each soil sample is correctly preserved and identifiable in the entire process from field to the laboratory, see (Huising and Mesele, 2022a). During a field campaign, it is advised to use sample chain-of-custody logs to track when samples change hands among field/laboratory staff as they move from field to lab. It helps with tracing what happened in cases of lost or damaged samples. In addition, soil surveyors and researchers should include sufficient replicates in the set of soil samples that are shipped to the laboratory for the error estimation, which should be anonymized and randomized before shipment (Van Leeuwen et al. (2022)), see also section 4.1.6.

Avoid as much as possible handwritten codes in the field to avoid errors in the code samples chain; pre-printed, waterproof labels with understandable codes or QR-codes can be used for this. It would be advisable to double bag samples in the field if a lot of transport or handling is expected. Bags may tear and samples may be lost or rendered useless.

3.4.2 Soil sample collection, sampling protocols and procedures

Once sampling locations are selected, soil samples can be taken. The collection of soil samples is often done:

- At **fixed depths**, irrespective of pedogenetic boundaries in the soil profile. Examples are the 0-30 cm depth interval used in LUCAS Soil 2022, 0-20 cm in LUCAS 2009, 2015, 2018, or the 0-20 cm and 20-50 cm depth intervals used for the AfSIS program in Africa.
- By **pedogenetic horizons**. Samples are taken from layers with varying depth and thickness based on uniformity of morphological soil properties as identified by an expert during soil profile description.

The choice of sampling depth strategy and other aspects of sampling (e.g., composite, or single samples) depends on the aim of the campaign. The usability of the results and success of the campaign depends on the consistency and quality of sampling and application of used standards and protocols.

[17] https://dickbrus.github.io/SpatialSamplingwithR/
[18] https://ejpsoil.eu/

For some soil properties direct measurement in the field or lab can be difficult or expensive, such as available water capacity or bulk density. In such cases pedotransfer functions can be derived and used. Pedotransfer functions are functions that predict secondary soil properties from (readily available) measured properties. There are various global and regional pedotransfer functions (e.g., Wosten et al ., 2013; Saxton and Rawls, 2006) that require consideration to determine their applicability in specific environment settings, as the performance of these functions depends on which it is developed and tested.

Standardized and consistent methods should be used for soil data collection and sampling that will result in a uniform sample set. Internationally several (de facto) standards for soil survey and soil sampling protocols exist:

National

Each country can have its own national soil sampling protocol. A well-known standard that is widely used internationally is the US *Soil Survey Manual*.

LUCAS Soil

The European Commission launched a soil assessment component in extension of the land cover survey named Land Use/Land Cover Area Frame Survey (LUCAS). It uses standard protocols for collecting soil samples and for performing field measurements in the soil survey.

Soils4Africa

The EU H2020 project *Soils4Africa* developed a protocol for field survey and standard operating procedures for soil sample collection and field observations for a continental survey of soil conditions in Africa's agricultural land (Huising et al., 2022, Huising and Mesele, 2022a, Huising and Mesele, 2022b), that builds on the principles of the LUCAS survey. The manuals are supplemented by several protocol videos in English, French and Arabic.

ICP Forest

The International Co-operative Program on Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forests) has developed harmonized and standardized methods for forest soils (ICP Forests Manual).

FAO Guidelines for soil description

A de facto standard for (international) soil description is provided by the FAO with the Guidelines for Soil Description FAO, 2006).

WRB Field guide

Field description guide made available by the IUSS WRB WG[19] to assist soil description for soil classification according to the WRB (World Reference Base), version 2022.

AfSIS field guide

In the framework of the AfSIS (African Soil Information System), a field guide has been developed and applied in several countries in Africa.

[19] https://www3.ls.tum.de/boku/wrb-working-group/, https://www.isric.org/explore/wrb

ISO

A range of soil descriptive and sampling standards exist for environmental or soil contamination data collection. These are often aimed at (human or environmental) risk assessment and are applied in built-up or urban and industrial areas. The main standardizing body for this internationally is the International Organization for Standardization (ISO).

- the ISO 18400-101:2017[20] is developed as a guidance on the design of sampling programs for soil in Europe. It is applicable to the sampling of soil and soil material for soil in the landscape, soil stockpiles, potentially contaminated sites, agricultural soils, landfills, and forest soils;
- the ISO 25177:2019[21] provides guidance on the description of soil in the field and its environmental context. It is applicable to natural, near-natural, urban, and industrial sites;
- the ISO 14688-2[22] is the basis of classification of those material characteristics most commonly used for soils for engineering purposes. In engineering, at what depth the layers sand, clay and peat lie are important, but the subdivision in each layer is less relevant. It is applicable for geotechnical or construction soil descriptions.

The best practices are developed by the domain and written down in (de facto) standards and protocols. The **best practices** identified by the soil science domain during sampling campaigns are often described in survey protocols and typically include logistical considerations and organization as described in chapter 3.1 Overall design steps for a field campaign and 3.2 Other tools and good practices for field data collection.

3.4.3 Full soil profile description and assessment of site and soil conditions, protocols, and procedures

In addition to sampling, it can be required to provide a full soil profile description and assessment of site and soil conditions. The description of the soil profile is the basis for soil classification and soil class mapping, which is grouping soils with a similar range of properties into units[23]. From those units governing soil processes, properties and functions can be derived. The occurrence and coherence of soil types, classes or units in an area can provide an understanding of the link between soils and land use and landscape. If needed more precise data on specific soil properties can be acquired by soil sampling (see chapter 3.1.1 Soil sample collection, sampling protocols and procedures) and property mapping (see chapter 7.1.2 Digital soil mapping concepts). More information on landscape based soil class mapping is provided in Chapter 5 of Van Egmond and Fantappiè Eds. (2021) which is Arrouays et al. (2021).

The most important guidelines for soil profile description and classification are:

FAO

The Food and Agriculture Organization of the United Nations (FAO) published the internationally recognized guidelines for soil description. The guideline provides a complete procedure for soil description and for collecting field data (Jahn et al, 2006).

[20] https://www.iso.org/standard/62842.html
[21] https://www.iso.org/standard/69585.html
[22] https://www.iso.org/standard/66346.html
[23] http://www.fao.org/3/a-az922e.pdf

WRB

The World Reference Base (WRB) is an international system for classification of soils[24] . The WRB field Guide of 2022, which is included in the 4th edition of the WRB, provides all field characteristics needed for WRB 2022 classification and some other general field characteristics.

USDA

The U.S. Department of Agriculture (USDA) provides a USDA Soil Taxonomy to interpret soil surveys in the USA. They also have a field guide for soil descriptions and sampling soils performed in the USA (Schoeneberger, 2012).

3.2 Other tools and good practices for field data collection

A complete list of field equipment can be found in the Protocol for Field Survey of the Soils4Africa project in Soils4Data logging. During field data collection itself, different computer-based software tools can be used for ease of registration. The choice of tool will depend on the capacities of the survey staff and infrastructure. The project aims to develop a decision support instrument for this and other decisions later on.

ODK Central (server) and ODK Collect (mobile app)

The Open Data Kit (ODK) central[25] can be used to manage users' accounts and permissions, store form definitions, and allow data collections with ODK collect. ODK collect is an open-source Android app that replaces paper forms used in survey-based data gathering. It can be used when collecting new data and metadata and has an excel editable form backend. An implementation of ODK using the FAO Guidelines for soil description is the Soil Description DevTool[26].

KoBotoolbox (server) and KoboCollect (ODK compatible mobile app)

KoBoToolbox[27] is a tool for collecting and managing data in challenging environments such as humanitarian emergencies (KoBoToolbox, 2023). The KoBotoolbox is comparable to ODK central: both can be used or a combination. KoBoCollect is the corresponding android-app based on ODK Collect. It also offers the possibility to protect privacy sensitive data upon collection.

Qfield

Qfield[28] is the professional mobile app for QGIS, allowing users to deploy their existing projects in the field. This can be used when you want to edit your data in QGIS in the field.

Commercial (subscription-based services) data logging

- **OnaData**
  OnaData[29] is a mobile data collection platform used for data collection and real-time monitoring. OnaData can be used in humanitarian work.

[24] https://www.isric.org/explore/wrb
[25] https://docs.getodk.org/central-intro/
[26] https://www.isric.org/news/soil-description-devtool
[27] https://www.kobotoolbox.org/about-us/
[28] https://qfield.org/
[29] https://ona.io/home/

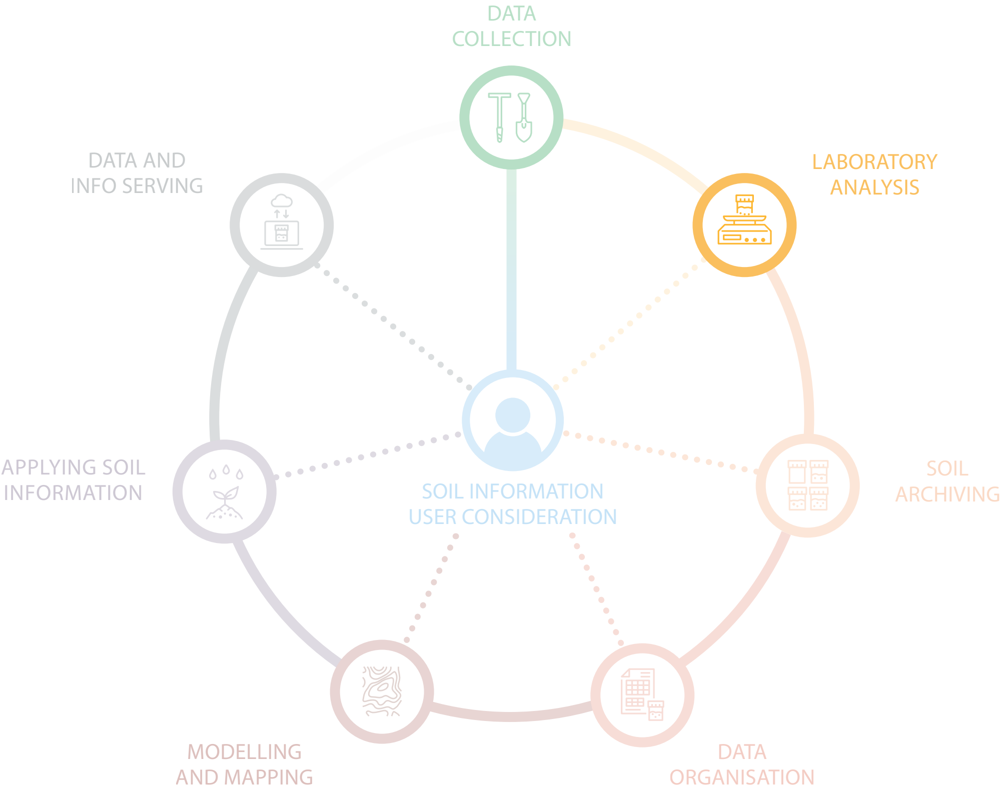
- **CommCare**
  CommCare[30] is a digital platform for frontline work for collecting data. It can track data, work offline.

- **SurveyCTO**
  SurveyCTO[31] is a mobile data collection platform for working in offline settings. SurveyCTO is based on ODK with improvements on hosting, documentation, and support. It can be used when working offline in the field.

- **ArcGIS Survey 123**
  ArcGIS Survey123[32] is a simple form for data gathering. It can be used to create, share, and analyze surveys.

Forms (digital or paper)
Paper forms following standards can be used for collecting data. These forms can later be transcribed to a digital format at the office. This method includes a risk of spelling errors, readability issues and non-valid field descriptions since the standards are not enforced by any software.

[30] https://www.dimagi.com/commcare/
[31] https://www.surveycto.com/
[32] https://survey123.arcgis.com/


DATA COLLECTION
LABORATORY ANALYSIS
SOIL ARCHIVING
DATA ORGANISATION
MODELLING AND MAPPING
APPLYING SOIL INFORMATION
DATA AND INFO SERVING
SOIL INFORMATION USER CONSIDERATION

4. Laboratory Analysis

After the collection of soil data and samples, the development of the SIS continues with laboratory analysis on the soil samples. Soil samples collected during field surveys or monitoring programs need to be analyzed in a consistent way to permit sound interpretations. This can be done in laboratories that analyze soil. The aim is to deliver consistent and comparable results of sufficient or high quality within and between labs in time and on a wide range of soil properties. To this end many lab methods have been and continue to be developed and standardized into standard operating procedures (SOPs). Most reputable labs have adopted quality assessment and control procedures, and several tools and communities exist that aim to standardize methods, improve quality, and provide capacity building.

This chapter provides an overview of categories of lab methods and introduces various standards and (digital) tools available to facilitate good quality soil lab analysis.

4.1 Methods for laboratory analysis

Laboratory methods for soil analysis can be subdivided into traditional laboratory methods ('wet chemistry') and spectroscopic measurements that characterize the samples' mineral/organic composition ('dry chemistry'). Estimates of concentrations of soil constituents derived from spectroscopy are, depending on the soil property estimated, often a bit less accurate than estimates obtained using wet chemistry data. Nonetheless, spectroscopy can be cost-effective when a large number of soil samples must be analyzed, despite the costs and limitations of necessary calibration against wet chemistry data (Shepherd et al, 2022). More information about infrared spectroscopy as a technique is detailed in chapter 4.1.4 Infrared spectroscopy lab analysis methods, in Annex IV: Common lab, proximal and remote (soil) sensing methods and in Shepherd et al. (2022).

Traditional, conventional lab methods are often referred to as wet chemistry methods, although not all require extractant solutions, some use (non-infrared) spectroscopic techniques and the term groups a wide range of methods together. A common way to subdivide the analytical methods is according to the soil properties they describe, that is whether these are chemical, physical, or biological properties. This chapter describes these three groups of wet chemistry lab methods, followed by a subchapter on soil spectroscopy.

4.1.1 Traditional chemical lab analysis methods

Soil is often analyzed to determine its ability to supply the necessary plant nutrients for a given crop, or to assess pollution levels. Soil analyses can be related to, amongst many others, potential nutrient uptake, supplementation of plant nutrients through fertilization and the target yield. Specific analyses will be needed depending on the question at hand, see for example the ISO standards for chemical and physical characterization, USDA-NRCS Kellog Soil Survey Laboratory Manual, ISRIC procedures for soil analysis, ICRAF standard operating procedures or GLOSOLAN Standard Operations Procedures (FAO, 2008). In general, the soil analyst will choose an extractant (e.g., KCl, oxalate, Mehlich solution, (hot) water) to dissociate the nutrient of interest from other components in the soil, so that this nutrient can then be measured. These pools differ from each other in plant availability or relevance depending on soil type and processes, see Fixen and Grove (1990) and Elrashidi (2010). Lab techniques that are used in chemical analysis of soils include XRD, XRF, mass spectrometry, AAs, ICP, AES, etc. For a detailed description see Soil Survey Staff (2022).

4.1.2 Traditional physical lab analysis methods

Physical soil tests are used to determine the texture and structure of soil, giving insight in their suitability for agriculture, forestry, or a foundation for construction. This information is important for making effective land-use planning and management decisions. Examples of lab techniques for physical soil property analysis are the pipette method or the laser diffraction method for clay content (Svensson et al., 2022), the (dry or wet) sieve method for grain size distribution, the pressure pan method for water retention characteristics, and soil core method for bulk density. Examples of protocols for these and other techniques are provided by ISO, ISRIC and USDA.

4.1.3 Traditional biological lab analysis methods

An active population of soil organisms is essential for a healthy soil. Together with the analysis of organic matter, a biological analysis provides a picture of a soil's overall health, and its response to soil management practices. Techniques include DNA analyses, basal respiration, PLFA, and earthworm counting. The science on methods for estimation and interpretation of soil biological properties is young and very much in development; novel advances are foreseen. An example of an innovative approach is the BIOSYS framework for selecting appropriate soil biological measurement methods. Examples of protocols are provided by ISO.

4.1.4 Infrared spectroscopy lab analysis methods

I
nfrared spectroscopy in the lab is increasing in popularity and quality due to significant advances in operationalization of the technique in recent years. The measurement principle is the same as in field sensor application (see Annex IV: Common lab, proximal and remote (soil) sensing methods), but because conditions in the lab are much more controlled and instrumentation can be designed for permanent desktop applications, the quality of soil property estimations using infrared spectroscopy in the lab is much better. Recent research shows that for some soil properties the uncertainty in the predictions does not significantly differ from the wet chemistry measurement uncertainty predictions in professional big labs (Reijneveld et al., 2022). This indicates that for several applications lab IR spectroscopy can be an interesting alternative or addition by nearing the quality of wet chemistry analyses at (much) lower operational cost after a higher initial investment.

The quality of the result is dependent on the spectral predictability of the soil property in the near and or mid infrared parts of the electromagnetic spectrum (Dangal et al., 2021/2020; Shepherd et al., 2022; Nocita, 2015), the soil spectral calibration library used, the data analysis method and the sensitivity and quality of the instrument and its operation. In general, measurements in the mid-infrared provide better estimates than measurements in the near infrared since more soil related absorption features are present in the mid infrared range. However, recent research has shown that the gap is diminishing with high quality instrumentation and libraries and/or depending on the soils (Ramifehiarivo et al. 2023).

A soil spectral library (SSL) is a calibration set where soil samples are measured both with wet chemistry and spectrally, allowing derivation of spectral calibration models for soil properties in the feature space (the set of all values for a target variable in a target universe) or region that is represented in the library.

Several open soil spectral libraries exist, for example Global Soil Spectral Calibration Library and Estimation Service (Shepherd et al., 2022); the Open Soil Spectral Library of the USDA Food and Agriculture; the Brazil Soil Spectral Library, the LUCAS Soil spectral library, the ICRAF-ISRIC spectral library (Terhoeven-Urselmans et al., 2010), the Swiss soil spectral library and more. In addition to the spectral libraries, also Estimation Services are emerging that facilitate the prediction of soil properties for new spectra online (Shepherd et al., 2022, soil-spectroscopy.org, BraSpecS, globeSpeC (Shen et al., 2022)).

Soil sample preparation is minimal but important. Samples need to be dried (crushed) and sieved to 2 mm, and if analyzed with mid-infrared instrumentation the sample needs to be fine-grinded as well. For near infrared measurements fine grinding is not needed. Typically, 20 to 200 samples can be measured per day. Well-known instrument suppliers are FOSS (NIR), Bruker Optics, Thermo Scientific, Agilent (MIR), but there are many other suppliers on the market.

More information on the measurement principle of soil spectroscopy, its use for lab applications and best practices is described in Annex IV: Common lab, proximal and remote (soil) sensing methods and in Shepherd et al. (2022).

4.1.5 Transfer functions for lab method data harmonization

There can be many reasons to select and use specific lab methods and often within a project or lab the same methods are used. When working with data from different labs, projects, institutes, or organization however, the lab methods are often not (exactly) the same and the resulting data can therefore not be combined easily without creating an error or increased uncertainty. An example of this is when legacy data collected using previous labs, lab methods or protocols is gap filled with newly collected data using the latest lab methods or harmonized SOPs. A way to mitigate this is to develop transfer (or harmonization) functions between different lab methods for the same soil property, or even transfer functions for the same lab methods and soil properties but analyzed in different labs. The easiest way to develop a **transfer function** is to analyze the same set of relevant samples with both lab methods or by both labs and derive the relationship between the two by means of linear regression or more complex statistical models. It should be noted that it is advisable to use soil samples from the domain of interest for derivation and application of the transfer function. Furthermore, it may not be possible to derive reliable transfer functions between soil chemical properties analyzed with different extractants, in particular soil phosphorus (see ElRashidi, 2010). The latter is due to the fact that with different extractants, different pools are measured in the soil (e.g., Fixen and Grove, 1990 and Elrashidi, 2010).

4.1.6 Other relevant aspects of lab analysis

There are additional crucial factors to consider in laboratory analysis:

- **Quality management gaps**: these are areas where a laboratory does not perform activities that support a Quality Management System (QMS). Gaps are closed by implementing the QMS procedures and developing, unifying, and completing the documentation of these procedures.
- **Quantification of laboratory measurement errors**: laboratory instruments have limited precision and non-zero detection limits. Instrument drift can also occur, which can be addressed by regularly recalibrating laboratory equipment with reference samples. Sample preparation also introduces variability and error. The combined effect of these error sources can be estimated with replicated measurements of the same soil sample. It is advised that soil surveyors and researchers include sufficient replicates in the set of soil samples that are shipped to the laboratory, which should be anonymized and randomized before shipment. For statistical modelling of contributions of different laboratory errors, see Van Leeuwen et al. (2022).
- **Between-lab variability**: Bias (i.e., systematic errors) of laboratories can best be assessed by comparing the results of multiple laboratories. For this, proficiency or interlaboratory tests are in place, such as the Wageningen Evaluating Programs for Analytical Laboratories (WEPAL). Quantification and reduction of between-laboratory variability is also addressed by GLOSOLAN.

4.2 Standards for laboratory analysis

There are different standards or Standard Operating Procedures (SOPs) that can be applied in laboratory analyses. In this section the different standards are described.

Standard Operating Procedures

ISO

- ISO standards are available for most chemical and physical laboratory analyses on soils. These have been standardized by groups of experts.

GLOSOLAN

- GLOSOLAN is the Global Soil Laboratory Network of the FAO- Global Soil Partnership and consists of over 800 lab experts from all over the world with the aim to improve the quality of soil laboratory analysis globally. GLOSOLAN SOPs are harmonized by groups of experts based on different versions or executions of existing lab methods for chemical and physical soil property analysis. Every year several wet chemistry standards are harmonized and approved by the GLOSOLAN community.

ISRIC

- ISRIC has published standard operating procedures for many chemical and physical soil properties in Van Reeuwijk (2002) which are used for example in the FAO Unesco Soil Map of the World.

USDA-KSSL

- The mission of the Kellogg Soil Survey Laboratory is to measure soil properties that are critical to soil survey and conservation efforts of the USDA Natural Resources Conservation Service and the National Cooperative Survey. They provide a detailed soil survey laboratory manual that is widely used/referred to internationally. For spectroscopy, a separate SOP has been published USDA-NRCS Kellog Soil Survey Laboratory.

National SOPs or lab standardization bodies

- A wide range of analytical procedures is being used at the national level for specific applications. Internationally, results of such analyses often are not comparable. Hence the need for harmonizing these methods to a common standard (SOP) in ISO or GLOSOLAN and to compare the results of these methods in the context of international comparative analyses (see proficiency testing). An example of such an effort is being undertaken by GLOSOLAN, the laboratory network of the Global Soil Partnership.

SOPHIE

- The Soil Program on Hydro-Physics via International Engagement (SOPHIE) network works on the harmonization and improvement of soil hydrophysical measurements in the lab. These properties determine the soil – water interaction, such as water retention, infiltration capacity etc.

Soil spectral analysis

- USDA-NRCS Kellog Soil Survey Laboratory and by ICRAF;
- The GLOSOLAN soil spectroscopy working group is working on publishing an SOP for MIR soil lab analysis;
- The IEEE P4005 WG is working on SOPs for field IR measurements;
- ICRAF provides standard operating procedures for spectra data analysis with the software R.

