Geografisk Tiddskrift - Danish Journal of Geography, 2006, 106(1) 71-85.

Soil Profile Analytical Database for Europe (SPADE): Reconstruction and Validation of the Measured Data (SPADE/M)

R. Hiederer[1], R.J.A. Jones[2], J. Daroussin[3]

[1] European Commission Joint Research Centre,
Institute for Environment and Sustainability, TP 262, 21020 Ispra, Italy
roland.hiederer@jrc.it (corresponding author)

[2] National Soil Resources Institute,
Cranfield University, Bedfordshire, Silsoe MK45 4DT, UK
r.jones@Cranfield.ac.uk

[3] Unité de Science du Sol, Institut National de la Recherche Agronomique, INRA,
Avenue de la Pomme de Pin, BP 20619 Ardon, F-45166 Olivet Cedex, France
Joel.Daroussin@orleans.inra.fr

Abstract

The Soil Profile Analytical Database of Europe of Measured profiles (SPADE/M) was created to provide a common structure for storing harmonized information on typical soil profile properties of European soils. The main difficulty encountered in constructing the database was the transfer of the source data from individual electronic spreadsheet pages to the more rigid structure of a relational database. The data in spreadsheet format had been collected more than 12 years earlier but pressure was mounting for the capability to link these data to the Soil Map of Europe. A semi-automatic process was implemented to transfer data from nominal

positions on the spreadsheet page to an intermediate structure highlighting any deviations from expected values. Conflicting situations were solved by manual intervention and expert judgement. Data in the intermediate structure were subjected to a validation procedure with the aim of storing uniform data in the database. The validation checks cover format authentication, restricting entries to permissible values and those passing plausibility tests. In cases where a horizon property could not be represented consistently following the field specifications, the database structure was adapted to accommodate those conditions. The database model was extended to allow data from multiple samples taken at the same plot and from the analysis of samples from different laboratories to be stored.

Keywords: *soil properties, soil profile data, database design*

Introduction

The idea to compile a Soil Profile Analytical Database of Europe (SPADE) was first discussed at a meeting with the European Commission Directorate-General for Agriculture (then DG VI) in the autumn of 1986. Following publication of the Soil Map of Europe at scale 1:1 mio. (CEC, 1985), Madsen (1991) formally outlined the principles of such a database at a meeting of the European Heads of Soil Survey in Silsoe (UK) in December 1989. The Soil Map of Europe had already been digitised under the programme Coordination of Information on the Environment (Corine) (Platou *et al*., 1989). The objective of compiling the SPADE database was to provide additional information on soil properties with European coverage in a standard form to enhance the legend of the original soil map.

In 1990, the project Monitoring Agriculture with Remote Sensing (MARS), based at DG Joint Research Centre (JRC), Ispra, Italy, commissioned a research project to update the spatial component of the European Soil Database (Jamagne & King, 1991; Jamagne *et al*., 1995). During the 1990s, the MARS Project became the main driving force for compiling soil data at European level, with the immediate aim of improving the modelling of the soil water balance in

the Crop Growth Monitoring System (CGMS) developed by DG JRC to forecast the yields of the major arable crops throughout Europe (Vossen & Meyer-Roux, 1995; Daroussin, 1999a).

The initial contract to compile SPADE began in 1992 with the design of the standard forms for the compilation of the profile data, but only for the EU-12 Member States (Madsen & Jones, 1995a, b). The intention was to collect representative soil profile data for all the main soil types distinguished on the published Soil Map of Europe (CEC, 1985). Consequential for the latter use of the data was that the data collection process started at a time when personal computers running 16-bit operating system were slowly being introduced into the research community, but were by no means universally available. Many of the data contributors did not have access to any type of personal computer and those who did were confronted with a number of different spreadsheet software packages for data capture and storage. The initial aim of collecting data for all the main soil types in Europe proved unattainable, because too large a proportion of the project resources was spent on data entry. The intricacies of data confidentiality were a further hindrance to achieving comprehensive European coverage.

For compiling the database, two different formats (Proformas) were defined (Breuning-Madsen & Jones, 1995):

- *Proforma I (estimated data)*: was designed to capture profile data representative of specific soil types, but not geo-referenced to any particular location. National experts were requested to provide the data from measured or estimated parameters according to the specified format and using harmonized analytical methods. Problems of data confidentiality were avoided because the data could be linked to spatial units (map units) and not to any specific point on the ground. This is important because most land (and thus soil) in Europe is in private ownership.

- *Proforma II (measured data)*: was designed to capture geo-referenced, measured data from sample points, for which the soil had been examined and analysed. The Proforma allows recording of the analytical methods applied, but not necessarily

standardized between samples. It was accepted that compiling a comprehensive profile database for Europe by this approach would only be possible in the long-term.