Quality control standards:

Within lab quality control:

- ISO defines Quality Control (QC) as ISO *“the operational techniques and activities that are used to satisfy quality requirements.”* An important part of quality control is the Quality Assessment (QA), an evaluation of the products themselves. QC is primarily aimed at the prevention of errors. The control system should have checks to detect potential errors.
- An extensive set of guidelines for quality management in soil and plant laboratories is provided by ISRIC, WEPAL and FAO (FAO and ISRIC, 1998).
- Lab can also have their own standard for quality control. This typically consists of using standard reference samples in all batches of soil analysis. If the result for the reference or control sample is outside limits sets, the system is out of control and the entire batch should be reanalyzed. Other control measures are consistency checks on results include checking the plausibility of results, e.g., in a soil with pH of 4 the presence of $CaCO_3$ is not expected, the texture fractions of a sample should add up to 100 %, etc. Logging of metadata and anomalies in the procedures, e.g., a sample dropped to the floor, a vial was broken, sample code barely readable, etc

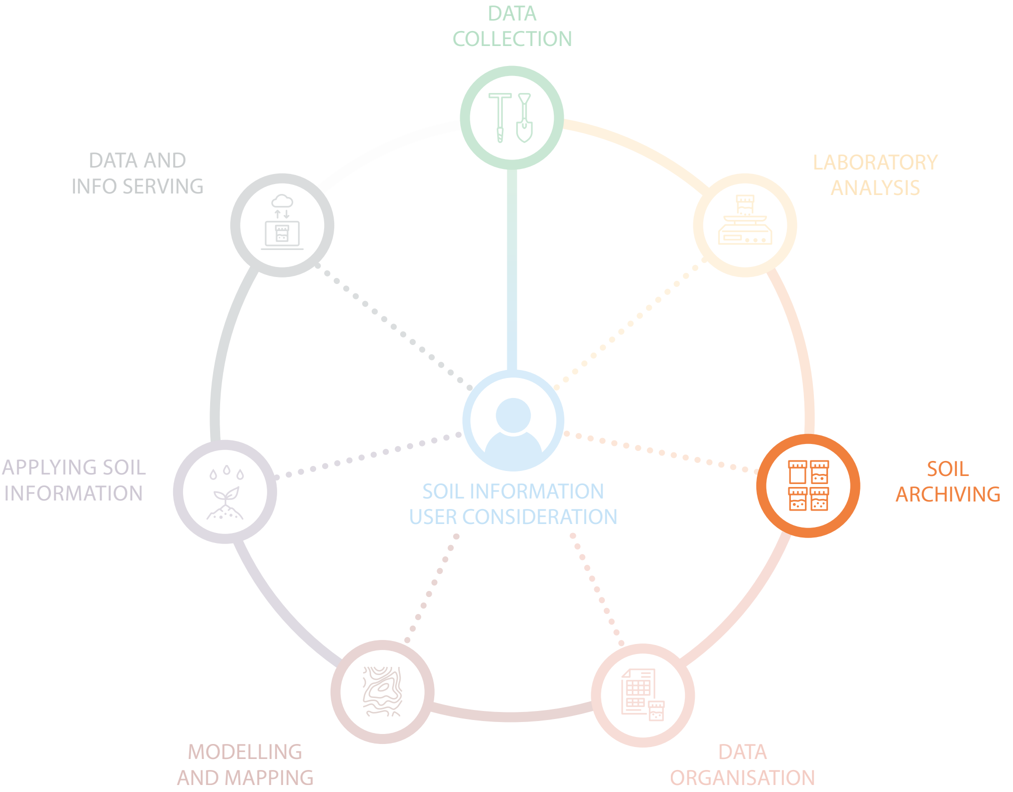
Proficiency testing

- Proficiency testing determines the performance of individual laboratories for specific tests or measurements and is used to monitor laboratories' continuing performance. As this term implies, proficiency testing (PT) compares the measuring results obtained by different laboratories. Examples of PT programs include WEPAL's (Wageningen Evaluating Programs for Analytical Laboratories) International Soil-Analytical Exchange Program (ISE) with global participation and The North American Proficiency Testing (NAPT).
- Participation in PT schemes provides participants with external quality control and helps potential clients in the selection of labs and procedures. For selection of a PT scheme see: https://www.eurachem.org/index.php/publications/guides/usingpt

4.3 Tools for laboratory data organization and analysis

A Laboratory Information Management System (LIMS) is software that allows you to effectively manage samples and associated data. In principle, a LIMS helps to plan, guide and record the passage of a sample through the laboratory, from its registration, through the program of analyses, the validation of data (acceptance or rejection), before the presentation and/or filing of the analytical results, and invoicing (FAO). Since the LIMS contains all soil analysis results, their analysis methods and quality evaluation in a structured way, it can be a direct data source for a SIS. In a SIS multiple data sources are combined. One of these sources can be the LIMS when LIMS data can be linked directly with field soil data through unique sample codes for example. This process can be automated when a secure connection between LIMS and SIS can be built. An example of using (the results of) a LIMS for a SIS is the SIS that is currently under construction in the Soils4Africa project[33] (Turdukulov et al., 2021).

[33] https://www.soils4africa-h2020.eu/


DATA COLLECTION
LABORATORY ANALYSIS
SOIL ARCHIVING
DATA ORGANISATION
MODELLING AND MAPPING
APPLYING SOIL INFORMATION
DATA AND INFO SERVING
SOIL INFORMATION USER CONSIDERATION

5. Soil Archiving

The standard practice of description of soil profiles and sampling of soils is done based on soil augerings or in soil pits. Physical soil samples may be taken for analysis in the laboratory which yields data for interpretation and storage in databases/soil information systems. In some cases, additional material is collected for future reference and/or for educational purposes. Physical soil sample archiving itself is the organized storage of sampling material (lab samples), soil specimens, soil documents and reports relevant to the (digital) data in the SIS. Soil archiving follows the soil laboratory analysis in the development of the SIS.

A soil sample represents a specimen of soil at a specific moment in time, a particular location, and a specific depth (range) from the surface. Soil sampling involves, among others, planning, human effort, travel, and laboratory analysis and is therefore expensive. It is not possible to take the same sample again because of the extractive nature of sampling and since soils change in time and space, it is often not possible to take a comparable sample later. Archiving physical soil material allows comparative analysis on various aspects, such as laboratory measurement method and calibration, land management impact, variation in time and space. Also, a soil archive may provide a reference for research, classification, and mapping of soils. Bergh et al. (2022) state that soil archives preserve a snapshot of soils from a specific time and location, allowing researchers to re-evaluate soils of the past in the context of the present for an improved understanding of long-term soil change (see for instance Karssies and Wilson, 2015). Many soil centers manage a soil archive for future research and reference.

Soil archives are often part of governmental organizations, experimental stations and research organizations for long-term research goals and policy related questions. Building and maintaining an archive requires investment in labor and facilities that does not fit the business case of commercial laboratories. A second limitation for commercial laboratories is that the samples are property of the clients that request analyses and consent may not be provided for storage or other uses of the sample. This is also clear when we look at the examples of national and international soil archives in table 5.1 and 5.2. The exception for commercial laboratories may be samples for quality control, calibration, and development of new measurement methods, for example building spectral libraries on samples that were analyzed with chemical and physical laboratory methods (Reijneveld et el., 2022).

Berg et al. (2022) made a literature review of soil archives and found that the age of soil archives across their compilation ranged from 5 to 160 years old, with mean and median archive ages of 48 and 37 year, respectively. Reliance on younger soil archives in publications was much more common, with the 25–34-year archive age range used most frequently for investigating long-term soil change. They conclude that soil archive use has increased since 1980.

This chapter will focus on physical soil (data) archiving, while chapter 6 describes digital soil (data) archiving. Physical specimens can be documents or objects (soil samples, sampling materials, thin sections, soil monoliths, hand pieces).

5.1 Methods for physical soil archiving:

The important aspects for archiving and preservation of physical soil collections, which includes reports, maps, soil samples are:

1) To formulate a collection policy that describes the principles for the management of the collection. The policy describes what is collected and for what purpose, hence defining the boundaries of what will be part of the collection and what not (and guiding acquisition of objects and samples).
2) Define procedures for collection management. Procedures in managing physical soil archives is about the 'how' of operational management, this includes activities and actions such as: acquisitions and disposal, sample, or object registry, cataloguing, movement of objects/samples, loan or use of objects/samples, care, and conservation. See for a full list of procedures that may apply, Spectrum 5.0 (2017).
3) Annual planning of activities related to collection/archive management. The annual planning formulates who does what, and the where and when of activities. This evolves from the procedures for collection management.

5.2 Archiving organizations

Soil archives are part of organizations that serve user groups with usually an emphasis on different regions and specific focus. These may be regional or local, national, and international or global.

5.2.1 National soil archives

Examples of national soil archives are described in table 3.

Table 3 National Soil Archives

Country	Curator
Australia	Australian National Soil Archive – CSIRO
China	Institute of Soil Science, Chinese Academy of Sciences
Denmark	Danish Institute of Agricultural Sciences
Ethiopia	Archive of the Ethiopia Soil Information System (EthioSIS) survey – Ministry of Agriculture
France	Institut National de la Recherche Agronomique (INRAE)
India	Punjab Agricultural University
New Zealand	National Soils Archive (NSA) – Landcare Research
The Netherlands	Wageningen Environmental Research, Wageningen
Switzerland	Swiss Federal Research Institute
United Kingdom	Rothamsted Experimental Station
United States	Agricultural Research Service, US Department of Agriculture
United States	Duke University
United States	National Ecological Observatory Network (NEON)
United States	Hubbard Brook Sample Archive
United States	U.S. Geological Survey; National Uranium Resource Evaluation archive
United Kingdom	National Soils Archive of the James Hutton Institute

5.2.2 International soil archives

International organizations that manage continental and global archives are listed in table 4.

Table 4 International soil archives

Country	Curator
The Netherlands	ISRIC World Soil Information, World Soil Reference Collection
The Netherlands	Wepal-Quasimeme, samples from laboratory exchange
Kenya	Archive of systematically collected soils samples, CIFOR-ICRAF
Philippines	International Rice Research Institute, IRRI
Italy	LUCAS soil sample archive for Europe, JRC
Italy	FAO soil legacy reports https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/soil-legacy-reports/en/
United Kingdom	World Soil Survey Archive and Catalogue (WOSSAC) at Cranfield University, UK https://www.wossac.com/#:~:text=WOSSAC-,World%20Soil%20Survey%20Archive%20and%20Catalogue,-Search%20the%20archive

5.2.3 Libraries, Reference institutes and Museums

Most soil institutes maintain a soil library with archived survey reports, maps, and associated documents. Examples of national soil libraries are: the National Library of Scotland and the library of New Zealand's' Manaaki Whenua Landcare Research. Examples of libraries with maps and reports with regional to global coverage are: the World Soil Survey Archive and Catalogue (WOSSAC); the world soil library and map collection of ISRIC – World Soil Information and the repository of the European Soil Data Centre (ESDAC).

Moreover, there are national soil reference centers, such as the National Soil Resources Institute of Cranfield University; the National Soil Survey Center of the USDA Natural Resources Conservation Service and ISRIC – International Soil Reference and Information Centre.

A global review counted 38 soil museums specifically dedicated to soils, 34 permanent soil exhibitions, and 32 collections about soils that are accessible by appointment (Richer-de-Forges et al., 2021). Important soil museums include the Central Museum of Soil Science in St Petersburg, the Emirates Soil Museum in Dubai, the Soil Museum of Thailand in Bangkok and the World Soil Museum in Wageningen.

There are also reference institutes for soil data:

- FAO soil legacy reports
- FAO soil legacy maps
- High-resolution DSSAT-compatible database for Africa based on AfSIS
- Global High-Resolution Soil Profile Database for Crop Modeling Applications

5.3 Tools for physical soil archiving and preservation

Few soil archives have published their procedures for development, management, and use of their archive. CSIRO - Australian National Soil Archive is a favorable exception. Other publications on archiving of soil samples are Quantitative Guidelines for Establishing and Operating Soil Archives (Ayres, 2019) and the chapter of Boone et al. (1999) on Soil sampling, preparation, archiving, and quality control.

The tools for physical soil archiving are described below:

- **Metadata document:**
  A metadata document (indicating the lifecycle status) describes the metadata of the objects or documents, for example where it has been collected, when, by whom, how it is processed, etc. The metadata document is preferably kept in a registry for the full lifecycle of soil information with a clear link to the actual object or document, e.g., id and location number, which is updated when changes occur. Access to the resource preferably exists at a specific location from collection until removal. If a resource is moved to a new location, the old location could provide a forward reference to the new location. More information on metadata can be found in chapter 6 Data Organization.

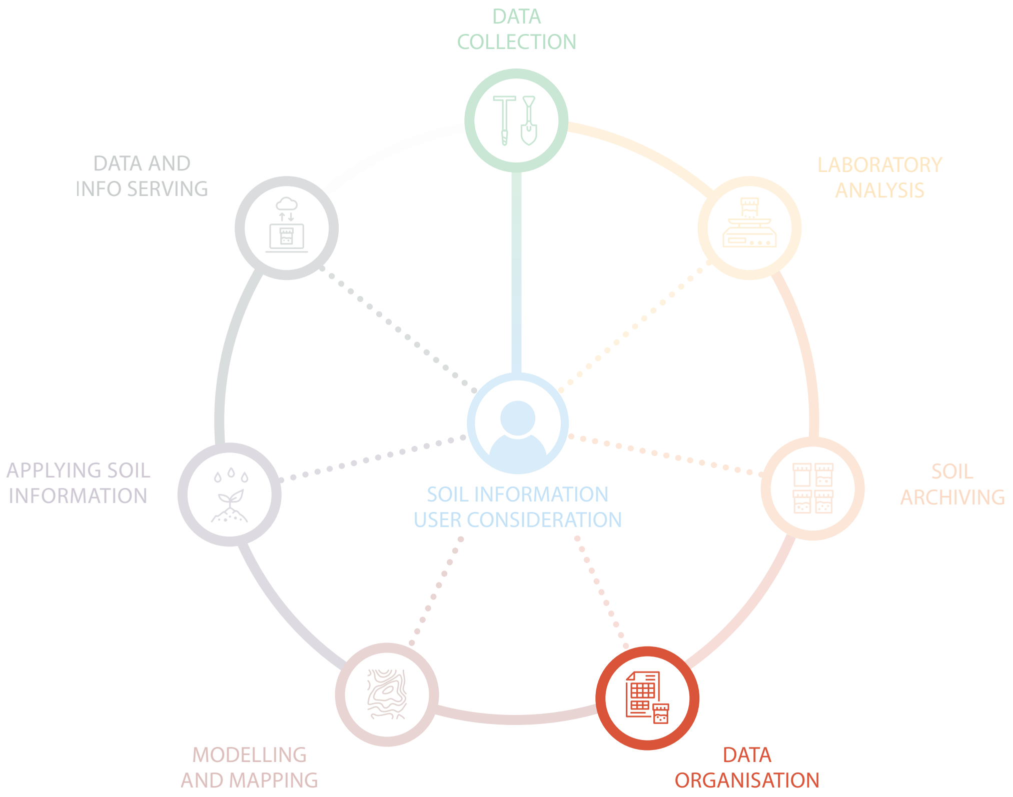
- **Collection management plan and policy:**
  A collections management policy is a set of policies that address various aspects of collections management. It defines the scope of a museum's collection and how the museum cares for and makes collections available to the public (American Alliance of Museums, 2012). The process of collection management is achieved by incorporating methods of organization and staffing, selecting, and deselecting, budgeting, marketing, and promoting, understanding electronic resources and the role of interlibrary cooperation, and evaluating and assessing success (American Alliance of Museums, 2012). An example is the ISRIC collection management policy for physical collections[34], where amongst other the SPECTRUM standards are discussed.

- **Storage Facilities:**
  Storage is crucial for the future of a museum as museums are representatives of our natural and cultural heritage. Inadequate storage is mostly due to lack of funds, or these funds are used for other goals. To properly care for the museum collection, (technical) knowledge on conservation, storage systems, record-keeping and security is needed. Storage facilities consists of service yard, loading dock, receiving area, washing area, registration and holding area, curatorial offices or laboratories, collection research area, photo area, conservation laboratory and collection storage area (Verber Johnson, 1979).
- **Digitization tools:**
  Digitization has become essential in the overall management of collections. It also helps to better understand and study the soil (Wadoux and McBratney, 2021). Museums increasingly connect all the information related to the objects in a digital repository, including images, history of the object, conservation reports, exhibition texts, related publications, and physical location of the objects using a form of barcode. Once created, museums can

[34] https://www.isric.org/management-policy-world-soil-reference-collections

easily reposition selected content to the Internet allowing remote access to information about the collections. Digitization allows museums to participate in the information economy, but it will also require a significant investment (Navarrete, 2020).

- **Soil monolith preparation:**
  The sampling and preparation of soil monoliths is explained in the technical paper Procedures for the collection and preservation of soil profiles Van Baren and Bomer (1979) and for lacquer peels by Stoof et al (2009).


DATA COLLECTION
LABORATORY ANALYSIS
SOIL ARCHIVING
DATA ORGANISATION
MODELLING AND MAPPING
APPLYING SOIL INFORMATION
DATA AND INFO SERVING
SOIL INFORMATION USER CONSIDERATION

6. Data organization

The **data organization** step manages the aspect of the adoption of field observation data, lab data, analyzed data and metadata into a central system. This can include new data but also existing, or legacy data. It prevents data loss, assesses the quality and harmonization of the data and preparation of the data for reuse. Data organization follows the soil archiving step in the development of the SIS and when the smaller definition of a SIS as a digital infrastructure for soil data is used, it can be the first step after the user needs assessment.

For choice of tools there is no one size fits all, the choice will depend on the human capacity and infrastructure (enabling environment) options in a country. Nevertheless, harmonization of the output of the systems in desirable, to allow a compiling system as for example GloSIS, INSPIRE geoportal/EUSO or WDC Soils, or other repositories to easily harvest data and metadata from the SIS on a transnational level. This means to use a metadata standard, common formats, license, etc. More information is provided in chapter 9.

Digital soil data, other than reports, maps, and documents, are stored in two main digital representations : attribute tables and raster formats.

The **attribute table**, if it includes a geometry, also referred to as vector data, typically contains the results of a soil data collection campaign or subsequent lab analysis. Raster format is a result of earth observation (aerial imagery) or a modelling effort. For every pixel on a grid, a numeric value(s) for the given variable is observed or predicted. The tools and procedures to work with these two types of data are quite distinct.

Raster data are often stored as flat files, two common formats being Geo Tagged Image File Format (GeoTIFF) and Network Common Data Form (NetCDF). These files are organized into catalogues (a folder structure). Rasters can also be stored in a database. This scenario is relevant if you use the raster analyses functions of the database to filter on or calculate cell values. Raster files can result in large files if they extend to a large area or are detailed. Large files led to long loading times. Various scenarios are common to optimize performance. Most common is adoption of an image pyramid, in which the data is subsampled at various resolutions. The client then reads data from the most relevant resolution only.

Attribute table data can be stored as spreadsheets, or in databases. Storing in databases has two main advantages. 1) Data are accessible for reading and writing by multiple users, and 2) the integrity and security of the data are better maintained:

- Integrity: databases provide options to set rules to maintain integrity. For example, a parent record cannot be removed if there is a related child record.
- Security: databases enable fine-grained configuration of access on table level, for Read, Update and Delete privileges.

In either format, metadata of data is typically registered as a file close to or embedded in the data file, or in a registry, catalogue and/or **Document Management System**.

This chapter introduces various standards and tools available to facilitate data organization.

6.1 Methods and standards for data organization

6.1.1 Methods

The methods described below are a combination of generic data practices endorsed by the FAIR principles (findability, accessibility, interoperability, and reusability) (Wilkinson et al., 2016) and data organization aspects specific to the soil domain.

Data management:

- Databases:
  Databases provide structured storage of data with efficient query capabilities. Some databases provide versioning mechanisms. Databases usually include an advanced authorization system to allow view and edit capabilities to be assigned to roles. Databases are less optimal for unstructured and array (grid) data. The data model in a database determines the way the data is structured. To facilitate interoperability of data with partners a data model based on common soil data standards can be used (see below). This model can then be extended to fit the specific purpose of the SIS and its use cases.

- Cataloguing and metadata:
  Cataloguing and metadata facilitate users (including yourself) to find resources within your organization and assess if these resources are of interest to their case. Metadata is data about the data and typically includes a description, date, location, usage constraints and contact of the resource.

- Data rescue:
  Data rescue is the process of preventing data loss at incidents, including the actual restoration of missing operational data. The process is a mix of data backup and data synchronization.

- Data archival/removal:
  Data archival/removal manages the proper storage or destruction of data at the end of life. Proper destruction of data is especially relevant for data with a privacy constraint.
  For data storage proprietary formats and systems are incidentally used. For interoperability and archiving purposes consider using instead open source and or standardized formats (netCDF, Tiff, JPG2000, GML, GeoJSON, SQLite) and systems (Web Coverage Service, STAC). Prevent use of formats for which compression leads to irreversible data loss.

- Data standardization:
  Data standardization captures data in a standardized domain model or ontology, including the use of common code lists. This can be implemented in a standardized data model in the database. Data in a common model facilitates data interoperability with partners. It also helps in identifying which elements to capture. Relevant common models in the soil domain are ISO28258, INSPIRE Soil, SoilML and GLOSIS. Implementations of ISO28258 and GloSIS as data models are available.

- Usage constraints:
  Aspects of access and usage constraints are typically mentioned in metadata accompanying a dataset. Make sure to apply a license to your data, regardless of if it is fully open or closed. If you publish your data as open data, preferably use a common open data license, such as ODBL or Creative Commons. For users it is inconvenient to combine data from multiple sources if they use a variety of access and usage standards.

- Repositories:
  Public or closed repositories are local or online collections of metadata or of data with metadata that aim to store and safeguard (meta)data, publish metadata to increase findability and if the license allows also publish the data.
  Persistent repositories are repositories of which the hosting organization has guaranteed to keep the records more than e.g., 10 or 20 years. Examples are Zenodo, implementations of DataVerse, etc. These are suitable for archiving.
  Good practices are to always indicate a data license and metadata.

Quality assessment:

- Quality assessment evaluates for each of the methods if they are effective for intended use. Quality assurance of data is described in the ISO19157 standard. There are many aspects of which you can assess the quality of data, such as positional accuracy, domain consistency, completeness. The outcomes of these assessments should be captured in a metadata document accompanying the dataset, see Annex V for the quality elements of the ISO19157.
  For ease of consumption, the GeoViqua[35] project designed a quality vignette for spatial data, which combines multiple quality aspects into a single graphic representation.

- Unique and Persistent identification of resources is an essential pre-condition for proper data management and reuse;

- Facilitate and publish User Feedback:
  The group on best practices of data on the web defined a list of best practices when publishing data. Two of their best practices related to data publication are often overlooked: (1) Use feedback from the audience for improvement, and (2) also publish the feedback because it may help others to better understand the data.

6.1.2 Metadata standards

There are various metadata standards. Examples of common metadata standards are given below.

- DublinCore:
  DublinCore[36] is a set of terms to describe resources originating from the library domain. The set forms the basis of DCAT and is also the minimal set required for Catalogue Service for the Web.

[35] https://cordis.europa.eu/project/id/265178
[36] https://www.dublincore.org/

- ISO 19115:

  ISO 19115-1:2014[37] is an ontology for dataset metadata common in the geospatial domain. It defines the schema required for describing geographic information and services by means of metadata. It is applicable to the cataloguing of all types of resources, clearinghouse activities, and the full description of (geographic) datasets and services.

- INSPIRE technical guidance:

  The technical guidance of INSPIRE[38] holds the Implementation specification for defining metadata for INSPIRE datasets and services in ISO/TS 19139 based XML format in compliance with the INSPIRE Implementing Rules for metadata.

6.2 Models and Tools for (meta)data management

In this section we define a model as a structured representation of concepts that reflects its interrelations. Examples are a data model, a domain model, or a metadata model.

6.2.1 Metadata models for data management:

- DataCite:

  DataCite[39] is a leading global provider of DOIs for research data and provides a Metadata Schema as a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes. The DataCite metadata schema has been adopted by a range of academic repositories, such as Zenodo and Dataverse. It is encoded as XML.

- DataPackage:

  DataPackage[40] is a metadata model and approach from the CKAN community, to place metadata files with the data.

- Schema.org:

  Schema[41] is a metadata model used by search engines to facilitate embedded structured data in websites. Modelled as RDF, usually encoded as JSON-LD or microdata.

[37] https://www.iso.org/standard/53798.html
[38] https://github.com/INSPIRE-MIF/technical-guidelines/tree/2022.2/metadata/metadata-iso19139
[39] https://datacite.org/
[40] https://github.com/frictionlessdata/ckanext-datapackager
[41] https://schema.org/

6.2.2 Tools for data management:

Relational databases

- PostgreSQL:
  PostgreSQL[42] is an Open-Source relational Database Server, including advanced spatial support via the PostGIS extension. PostgreSQL can be used when dealing with large datasets.

- Oracle:
  Oracle[43] is a proprietary relational database server including spatial support. Oracle database products offer customers cost-optimized and high-performance versions of Oracle Database, the world's leading converged, multi-model database management system, as well as in-memory, NoSQL, and MySQL databases.

- SQL server:
  SQL server[44] is a proprietary relational database server including spatial support. SQL server offers enhanced performance and provides efficient permission management tools.

- MS Access:
  MS Access[45] is a proprietary file based relational database of Microsoft Office. The Access software is optimal to interact with the database. Access can help create appealing and highly functional applications in a minimal amount of time, especially when working in Microsoft Office.

- SQLite / GeoPackage:
  GeoPackage[46] is a file based relational database. Tools like Dbeaver provide a user interface to the database. GeoPackage is a standardized format to store spatial data by the Open GeoSpatial Consortium as an extension to SQLite.

- Virtuoso:
  Virtuoso[47] aims to support all major paradigms of data storage be a relational database management system (RDBMS), an object-relational database, an RDF triple-store and SPARQL engine, store XML, full-text, and other file-based formats. Virtuoso also implements GeoSPARQL, making it a geo-spatial triple-store. Virtuoso is written in C programming language and designed to run as a multi-threaded server, it is therefore a fast and lightweight server, requiring few resources and easy to manage in containerized environments. Virtuoso is much more than a triple-store, with data provision and browsing functionalities that are especially useful to data providers in the Semantic Web.

- Jena:
  Jena[48] is an umbrella term for a software complex developed by the Apache Foundation, with two broad functions: (i) management and analysis of knowledge graphs, and (ii) a

[42] https://www.postgresql.org/
[43] https://www.oracle.com/database/what-is-a-relational-database/
[44] https://www.sqlservertutorial.net/getting-started/what-is-sql-server/
[45] https://www.microsoft.com/en/microsoft-365/access
[46] https://www.geopackage.org/
[47] https://virtuoso.openlinksw.com/
[48] https://jena.apache.org/documentation/tdb2/tdb2_admin.html

triple store, both SPARQL and GeoSPARQL enabled (the latter is known as Jena Fuseki). In recent versions, various APIs have been introduced to enable automated interaction. As a triple store, Jena is a lean piece of software, easy to learn and deploy, particularly useful to serve knowledge graphs created externally. On the other hand, Jena presents a narrow set of functionalities (when compared with Virtuoso, for instance) and can be demanding on resources within production environments.

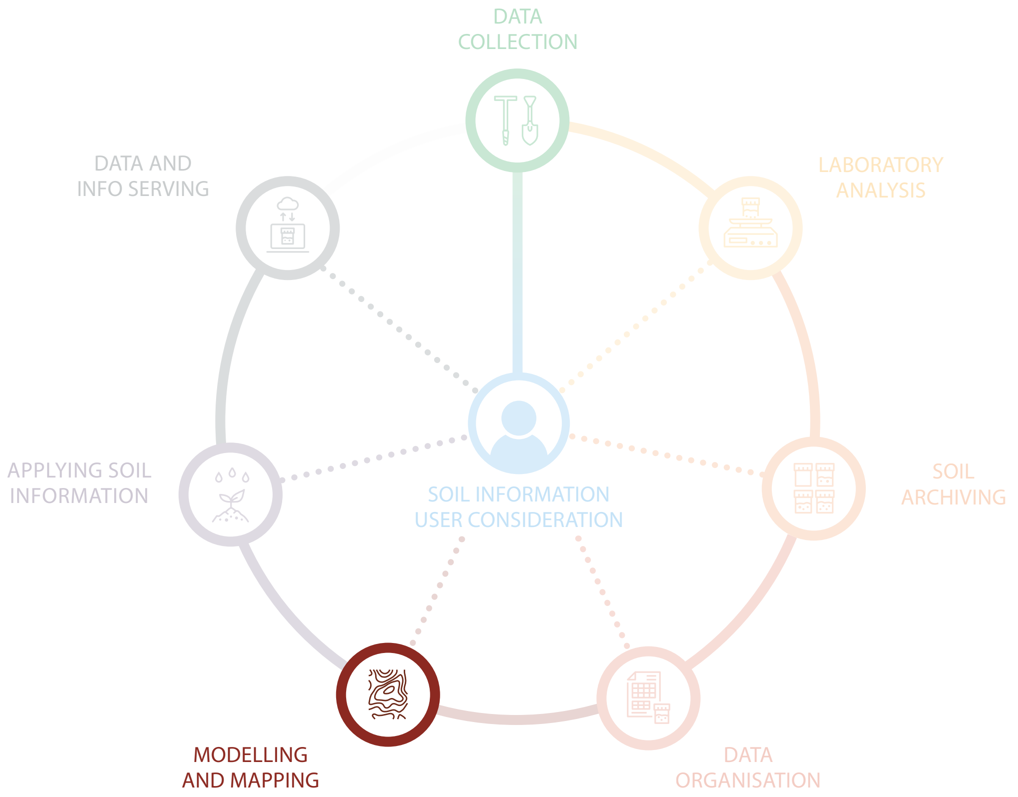
Object Oriented / NoSQL databases

- MongoDB:
  MongoDB[49] is an open-source document-oriented database server with geospatial support. For example, GeoServer and pygeoapi can use MongoDB as a backend. MongoDB provides the services and tools necessary to build distributed applications fast, at the performance and scale user's demand.

- CouchDB:
  Apache CouchDB[50] is an open-source document-oriented NoSQL database, implemented in Erlang with spatial capabilities via the GeoCouch plugin. CouchDB lets you access your data where you need it from mobile phones to web browser.

- Repositories:
  See chapter 9.2.1 Catalogue services: and chapter 9.2.2 Host data on a repository

[49] https://www.mongodb.com/
[50] https://couchdb.apache.org/


DATA COLLECTION
LABORATORY ANALYSIS
SOIL ARCHIVING
DATA ORGANISATION
MODELLING AND MAPPING
APPLYING SOIL INFORMATION
DATA AND INFO SERVING
SOIL INFORMATION USER CONSIDERATION

7. Modelling and mapping

Once soil data is collected, analyzed, archived, and organized, the soil data can be used for modelling and mapping in the next step of the development of a SIS. Soil databases typically contain observations and measurements taken at sampling (point) locations. Many users though, require information in the form of soil maps for their applications. Therefore, developing **soil maps** from soil (point) data stored in a database is a logical next step in the soil information workflow. For soil mapping, two mapping approaches are often used: **conventional soil mapping** (also known as landscape-based soil class mapping) or **digital soil mapping**.

Soil maps used to be drawn by soil surveyors who would take observations in the landscape and then, often supported by aerial photography, delineate soil bodies that are homogeneous in terms of morphology and composition. These soil bodies are subsequently classified s based on a soil classification system. The result is a traditional soil class map that is often accompanied by a survey report that provides descriptive information and quantitative data on the soil bodies.

Since the beginning of the 21st century, more maps are developed using statistical methods for modelling and mapping soil spatial variation. These methods are referred to as '**digital soil mapping**' (DSM). Nowadays, DSM is an accepted practice for soil mapping and more often used than conventional soil mapping. The widely used Soil Survey Manual of the U.S. Department of Agriculture now contains a chapter dedicated to DSM. With the advent of DSM, not only the way soil maps were produced changed, but also the type of maps. DSM typically produces gridded maps of quantitative soil properties at a specific spatial resolution (grid cell size), while conventional survey typically produces polygon maps of soil types and associated soil properties at a certain (cartographic) scale level.

This chapter presents an overview of methods, tools, and standards for developing soil maps from soil observational data. Other type of models, such as **dynamic or process models**, that use soil data as input to predict functional properties of the soil (that often cannot be directly measured such as carbon sequestration potential) or other type of agricultural or environmental variables (such as yield potential based on soil nutrient status and soil depth) are considered in Chapter 8 on 'Applying soil information'.

7.1 Methods and standards for modelling and mapping

As mentioned above, **conventional soil survey** and **digital soil mapping** are two distinct types of soil mapping. Although quite different in execution, both approaches are based on the same operational paradigm: the *soil-landscape model* (Hudson, 1982). This model assumes that the spatial distribution of soil classes and properties can be inferred from their position in the landscape as well as effects of other soil forming factors, such as parent material, climate, vegetation, fauna (incl. human activities) and (geological) time. Thus 'modelling' in the context of this chapter refers to 'soil-landscape' modelling with the aim to predict (i.e., map) the soil spatial distribution from a set of soil observations and environmental properties.

7.1.1 Conventional soil mapping

In **conventional soil survey**, the soil-landscape model is a conceptual model based on tacit knowledge of the soil surveyor. The surveyor infers soil-landscape relationships from observations in the field (incl. soil observations from pits or augers) supported with aerial photography or (nowadays) remote sensing imagery, such as for instance digital elevation models or land cover maps. Based on these observations and relationships the surveyor is able to delineate (map) and classify soil bodies according to a soil classification system. The value of soil class maps is that based on the soil class, a set of soil properties can be derived and the soil processes leading to those properties are understood. This allows the user to understand the soil landscape system, its possibilities, and drawbacks on one map. The downside is that interpretation requires an understanding of soil science and classification, and there are many soil classification systems.

Soil classification standards are described in chapter 3.1.2 Full soil profile description and assessment of site and soil conditions, protocols, and procedures. Methods for conventional soil survey are provided by the NRCS Soil Survey Manual and in Arrouays et al. (2021). For conventional soil maps uncertainty is evaluated as percentage of correctly classified, or map purity, although this is not always reported.

Conventional soil maps can be hosted and served by a SIS given that the maps are converted in GIS format which requires georeferencing and digitization of the map units. To support interoperability (the 'I' in 'FAIR' data management), file formats should preferably be open formats such as 'geopackage' for vector data and 'geotiff' for raster data. Some established soil information systems provide conventional soil maps as images (in PNG or JPG format). However, the usefulness of such files in a digital environment is extremely limited. However there has been supervised and unsupervised methodological development efforts to disaggregate and update legacy soil maps there by enhance usability and compatibility issues e.g., Disaggregation and Harmonization of Soil Map Units Through Resampled ClassifiCation Trees (DSMART) , fuzzy c-means (FCM) , and k-means (KM) clustering techniques.

7.1.2 Digital soil mapping concepts

Digital soil mapping (DSM) is also based on the soil landscape model but uses (geo)statistical models, instead of tacit models, to relate observations or measurements of soil properties to maps or spatial data products (like satellite imagery) of environmental variables that represent the soil forming factors to derive predictive relationships. Once such relationships are quantified (calibrated), these can be applied to predict the soil properties of interest across the mapping area from a set of environmental variables, available in the form of (digital) maps that cover that area.

Advantages of DSM are that it provides quantitative, gridded soil maps at a user-defined resolution that can be input to process models and are often easier to understand by non-soil scientists. DSM often provides pixel-specific assessments of prediction uncertainty that allows to carry out uncertainty propagation studies or to determine the fitness-for-intended-use of the digital soil map. The DSM workflow is fully transparent and reproducible when coded workflows are used for modelling and mapping. The disadvantage of DSM is that little soil system understanding can be gathered from a single soil property map, requiring multiple maps of soil properties and soil science expertise for interpretation. Also, DSM has a steep learning curve and application of DSM can be limited by available computational capacity, especially when large (high resolution) datasets are used as inputs.

Below we explain several key components of DSM:

Point data are the main input for a DSM model and must be georeferenced.

- Input point data can refer to observations taken directly in the field (e.g., soil depth, morphological characteristics, soil classes) or observations (measurements) from soil samples analyzed in a laboratory or indirect measurements from e.g., proximal soil sensing methods. Point data are typically derived from the SIS database or other sources (see chapters 3, 4 and 6).

- A georeferenced point data is essential in harmonizing legacy soil data. Most important are the position, the date, and the coordinating system of the data point: these are essential for DSM modeling. The more accurate the point, the better results you will get from your model. The accuracy depends on the resolution you would like to model:

Table 5 Relationship between the goal of soil survey, sampling density and scale of derived soil maps (Tóth, et al. 2013).

Kind of survey or map and level of intensity	Purpose and use of the survey results	Area represented by one sample (ha)	Indicative scale of published maps
Precision farming (intensive, level 1)	Special; executive purpose – within parcel	< 1	> 1:1000
Detailed (field scale, level 2)	Special; executive purpose – for parcel	1 - 50	1:1000 – 1:10.000
Semi-detailed (farm to regional scale, level 3)	General and special; planning purpose	50 – 1000	1:10.000 – 1:100.000
Reconessaince (regional scale, level 4)	General; planning purpose	1000 – 5000	1:100.000 – 1:250.000
Reconessaince (regional to national scale, level 5)	General; orientation purpose on national scale	5000 – 20.000	1:250.000 – 1:500.000
Exploratory surveys and compilations (national to continental scale, level 6)	General; orientation purpose on continental and global scale	> 20.000	< 1:500.000

- Detailed information on point data description is given in LUCAS TOPSOIL SURVEY: Methodology, data and results.

Auxiliary information or 'covariates' are spatial datasets of environmental variables that are related to the soil forming factors. Typically, these include terrain parameters derived from a digital elevation model, land cover maps, vegetation indices or spectral reflectance obtained from Earth observation imagery, maps of climate variables and soil class, geological or geomorphological maps. A requirement for selecting covariates for a mapping area is that these cover the entire area. It is important that the resolution of the selected covariates is appropriate for the target resolution of the digital soil map. There are numerous sources where covariate layers can be obtained for free. A reliable source is the Earth Engine Data Catalogue. It hosts a large suite of Earth observation imagery and products derived from them.

Another data gathering platform is Microsoft Planetary Computer. Many of these products can also be found elsewhere from gathering services (e.g., Amazon, google STAC catalogue) and from the original data provider. For instance, the layers from the Copernicus Global Land Service are also available via the Land Copernicus library, the SRTM DEM is available from the CGIAR Consortium for Spatial Information platform. Sentinel products are available via the sentinel hub and NASA products are available via the NASA data platform. There are also sources with ready-to-use analysis of the covariates such as SoilGrids or Digital Earth Africa.

Uncertainty:

- **What is uncertainty**: Soil data are often contaminated by various error sources, such as measurement, sampling, classification, and mapping errors. These errors will propagate through models and analyses and can affect decision making. In practice we represent errors by probability distributions, where the width of the distribution signifies the uncertainty and can be characterized for instance by the standard deviation.

- **Sources of uncertainty**: The main sources of uncertainty in soil point and profile data are field estimation error, laboratory measurement error, and errors in recording the geographic position of sampling locations. Further errors will be introduced if field protocols, and analytical methods are not standardized and need to be harmonized using transfer functions (chapter 4.1.5 Transfer functions for lab method data). Mapping soil classes and soil properties from point observations and maps of explanatory environmental variables brings along other error sources. The main causes of these spatial prediction errors are that the explanatory variables do not explain all spatial variation of the target soil properties, which mapping models are not flexible enough to capture all information contained in explanatory variables, or that the training data set is too small to estimate model parameters optimally.

- **Quantifying uncertainty**: DSM models can quantify the uncertainty associated with the predicted soil value or class *for each pixel* in the mapping area. Uncertainty is typically expressed through a prediction error variance or a prediction interval width. The latter is expressed as the 90% prediction interval that is calculated from the 95% and 5% quantiles of the prediction distribution. Important examples of methods that can do this are kriging and quantile regression forests.

Accuracy and validation:

- While uncertainty assessment gives a measure of prediction uncertainty for each pixel, statistical **validation** gives a measure of accuracy of the soil map as a whole. Here, the map values are compared with independent observations, which are ideally obtained by probability sampling from the area of interest. Another option is cross-validation, where the input dataset is iteratively divided into calibration and validation datasets. A third option is data splitting where a dataset is split in two parts, one of which is used to calibrate a prediction model (typically containing - an arbitrary - 70% of the data points) and the other for validating the predictions with that model. Common **validation metrics** for quantitative soil properties are the Mean Error, the Root Mean Squared Error (RMSE) and the Model Efficiency Coefficient that quantifies the variation in the soil input data explained by the DSM model. In case of categorical variables such as soil type, the most important metric is the map purity, defined as the proportion of points in the area of interest that are correctly classified. Additional probability sampling is preferred for validation. If collection of additional validation data with probability sampling is not possible, then cross-validation is a suitable alternative. Data splitting is sub-optimal and should be avoided. More details can be found in Brus et al (2009) who review validation of digital soil maps.

Resolution or scale:

- Traditionally, the 'scale' of a soil map has a cartographic definition, namely the ratio of a distance on Earth compared to the same distance on a paper map. In the digital era the term 'scale' more often refers to the spatial extent or resolution of a soil map. Here, 'extent' refers to the size of the area covered by the map, while 'resolution' is the distance between predictions displayed in a raster map. For instance, SoilGrids predicts soil properties on a regular grid of points covering the globes that are approximately 250 m apart. SoilGrids also predicts at six standard depths within the top 2 m of the soil, and hence has much higher vertical resolution than horizontal resolution. It is important not to confuse spatial resolution with spatial accuracy. For instance, if independent validation data show that a fine scale soil map has a larger RMSE than a coarse scale soil map, then the fine scale soil map has a higher resolution but a lower accuracy. This is an example of a map that is considered to provide a false sense of accuracy and should be avoided.
- The recommended resolution and scale for soil classes and/or attributes are provided by McBrateny et al. (2003) and Rossiter (2008). These are widely used references for resolution and scale.

Figure 3 shows how these concepts come together in the DSM workflow. Soil point data and covariate layers are used as input. If multiple sources of input point data are used, then these need to be merged which will require some form of standardization (see Chapter 4 and 6). Covariate layers typically come from different repositories and need to be brought to a common spatial extent and resolution that results in a stack of raster layers. Covariate values are extracted at the data points and combined with the observed and/or measured values of the target (soil) variables at these points using a spatial overlay operation. This results in a **regression matrix** that is subsequently used as input for the calibration (training) of a (geo)statistical prediction model. The calibrated model is applied to the covariate stack that results in a prediction of the target soil variable for each pixel in the area of interest, including an estimate of the prediction uncertainty. The predicted values are compared with the observed values, for instance using cross-validation, resulting in accuracy statistics. Note that this workflow is a general representation and can be elaborated further. For instance, with a 'recursive feature elimination' step to reduce the number of (redundant) covariates offered to the model to increase computational efficiency.

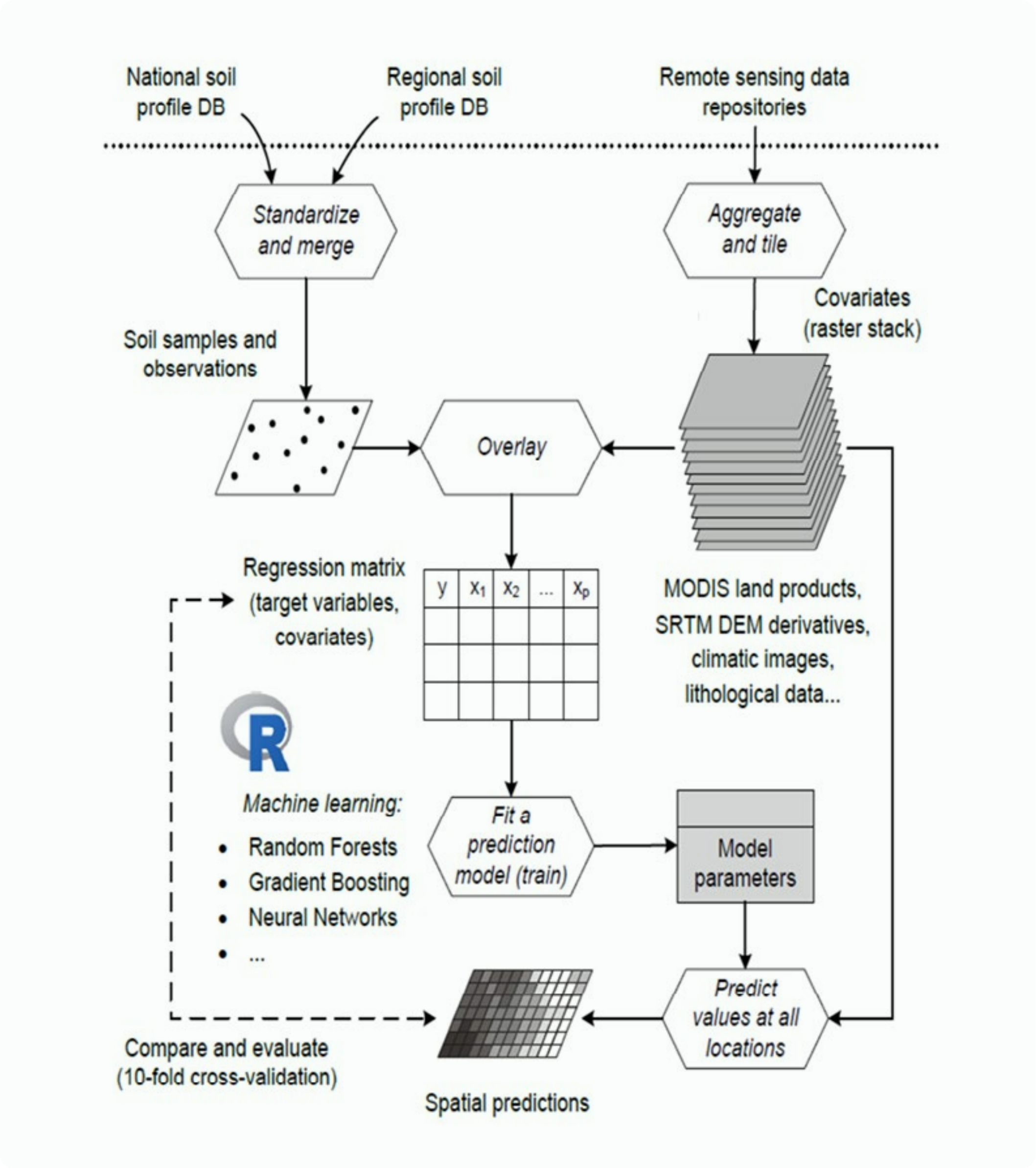

National soil profile DB
Regional soil profile DB
Remote sensing data repositories
Standardize and merge
Aggregate and tile
Soil samples and observations
Covariates (raster stack)
Overlay
Regression matrix (target variables, covariates)
y
x1
x2
...
xp
MODIS land products, SRTM DEM derivatives, climatic images, lithological data...
Machine learning:
Random Forests
Gradient Boosting
Neural Networks
...
Fit a prediction model (train)
Model parameters
Predict values at all locations
Compare and evaluate (10-fold cross-validation)
Spatial predictions

Figure 3 Schematic representation of the DSM workflow

7.1.3 Statistical Models

DSM models predict the soil type or soil properties at all locations in an area of interest from soil measurements at point locations and maps of environmental covariates. Two types of models are often used: geostatistical modelling and machine learning.

Geostatistical modelling:

- **Geostatistics** is founded on the First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things". Thus, it exploits the fact that environmental variables, including soil type and soil properties, are spatially correlated. Spatial correlation is quantified by the semivariogram, which is a mathematical function that shows how the variation between two data points changes as the distance between them increases. Geostatistics was originally developed in the 1950s in mining and geology but is nowadays widely used in many fields in the Earth and environmental sciences.

- Any geostatistical analysis starts with an exploratory data analysis and modelling a semivariogram. Next, a map is made using a spatial interpolation technique known as kriging. There are many variants of kriging. Ordinary kriging is the most basic variant and predicts the value of a soil property at an unmeasured location as a weighted linear combination of the soil measurements derived from the semivariogram. Nearby measurements typically get larger weights than more distant measurements because they have a stronger correlation. An attractive property of kriging is that it not only makes predictions, but that it can also quantify the spatial interpolation error, by means of the kriging standard deviation.

- Ordinary kriging solely relies on soil measurements and does not benefit from covariate information. In the 1990s it was therefore extended to other variants such as kriging with external drift and regression kriging. These variants also incorporate the multiple linear regression correlation between observations on the environmental covariates. Regression kriging tends to have a higher prediction accuracy than ordinary kriging, particularly when the covariates explain a large part of the soil spatial variation.

- There are many geostatistical textbooks. Isaaks and Srivastava (1989) focuses on applications and does not require a strong mathematical background. Webster and Oliver (2007) provides more technical-statistical detail while Chilès and Delfiner (2012) provides an in-depth treatment of the subject. Wikle et al. (2019) extends geostatistics to the space-time domain and illustrates the theory with examples and R scripts. Bivand et al. (2013) explains how spatial-statistical analyses are done in R and also includes chapters on geostatistics.

Machine learning models (or Artificial Intelligence):

- **Machine learning models** are extremely flexible statistical models that fit a relationship between a dependent variable (i.e., the soil type or a soil property) with explanatory variables (i.e., environmental covariates) and use that relationship to predict the dependent variable from the explanatory variables. These data-driven models work best if large training datasets and many covariates are available and, in such cases, they outperform multiple linear regression and kriging. For this reason, machine learning has replaced geostatistical modelling as a main DSM modelling approach. For an excellent introduction to machine learning, see James et al. (2013). A more statistical in-depth treatment is provided in Hastie et al. (2008). Malone et al. (2017) explains the use of machine learning and geostatistics for DSM with examples and R scripts.

- There are many different machine learning algorithms. The one most often used in DSM is random forest. Random forest creates multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It works by randomly selecting subsets of covariates and measurements and using these to train individual trees. It can also quantify the prediction error using a variant known as quantile regression forest. Another often used method is Artificial Neural Networks (ANNs), which are a type of machine learning algorithm that mimics the way the human brain works to recognize patterns in data. Convolutional neural networks (CNNs) are a specific type of ANN commonly used in image processing tasks. CNNs are designed to capture spatial contextual information, this can help to improve the accuracy of soil mapping.

- While machine learning is usually used in the spatial domain to derive static soil maps, it can be extended to the space-time domain. The machine learning algorithm and workflow remain the same, the only difference is that part of the environmental covariates that explain soil variation are dynamic. Examples of such covariates are land use, vegetation indices and climate. The machine learning model is trained in the usual way, using paired observations of the soil property and environmental covariates. But extra care has to be taken to ensure that the soil observations are paired with covariate values of the year of soil sampling. An elaboration is to incorporate temporal delay functions to allow that the soil property value in a certain year also depends on covariate values in previous years. For instance, it is well known that the effect of a land use change on soil organic carbon can span more than 20 years. See Heuvelink et al. (2022) for an example of machine learning for space-time soil mapping.

- Machine learning algorithms can be used in digital soil mapping to model complex relationships between soil properties and environmental factors. Advantages include their ability to handle nonlinear relationships and adapt to complex data, while disadvantages are the need for large amounts of data for effective training and potential overfitting. Additionally, machine learning algorithms may be difficult to interpret and may require significant computational resources for training and application.

7.1.4 Digital Soil Mapping standards

GlobalSoilMap

The IUSS Working Group Global Soil Map, formed from the GlobalSoilMap initiative, has published technical specifications for developing national digital soil maps for a set of key soil properties in a standardized way. These specifications have been adopted as a *de facto* (community) standard for developing digital soil maps and many countries have adopted these standards to produce their (national) soil maps. The guidelines also form the basis of the global SoilGrids system.

SOTER

The Soil and Terrain (SOTER) database programme produced soil databases for large parts of the world. A SOTER database consists of a GIS layer with SOTER units (that are a combination of a soil class and terrain elements; originally in shapefile format) and a separate database (originally in MS Access format). Though inactive now, the SOTER program ran for 30 years during which it became a de facto standard for developing national, regional, and continental soil databases. Procedures have been documented in a procedures manual (Van Engelen and Dijkshoorn (Eds), 2013).

Besides these community standards, there are several initiatives that aim to produce harmonized soil information products, also at continental and global level based on defined guidelines, procedures, and technical specifications:

Global Soil Partnership

Global Soil Partnership is a network of stakeholders in the soil domain established in 2012 by members of the Food and Agriculture Organization of the United Nations (FAO). Through the FAO member countries, the GSP develops global soil data products compiled from country contributions, using a bottom-up approach supported by an extensive capacity building programme. Technical specifications and guidelines are established for this purpose to ensure countries map the target soil properties in a standardized way.

Harmonized World Soil Database

The Harmonized World Soil Database is a global soil database that is compiled from a set of regional and continental soil databases, including SOTER, that are harmonized using a standard set of procedures.

EJP SOIL

The European Joint Program on Soil produced a report on harmonized procedures for the creation of both conventional and digital soil maps (Van Egmond and Fantappiè Eds., 2021) and is undertaking research on the optimal combination of bottom-up or country-driven and top-down mapping approaches by testing different combinations of harmonized European input point data and covariates, versus national input point data and covariates. Results are expected in 2024.

USDA

United States Department of Agriculture (USDA) created a soil survey manual that has a wider implementation than other national standards. It provides the major principles and practices needed for making and using soil surveys and for assembling and using related data (Ditzler et al., 2017).

FAO technical manuals

The FAO provides technical manuals for mapping of salt-affected soils and a Cookbook for soil organic carbon mapping. These provides generic methodologies and the technical steps to produce maps.

7.2 Tools and workflows

Digital soil maps are typically produced using (coded) workflows. These workflows are a compilation of computer scripts that use various low level code libraries and (sometimes) higher level tools. These scripts execute a sequence of tasks that result in maps and cross-validation metrics. It is important to document each step and work in a reproducible way. This entails using programming tools instead of point and click solutions. Use of containers, see below, is highly recommended. Both for reproducibility and ease of use. The main tools used in modern DSM will be described below. All are open source if not indicated differently in the description.

7.2.1. Tools

Programming languages for code development of DSM workflows:

R

R is a programming language with numerous packages to work with spatial data and run geostatistical and machine learning models. It works across the major operative (operating) systems. It is used in the academic community. R is typically used to develop DSM workflows and is particularly powerful for data processing and analysis and statistical modelling.

Python

Python is a programming language with numerous packages to work with spatial data and many more user cases. It works across the major operating systems.

Bash

Bash (Bourne Again Shell) is a free and enhanced version of the Bourne shell distributed with Linux and GNU operating systems. A shell program provides access to an operating system's components. The shell gives users (or other programs) a way to get "inside" the system; it defines the boundary between inside and outside. It is used to manage tasks in parallel, work with substantial number of files and low-level tasks.

Libraries:

A library is a merged collection of code scripts that can be used iteratively to save time. It is similar to a physical library in that it holds reusable resources, as the name implies. It contains code bundles that can be reused in a variety of programs[51].

GDAL

GDAL is a translator library for raster and vector geospatial data formats that is released under an MIT style Open-Source License by the Open-Source Geospatial Foundation. As a library, it presents a single raster abstract data model and a single vector abstract data model to the calling application for all supported formats. It also comes with a variety of useful command line utilities for data translation and processing. It works from the command line, python, and R code, among others.

GEOS

GEOS is a C/C++ library for computational geometry with a focus on algorithms used in geographic information systems (GIS) software. It implements the OGC Simple Features geometry model and provides all the spatial functions in that standard as well as many others. GEOS is a core dependency of PostGIS, QGIS, GDAL, and Shapely.

PROJ

PROJ is a generic coordinate transformation software that transforms geospatial coordinates from one coordinate reference system (CRS) to another. This includes cartographic projections as well as geodetic transformations.

[51] https://www.javatpoint.com/library-in-python#:~:text=A%20Python%20library%20is%20also,same%20code%20for%20different%20programs .

Code development and data processing environments:

For **code development** (R as well as python), RStudio Desktop is a popular software solution that can be installed on a local computer, while RStudio Server allows user to run RStudio from a centralized server-based environment that can be accessed from a web browser.

Containers can be used to make DSM workflows portable. A container is a standard unit of software that packages code and all its dependencies so that the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, system tools, system libraries and settings. It is used to install all dependencies and keep track of versions for reproducibility and ease of use. A popular tool for creating and running containerizing applications locally is Docker. Containers can be run on a number of tools and platforms.

To ensure transparent and reproducible DSM workflows, proper code management. There is various tool for this:

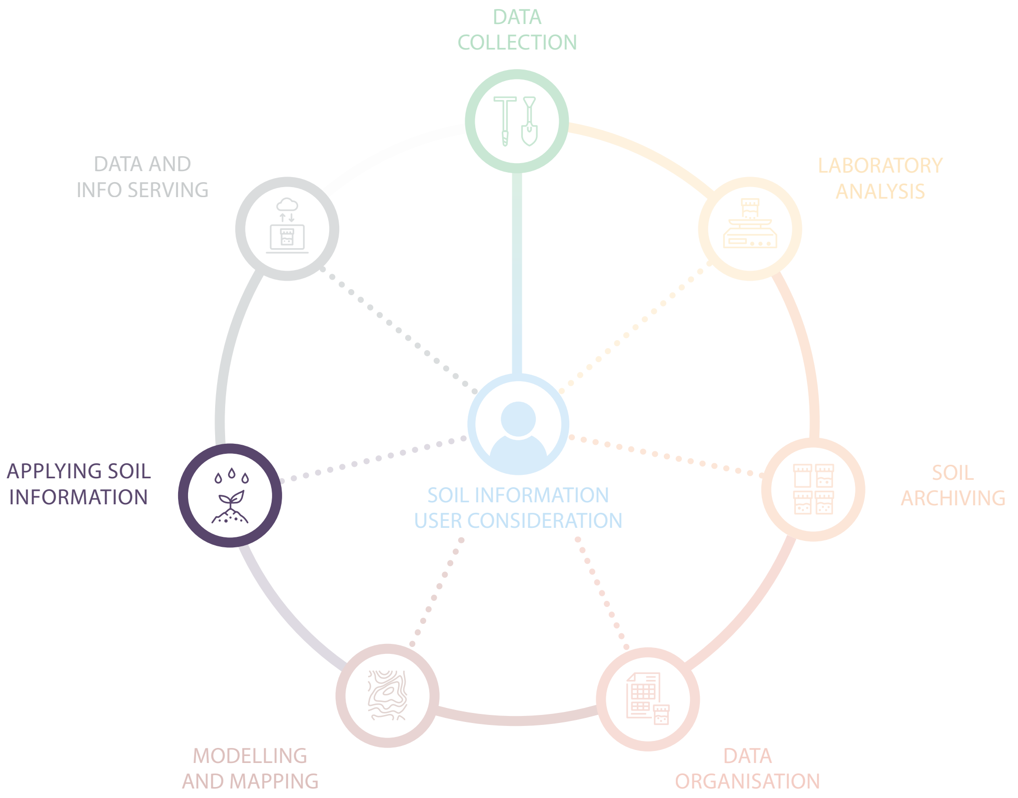
- Git: git is a distributed version control system. It is used to keep track of code versions in DSM workflows and its management during development.
- Though Git is widely used, other tools include Mercurial and Bitbucket.

Other tools are often used to manage the data or to prepare and pre-process covariates. In particular, the following are useful to mention:

- GRASS-GIS for its capabilities in data management and processing and the ease of creating tiles for parallel processing of the data.
- GEE (Google Earth Engine) and its python API is useful to create covariates. However, it requires an account with Google.

7.2.2 Workflows

The tools above can be combined in "workflows", i.e., a series of steps coded in a programming language that can be run by less experienced users. There are no official DSM workflows, but a range of approaches integrating R, Python and other tools are often used. Workflows can have different complexity depending on the results required (maps, uncertainty, validation statistics), on the data size (larger data require parallelization and tiling in the workflow) and on the level of coding of the user (workflows can be a series of scripts to be modified or a semi-automatic tool where changing a configuration file is enough to run the workflow). Several workflows are or will be made available by EJP SOIL and others in 2023 or 2024.


DATA COLLECTION
LABORATORY ANALYSIS
SOIL ARCHIVING
DATA ORGANISATION
MODELLING AND MAPPING
APPLYING SOIL INFORMATION
DATA AND INFO SERVING
SOIL INFORMATION USER CONSIDERATION

8. Applying soil information

After soil modelling and mapping, is the application of soil information the next step in the development of the SIS. Soil information is applied at different **scale levels**: from field to continental and even global. In general, soil information is applied to inform **decision-making processes** related to policy development, planning, and monitoring of the environment at these various scales. Therefore, this chapter is closely related to users and user requirements as addressed in Chapter 2.

At a global level, processes and trends are studied to allow assessments of the state of the soil at coarse spatial scales. Such information is for instance used for status reports on the world's soil resources.

The scale level of most agricultural and environmental development projects typically ranges from farm to country. In such projects, soil information is often the basis for investment planning, for instance for interventions in agriculture (e.g., soil fertility recommendations), land (e.g., improved soil and land management, spatial planning) or landscape (e.g., restoration of degraded environments).

From the decision-making perspective, soil is just one of many domains to be considered. **Integration** with other domains is often a pre-requisite to ensure soil data are used properly and to their fullest.

Soil information can be used in various **applications** such as:

- soil fertility assessment and food security studies;
- soil water conservation;
- carbon stock change assessments and carbon sequestration potential;
- land quality assessment, land evaluation and land use planning;
- assessment and mitigation of soil threats;
- infrastructure construction (roads, cables, bridges, buildings);
- archaeology;
- precision farming;
- soil health assessments;
- teaching and studies on soil variation at different spatial scales and soil formation.

This chapter introduces the tools, challenges, and examples for applying soil information.

8.1 Tools for applying soil information

Soil information can be visualized, summarized, and applied in different formats. In this section, tools for applying Soil Information are discussed.

- Decision Support Systems:
  A Decisions Support System (DSS[52]) for soil management guides users in assessing soil functions on their land and helps to optimize sustainable land management practices (Debeljak, 2019).

- Viewers / Dashboard:
  A dashboard is an online platform for soil information to view and summarize data on soil-related issues. It provides a quick insight into the data and information provided by the SIS. An example is the WoSIS dashboard.

- Scenario models:
  Scenario modelling analyses and evaluates potential future events in order to make better-informed decisions based on the soil information. Scenario modelling is for example used to support soil and water conservation interventions, to model soil conservation function, to model soil organic carbon changes and to model soil health scenarios. These models are of different complexities. More information is in chapter 8.2 Dynamic process and balance models.

- Handbooks:
  Soil information is used to create agriculture-related handbooks to assist with farming needs. An example is the agriculture handbooks from the USDA.

- Planning:
  Soil information can be used for a variety of planning purposes including land use planning, environmental farm planning and watershed management planning[53]. A soil management guide can be used as a guideline.

- Fertilizer recommendations:
  Soil information such as georeferenced crop trail responses can be linked to information on land- and soil-based characteristics to make recommendations and solutions for balanced fertilization at regional as well as farm level[54].

- Apps:
  Apps can be used on mobile devices all over the world. The Soil Quality App (Sqapp[55]) is an example providing location-specific soil quality information and sustainable land use management options. It uses soil information to make recommendations for agricultural management practices.

[52] https://www.frontiersin.org/articles/10.3389/fenvs.2019.00115/full

[53] https://www.gov.mb.ca/agriculture/environment/soil-management/soil-management-guide/soils-information-for-planning-purposes.html

[54] https://ifdc.org/soils-consortium/

[55] https://www.isqaper-is.eu/sqapp-the-soil-quality-app

8.2 Dynamic process and balance models

There are various models for applying soil information. The main topic groups are:

I. **Soil Carbon models:**
soil carbon models are used to model the relationship between soil carbon and other soil parameters, such as soil temperature, moisture, nutrients, soil organic matter. Models are process based, attempting to model the soil processes by describing e.g., interactions between carbon pools, or they are balance models, accounting for inputs and outputs of a system, thereby calculating the resulting balance of e.g., soil organic carbon. The application of these models is often in the field of climate change or in combination with other models. Examples of models are Roth-C, CENTURY, DAYCENT, YASSO, SOMM;

II. **Soil water models:**
Soil water models simulate water flow and/or transport through soils. The application of these models is in the field of agriculture, water management and environmental protection. Examples of models are SWAP, APSIM, WATBAL (Ranatunga et al., 2008);

III. **Soil erosion models:**
soil erosion models are used to simulate erosion, transport, and deposits of soil over land surface. The application of these models is in the field of land management, agricultural management practices and land use. Examples of models are RUSLE, SWAT, MMF, WEPP, PESERA;

IV. **Nutrient Transport Models:**
nutrient or e.g., pollutants models simulate the pathways and extent to which nutrients move through soil, water, watersheds. Examples of models are: ANIMO, VEMALA, INCA;

V. **Crop response models:**
crop response models are used to estimate crop yield, growth and/or production from soil properties. The application of these models is in the field of fertilizer recommendations, precision management, regional assessment of climate variability and climate change. Examples of models are QUEFTS, WOFOST, DSSAT, SUCROS, AQUACrop.

Combinations of models from these groups are also used, an example of this is the SWAP-WOFOST model or the MITERRA model, which was lately enriched with the Roth-C model.

A more extensive but non-exhaustive list of some of the models mentioned here is provided in Annex VI.

8.3 Challenges

A SIS is an innovative tool which provides important baseline information, can monitor change, be a basis for or provide recommendations and advice for further applications. The relevance of a SIS to users, and therefore the success of a SIS, depends to a substantial extent on the presence and accessibility of relevant, up-to-date, reliable, and FAIR (Findable, Accessible, Interoperable, Reusable) data. In addition, aspects such as functionality, user-friendliness, and ease of use of the system are factors for success. Although a SIS should be designed with the user needs in mind, a SIS is nonetheless not always able to function optimally due to several reasons related to the data, or content:

1. Lack of available data:
   Available data is essential in a SIS to provide recommendations and advice for further applications by users. A challenge to a SIS not containing all or sufficient relevant data to answer the information questions of the users. This can either be because the data was never collected and can therefore not be made available in the SIS. Or the data is present, but is not made available, either open or after registration or payment. The first can be due to the soil collection policy or incentives in the country in past years or the present. The latter can be due to privacy sensitivity or financial restrictions of the data. Privacy sensitivity can be due to personal information in the data such as names and addresses. A financial restriction can be incurred because some institutes depend on the revenue of data provisioning for new data collection. Remedies to this challenge are in many cases related to soil data collection and provisioning policies, political choices with respect to the openness of (soil) data (Dutch BRO[56], USA OPEN Government Data Act[57]), viable business models that address both data producers' and data users' needs.

2. Lack of reliable data:
   Reliable data in the SIS is at the heart of science and a requirement for results to be accepted as factual. Too often, however, details, metadata and data are not publicly available to repeat a study, i.e., perform it again in a comparable manner (Bond-Lamberty, 2016) and therefore to check the reliability of the data in the SIS. This influences the use of the data for users. Note that reliable data is not the same as accurate or harmonized data but indicates if the data can be trusted and its origin and derivation is clear and transparently described. The lack of reliable and harmonized soil data has hampered the use of the SIS for global assessments and environmental impact studies, land degradation assessments and adapted sustainable land management interventions. Reliability of maps and harmonized data can be judged better, and reproducibility is enhanced when coded workflows are used that are managed via code repositories such as Github or Gitlab. Using coded workflows should be the norm when producing soil information for SIS.

3. Lack of structured or harmonized data:
   To enable optimal use of the soil data in the SIS it should be at least findable and accessible, and ideally interoperable and (easily) reusable. The findability of data in a SIS is greatly enhanced by proper, and standardised, description of its metadata. A SIS can sometimes be used as a data-dump where all available data is added. However, when the data is not described systematically (i.e. the metadata), for the users the relevant data is not easy to find and to use. It's interoperability and reuse is greatly enhanced by a structured, and if possible standardised (data model standardisation according to an ontology) form of the data, where the structure or standard used is described in the metadata. This representation in machine readible format allows easier exchange of data. In the SIS, it is essential to structure the information to create an optimal use for

[56] https://basisregistratieondergrond.nl/english/about-key-registry/
[57] https://www.congress.gov/bill/115th-congress/house-bill/1770

the information users. Another step that increases the interoperability and reuse of the data is the harmonisation of data ittself beyond the harmonised description of the data. Preferably this is described in the metadata and is transparently done.

4. Lack of up-to-date and dated data:
   Soils differ not only in space but also in time. The date of acquisition of soil data in the SIS should therefore always be provided in the metadata belonging to the data. The most recent data should be presented first and when not up-to-date, this should be clearly indicated. Lack of up-to-date data in the SIS results in bad-informed recommendations and advices for the information users. The inclusion of older data in the SIS, when properly described, allows for time series analyses and can be of great value. Another aspect of up-to-date data is keeping the accessibility (e.g. format, persistence, repository) and description of the data up-to-date. This activity needs to be organised. Too often data is not kept up-to-date – especially after a project ends. The hosts of the SIS should ensure that the provided data is updated and checked.

5. Lack of integration of soil properties data with other relevant data :
   Information services (e.g., models, apps, websites) combine spatial soil information with spatial data on other relevant aspects of land, crops, and climate in operational platforms. Some SISs lack up-to-date technology in these aspects and, therefore, cannot integrate these aspects. The SIS is therefore unable to provide meaningful interpretations of soil data in relation to data from adjacent domains for information users.

6. Lack of model compatible soil data:
   Soil data that is useable in models is in most SISs a challenge. More attention should be paid on the compatible soil data for models which is required for further processing. For example, there is a large demand for DSSAT and SWAT compatible soil grids, but without compatible data, these grids cannot be created. Protocols for developing model compatible soil data/information have been developed by Han, et al. 2019 and Dalgliesh, et al. 2016.

7. Lack of human technical resources:
   Maintaining, operating and usage of a SIS depends greatly on the available human (technical) resources. A SIS only has value if it is kept up to date by the hosting institute. If the human technical resources are lacking to do this continuously or at least regularly (weekly), the sustainability and therefore impact and usability of a SIS to end users will be in danger. It is therefore always important to ensure capacity building of the hosting institute as part of setting up a SIS.

8. Lack of technical security measures:
   A SIS is a digital product and can be subjected to cyber-attacks. If the security of the IT systems that supports a SIS is not present, this can affect the sustainability of a SIS. It is important that the security is in place to avoid cyber-attacks resulting in the SIS going offline. Secondly, good security measures also ensure that the data, knowledge and procedures that make up the SIS are secure against unwanted edits or deletion. Therefore, securing the SIS with proper measures and systems against cyber-attacks is vital for the SIS.

8.4 Examples

Soil information in a SIS can be applied for various purposes. Below, examples are given on how information in a SIS can be used.

- Integrated Soil Fertility Management (also described in chapter 2):
  integrated Soil Fertility Management (ISFM) uses soil, crop and climate data as well as soil and crop growth models to provide soil and crop specific fertility advice. It has the potential to improve effectivity and efficiency of agronomic practices, including fertilizer recommendations and organic matter management, and thereby boost crop production and farm income. Land users and their intermediaries can use the system to obtain advice on the soil fertility status and improvement of soil fertility using spatial nutrient gap analysis based on yield response data. The soil data is retrieved from or stored in a SIS.

- Soil and Water Conservation (also described in chapter 2):
  soil and water conservation (SWC) uses soil, landscape and hydrological data and models and focusses on engaging stakeholders in more sustainable land use and land management practices. Catchment managers, authorities, extension staff and farmer organizations can use the system to obtain information on land use and land management practices and their suitability with regards to soil and water conservation and climate change adaptation. The soil data is retrieved from or stored in a SIS.

- Land Use Planning:
  Land Use Planning is the process of deciding which land use is allowed or adviced where in an area or landscape. This multi-sectoral decision making process needs to balance the demands for use of space for e.g. agriculture, nature, urban and industrial land use with the available area and its suitability for these different land use types as governed by the landscape, soil and water characteristics of the area. Relevant, up-to-date and sufficient quality soil data is vital for better informed decision making. Land evaluation, the evaluation of the suitability of an area for a specific land use or crop, can be used to assist in this process. The SIS can be used as a platform for information on land use and land use change processes[58] when adding data and information on land use, landscape, water characteristics.

- Integrated Landscape Management:
  Integrated Landscape Management (ILM) uses soil and other data at local scale. ILM is the management of production systems and natural resources in an area large enough to produce vital ecosystem services such as food production, water storage, providing biodiversity, etc. and small enough to be managed by the people using the land and producing those services[59]. The soil data needed for this can be retrieved from or stored in a SIS.

- Soil quality (health) indicators:
  For decision making or policy development and evaluation purposes it is often convenient to 'summarize' the status of a soil by looking at one or several indicators that are considered to be indicative for the overall quality or health of that soil. Although soil quality

[58] https://www.isric.org/projects/laurel-land-use-planning-enhanced-resilience-landscapes
[59] https://www.soils4africa-h2020.eu/serverspecific/soils4africa/images/Documents/UseCases.pdf

and soil health are often used as synonyms, they have a distinctly different meaning. Soil quality is usually relative to a soil use, e.g., the quality of a soil to perform certain soil functions or ecosystem services. Soil health is the healthiness of a soil compared to its optimal state, often considered to be under permanent grassland or nature. It is therefore also an indication of soil degradation. Indicators are usually considered in bundles of physical, chemical, and biological indicators that can be combined to form an index. Often soil indicators are assessed based on measurements of soil parameters at a baseline point in time or situation and their change over time is measured in soil monitoring systems, repeated measurements over time at the same locations. Apart from monitoring change, the indicators are also used to evaluate the status of the soil against threshold values. Useful thresholds are typically dependent on the soil type, land use and climate of the soil[60], although general rules of thumb are also used. Various threshold systems exist, each with their pros and cons[61]. Examples are the Land Degradation Neutrality[62] indicators, the EU Soil Strategy[63] indicators, national soil indicator systems (an overview of European indicators and soil monitoring systems is reported by the EJP SOIL project[64],[65]), the overview and analysis by the EEA[66], the Soil Health Institute[67] indicators. The Global Soil Partnership is working on a set of global soil health indicators as well.

- MRV systems

  Consistent monitoring of changes in soil organic carbon (SOC) stocks ––and net GHG (greenhouse gas) emissions)–– reporting, and their verification (MRV), is key to facilitate investment in sustainable land use practices that maintain and increase soil carbon, as well as to incorporate soil carbon sequestration in GHG emission reduction targets at the international and national level. An MRV framework provides a theoretical description or concept of a comprehensive MRV system. The framework is defined by the context of the MRV, for example assess changes in SOC over time in croplands subject to defined land use/management interventions or changes in policies. The framework itself consists of various components (e.g., monitoring, modelling, and reporting) aimed at quantifying and verifying SOC change over time vis a vis a baseline and intervention scenario. Each of these components is characterized by a set of methodologies (e.g., field sampling protocols, type of model used, and verification procedures); these are described in protocols that provide a step-by-step procedure on how to solve an issue, following a uniform set of standards. The soil information used and produced in MRV systems can be stored and accessed in a SIS or multiple SISs. As such they are considered a vital component of an MRV system. For additional information see, for example, GSP SOC MRV or the WB MRV sourcebook as well as the state-of-the-art paper by Smith et al (2020).

[60] https://soilhealthbenchmarks.eu/

[61] https://ejpsoil.eu/science-to-policy/workshop-carbon-farming-1

[62] https://www.unccd.int/land-and-life/land-degradation-neutrality/ldn-principles

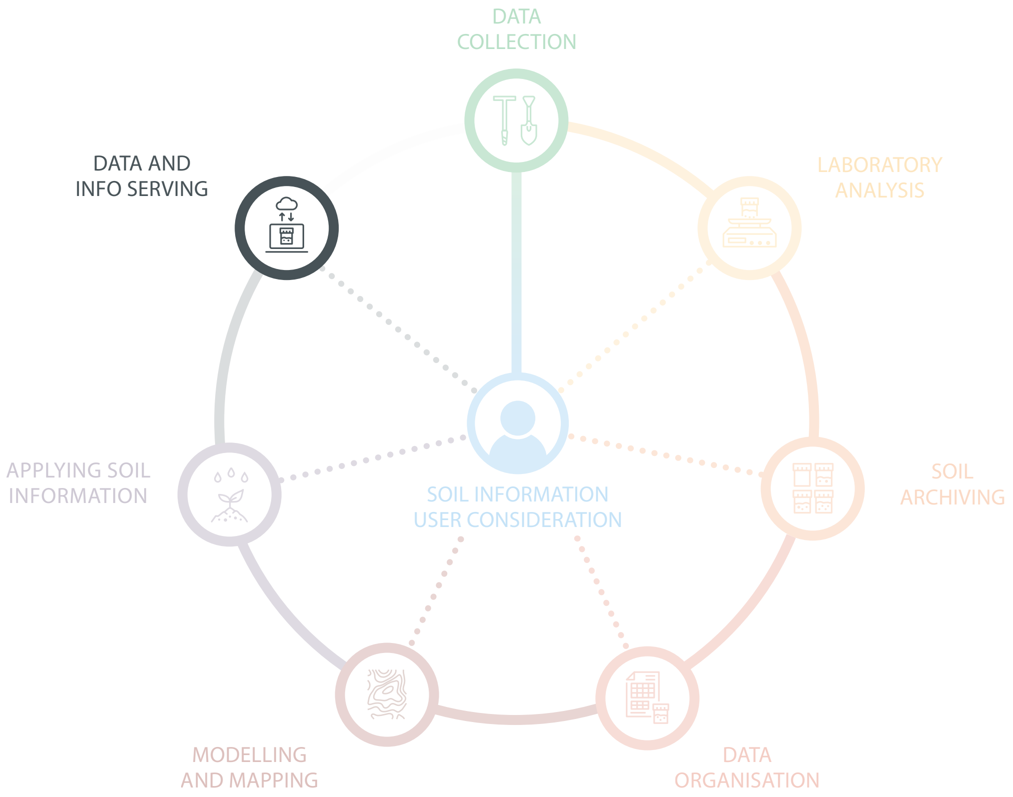
[63] https://environment.ec.europa.eu/topics/soil-and-land/soil-strategy_en

[64] https://ejpsoil.eu/fileadmin/projects/ejpsoil/Policy_briefs/SIREN/SIREN_Policy_brief.pdf, https://ejpsoil.eu/fileadmin/projects/ejpsoil/WP6/EJP_SOIL_Deliverable_6.3_Dec_2021_final.pdf

[65] https://ejpsoil.eu/science-to-policy/workshop-carbon-farming-1

[66] https://www.eea.europa.eu/publications/soil-monitoring-in-europe

[67] https://soilhealthinstitute.org/our-work/initiatives/measurements/


DATA
COLLECTION
LABORATORY
ANALYSIS
SOIL
ARCHIVING
DATA
ORGANISATION
MODELLING
AND MAPPING
APPLYING SOIL
INFORMATION
DATA AND
INFO SERVING
SOIL INFORMATION
USER CONSIDERATION

9. Data and information serving

The last step of the development of a SIS is serving the data and information present in the SIS in an online environment to make it accessible for users. **Data and information serving** is the process of making soil data available within the organization as well as to partners and to the general public (i.e., publishing soil data). As part of soil data publication, it is important to consider aspects such as **Findability, Accessibility Interoperability, and Reusability of data (FAIR).** Because soil data are inherently spatial, conventions around data publication of the spatial data community are essential. The spatial data community with its standardization body, the Open Geospatial Consortium (OGC), has been successful in its definition and adoption of standards. A relevant concept from the geospatial data community is the **Spatial Data Infrastructure (SDI)**, a technical infrastructure to facilitate data sharing based on distributed components connected via standardized API's.

This chapter introduces various standards and tools available to facilitate data and info sharing. More detailed technical information for developers on this topic can be found at the Soil Data Assimilation wiki[68] developed by ISRIC in the context of the EJP SOIL project.

9.1 Standards

When choosing and building the data and information provisioning services to provide the SIS data online it is highly advisable to use existing standards for spatial data exchange on the web where the choice is between (or a multitude of) the standards described in chapter 9.1.1, 9.1.2, 9.1.3, 9.1.4. For each of these standard's tools are available and described in chapter 9.2.

To encode or structure the data that is exchanged in a data model it is advisable to provide the data in one of the ontologies described in chapter 9.1.5. Alternatively, the data can be provided in the existing data model structure (e.g. the one used in the SIS backend) but a mapping to one of the main ontologies is desirable to make the data interoperable and enable reuse.

A vital last step in which it is relevant to consider standards is the determination of the data and software policy and licenses that are assigned to respectively the data in the SIS and possibly the software developed to provide specific functionality (e.g. advice as specified in chapter 8). Main standards for licenses are described in chapter 9.1.6.

9.1.1 Traditional OGC Standards

Modern data infrastructures where the data is maintained at the source are connected via web services. Web services facilitate machine-to-machine interaction via the Internet. Using standardized Application Programming Interfaces (API's) lowers the effort required to connect to web services. Web services for exchange of geospatial data are commonly based on Open Geospatial Consortium (OGC) standards.

[68] https://ejpsoil.github.io/soildata-assimilation-guidance/

Below are examples of OGC standards:

- Catalogue Services for Web (CSW) defines a common interface to discover, browse and query catalogues;
- Web Map Service (WMS) facilitate map visualizations of data;
- Web Feature Service (WFS) used for accessing and manipulating vector data;
- Web Coverage service (WCS) used for accessing and manipulating raster data;
- Web Processing Service (WPS) facilitates spatial processing functionality;
- Sensor Web Enablement (SWE) used for accessing and manipulating sensors and instruments.

These standards are established in the geospatial industry. OGC and World Wide Web Consortium (W3C) initiated the design of a new generation of spatial data exchange standards. This resulted in a new range of OGC API standards.

Many soil data fit the model of Observations and Measurements (O&M) created by OGC. Both field observations, soil profile data and laboratory analyses fit in the model. The soil standard ISO28258, therefore, has Observations and Measurements in its core. Sensor web standards are an optimal fit to share and use observation data. Sensor web has advanced query capabilities to filter the relevant observations.

9.1.2 OGC API

OGC is following an agile approach in developing a new generation of spatial data exchange standards, called OGC API. OGC API builds on existing standardization efforts, such as REST and Open API. Every OGC API standard has a fairly minimal starting point and additional functionality can be added using extensions. Many of the existing tools can be used with little to no effort. An additional benefit of OGC API standards is its ability to be found by search engines. This is an important driver of the development of OGC API and the reason for adoption by a wider audience.

Some of the OGC API standards currently being developed and adopted:

- OGC API – Maps & Tiles. Draft specification describes an API that can serve spatially referenced and dynamically rendered electronic maps and tiles;
- OGC API – Features. Offers the capability to create, modify, and query spatial data on the Web;
- OGC API – Coverages. Specification that defines a Web API for accessing coverages and grids;
- OGC API – Records. Multi-part draft specification that offers the capability to create, modify, and query metadata on the Web.

9.1.3 GraphQL

Use of OGC API is a useful mechanism to share basic soil data products, such as soil property and soil class maps. However, OGC API is less optimal to share a more complex system of for example measurements on soil samples taken at a certain depth of a profile at a plot location. OGC APIs have not been designed around systems with hierarchical relations between features. The linkage mechanism between features is limited and query capabilities over these links is non-existent.

For more complex systems GraphQL is a de facto standard. GraphQL is a self-describing API providing hierarchical query capabilities. The GraphQL community has made it extremely easy to set up a GraphQL API on any database using the PostGraphile extension.

9.1.4 Semantic web

Even more advanced are the expression and query options of the semantic web. These describe data using common ontologies, such as SOSA. Users can query the data following the conventions from those ontologies using SPARQL.

The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable IT specialists to create data stores on the Web, build vocabularies, and write rules for handling data.

9.1.5 Ontologies

Data ontology is a way of structuring or linking data in various formats. It shows relations that exist between entities. The following soil data ontologies are in use by the community:

ISO 28258
The international standard ``Soil quality — Digital exchange of soil-related data'' (ISO 28258:2013) is meant to provide a general framework for the recording and unambiguous exchange of soil data, consistent with other international standards and independent of particular software systems. Aligned with the Observations & Measurements standard (Cox 2011), ISO 28258 captures a wide range of concepts from soil surveying and physio-chemical analysis, including Site, Plot, SoilProfile, ProfileElement and SoilSpecimen.

INSPIRE Soil
A detailed data specification for the soil domain was published by the European Commission in 2013 (INSPIRE Thematic Working Group 2013), as part of its aim to create a spatial environmental data infrastructure. The INSPIRE domain model targets inventories of soil conditions and soil properties with soil monitoring over time in mind, but also soil mapping, primarily derived from soil inventory data.

GLOSIS
The Global Soil Partnership launched a call for an international consultancy to assess the state-of-the-art in soil information exchanges and propose a path towards its operationalization as the backbone of Global Soil Information System (GloSIS). The GloSIS web ontology has been successfully demonstrated as a vehicle to exchange soil information as Linked Data (Palma et al., 2022).

AnzSoilML
The Australian and New Zealand Soil Mark-up Language (AnzSoilML) (Simmons et al., 2013) was the first application of O&M to the soil domain. Its domain model targets soil properties and related landscape features specified by the institutional soil survey handbooks used in Australia and New Zeeland (National Committee on Soil and Terrain (Australia), 2009; Milne et al. 1995).

9.1.6 Soil data licenses

Releasing data without making the terms of use clear can be confusing and counterproductive, as it may deter potential users from using and citing the data. One of the most effective ways to communicate permissions to potential users of data are data licenses. A data license is a legal arrangement between the creator of the data and the end-user, or the place the data will be deposited, specifying what users can do with the data. Creative common (CC) is the most commonly and widely used suite of data licenses. They clearly describe how data can and cannot be reused; CC licenses are irrevocable. This means that once you receive material under a CC license, you will always have the right to use it under those license terms, even if the licensor changes his or her mind and stops distributing under the CC license terms. There are six different CC license types, ranging from most (CC0) to least permissive (e.g., CC BY NC SA), for details see https://creativecommons.org/about/cclicenses/.

Alternatively, for software a different suite of licenses is required (e.g., MIT or GPL). Software licenses typically provide end users with the right to one or more copies of the software without violating copyrights. For additional details and options please see: https://choosealicense.com/licenses/.

9.2 Tools for data and information serving

Soil datasets and services must be made findable, accessible, and reusable. This can be done by describing and serving the metadata online. Metadata is data that describes other data and is a mechanism that allows users to find, analyze and evaluate a given soil data resource. The metadata also includes an assessment of potential access constraints to the data. Data with access constraints may require measures to be put in place to limit access to the resource. A metadata resource is typically a registry or catalogue to find the data ("is this dataset suitable for my use case?"). Below are examples of catalogue services that make soil datasets and services discoverable. Most to all SISs use a catalogue service to display their metadata online.

9.2.1 Catalogue services

GeoNode

GeoNode[69] is an open-source platform for creating and sharing geospatial maps and data, with a focus on collaboration and the creation of interactive web maps. It was designed to provide organizations with a one stop SDI solution to publish spatial data on the web from an easy-to-use interface. It is built using Django and provides a web-based interface for managing geospatial data, as well as tools for creating and publishing maps. GeoNode is a spatial content management software, so it is not a catalogue, which can import and link to external resources. Also, GeoNode is complex to setup and maintain due to its wide range of components and technologies used (Java, Django, GeoServer). The ease-of-use comes with a limitation of functionality: it is difficult to customize the platform for user specific needs (i.e., setting up a customized metadata profile) and security aspects are challenging.

[69] https://geonode.org/

GeoNetwork

GeoNetwork[70] is a platform for managing and sharing metadata about geospatial data. It is built using Java and provides a web-based interface for managing metadata records, as well as tools for searching and discovering geospatial data. GeoNetwork provides support for a range of metadata standards, including ISO 19115 and FGDC, and can be used to create a centralized repository for metadata about geospatial data. It provides powerful metadata editing and search functions as well as an interactive web map viewer. GeoNetwork is more focused on enabling users to store and search spatial metadata and access the corresponding data.

CKAN

Comprehensive Knowledge Archive Network (CKAN[71]) is an open-source data portal and data management system. CKAN is a more general-purpose data management platform. It is often used for managing non-spatial data, but also supports geospatial data through extensions. CKAN provides a wide range of data management functionality, including data storage, discovery, and visualization. CKAN is more flexible and customizable than GeoNetwork and is often used as a central data portal for organizations with a wide range of data types. CKAN is more focused on enabling users to find and use more generic data types, not necessarily geospatial data.

Dataverse

Dataverse[72] is another general-purpose open-source platform for storing, managing, and sharing research (spatial) data. It is specifically designed for social science research data, and it provides features specifically tailored for this type of data, including support for versioning and access control, data citation, and the ability to publish and discover datasets. Just like CKAN, geospatial data support can be enabled through extensions. Dataverse is focused more on allowing users to share, preserve, cite, and explore research data.

Esri Geoportal Server

Esri Geoportal Server[73] is a free, open-source product from ESRI that enables discovery and use of geospatial resources including datasets, rasters, and Web services. It helps organizations manage and publish metadata for their geospatial resources to let users discover and connect to those resources. It is well linked with the proprietary GIS products that ESRI develops such as ArcMap/ArcCatalog and ArcGIS Online.

9.2.2 Host data on a repository

A minimal approach to data sharing is to distribute the dataset as a single file. This approach is valid if the dataset does not receive daily updates and is of moderate size. It is important to verify that the data file is accompanied with relevant metadata, so that the user is aware of the contents. The metadata will also help in the findability of the dataset.

[70] https://geonetwork-opensource.org/
[71] https://ckan.org/
[72] https://dataverse.nl/
[73] https://www.esri.com/en-us/arcgis/products/geoportal-server/overview

Many repositories exist via which one can distribute or access data. Some examples are:

Zenodo

Zenodo[74] is a very persistent (> 25 years guaranteed) repository hosted by CERN for academic data (and other resources). Zenodo provides an extensive mechanism to create metadata and has a DOI printing capability.

Github

Github[75] is traditionally a platform for sharing source code and software. But it is increasingly used to work collaboratively on documentation, but also datasets. Github is able to efficiently store changes on the dataset. Github includes a wiki and issue tracker to facilitate documentation and user feedback. File upload is preferably in a text-based format, such as CSV or GeoJSON. Github presents these formats as tables and maps.

Re3data

Re3data[76] is a global registry of research data repositories that covers research data repositories from different academic disciplines. It includes repositories that enable permanent storage of and access to data sets to researchers, funding bodies, publishers, and scholarly institutions. Re3data promotes a culture of sharing, increased access, and better visibility of research data. The registry has gone live in autumn 2012 and has been funded by the German Research Foundation (DFG).

GEE

Google Earth Engine[77] (GEE) combines a multi-petabyte catalogue of satellite imagery and geospatial datasets with planetary-scale analysis capabilities. Scientists, researchers, and developers use Earth Engine to detect changes, map trends, and quantify differences on the Earth's surface. GEE is available for commercial use and remains free for academic and research use.

WRI

The World Resource Institute[78] (WRI) believes that good data is the foundation of good decision-making. They produce data sets, data platforms and data-based tools, which they make freely available through their open data commitment.

GBIF

The Global Biodiversity Information Facility[79] (GBIF) is an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth.

ESDAC/EUSO

EU Soil Observatory[80] (EUSO) is a dynamic and inclusive platform that provides the relevant European Commission Services and the broader European soil user community, with knowledge and data flows needed to safeguard and restore soils.

[74] https://zenodo.org/
[75] https://github.com/
[76] https://www.re3data.org/
[77] https://earthengine.google.com/
[78] https://www.wri.org/
[79] https://www.gbif.org/
[80] https://joint-research-centre.ec.europa.eu/eu-soil-observatory-euso_en

National data portals
Governments host data portals containing relevant national datasets on soil. Aggregated portals, such as the INSPIRE Geoportal for Europe, exist which aggregate the content of these portals to a continental level.

GEOSS
Global Earth Observation System of Systems[81] (GEOSS) is a set of coordinated, independent Earth observation, information and processing systems that interact and provide access to diverse information for a broad range of users in both public and private sectors. GEOSS links these systems to strengthen the monitoring of the state of the Earth.

Relevant software to implement an online file repository locally are:

Apache Webdav
Apache webdav[82] is a module on the popular apache webserver providing webdav access. With webdav users can easily work with your data. One can for example mount a remote webdav folder as a drive-in window.

Gitlab
Gitlab[83] is an open-source implementation of a sharing platform based on the git protocol.

DataPackage
DataPackage[84] is a set of conventions and tools for working with data files and metadata on a file repository.

9.2.3 Spatial Convenience APIs for the web

Instead of downloading a full dataset from a data repository, OGC defined a series of standardized APIs through which users can request filtered subsets of the data up to rendered maps of sections of the data. Various software implementations exist which provide these convenient APIs for spatial data.

Various strategies are available to host data through APIs.

- Host your data in the cloud (SAAS);
- Host data within your organization, but maintained by a service provider (on-premises-managed);
- Set up and maintain the infrastructure yourself (on-premises-self);

Maintaining an infrastructure on-premises gives you full control of the data and functionality. However, it is costly because it requires trained staff with 24x7 availability.

[81] https://www.earthobservations.org/geoss.php
[82] https://httpd.apache.org/docs/2.4/mod/mod_dav.html
[83] https://about.gitlab.com/
[84] https://testing.datahub.io/docs/data-packages

This section lists software which provides OGC compliant service. These tools can be installed by the user, or SAAS providers will offer it as a service.

Mapserver:
Mapserver[85] is an open-source C server application providing OGC Data services on various data backends. MapServer is an efficient, scalable, and lightweight application, making it ideal for serving large or complex data sets. However, it does not have user friendly web interface to manage and style the datasets and has a smaller user community compared to, for instance, GeoServer.

GeoServer:
GeoServer[86] is a java based open-source GIS server application providing OGC Data services on various data backends. It is a full featured application with web administration and authorization making it easy to use even for users with limited technical skills. However, it is a resource intensive application and is not optimal for serving complex soil datasets.

QGIS server:
QGIS server[87] is an open-source GIS server application providing OGC Data services on various data backends. QGIS Server provides a wide range of features and capabilities for managing and serving GIS data and services, including support for rendering maps, providing data access, and performing analysis. QGIS Server integrates with the QGIS Desktop GIS software, which provides a powerful and comprehensive GIS software environment.

Pygeoapi:
Pygeoapi[88] is an open-source Python library that implements the OGC API suite of standards for geospatial data access. It provides a RESTful API for accessing and manipulating geospatial data and services, including data sources such as vector data, raster data, and time series data. Pygeoapi is a lightweight and flexible option for serving and accessing geospatial data and services, with limited features for visualization and integration with a GIS software environment.

ArcGIS Server:
ArcGIS Server[89] is a GIS server software that allows users to publish, manage, and access GIS services and data. ArcGIS Server is typically installed on-premises or in a private cloud and provides a range of tools for creating, managing, and sharing GIS services, such as maps, geospatial data, and analysis tools. The SAAS alternative to ArcGIS Server is called ArcGIS Online.

9.2.4 Map Visualization

The tools in this section can be used to create map visualizations in a web environment. This is optional in a SIS. Consider that most of the catalogue software in the section above includes a map or graph visualization option. But in many cases, you require a dedicated map viewer for a dedicated use case.

[85] https://mapserver.org/
[86] https://geoserver.org/
[87] https://docs.qgis.org/2.14/en/docs/user_manual/working_with_ogc/ogc_server_support.html
[88] https://pygeoapi.io/
[89] https://enterprise.arcgis.com/

Various JavaScript libraries exist which can enable an interactive map viewer on a website. On one hand there are the commercial SAAS providers such as Google Maps, Bing Maps, mapbox.com, maptiler.com, and arcgis.com which typically offer a range of background maps and an SDK to create interactive map applications. On the other hand, there are software libraries, like LeafletJS, OpenLayers, terra.js, which you can use to build a map application, while using data services provided by yourself or others. Overall, online map applications are great for creating web-based map applications, but they require developer skills to be set up.

9.2.4.1 Map applications

Many applications exist that can be used to create map applications for the web, also called WebGIS. This is optional in a SIS. Some examples are given below:

QGIS2web:
QGIS2web[90] is a plugin for QGIS which can prepare a web application from the current map view in QGIS Desktop;

MapStore:
MapStore[91] is a WebGIS framework to create, manage and securely share maps and mashups;

Oskari:
Oskari is a framework for easily building multipurpose web mapping applications utilizing distributed Spatial Data Infrastructures;

Wegue:
Wegue[92] (WebGIS and Vue) combines the power of Vue.js and OpenLayers to make lightweight webmapping applications;

Vertigis Studio:
Vertigis Studio[93] (previously Geocortex) is a platform for creating maps and reports on distributed sources.

A number of tools exist which enable users to create map visualizations on their desktop. The tools can download the data or connect to a remote service and create the visualization. These tools can advertise be advertised as part of your dissemination strategy.

QGIS:
QGIS[94] is a free and open-source desktop GIS (Geographic Information System) application. QGIS provides a wide range of tools for working with geographic data, including data import, manipulation, analysis, and visualization. QGIS provides more advanced features than Leaflet/OpenLayers, such as raster and vector analysis, spatial SQL queries, and support for a wide range of data formats, however it does not have web-based deployment.

[90] https://www.qgistutorials.com/en/docs/web_mapping_with_qgis2web.html
[91] https://github.com/geosolutions-it/MapStore2/
[92] https://github.com/wegue-oss/wegue
[93] https://www.vertigis.com/vertigis-studio/
[94] https://www.qgis.org/en/site/

ArcGIS Pro:
ArcGIS[95] Pro is a proprietary product, part of the larger ArcGIS platform developed by Esri. It is powerful GIS software for data visualization, analysis, and management. ArcGIS/ArcMap provides a wide range of tools for working with geographic data, including data import, manipulation, analysis, and visualization. ArcGIS/ArcMap also integrates with the larger ArcGIS platform, which provides additional tools and services for working with GIS data and provides commercial support.

Golden Software Surfer:
Surfer[96] provides an extensive set of modelling tools to display data while maintaining accuracy and precision.

9.2.4.2 Dashboarding software

Providing an overview of various aspects of a dataset through a series of diagrams on a dashboard is a popular visualization strategy, leading to improved decision-making and more effective data analysis. This is optional in a SIS. Diagram visualization is increasingly embedded in catalogue software, but for advanced analyses options, you can better use or provide dedicated software. Some examples of dashboarding software include:

Kibana:
Kibana[97] is a dashboarding platform that provides a powerful set of tools for visualizing and exploring data. Being part of Elastic Stack, a key strength of Kibana is its ability to work with substantial amounts of data and provide real-time insights and analysis. At the same time Elastic is resource intensive and requires expertise to set up.

Apache Superset:
Apache Superset[98] is an open-source data visualization and business intelligence platform that allows users to create and share interactive dashboards and reports. Superset provides a wide range of visualization options, including charts, tables, maps, and pivot tables, and allows users to easily create custom visualizations with its built-in SQL editor. Additionally, Superset supports multiple data sources, including databases and big data platforms, and provides robust security and access control features. However, Superset offers a less flexible and customizable user interface, and interactive functionality of some visualizations is limited compared to Tableau and PowerBI.

PowerBI:
PowerBI[99] is a proprietary data visualization and business intelligence software part of Microsoft Office family products. PowerBI allows users to connect, visualize, and share data with interactive dashboards, charts, and reports. Power BI has a more limited range of data source options (compared to Tableau) but has a strong focus on integration with Microsoft products and services.

[95] https://www.arcgis.com
[96] https://www.goldensoftware.com/products/surfer
[97] https://www.elastic.co/kibana
[98] https://superset.apache.org
[99] https://powerbi.microsoft.com

Tableau:
Tableau[100] is proprietary data visualization and business intelligence software that allows users to connect, visualize, and share data with interactive dashboards, charts, and reports. It is designed to help people see and understand their data and make informed decisions. Tableau offers a wider variety of visualizations (compared to other dashboarding tools) and features flexible and customizable user interface. However, it could be less cost-efficient and more difficult to publish Tableau non-public dashboards.

9.2.5 Validation and service quality

Once soil data is published on the Internet, there is a need to validate the configuration and setup of the published web service to ensure that they are operating as expected and meet specific standards and quality requirements. This is highly advisable in a SIS. In case a user hosts data at a service provider, these quality aspects are usually verified by the service provider. When selecting a service provider, one should validate if the provider monitors and reports on these aspects. There are four aspects which are relevant to monitor the infrastructure:

- OGC Standards compliance;
- Service availability, capacity & performance;
- Service Usage;
- Security incidents and risk of data loss at incidents.

This is the task of validation tools that include (but not limited to):

1. **Tools that validate standards conformance of the service:**
   - GeoHealthCheck:
     GeoHealthCheck[101] is an open-source tool from the geopython community. GeoHealthCheck performs tests on running geo web services at intervals to test on availability and conformance of a service to OGC standard. For example, on a WMS it drills down from GetCapabilities to GetMap and GetFeatureInfo requests on each of the layers advertised.
   - Team engine:
     Team Engine[102] is the utility used by Open Geospatial Consortium to check standards compliance of software products. You can use the online service or install the utility to evaluate any software yourself.

2. **Tools that validate the availability of the service:**
   - Zabbix:
     Zabbix[103] is an open-source framework to assess availability of a service and monitor hardware resources.

[100] https://www.tableau.com
[101] https://geohealthcheck.org
[102] https://cite.opengeospatial.org/teamengine
[103] https://www.zabbix.com/community

- GeoHealthcheck, see the above, it can also be used to assess the availability.
- Uptimemonitor:
  Uptimemonitor[104] is one of many SaaS providers which can assess your service on availability at intervals.

3. **Tools that monitor the usage of the service:**
   - Elastic Search/logstash[105], parses webserver access logs and shows in interactive dashboards (steep learning curve/heavy setup);
   - AWStats[106], parses webserver access logs and provides reports (minimalistic);
   - Matomo[107], parses webserver access logs and evaluates actual website visits via a browser script and cookie;
   - Splunk[108] parses webserver access logs and shows in interactive dashboards (steep learning curve/heavy setup).

[104] https://uptimerobot.com
[105] https://www.elastic.co
[106] https://awstats.sourceforge.io
[107] https://matomo.org
[108] https://www.splunk.com

References and Resources

Apache (2023). Apache Jena: A free and open-source Java framework for building Semantic Web and Linked Data applications. The Apache Software Foundation. https://safaritrail.nl/programa/ (accessed 8 Feb 2023).

AMICE (1993). CIMOSA: Open System Architecture for CIM. 2nd edition. Springer-Verlag, Berlin.

Arrouays, D., M. Fantappiè, L. Borůvka, C. Piccini, D. Walvoort, B. Stenberg, J. Wetterlind, C. Calzolari, V.L. Mulder, S. Madenoglu, V. Penížek, G. Aust, E. Leitgeb, L. Poggio, R. Skalsky, A.B. Møller, F. van Egmond, F. Ungaro., (2021). Harmonized procedures for creation of soil maps. In: Van Egmond and Fantappiè (Eds), Report on harmonized procedures for creation of databases and maps. EJP Soil Deliverable 6.1. Available at: https://ejpsoil.eu/fileadmin/projects/ejpsoil/WP6/EJP_SOIL_D6.1_Report_on_harmonized_procedures_for_creation_of_databases_and_maps__final.pdf

Bergh, E. L., Calderon, F. J., Clemensen, A. K., Durso, L., Eberly, J. O., Halvorson, J. J., Jin, V. L., Margenot, A. J., Stewart, C. E., Van Pelt, S., & Liebig, M. A. (2022). Time in a bottle: Use of soil archives for understanding long-term soil change. Soil Science Society of America Journal, 86, 520–527. https://doi.org/10.1002/saj2.20372

Bispo, A., D. Arrouays, N. Saby, L. Boulonne and M. Fantappiè. (2021) Proposal of methodological development for the LUCAS programme in accordance with national monitoring programmes, EJP SOIL deliverable D6.3, Available at: https://ejpsoil.eu/fileadmin/projects/ejpsoil/WP6/EJP_SOIL_Deliverable_6.3_Dec_2021_final.pdf

Bjarnason, E., Wnuk, K., & Regnell, B. (2011, July). A case study on benefits and side-effects of agile practices in large-scale requirements engineering. In proceedings of the 1st workshop on agile requirements engineering (pp. 1-5).

Bond-Lamberty, B., Smith, A. P., & Bailey, V. (2016). Running an open experiment: transparency and reproducibility in soil and ecosystem science. Environmental Research Letters, 11(8), 084004.

Boone, R.D., Grigal, D.F., Sollins, P., Ahrens, R.J., and Armstrong, D.E., (1999). Soil sampling, preparation, archiving, and quality control. In: Robertson, G.P., Coleman, D.C., Bledsoe, C.S., and Sollins, P., editors, Standard soil methods for long-term ecological research. Oxford Univ. Press, Oxford, UK. p. 3–28.

Bucher, T. Fischer, R., Kurpjuweit, S. and Winter R. (2006). “Enterprise architecture analysis and application: An exploratory study,” in EDOC Workshop TEAR, Hong Kong.

Bünemann, E. K., Bongiorno, G., Bai, Z., Creamer, R. E., De Deyn, G., De Goede, R., ... & Brussaard, L. (2018). Soil quality–A critical review. Soil Biology and Biochemistry, 120, 105-125.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., (2009). Sampling for validation of digital soil maps. European Journal fo Soil Science 62, 394-407. doi: https://doi.org/10.1111/j.1365-2389.2011.01364.x

Brus, D.J. (2022). Spatial Sampling with R. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781003258940. Available at: https://dickbrus.github.io/SpatialSamplingwithR/

Chiles, J. P., & Delfiner, P. (2009). Geostatistics: modeling spatial uncertainty (Vol. 497). John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118136188

Cox, S. (2011). OGC Abstract Specification Geographic information — Observations and measurements. Technical report.

Dalgliesh, N.; Hochman, Z.; Huth, N.; & Holzworth, D. 2016. A protocol for the development of APSoil parameter values for use in APSIM - Version 4, CSIRO, Australia. 24 p. International Research Institute for Climate and Society (IRI); Michigan State University (MSU); HarvestChoice, International Food Policy Research Institute (IFPRI), 2015, "Global High-Resolution Soil Profile Database for Crop Modeling Applications", https://doi.org/10.7910/DVN/1PEEY0 , Harvard Dataverse, V2

Debeljak, M., Trajanov, A., Kuzmanovski, V., Schröder, J., Sandén, T., Spiegel, H., ... & Henriksen, C. B. (2019). A field-scale decision support system for assessment and management of soil functions. Frontiers in Environmental Science, 7, 115.

Ditzler, C., Scheffe, K., and Monger, H.C. (eds.). 2017. Soil survey manual. Soil Science Division Staff USDA Handbook 18. Government Printing Office, Washington, D.C.

De Gruijter, J. J., Bierkens, M. F. P, Brus, D. J., Knotters, M. (2006). Sampling for natural resource monitoring. Springer. DOI: https://doi.org/10.1007/3-540-33161-1

De Bruin, S., Bregt, A. and Van de Ven, M. (2001). Assessing fitness for use: the expected value of spatial data sets. International Journal of Geographical Information Science 15, 457471.

De Vries, M., Gerber A. and van der Merwe, A. (2014) In: Aveiro D., Tribolet J., Gouveia D. (eds) "The Nature of the Enterprise Engineering Discipline." Advances in Enterprise Engineering VIII. Springer International Publishing, p. 1-15.

Dietz, J. (2006). "Enterprise Ontology – Theory and Methodology". Springer-Verlag Berlin Heidelberg.

Dietz, Jan L. G., Mulder, Hans B. F. (2020) "Enterprise Ontology". Springer Publishing Company, Incorporated. ISBN 978-3-030-38854-6.

Dick, J., Hull, E., Jackson, K. (2017). Requirements Engineering, Fourth Edition. Springer Publishing Company, Incorporated. ISBN 978-3319869971.

Ehmke, T. (2018). Precision soil sampling harnessing deeper data for nutrient management. *Crops & Soils*, *51*(2), 16-23.

Elrashidi MA (2010). *Selection of an appropriate phosphorus test for soils*, Soil Survey Laboratory, USDA, Lincoln (NE). http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_051918.pdf

Engelsman, W., Iacob, M. E. and Franken, H. M. (2019). “Architecture-driven requirements engineering,” in Proceedings of the 2009 ACM Symposium on Applied Computing (SAC ’09), Honolulu, Hawaii. Pp. 285-286.

Erling, O. (2012). Virtuoso, a Hybrid RDBMS/Graph Column Store. IEEE Data Eng. Bull., 35(1), 3-8.

Fantappie,. M., G. Peruginelli, S. Conti, S. Rennes, F. van Egmond, and C. Le Bas, (2021), Report on the national and EU regulations on agricultural soil data sharing and national monitoring activities. Available at https://ejpsoil.eu/fileadmin/projects/ejpsoil/WP6/EJP_SOIL_D6.2_Report_on_national_and_EU_regulations_on_agricultural_soil_data_sharing_v2.pdf

FAO and ISRIC (1998). Guidelines for Quality Management in Soil and Plant Laboratories. (FAO Soils Bulletin – 74). Available at: https://www.fao.org/3/W7295E/W7295E00.htm

Fixen FE and Grove JH (1990). Testing for soil phosphorus. In: Westerman RL (editor), *Soil testing and plant analysis*. Soil Science Society of America, Inc., Madison (WI), pp 141-180.

Giachetti, R.E. (2010). Design of Enterprise Systems: Theory, Methods, and Architecture. CRC Press, Boca Raton, FL.

Global Soil Partnership (2017a). “Plan of Action for Pillar Five of the Global Soil Partnership.” GSP – Global Soil Partnership.

Global Soil Partnership. (2017b). “Plan of Action for Pillar Four of the Global Soil Partnership.” GSP – Global Soil Partnership.

Guest G, Namey E, Chen M (2020) A simple method to assess and report thematic saturation in qualitative research. PLOS ONE 15(5): e0232076. https://doi.org/10.1371/journal.pone.0232076

Han, Eunjin, Amor VM Ines, and Jawoo Koo. “Development of a 10-km resolution global soil profile dataset for crop modeling applications.” Environmental Modelling & Software 119 (2019): 70-83. https://www.sciencedirect.com/science/article/pii/S1364815218313033

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer. https://hastie.su.domains/Papers/ESLII.pdf

Hennink, M., Kaiser, B.N. (2022) Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Social Science & Medicine 292, 114523. https://doi.org/10.1016/j.socscimed.2021.114523

Heuvelink, G.B.M., Angelini, M.E., Poggion, L., Bai, Z., Batjes, N.H., van den Bosch, R, Bossion, D., Estella, S., Lehmann, J., Olmedo, G.F., Sanderman, J, (2021). Machine learning in space and time for modelling soil organic carbon change. European Journal of Soil Science 72 (4), 1607-1623. doi: https://doi.org/10.1111/ejss.12998

Hudson, B.D. The Soil Survey as Paradigm-based Science (1992). Soil Science Society of America Journal 56 (3), 836-841. DOI: https://doi.org/10.2136/sssaj1992.03615995005600030027x

Huising J, Leenaars J, Csorba A, and da Graca Silva VF (2022). *Detailed guidance for field work* Wageningen, 31 p. https://soils4africa-h2020.us7.list-manage.com/track/click?u=34389e07a98c07e3fba9720d0&id=b0fb5d0b8f&e=712da1e881

Huising, J. M. and Mesele S (2022a). *Protocol for Field Survey.* Soils4Africa Deliverable 4.2. Available at: https://www.soils4africa-h2020.eu/serverspecific/soils4africa/images/Documents/protocolfieldsurveyENG.pdf

Huising, J. M. and Mesele S (2022b). *Standard Operating Procedure for soil sample collection and field observations.* Soils4Africa. Available at: https://www.soils4africa-h2020.eu/serverspecific/soils4africa/images/Documents/SOPfieldsurveyenglish.pdf

INSPIRE Thematic Working Group Soil (2013). “D2.8.iii.3 INSPIRE data specification on soil – draft guidelines”. Standard, European Commission Joint Research Centre.

Isaaks, E. H., & Srivastava, R. M. (1989). Applied geostatistics (Vol. 561). New York: Oxford universitypress.

ISO 12207:2017 (2017). ISO/IEC/IEEE 12207:2017 (2017), “Systems and software engineering– Software life cycle processes”. Standard, International Organization for Standardization, Geneva, CH.

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

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

Jahn, R., Blume, H. P., Asio, V. B., Spaargaren, O., & Schad, P. (2006). *Guidelines for soil description*. FAO.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. Springer.

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

Janssen, B.H., Guiking, F.C.T., van der Eijk, D., Smaling, E.M.A., Wolf, J. and H. van Reuler (1990). A system for quantitative evaluation of the fertility of tropical soils (QUEFTS). Geoderma 46(4), 299-318. DOI: https://doi.org/10.1016/0016-7061(90)90021-Z.

Karssies, L. and P. Wilson (2015). “Soil Archives: supporting Research into Soil Changes.” IOP Conference Series: Earth and Environmental Science 25: 012021.

Lankhorst, M. (2017). “Enterprise Architecture at Work”. Springer Publishing Company, Incorporated. ISBN 978-3662539323.

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). Using R for digital soil mapping (Vol. 35). Cham, Switzerland: Springer International Publishing. https://link.springer.com/book/10.1007/978-3-319-44327-0

Marshall, C. (2000). “Enterprise Modelling with UML”. Addison-Wesley, MA. P. 313-3. ISBN 0-201-43.

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

Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. W3C recommendation.

Motsara, M. R. (2015). Guide to laboratory establishment for plant nutrient analysis. Scientific Publishers.

National Committee on Soil and Terrain (Australia) (2009). “Australian soil and land survey field handbook”. Number 1. CSIRO PUBLISHING, third edition.

Open Geospatial Consortium (2011). “OGC GeoSPARQL – A Geographic Query Language for RDF Data”. Document 11-052r3. URL http://www.opengeospatial. Org/standards/requests/80.

Palma, R., Janiak, B., de Sousa, L. M., Schleidt, K., Rezník, T., van Egmond, F., Leenaars, J., Moshou, D., Wilson, P., Medyckyj-Scott, D., Ritchie, A., Yigini, Y. and Vargas, R. (2022). “GloSIS: The Global Soil Information System Web Ontology”, Semantic Web Journal. Under review. Tracking #: 3325-4539.

PAS197:2009 Code of Practice for Cultural Collections Management (2009), bsi.knowlegde, https://knowledge.bsigroup.com/products/code-of-practice-for-cultural-collections-management/standard

Pohl, K. (2010). Requirements engineering fundamentals, principles, and techniques. Springer Publishing Company, Incorporated.

QUDT (2011). “Quantities, Units, Dimensions and Types (QUDT)”. DOI https://doi.org/10.25504/FAIRsharing.d3pqw7.

Ramifehiarivo, N., B.G. Barthes, A. Cambou, L. Chapuis-Lardy, T. Chevalier, A. Albrecht, T. Razafimbelo (2023). Comparison of near and mid-infrared reflectance spectroscopy or the estimation of soil organic carbon fractions in Madagascar agricultural soils. Gerderma Regional 33, https://doi.org/10.1016/j.geodrs.2023.e00638

Ranatunga, K., E.R. Nation, D.G. Barratt (2008). Review of soil water models and their applications in Australia. Environmental Modelling and Software 23-9. https://doi.org/10.1016/j.envsoft.2008.02.003

Reijneveld, J. A., van Oostrum, M. J., Brolsma, K. M., Fletcher, D., & Oenema, O. (2022). Empower Innovations in Routine Soil Testing. Agronomy, 12(1), [191]. https://doi.org/10.3390/agronomy12010191

Richer-de-Forges, A. C., et al. (2021). Chapter Five – A review of the world's soil museums and exhibitions. Advances in Agronomy. D. L. Sparks, Academic Press. 166: 277-304.

Řezník, T. and Schleidt, K. (2020). Data Model Development for the Global Soil Information System (GloSIS). Technical report, GSP – Global Soil Partnership.

Ross, J. W. and Weill, P. (2005). "Understanding the Benefits of Enterprise Architecture," CISR Research Briefings.

Schoeneberger, P. D. (2012). Field book for describing and sampling soils, version 3.0. Lincoln, NE: Natural Resources Conservation Service, National Soil Survey Center.

Schoeneberger PJ, Wysocki DA, E.C. Benham and Soil Survey Staff (2012). Field book for describing and sampling soils (ver. 3.0). National Soil Survey Center Natural Resources Conservation Service, U.S. Department of Agriculture, Lincoln (NE) https://www.nrcs.usda.gov/sites/default/files/2022-09/field-book.pdf

Sen, M. and Duffy, T. (2005). "GeoSciML: development of a generic geoscience markup language". Computers & geosciences, 31(9):1095–1103.

Shepherd, K. D., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J., & Ge, Y. (2022). A global soil spectral calibration library and estimation service. *Soil Security*, *7*, 100061.

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

Soil Survey Staff (2022). Kellogg Soil Survey Laboratory methods manual. Soil Survey Investigations Report No. 42, Version 6.0. U.S. Department of Agriculture, Natural Resources Conservation Service

Spectrum 5.0 (2017). Collections Trust. https://collectionstrust.org.uk/Spectrum/. Retrieved June 3, 2020.

Stoof, C. R., et al. (2019). "Soil lacquer peel do-it-yourself: simply capturing beauty." SOIL 5(2): 159-175.

Svensson, D.N., I. Messing, J. Barron (2022). An investigation in laser diffraction soil particle size distribution analysis to obtain compatible results with sieve and pipette method. *Soil and Tillage Research 223, 105450.* https://doi.org/10.1016/j.still.2022.105450.

Terhoeven-Urselmans, T., Vagen, G., Spaargaren, O., Shepherd, K.D. (2010). Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library Soil Sci. Soc. Am. J., 74 (5) (2010), pp. 1792-1799, 10.2136/sssaj2009.0218

Teuling, K., Kempen, B., Knotters., M., Saby, N., Brus, D., Vašát, R., van Egmond, F., Bispo, A. (2021). Sampling theory for mapping and monitoring purposes. In: Van Egmond and Fantappiè (Eds), Report on harmonized procedures for creation of databases and maps. EJP Soil Deliverable 6.1. Available at: https://ejpsoil.eu/fileadmin/projects/ejpsoil/WP6/EJP_SOIL_D6.1_Report_on_harmonized_procedures_for_creation_of_databases_and_maps__final.pdf

Thayer, Richard H. and Dorfman, Merlin, eds. (1997). Software Requirements Engineering (2nd ed.). IEEE Computer Society Press. ISBN 978-0-8186-7738-0.

TOTH, G., JONES, A., MONTANARELLA, L., ALEWELL, C., BALLABIO, C., CARRE, F., ... & YIGINI, Y. (2013). LUCAS Topoil Survey-methodology, data and results.

Turdukulov, U., B. Kempen, J.S. Mendes de Jesus, L. Calisto, P. van Genuchten, L. Poggio (2021). D6.1 Technical design of the Soil Information System. Available at https://www.soils4africa-h2020.eu/serverspecific/soils4africa/images/Documents/Soils4Africa_D6.1_TechnicaldesignoftheSIS_v01.pdf

Van Baren, J. H. V., and W. Bomer (1979). Procedures for the collection and preservation of soil profiles. Wageningen, International Soil Museum: p. 22.

Van Egmond, F. and Fantappiè, M. (Eds) (2021). Report on harmonized procedures for creation of databases and maps. EJP Soil Deliverable 6.1. Available at: https://ejpsoil.eu/fileadmin/projects/ejpsoil/WP6/EJP_SOIL_D6.1_Report_on_harmonized_procedures_for_creation_of_databases_and_maps__final.pdf

Van Egmond, F., R. Koomans, K. Teuling, M. Tijs, G. Staats, K. Pepers, J. de Haan, G. van Os (2022). Validating a new in-situ soil bulk density sensor. Poster presentation at World Congress of Soil Science 2022 in Glasgow.

Van Engelen V.W.P. and Dijkshoorn, J.A. (eds.), 2013. Global and National Soils and Terrain Databases (SOTER). Procedures Manual, Version 2.0, ISRIC – World Soil Information, Wageningen. 198 pages, 10 figures and 9 tables. Available at: https://www.isric.org/sites/default/files/isric_report_2013_04.pdf

Van Leeuwen, C.C.E., V.L. Mulder, N.H. Batjes and G.B.M. Heuvelink (2022), Statistical modelling of measurement error in wet chemistry soil data. *European Journal of Soil Science* 73, e13137.

Van Reeuwijk LP (2002). Procedures for soil analysis (6th ed.). Technical Paper 9, ISRIC, Wageningen, 81 p. https://www.isric.org/sites/default/files/ISRIC_TechPap09.pdf

Wadoux, A., B. Malone, B. Minasny, M. Fajardo, A. McBratney (2021. Soil Spectral Inference with R, Analysing Digital Soil Spectra using the R Programming Environment. Springer Nature Switzerland AG. ISSN 2352-4774. https://doi.org/10.1007/978-3-030-64896-1

Wadoux, A. and A. McBratney (2021). “Digital soil science and beyond.” Soil Science Society of America Journal.

Watkins, R., Meiers, M. W., & Visser, Y. (2012). *A guide to assessing needs: Essential tools for collecting information, making decisions, and achieving development results*. World Bank Publications.

Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists.
John Wiley & Sons. https://www.wiley.com/en-us/Geostatistics+for+Environmental+Scientists%2C+2nd+Edition-p-9780470028582

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). Spatio-temporal statistics with R. CRC Press. https://spacetimewithr.org/Spatio-Temporal%20Statistics%20with%20R.pdf

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al (2016), 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3, no. 1 (15 March 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Annex I: Example questions for a technical capacity assessment created by ISRIC

I.I Related to data registry

1. Are you sharing digital datasets via the web?
   a. *If yes how?:*
   b. *% of datasets send in person on request (email, Wetransfer, Dropbox, etc)*
   c. *% of datasets as file download*
   d. *% of datasets as WMS/WFS/WCS*
   e. *% of datasets via alternative API's (for example, ArcGIS online)*
   f. *Other (specify)*

2. Is metadata available for datasets?
   a. *% of datasets have metadata*
      i. *Specify metadata formats: iso19139, iso19115-2, FGDC, DCAT*

3. Are you publishing the metadata on the web?
   a. *In a web catalogue (specify URL)*
   b. *In a file repository*
   c. *On a webpage (specify URL)*
   d. *Other (specify)*

4. Which tools do you use for metadata:
   a. *GeoNetwork/CKAN/GeoNode/pyCSW/...*

5. Do you measure usage of datasets?
   a. *Web portal has a statistic module*
   b. *Access logs are captured and evaluated*
   c. *Data download requires registration*
   d. *Evaluation of search behavior on search engines*
   e. *Other (specify)*

I.II Related to IT and SDI infrastructure

1. What kind of GIS software is used in your organization?
   a. *Specify: QGIS/ArcGIS/ILWIS/...*

2. What (central) database management systems are used for storing spatial data in your organization?
   a. *Oracle,*
   b. *SQL-server,*
   c. *PostgreSQL*
   d. *Access*
   e. *FGDB*
   f. *Other (specify)*

3. What file systems/formats are used for storing spatial data?
    a. *Vector data (Shapefile, Geopackage, CSV, ...)*
    b. *Raster data (TIFF, mrSid, laz, ASCII, ...)*
    c. *Other (specify)*

4. Does your organization publish spatial maps on the internet?
    a. *If yes, specify web mapping software used for publishing geospatial data on the web:*
        i. *QGIS server*
        ii. *ArcGIS Server/ArcGIS Online*
        iii. *GeoServer*
        iv. *Mapserver*
        v. *Other (specify)*
    b. *How are web sites / web applications maintained and hosted?*
        i. *On Premise*
            1. *Physical servers, internal computers, (specify Operating Systems: Windows, Linux)*
            2. *Virtual servers (e.g., Vmware, VirtualBox)*
            3. *Container technology, (e.g., docker, Kubernetes, OpenShift)*
        ii. *Off premise*
            1. *Cloud provider ... (e.g., Azure, AWS, CloudAfrica)*
            2. *Software as a Service ... (eg, ArcGIS Online / CartoDB / Hale Connect)*
        iii. *Who maintains the infrastructure?*
            1. *Self*
            2. *External*
    c. *Do you use any cloud infrastructure such as:*
        i. *Amazon Web Services*
        ii. *Microsoft Azure*
        iii. *Google Cloud Platform*
        iv. *AfricaDataCentres.com*
        v. *CloudAfrica.net*
        vi. *HostAfrica.co.za*
        vii. *Other, please specify which one.*

5. Do you have high performance machines (PC/HPC hardware) for highly intensive computational tasks?
    a. *On premise ... (e.g., Nettapp, HPE)*
    b. *Cloud computing ... (e.g., Azure, Google)*

6. Is there a central authentication available for hub members?
    a. *LDAP*
    b. *Single sign on (Oauth)*
    c. *Federation (SAML)*

7. Use of mobile apps for data acquisition
    a. *Mobile app for staff, ... (e.g., ODK, ESRI based, Tailored)*
    b. *Mobile app for citizens, ... (e.g., ODK, ESRI based, Tailored)*

8. **Community tools, accessible by hub members**
    a. *Mailing list software, ... (e.g., mailchimp)*
    b. *Chat / Forum functionality (Discourse, Gitter, Slack)*
    c. *Documentation / Wiki / CMS functionality, ... (mkdocs, Hugo, GitHub)*
    d. *Code repository / Issue tracking .... (GitHub, GitLab, Jira, Trac)*

I.III Related to sources for the project

1. Who will be the technical contact (GeoICT) for the project?

2. Can a test environment be made available to the project, which is accessible remotely from outside (i.e., from ISRIC)?

3. Can you describe the team participating in the LSC Hubs project (i.e., number of participants), including their skills on:
    a. *Data curation,*
    b. *Programming (indicate languages),*
    c. *Databases and SQL,*
    d. *DevOps,*
    e. *Web mapping technology*
    f. *Front end development (JavaScript frameworks)*
    g. *App development*
    h. *community building,*
    i. *training and/or guidance?*

4. What areas of staff skills would you like to strengthen through training in the context of this project?

5. What would be the best suitable timeline for training and building capacity of LSC Hubs?

6. What IT or technology-oriented projects are currently ongoing within your organization that could help advancing the project? (Improve network access, Improve continuity of power supply, Update hardware infrastructure etc.)
    a. *Specify:...*

Annex II: Example Outline Stakeholder Workshop based on LSC-IS project

Day 1	
8.00 – 8.30 am	Arrival and Registration
8.30 – 9.15 am	Welcoming Remarks. Objectives of the Workshop and overview of the agenda Expectations of the day
	Introduction of the participants
	Participants Expectations, Questions and Comments
9.15 – 10.30 am	**Activity 1.** Presentations to provide a background of the project
10.30 - 11.00	**HEALTH BREAK**
11.00 am – 12.00 pm	**Activity 2.1** Group work on Identification of key stakeholders/partners, their roles (users, suppliers, or both -intermediaries), and challenges and opportunities in producing or use of SIS
12.00 - 12.30 pm	**Activity 2.2** Session Reflection activity and Group plenary presentations. Max. 5 min. with key points per group Summarize the points
12.30 - 1.30 pm	**LUNCH**
1.30 - 4.00 pm	**Activity 3** Group work – Specifying SIS needs and SIS users- identify data sets for specified use cases.
4.00 - 4.45 pm	**Activity 3.2** Session Reflection activity and Group plenary presentations Max. 5 min. with key points per group Summarize the points
4.45 - 5.00 pm	Wrap-up and closure of the day: Review, summary, and capture emerging questions. Expectations for tomorrow
	Announcements
5.00 pm	**Departure**
5.00 - 6.00 pm	**Core-team reflections**

Day 2	
8.00 – 8.30 am	Registration and Day–2 Agenda Overview
8.30 – 9.00 am	Recap of Day–1 Expectations for today
9.00 - 10.30 am	**Activity 4.1** Group work – Identifying capacity requirements for SIS use and users to inform hub development.
10.30 - 11.00 am	**HEALTH BREAK**
11.00 - 11.30 am	**Activity 4.1** Continuing Group work – Identifying capacity requirements for SIS use and users to inform hub development.
11.30 - 12.30 pm	**Activity 4.2** Group plenary presentations and Reflection Activity Max. 5 min. with key points per group Summarize the points
12.30 - 1.30 pm	**LUNCH**
1.30 - 2.30 pm	**Activity 5.1** Participants identify policies/initiatives related to the use cases
2.30 - 3.00 pm	**Activity 5.2** Group plenary presentations and Reflection Activity Max. 5 min. with key points per group Summarize the points
3.00 - 3.30 pm	Summary, next steps, and meeting closure
3.30 - 4.00 pm	Workshop evaluation
4.00 pm	**Departure**
4.00 - 5.00 pm	**Core-team reflections**

Annex III: Example of Key Informants interviews for SIS created by ISRIC

III.I Users

Nr	Question – users (incl. hosting institute)	Category / theme	Intention
1	What are typical uses cases around climate-smart agriculture that you have developed/are using (that addresses the use cases)? *For example: soil fertility management, drought management, erosion risk control, etc.*	Use cases	To understand what use cases the stakeholders have developed or using
2	Which data or derived information do you need for your decision-making or decision-support processes or for use case development? (Answer as detailed as possible. For instance, not “soil data”, but what type of soil data but what type of soil data [nutrient content, pH, soil depth, etc.]; not “climate data” but what type of climate data [rainfall, temperature, annual means, or time series, real-time etc.]; not” crop data” but what type of crop data [yield for a specific crop, crop calendar, production statistics, etc.].)	Use cases / data	To obtain a more detailed overview of the type of datasets needed.
3	Describe the process or how you developed or provide the advisory services around the use cases.	Use cases	To understand the adequacy or gaps of the use cases
4	What products, services or advisories do you produce or provide that require data/information as input? *For example: policies, development plans, management plans, operational procedures, advisory, formulation of regulations, suitability assessments, forecasts, (functional) maps, etc.*	Use cases	To understand the type of outputs produced from data/information
5	For whom do you develop or provide these products, services, or advisories?	Use cases	To gain insight in the user groups served the data user that is interviewed.
6	What are the various levels through which the use case (products, services or advisories) is applied? Example: National, County, sub-county, ward, village, farm	Use cases	To understand the scale or level of application of the use case
7	Which applications or tools (such as models) related to the soil fertility and soil water conservation are employed in the use cases? *Applications/tools/models may include but are not limited to:* • *statistical (spatial extrapolation) models, for instance for mapping* • *models serving agricultural decision making/ support or advisory:* ○ *erodibility* ○ *water sufficiency* ○ *nutrient sufficiency* ○ *crop productivity* ○ *other (please specify)* • *other (please specify)*	Use cases	To understand the range of applications, models and tools used in use case developed or service provision

8	What are currently the main sources of the data?	Data	To understand the available data sources
9	What are current constraints in accessing data for your applications, use case or services? • *data are not available.* • *data are available but not accessible or not in the right format.* • *lack of technical resources (computers, software)* • *lack of human resources (staff with specific skill sets: data analysts, GIS technicians, modelers, etc.)* • *lack of financial resources (purchase of relevant hardware/software, tools, data, e.t.c.)*	Data	To gain insight in constraints that hamper access of available data and to identify gaps between data availability and accessibility
10	Are there any requirements with respect to the format of the data that you use?: • *GIS vector files (e.g., shapefile or geopackage),* ○ Point data (observational data). ○ Polygon (aerial data; what is the preferred scale level?). • *GIS raster files (e.g., GeoTiff, ASCII, ESRI raster; if raster, what is the preferred spatial resolution?)* • *Plain tables (e.g., MS Excel, csv)*	Data	To understand which data formats are required by the users, as well as scale level and/or resolution
11	If you are using spatial data, what are requirements do you have with respect to georeferencing of the data? Is there for instance a preferred coordinate system(s)?	Data	To understand requirements with respect to coordinate systems used
12	How would you like to obtain data for your application or decision-making or decision-support process or use case development or service provision?: • *data download* • *web service (WMS, WFS, WCS)* • *API* • *other (specify)*	Data	To understand how data are consumed by the users
13	What would you like to be done differently to ensure data: • *Availability* • *Accessibility* • *Usability* • *Scalable* • *Impactful*	Data	To understand recommendations on the 5 thematic areas (Availability, Accessibility, Usability, Scalability, and Impact)
14	What type of information or services or advisories do you miss/lack to address soil fertility and soil and water conservation challenges	SIS (proposed solution)	To gain insight on the information services or advisories that should be developed into the SIS
15	What do you expect from the SIS in terms of functionality? What type of functionality (e.g., data download, data viewer, data catalogue, dashboards (presenting what type of information?), user stories, data interpretation (translating data to advisory), etc.) would be helpful for you?	SIS (proposed solution)	To gain insight in requirements for the SIS
16	How would you like to access the SIS: laptop or desktop computer or mobile device (tablet, phone)?	SIS (proposed solution)	To understand the preferred way of accessing information.
17	Are you willing to pay for the information services?	SIS (proposed solution)	To understand the sustainability of the hub beyond the project life
18	What, from your view as a data user, is a critical factor to ensure sustainability of the SIS?	SIS (proposed solution)	To understand recommendations to ensure the SIS is effectively put in use and remains scalable and sustainable

III.II Providers

Nr	Question – suppliers (incl. hosting institute)	Category / theme	Intention
1	Which data, services, or advisories do you provide?	Data	To obtain more detailed information on the type of datasets currently supplied
2	What is the data format of the data you provide?: • GIS vector files (e.g., shapefile or geopackage), ○ Point data (observational data). ○ Polygon (aerial data). • GIS raster files (e.g., GeoTiff, ASCII, ESRI raster; if raster, what spatial resolutions?), • plain tables (e.g., MS Excel, csv)	Data	To obtain insight in the data format of the supplied data
3	How do you provide data to users: data download, web service (WMS, WFS, WCS), API, etc.?	Data	To understand current data provision channels
4	How is the information from data presented to the users (map portal, GIS system, mobile app)? And is this the best way to present data to users?	Data	To understand how the information from data is currently presented
5	Is there a digital data repository of datasets in your organisation? Are these datasets accessible as 1) catalogue service 2) web mapping service 3) APIs?	Data	To know if the organization has a data repository in place
6	Do you have metadata available for all published datasets?	Data	To understand if metadata exist
7	Are there metadata standards in place in your organisation?	Data	To ensure the quality of the metadata
8	What are the cost of hosting and maintenance of the SIS beyond the project life?		
9	What type of information or services or advisories do your users need to address the use cases?	SIS (proposed solution)	To gain insight on the information services or advisories that should be developed into the SIS
10	What do you expect from the SIS in terms of functionality? What type of functionality (e.g., data download, data viewer, data catalogue, dashboards (presenting what type of information?), user stories, data interpretation (translating data to advisory), etc.) would be helpful for you?	SIS (proposed solution)	To gain insight in requirements for the hub design
11	How would you like to access the SIS: laptop or desktop computer or mobile device (tablet, phone)?	SIS (proposed solution)	To understand the preferred way of accessing information.
12	What are the key security and privacy data requirements that the SIS should consider in its design?	SIS (proposed solution)	To understand security and privacy requirements
13	Do you think users are they willing to pay for the information services?	SIS (proposed solution)	To understand the sustainability of the SIS beyond the project life
14	What, from your view as a data provider, is a critical factor to ensure sustainability of the SIS?	SIS (proposed solution)	To understand recommendations to ensure the SIS is effectively put in use and remains scalable and sustainable

Annex IV: Common lab, proximal and remote (soil) sensing methods

All soil sensing methods rely on the interaction of energy or light of a part of the electromagnetic spectrum with matter, in this case mostly soil. The signal is either emitted by the soil (gamma-ray spectrometry) or is changed as a result of the interaction with soil. Usually, part of the light that interacts with the soil is absorbed and the rest is reflected back. Which wavelengths and how much is absorbed and reflected mostly depends on the chemical properties of the soil and the roughness of the surface. This is why the methods are sometimes referred to as chemometrics. Because various parts of the spectrum interact differently with the soil depending on its properties, we can use this principle and our understanding of the interaction to measure soil properties.

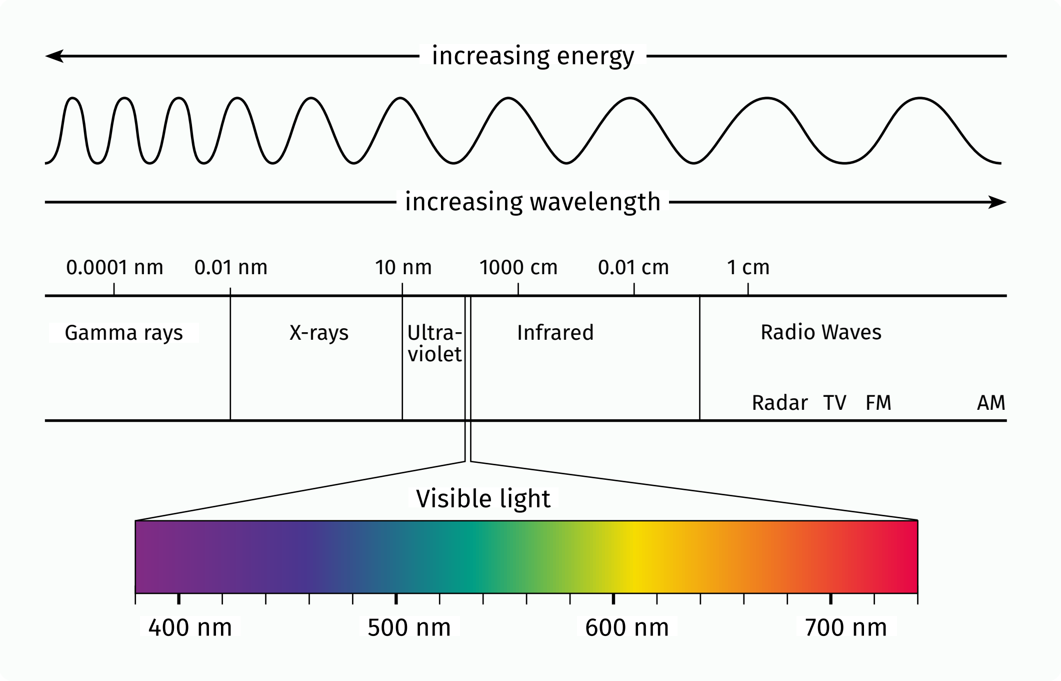


Figure 4 The electromagnetic spectrum

IV.I Measurement principle of infrared soil spectroscopy, measurement in the lab

A very suitable part of the electromagnetic spectrum to measure soil properties is the visible and near-infrared (NIR: 400-2500 nm) with an emphasis on the short-wave infrared (SWIR: 2000-2500 nm) and the mid-infrared (MIR: 600-4000 cm-1) range. This is where a lot of soil properties have absorption features, where they absorb light, to a large part due to organic bonds. The MIR has more relevant absorption features than the NIR, and therefore often provides better predictions of soil properties. But because MIR instruments are at present (2023) still more expensive and more used in labs than in the field and various NIR field instruments exist, both spectral ranges are used for estimation of soil properties.

To estimate or predict the soil properties with infrared spectroscopy first a **calibration soil spectral library** is built or an existing one is found that has samples of the relevant geographic area. A soil spectral library (SSL) is a calibration set where soil samples are measured both with wet chemistry and spectrally. From this SSL the relation between the wet chemistry and spectral soil properties is derived with statistical or machine learning models such as Partial Least Squares Regression (PLSR), locally weighed PLSR, cubist, random forest, and other models. This model can then be applied to the spectra of newly measured samples to determine their soil properties.

For good **quality predictions** a number of aspects are important:

- Use a soil spectral calibration library that covers the feature space (the set of all values for a target variable in a target universe) or region of your new samples.
- Use a spectral library that has ideally analyzed all wet chemistry and spectral measurements in the same lab(s), with the same procedures and of sufficient to good quality. If this is not possible, make sure lab methods are transformed to similar results using pedotransfer functions and understand the uncertainty or accuracy of the analyses.
- Use accepted standard operating procedures (SOP) for carrying out both wet chemistry but also the spectral measurements.
- Verify that instrumentation is working properly.
- Make sure sample preparation and presentation (to the instrument) is consistent and of good quality.
- Be wary of overfitting the spectral models. This will result in unstable results and a much lower accuracy in the validation dataset compared to the calibration dataset.
- Verify the quality of the predictions and report this with the results.

Soil sample preparation is minimal but important. Samples need to be dried (crushed) and sieved to 2 mm, and if analyzed with mid-infrared instrumentation the sample needs to be fine-grinded as well. For near infrared measurements fine grinding is not needed. Typically, 20 to 200 samples can be measured per day. Well-known instrument suppliers are FOSS (NIR), Bruker Optics, Thermo Scientific, Agilent (MIR), but there are many other suppliers on the market.

For **standard operating procedures** for soil spectroscopy these are relevant resources:

i. Standard operating procedures for spectral analysis are provided by ICRAF and the USDA-NRCS Kellog Soil Survey Laboratory;
ii. The GLOSOLAN soil spectroscopy working group is working on publishing an SOP for MIR soil lab analysis;
iii. The IEEE P4005 WG is working on SOPs for field IR measurements;
iv. ICRAF provides standard operating procedures for spectra data analysis with the software R;
v. Soil spectral inference with R (Wadoux et al., 2021).

With a relevant soil spectral library, SOPs and instrumentation in place, another challenge is the **data analysis**, the derivation of soil property estimations from spectra by deriving and applying the spectral models from the spectral library. Several good instruction manuals, R scripts (Wadoux et al., 2021), ICRAF: standard operating procedures for spectra data analysis with the software R) and proprietary software is available for this purpose.

Several **open soil spectral libraries** exist, for example Global Soil Spectral Calibration Library and Estimation Service (Shepherd et al., 2022); the Open Soil Spectral Library of the USDA Food and Agriculture; the Brazil Soil Spectral Library , the LUCAS Soil spectral library, the ICRAF-ISRIC spectral library (Terhoeven-Urselmans et al., 2010), the Swiss soil spectral library and more.

IV.II Innovations in soil spectroscopy

The initial investment in instrumentation (20 to 90 kUSD) and in building a soil spectral calibration library (SSL) is high for a single lab, depending on the expected throughput and the availability of a soil sample archive. This has triggered initiatives (Shepherd et al., 2022) to build an **open global soil spectral calibration library** (GSCL) and research into **combination of local soil spectral calibration libraries**. For the first, the principle is to start with a high-quality soil spectral library, based on data derived with the same spectral and wet chemistry SOPs (see Standards for laboratory analysis) and extend its feature space by adding representative samples from other countries or regions to increase its coverage. For the second, research is conducted if, and if so, how **spectral calibration transfer models** are needed to combine spectral measurements from different spectrometers. Initial results show that for high quality instruments spectral calibration transfer may not be needed between instruments of the same manufacturer. In this option, transfer functions to align wet chemistry methods need to be derived and applied as well.

It should be noted that predictions based on local soil spectral libraries are always better than predictions based only on a global soil spectral library. This can be mitigated to some extent by spiking the global library with local samples (thus extending its feature space). However, the global libraries allow labs without a current library to get started with this technique and improve the library with local samples during operation, thus significantly decreasing initial costs. A condition for both is an open data license on soil spectral libraries, allowing its re-use by other entities.

An alternative to spectral modelling by the user or the lab that performs the measurements that is upcoming is the use of an **estimation service** (Shepherd et al., 2022, soilspectroscopy.org, BraSpecS, globeSpeC (Shen et al., 2022)). An estimation service hosts relevant (global) spectral libraries and derives spectral models for soil properties either on demand or provides pre-calculated ones. A user can upload a new spectrum to the service and receives a soil property estimation in return, including an uncertainty estimate and the prediction method. This reduces the burden of data analysis for users/lab operators but is only useful when for the user relevant libraries are included.

IV.III Overview of proximal soil sensing methods

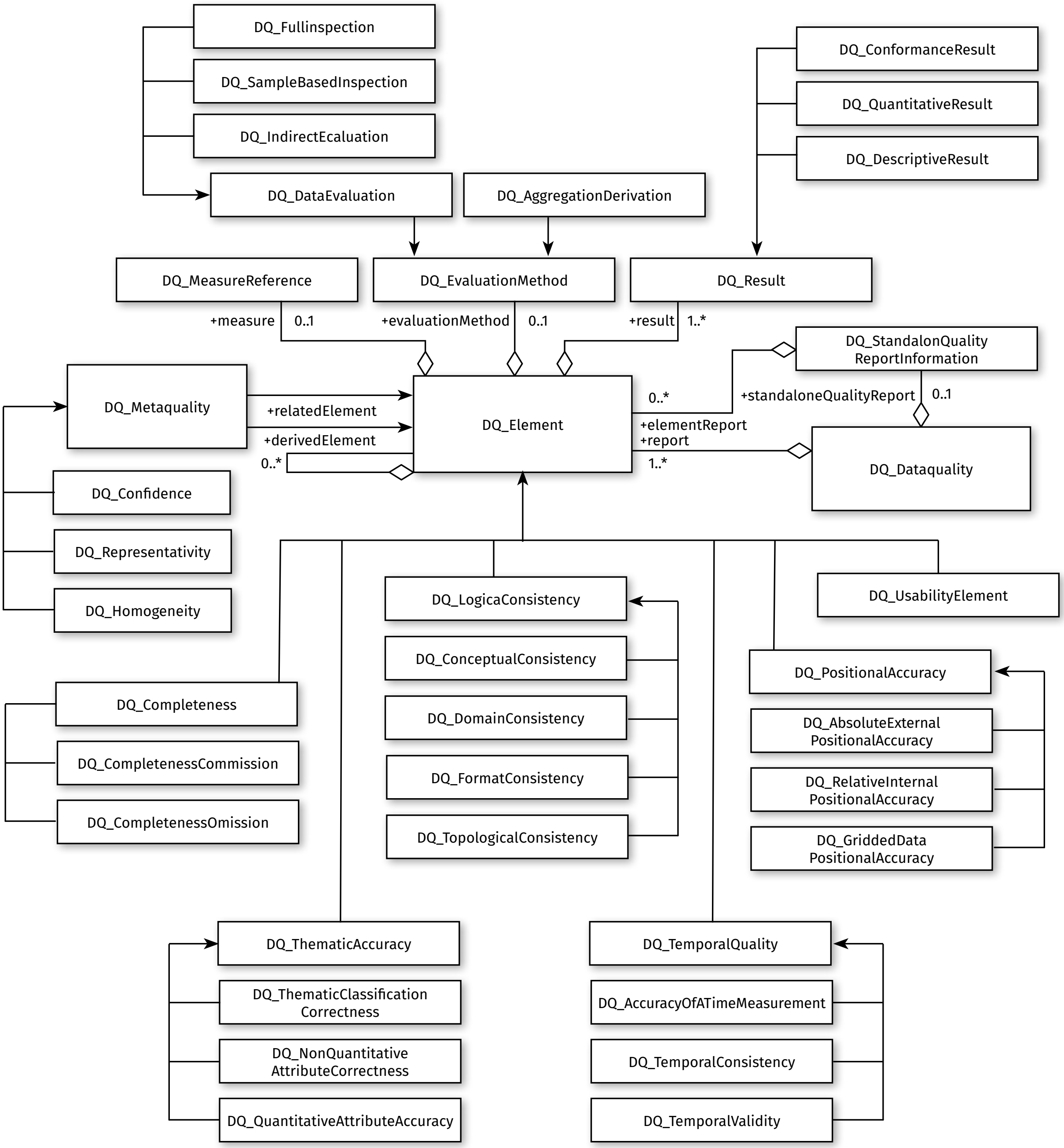
Proximal soil sensing methods	Measurement principle	Results	Platform
Electromagnetic Induction (EMI)	Magnetic field is created between two coils, this generates a secondary magnetic field in the soil which is dependent on the soil's conductivity. Coil distances determine measurement depth.	Combined value for soil conductivity which is related to clay, water and salts contents and porosity. Calibration samples/profile descriptions are needed to separate these contributions. Upon inversion of the signal a depth profile can be generated from multiple coils. Useful for delineation of management zones, rapid stratification, and characterization of an area. Be careful with (former locations of) compost heaps, this will result in high values.	Vehicle, helicopter
Visible and Near Infrared (VNIR) and Mid Infrared(MIR) spectroscopy	The chemical composition of a soil determines the amount and wavelengths of absorption of infrared light. Statistical models derived from a relevant spectral calibration library (consisting of wet chemistry and spectral measurements of soil samples) are used to predict the soil properties of newly measured samples. Penetration depth is max a few mms.	Quantitative estimates of soil property concentrations. The accuracy of the prediction is dependent on the quality/detail of the measurement, the relevance and quality of the library and the strength and distinctness of the spectral response of the chemical bonds that determine the spectral signature and correlation between spectrum and soil property.	Handheld, vehicle, UAV, airplane, satellite, lab
Ground Penetrating Radar (GPR)	Measures the reflection of an active electromagnetic radar signal sent into the soil (100 MHz – 2 GHz) on changes in di-electric constant, indicative of (sudden) changes in soil moisture, texture, density. Penetration depth and depth resolution is dependent on the wavelength(s) used, from 30 cm to 10 m.	Indication of layers in soil with different (texture) composition or structure or sudden substantial changes in moisture content and density. E.g., for fluvial soils, artefact location, cables. Depth resolution is more precise than EMI.	Vehicle
Gamma-ray spectrometry	Measures the radiation emitted by (natural) radionuclides in the soil (K, U, Th, Cs). The measured spectrum is deconvoluted to concentration of radionuclides. Their ratios and concentration are related to mineralogy and texture of the soil. Measurement depth is about 30 cm.	Quantitative estimates of soil texture of the tillage layer (30 cm) and of strongly with texture correlated soil properties. In larger areas an indication of provenance or parent material types.	Handheld, vehicle, UAV, airplane, lab
X-Ray Fluorescence	Excitation of electrons with radiation. Penetration depth is 1 to 2 mm	Total concentration of chemical elements (Zn, M, S, Cu, Si, Fe, Al, etc.)	Handheld, lab
Magnetics	Measures the magnetic susceptibility of soils. Measurement depth is dependent on coil distance	Indication of ferrous elements present in the soil, e.g., used in archaeology and geology/mining	Vehicle, airplane
Penetrologger	A pin with pressure sensor is pushed into the soil at constant speed and pressure. The pressure sensor measures penetration resistance	Penetration resistance of soil (MPa) per cm depth, often 0-80 cm range	Handheld
Gamma-ray attenuation -Soil bulk density	Gamma-ray attenuation: measure the absorption of emitted gamma radiation by soil and water surrounding the sensor	Field soil bulk density, with moisture correction dry bulk density can be obtained. Various sensor models exist. (RhoC (van Egmond et al., 2022), Australia, USA) Depth and range is model dependent.	Handheld
Soil moisture sensors	TDR, FDR, EC systems	Various models and qualities exist	Handheld, in situ, lab

For most if not all the above detailed sensing methods calibration data are needed to derive the final result. These are often soil point observations or lab results or soil profile descriptions that are modelled with the measured sensor values to derive the (cor)relation between sensor measurement and soil parameter or target variable.

IV.IV Overview of remote sensing methods

Remote sensing methods	Platform/method	Derived measurements
Multispectral VISNIR	Broader bands (nm) in 2 to 5 parts of the VISNIR spectrum	Vegetation /landcover monitoring, patterns in surface soils, aerial pictures, change detection, etc.
Hyperspectral VISNIR	Narrow (nm) and many bands covering an entire range in the VISNIR spectrum. Usually, smaller field of view	Soil and plant property measurements. E.g., SOC, clay, etc. and plant species
Thermal IR	Broader bands in the thermal infrared spectrum (between infrared and radar)	Temperature of the earth surface, plant stress
Synthetic Aperture Radar (SAR)	Several bands in the radar part of the spectrum (longer wavelengths). Sees through clouds	Moisture content, elevation
Gamma-ray (not on satellite)	Passive measurement of radioactivity (see table above)	Texture, parent material/provenance
LiDAR (not on satellite)	Laser or light-based reflection method to determine ranges	Detailed elevation maps, derived products (subsidence, slope, etc.) , vegetation structure

Annex V: quality elements defined by ISO 19157



Annex VI: Soil application models

VI.I Soil Carbon Models

Roth-C Model

The Roth-C model is a processed model used to simulate turnover of SOC in non-waterlogged topsoil that allows for the inclusion of the effects of soil type, temperature, moisture content and plant cover on the turnover process. It calculates total SOC (t ha-1), microbial biomass carbon (t ha-1) and Δ14C with a monthly time step on a year to centuries timescale at plot, field, regional or global scales with data from many long-term experiments, different regions, and counties throughout the world.

Pros	Cons
It is a widely used model with a long history of development and testing.	The model has limitations in simulating carbon dynamics in non-agricultural soils, such as forests and wetlands.
It can be used to simulate carbon dynamics in various agricultural soils and management practices.	The model requires detailed information on soil properties and management practices, which can be time-consuming and expensive to collect.
It considers the effects of soil temperature, moisture, and organic matter quality on carbon dynamics.	It assumes that carbon inputs and outputs are in equilibrium, which may not always be the case in agricultural systems.
It can be used to explore the impacts of different management scenarios on soil carbon sequestration and greenhouse gas emissions.	The model does not account for the effects of soil biota and their interactions with soil organic matter dynamics.

CENTURY Model

The CENTURY model is a process-based model used to simulate plant-soil carbon and nutrient cycling at monthly step for several types of ecosystems including grassland, agricultural land, forest, and savannas. It is composed of a SOM/decomposition sub-model, a water budget model, a grassland/crop sub-model, a forest production sub-model, and management and events scheduling functions. It computes the flow of C, N, P and S through the model's compartments.

Pros	Cons
The CENTURY is a well-established model with a long history of development and application.	The CENTURY model requires a large number of input parameters, some of which may be difficult to measure or estimate with precision.
It can simulate both the short-term and long-term effects of management practices on SOM and nutrient cycling.	The model is complex, and it may be challenging for users with limited modeling experience to set up and calibrate the model correctly.
The model can be used to evaluate the impacts of different land use and land management scenarios on soil carbon and nutrient dynamics, as well as on crop productivity and ecosystem services.	The model is sensitive to input parameters and model assumptions, which may introduce uncertainties into the results.
CENTURY has been applied in a wide range of agroforestry and land-use systems, including tropical and temperate forests, grasslands, croplands, and agroforestry systems	The CENTURY model does not explicitly simulate water balance, so it may not be suitable for assessing the impacts of land use and management practices on water resources.
The model has a user-friendly interface and can be run on a variety of platforms, including personal computers.	

DAYCENT Model

DAYCENT model is a process-based biogeochemical model that simulates fluxes of C, N and water cycles between the atmosphere, vegetation, and soil in terrestrial ecosystems over time periods ranging from hours to centuries. Model inputs include daily maximum/ minimum air temperature and precipitation, surface soil texture class, and land cover/use data. Model outputs include daily fluxes of various N-gas species (e.g., N2O, NOx, N2); daily CO2 flux from heterotrophic soil respiration; soil organic C and N; net primary productivity; daily water and nitrate (NO3) leaching, and other ecosystem parameters.

Pros	Cons
DAYCENT can simulate a range of ecological processes, including vegetation dynamics, nutrient cycling, and soil biogeochemistry.	DAYCENT requires a substantial amount of input data, including soil properties, climate data, and vegetation parameters, which can be time-consuming and challenging to collect.
The model is designed to be flexible and can be adapted to a wide range of ecosystems and management scenarios.	The model is computationally intensive, which can limit its application to large-scale spatial and temporal domains.
DAYCENT has been widely used in research and has a large user community, which provides resources and support for model application and development.	DAYCENT's complexity may make it challenging for non-experts to use and interpret the model outputs.
The model has been extensively validated against field data, which enhances the reliability of its predictions.	The model has some limitations in representing certain processes, such as nitrogen cycling in arid and semi-arid environments.

YASSO Model

YASSO model is a dynamic model of SOC cycling that calculates the amount and changes of SOC and heterotrophic soil respiration. The application of the newest version of the mode (YASSO20) includes earth system modeling, global climate simulations, greenhouse gas inventories and research on land ecosystems and climate change. The advantages of the model is globally applicable (various climatic conditions, ecosystem types and litter types; mineral soils down to 1 meter), easy to adopt (input data commonly available, computationally efficient and no project-specific calibration needed), transparent modelling process (assumptions, data and methods published scientifically); disadvantages are 1) the model has limitations in representing certain processes such as the effect of soil texture and soil fauna on decomposition rates; 2) the model does not account for spatial heterogeneity in soil properties, which can be an important factor in soil carbon dynamics; 3) the model has limited capability to simulate the effects of changes in land use and management practices on soil carbon dynamics.

SOMM Model

SOMM (Soil Organic Matter Model) is a dynamic model that simulates SOM mineralization, humification and nitrogen release including rate of the processes depending on litter fall's nitrogen and ash content, temperature, and moisture. Pros of the model that it can reflect the functioning of the main groups of soil decomposers and represent a system of linear differential equations with variable coefficients and can be used for modelling soil system and natural ecosystems' dynamics mostly in a wide range of environmental conditions from tundra to tropical rain forest; however the model are complex and can require specialized knowledge to use and interpret the results and may have limitations in accurately predicting changes in SOM due to factors such as variability in soil properties and climate.

VI.II Soil Water Models

SWAP Model

SWAP (Soil, Water, Atmosphere and Plant) model is designed to simulate flow and transport of water, solutes, and heat in unsaturated/saturated soils vertically and horizontally at field scale level, during growing seasons and for long term time series. It offers a wide range of possibilities to address both research and practical questions in the field of agriculture, water management and environmental protection. The strength of the model is that it can simultaneously simulate water flow, solute transport, heat flow, macropore flow and crop growth at field level and the SWAP adheres to the open-source philosophy that allows other research teams to integrate the model into all kinds of Decision Support Systems; the model could include the WOFOST version 7.1 as a special case for considering crop growth. However, the implementation of the combination would have limitations in the soil (e.g., drought stress, oxygen stress, salinity stress) result in diminished (actual) crop production. This results in a different prediction of actual crop growth compared to that predicted by WOFOST including its own simple soil module.

VI.III Soil Erosion Models

SWAT Model

SWAT (Soil and Water Assessment Tool) model is a basin-scale model which is able to quantify the impact of land management practices in large, complex watersheds by dividing a catchment into smaller discrete calculation units by combination of soil and land cover – namely Hydrological Response Unit (HRU). The total catchment behavior and water balance is a net result of manifold small HRUs. SWAT could also evaluate the impact of crop-land-soil management on downstream water and sediment flows. SWAT inputs are scalable, detail depends on the objective of the project.

Pros	Cons
Free, large user community: already used in some WATDEV study areas	Not raster-based: low spatial detail, limited landscape structure
Can be applied anywhere with free data	Simplistic simulation of groundwater
Focus on agricultural management: crop growth vs. water and nutrient limitations	Indirect simulation of the carbon cycle

RUSLE Model

RUSLE/USLE (Revised Universal Soil Loss Equation) model is an empirical model that calculates the potential erosion on a plot scale and calculates soil loss in tons per hectare using rainfall intensity, terrain, and soil characteristics, i.e., Rainfall (R), Slope length and steepness (LS) and soil erodibility (K) factors, soil cover factor (C) and the erosion control practices factors (P).

Pros	Cons
Easy to use: The RUSLE model is easy to use compared to other models, and it requires only a few input parameters.	Limited applicability: RUSLE is designed to estimate sheet and rill erosion on cropland and forestland, and it may not be suitable for estimating erosion rates in other types of landscapes.
Useful for conservation planning: RUSLE can be used to evaluate the effectiveness of different conservation practices for reducing soil erosion rates	Simplistic: RUSLE assumes that erosion rates are proportional to slope and rainfall intensity, which may not always be accurate.
Provides spatially explicit results: RUSLE produces spatially explicit estimates of soil erosion rates, which can be useful for identifying areas that are most vulnerable to erosion.	Lack of consideration for actual erosion processes: RUSLE does not consider the actual erosion processes (such as detachment, transport, and deposition), and it assumes that soil is eroded uniformly across a landscape.
Widely used and well-tested: RUSLE has been widely used and tested in different regions of the world, and it has been shown to provide reasonable estimates of soil erosion rates in many cases.	Limited consideration of vegetation: RUSLE considers only the cover management factor (C) to account for the effects of vegetation on erosion rates, and it does not account for the effects of plant roots or the spatial distribution of vegetation on soil erosion.

Overall, the RUSLE model is a widely used and easy-to-use model for estimating soil erosion rates, but it has some limitations in terms of its applicability, simplifications, and lack of consideration for erosion processes and vegetation.

MMF Model

MMF (Morgan-Morgan-Finney Erosion) model is used to predict annual soil loss from field-sized areas on hill slopes. The MMF separates soil erosion processes into a water phase and a sediment phase. It considers soil erosion to result from the detachment of soil particles by raindrop impact and runoff and the transport of those particles by overland flow and reveal a realistic image of soil erosion hotspot sites in catchment scale while introducing suitable soil conservation and/or management practices. The model could be deemed a simplicity of USLE but more flexible as it has a stronger physical base than USLE.

Pros	Cons
Simple and easy to apply (a few days)	Erosion only; no sediment transport
Well tested	Simple, easy, and fast model
Focus on SLMs for erosion control: tillage, terracing/bunds, intercropping.	Good for quick screening: identify hotspots, assess effectiveness of SLMs
	Not good for detailed erosion assessments

WEPP Model

WEPP (Water Erosion Prediction Project) model is a process-based model used to simulate erosion and sediment yield under different land uses, soil types, and management practices. It provides distinct types of outputs: water balance (surface runoff, subsurface flow, and evapotranspiration), soil detachment and deposition at points along the slope, sediment delivery and vegetation growth through main interface of a standalone Windows application.

Pros	Cons
WEPP is capable of modeling complex hydrologic processes, including infiltration, runoff, and sediment transport, making it a valuable tool for evaluating the effects of different land uses and management practices on erosion and sedimentation.	The model requires a significant amount of input data, including detailed soil and topographic data, which may not be available in all locations.
The model can be used to simulate the effects of climate change on soil erosion and sedimentation.	WEPP is computationally intensive, and simulations can be time-consuming and resource intensive.
WEPP includes a comprehensive user interface, making it accessible to users without a strong technical background in soil science or hydrology.	WEPP is primarily focused on water erosion processes and does not incorporate other processes such as wind erosion or gully erosion.
The model is widely used and has been applied in a variety of geographic regions, allowing for cross-comparisons and benchmarking.	There is a degree of uncertainty associated with any model, and the accuracy of WEPP predictions is dependent on the accuracy of input data and model assumptions.

PESERA Model

PESERA (The Pan European Soil Erosion Risk Assessment) model is a process-based and spatially distributed model to quantify soil erosion by water and assess its risk across Europe. The model can also be extended to include estimates of tillage and wind erosion. The model is biophysical model, grid-based with spatial resolution of usually 30-500m (depending on area) and month-step. Outputs of the model include above-ground biomass, erosion, soil humus content, runoff (per grid cell – no routing), soil water deficit through two simulation phases: equilibrium phase & simulation phase.

Pros	Cons
Comprehensive assessment: PESERA model considers varied factors such as climate, soil, topography, land use, and management practices to assess soil erosion and land degradation risks.	Data requirements: The PESERA model requires a lot of input data, which may not be readily available in some areas.
Easy to use: The model has user-friendly interfaces that allow users to easily input data and get the results.	Limited to Europe: Although the model has been successfully applied in Europe, it may not be as effective in other regions, especially in regions with different soil and climatic conditions.
Applicable at different scales: PESERA can be used at different scales, from local to regional, to assess soil erosion risks.	Not suitable for detailed local analysis: While PESERA is useful for regional assessments, it may not be suitable for detailed local analysis due to the limitations in spatial resolution.
Customizable: The model allows for customization by users, who can input their own data, such as soil characteristics and management practices, to tailor the results to their specific location.	Reliance on assumptions: PESERA relies on certain assumptions that may not always hold true, such as the assumption of uniform rainfall distribution over a given area.

EUROSEM Model

EUROSEM (The European Soil Erosion Model) is a dynamic distributed model for simulating erosion, transport, and deposit of sediments over the land surface by interill and rill process. It is designed as an event-based model for both individual fields and small catchments. Model output includes total runoff, total soil loss, storm hydrograph and storm sediment graph. The model provides for explicit simulation of interill and rill flow; the effects of plant cover on rainfall interception, infiltration, rainfall energy and flow velocity; and the effects of rock fragment cover on infiltration, flow velocity and splash erosion. However, when operating EUROSEM, it may be sensitive to parameter values, which can be difficult to determine accurately; assumes a constant soil erodibility factor over time, which may not reflect actual changes in soil properties due to land use and management practices; the model may not be suitable for simulating erosion under extreme climate events or complex terrain conditions.

OpenLISEM Model

OpenLISEM (Open Limburg Soil Erosion Model) is a spatial hydrological model that simulates runoff, sediment dynamics and shallow floods in rural and urban catchments. It is an event-based model, which can be used for catchments from 1 ha to several 100 km2. The model is designed to simulate the effects of detailed land use changes or conservation measures during heavy rainstorms and disaster risk management. The model is grid based with spatial resolution of 5-25m, event based (hours – seasons). Model outputs include summary of outlet values, hydrographs/sedigraphs at multiple outlet points, spatial maps, timeseries.

Advantages of using LISEM is that it can calculate the effects of land use changes and soil conservation scenarios. Driven by hypothetical rainstorms of known probability of return, LISEM is a valuable tool for planning cost-effective measures to mitigate the effects of runoff and erosion. LISEM produces detailed maps of soil erosion and overland flow that are useful for planners. The integration of LISEM in a raster-based GIS, which holds the many data on the distributions of land attributes, is especially useful. Other advantages of LISEM are the use of physically based mathematical relationships, the ease with which newly developed relationships can be incorporated and the incorporation of information about the spatial variability of land characteristics. However, it is clear that, although the model has several advantages over other models, the preliminary results of LISEM are far from perfect.

VI.IV Nutrient Transport Models

VEMALA Model

The VEMALA model is an operational, national scale nutrient loading model at a watershed scale. It simulates nutrient processes, leaching and transport in soil and in rivers and lakes. It includes two main sub-models, the WSFS hydrological model and the VEMALA water quality model. Advantages of the model has successive versions of the model that have been developed leading to a more process-based nutrient loading model including 1) VEMALA-N (simulates NO3-, organic N and total nitrogen leaching and load formation at a catchment scale); 2) VEMALA-ICECREAM (simulates particle bound and dissolved phosphorus load and erosion from agricultural areas; 3) VEMALA-ICECREAM-N (simulates the daily balance of organic matter, organic N, ammonium (NH4-N) and NO3-N pools by accounting for input of plant residues, organic and mineral fertilizer, atmospheric deposition, fixation by plants and decay of organic matter); 4) TOC model (simulates TOC processes in the soils in more process-based manner).

AnnAGNPS Model

AnnAGNPS (ANNualized AGricultural Non-Point Source Pollution) Model is a distributed parameter, physically based, continuous-simulation watershed-scale model that simulates quantities of surface water, sediment, nutrients, and pesticides leaving the land areas (cells) and their subsequent travel through the watershed on a daily time step. Model output is expressed on an event basis for selected stream reaches and as source tracking (contribution to outlet) from land or reach components over the simulation period. Model performance in predicting sediment yields need to be increased by improving the input parameters for both the RUSLE and HUSLE.

INCA Model

INCA (Integrated Nutrients in CAtchments) model is an eco-hydrological and catchment-scale nutrient model that simulate movement of N, P, C, pathogens, and harmful substances from agricultural systems to water bodies. The model is semi-distributed, based on hydrological response units (HRUs), dynamic, daily step, process based. Model outputs include nutrient and water processes in terrestrial environment, fluxes to receiving waters. The model is flexible and can be adapted to different catchment and land use types, allowing for customized assessments; it can help identify cost-effective interventions to improve water quality and reduce nutrient pollution. However, the model requires a significant amount of data, including information on land use, soil characteristics, and weather conditions, which can be challenging to collect in some cases; also, the model is primarily designed for agricultural systems, and its applicability to other systems may be limited.

ANIMO Model

Animo is a detailed process-oriented simulation model to simulate the transport of nutrients to groundwater and surface water systems and the emission of greenhouse gasses as a function of fertilization level, soil and water management and land use and for a wide range of soil types and land management practices hydrological conditions. The model comprises descriptions of the C, N and P cycle in both unsaturated and saturated soil.

VI.V Crop Response Models

QUEFTS Model

QUEFTS (The Quantitative Evaluation of the Fertility of Tropical Soils) model is a rule-based model that can be used to estimate crop yield from soil properties, the amount of fertilizer applied, and an estimate of the yield that could be obtained when soil nutrients are in ample supply. It can also be used to estimate the amount of fertilizer needed to reach a particular yield. The outputs of the model are nutrient-limited crop yield and soil nutrient contents (N,P,K). The model can be widely used as a simple model to recommend nutrient management practices. However, the model considers only the mean meteorological parameters irrespective of their daily variations, and also considers the soil as a homogeneous medium, which influences the crop yield estimation and modifications are needed to incorporate different agro-climatic and management conditions and different interactive processes, including other micronutrients, irrigation practices, and sub-soil properties as this model assumes that nutrients are the only limiting factors in crop yield.

Pros	Cons
Easily set up	Empirically derived (no crop growth simulation and no integration of water- and nutrient limited yield)
Possible to consider nutrient limitations for crop yield (min of water- and nutrient-limited yield)	Dyna-QUEFTS soil nutrient degradation and restoration effects not validated
Focus on agricultural management: level of fertilization; nutrient recovery fraction	
Calibration data available for many crops	

WOFOST Model

WOFOST (WOrld FOod STudies) is a simulation model for the quantitative analysis of the growth and production of annual field crops. It is a mechanistic, dynamic model that explains daily crop growth on the basis of the underlying processes, such as photosynthesis, respiration and how these processes are influenced by environmental conditions. WOFOST can be used to calculate attainable crop production, biomass, water use, etc. for a location given knowledge about soil, crop, weather, and crop management (e.g., sowing date). The model can recognize three levels of crop production: potential, attainable (limited) and actual (reduced) production. However, WOFOST simplifies the reality that users always have to be cautious when drawing conclusions from the simulation results that cannot surpass the quality of the input data.

DSSAT Model

DSSAT (Decision Support System for Agrotechnology Transfer) is a software application program that comprises dynamic crop growth simulation models for over 42 crops as well as tools to facilitate effective use of the models. The tools include database management programs for soil, weather, crop management and experimental data, utilities, and application programs. The crop simulation models simulate growth, development, and yield as a function of the soil-plant-atmosphere dynamics. DSSAT has been applied to address many real-world problems and issues ranging from genetic modeling to on-farm and precision management, regional assessments of the impact of climate variability and climate change, economic and environmental sustainability, and food and nutrition security.

The limitations of the model are 1) only a few crops are included in the system and the models do not respond to all environmental and management factors; 2) missing components to predict the effects of tillage, pests, intercropping, excess soil, water, and other factors on crop performance; 3) most useful in regions of the world where weather, water, and nitrogen are the factors that affect crop performance; 4) may not be good under severe environmental stress; 5) simulate the potential, and water and nitrogen-limited productions, but do not consider many factors that determine yield limitations in many agricultural fields, for example, Phosphorus availability; 6) soil, water balance model is limited to well-drained soils; 7) limited capability for handling impact of biotic stresses caused by insect pests, diseases, and weeds also the model currently has a static system that allows a user to define biotic stressors based on field damage observations but there is no coupling with dynamic pest and disease models.

AQUACROP

AQUACrop is a crop growth model that simulates yield response of herbaceous crops to water and is particularly well suited to conditions in which water is a key limiting factor in crop production and assist in management decisions for both irrigated and rainfed agriculture to understand the crop response to environmental changes; compare attainable and actual yields in a field, farm, or a region; identify constraints limiting crop production and water productivity; develop strategies under water deficit conditions to maximize water productivity; to study the effect of climate change on food production; analysis irrigation scenarios useful for planning.

The model has some limitations: 1) it can only simulate daily biomass production and final crop yields for herbaceous crops with single growth cycles; 2) it is designed to predict crop yields at the single field scale (point simulations) which is assumed to be uniform without spatial differences in crop development, transpiration, soil characteristics or management; 3) only vertical incoming (rainfall, irrigation and capillary rise) and outgoing (evaporation, transpiration and deep percolation) water fluxes are considered.

EPIC Model

EPIC (Environmental Policy Integrated Climate) Model is a cropping systems model that was developed to estimate soil productivity as affected by erosion as part of soil and water conservation. EPIC simulates approximately eighty crops with one crop growth model using unique parameter values for each crop on a daily time step and can simulate hundreds of years. It can be configured for a wide range of crop rotations and other vegetative systems, tillage systems, and other management strategies. It predicts effects of management decisions on soil, water, nutrient and pesticide movements, and their combined impact on soil loss, water quality, and crop yields for areas with homogeneous soils and management.

SUCROS Model

SUCROS (Simple and Universal Crop growth Simulator) model is a mechanistic crop growth model that simulates both potential and water limited growth of a crop, i.e., its dry matter accumulation under resp. ample and rainfed supply of water and nutrients in a pest-, disease- and weed-free environment under the prevailing weather conditions. The model can simulate different crops with crop specific input parameters.

InfoCrop Model

InfoCrop is a process based dynamic simulation model for simulating growth, development and yield of rice, wheat, maize, sorghum, pearl millet, mustard, soybean, chickpea, pigeon pea, potato, and cotton. It simulates the effects of weather, soil, and crop management (sowing, seed rate, organic matter nitrogen and irrigation) and pests. It provides daily and summary outputs on various growth and yield parameters, nitrogen uptake, greenhouse gas emissions, soil water and nitrogen balance. It is used for several applications including yield forecast and climate change studies.

Together with our partners we produce, gather, compile and serve quality-assessed soil information at global, national and regional levels.

We stimulate the use of this information to address global challenges through capacity building, awareness raising and direct cooperation with users and clients.

CABI

ISRIC
World Soil Information

BILL & MELINDA GATES foundation