Proformas I and II do not conform to the relational model of Codd (1970), whereby 'relations' are clearly defined at the design stage of the database construction and data are fully 'normalised' to avoid redundant storage. However, these Proformas were intended as the first stage in the construction of a database using a relational database management system (RDBMS). They did provide a standard view of the data familiar to the experts who were compiling the soil profile data sets and facilitate as much as possible the task of extracting data from mostly paper archives and their subsequent capture in electronic form. The quality of the soil data rested entirely with these national soil experts. The standard Proformas had the advantage of allowing data capture without the need for relational database management specialists to create readable views of the data from a fully relational system. This paper then describes the second stage of constructing a relational structure for the data.

The Food and Agriculture Organisation (FAO) and the International Soil Reference and Information Centre (ISRIC) had already been working on a soil database for storing and manipulating soil profile data in the late 1980s (FAO-ISRIC, 1989). This work continued in the 1990s, leading to a significant expansion in the availability of digital soil profile data for environmental research (Batjes, 1995, 1997; Batjes *et al*. 1995; Van Engelen & Wen, 1995). The compilation of the SPADE data sets was conducted in parallel and good contact was maintained with FAO and ISRIC throughout the project.

The purpose of Proforma I data was primarily to support modelling at scale 1:1 mio. with complete European coverage of soil types. By contrast, the geo-referenced Proforma II data were intended to form the basis for a European database of directly measured soil profile properties. A comprehensive coverage of all soil types was not the primary objective for the compilation of this database and it was assumed that given time a complete set of measured data for soil profiles in Europe would be collected.

In 1993, the Proformas were distributed to national experts in the EU-12 Member States working at institutions involved with the GIS Support Group to the MARS project (King, 1995). By the end of 1993, Proformas were returned, mostly in paper form, to Silsoe and Copenhagen for the first stage of data entry. At this time, a decision was taken to extend the geographical coverage of the European Soil Database to include Central and Eastern European countries (Jamagne & King, 1991; Jamagne *et al*., 1995) and thereafter the Proformas were sent to institutions in these countries for capturing data (Breuning-Madsen & Jones, 1998).

In 1999, version 1.0 of the European Soil Database was released on compact disk (CD) by the European Commission (Jones *et al*., 1998). It comprised the Soil Geographical Database of Europe (SGDBE) (King *et al*., 1995), the Soil Profile Analytical Database for Europe (SPADE) as spreadsheet files (Breuning-Madsen & Jones, 1995), and the Pedo-transfer Rules Database for Europe (Van Ranst *et al*., 1995). In the first version of SPADE, there were many missing data for some soil types and analytical data for several properties were totally absent. Subsequently, the Institut National de la Recherche Agronomique (INRA) Orleans compiled a relational database structure for the estimated profile data (Proforma I). The information provided by the measured profiles is linked to a specific geographic location, but the soil at the sample point is not necessarily representative for an area or soil type. Therefore, the compilation of a structured database for the measured profile data was not attempted by INRA.

SPADE and SPADE/M Data and Models

Not until recently was the measured data in SPADE recognized as a valuable source of information to support thematic analysis and modelling. In order to use the information provided for the measured profiles, the data had to be validated and put into a format, which would allow all data to be readily accessible to any interested user. This demand occasioned the development of the Profile Analytical Database of Europe of Measured Data (SPADE/M).

SPADE Measured Data

SPADE/M is based on Version 2.1.0.0, 29/03/1999 of SPADE, which is available on CD ROM, under licence from the European Commission (Jones *et al*., 1998). SPADE contains site specific information on FAO soil type (FAO-UNESCO, 1974 - legend soil name, modified CEC, 1985), land use, parent material and ground-water level, and analytical measurements on soil horizons, such as texture, organic carbon, pH and soil water retention, usually from single soil profiles. All profile data are recorded on a single spreadsheet page. The storage of the data within the cells of the spreadsheet page follows the general layout defined by Breuning-Madsen & Jones (1995). The standard format is given in Figure 1.

Figure 1: Data Entry Form for Plot and Measure Soil Profile Data (Proforma II)

The standard form consists of 3 main parts. The top part (cells A3 to AA8) contains information on the plot or site. The measurements for the horizon are split into two parts (A11 to AA18 and A22 to AA29). Additional information on depth to rock, other observations and the origin of the data are stored below the horizon data.

The SPADE dataset (v 2.1.0.0) contains measured data on soil profiles for 16 European countries. A total of 496 profiles are recorded in the files, with the number per country given in Table 1 and a geographical distribution presented in Figure 2. The location of 86 plots cannot be mapped, because geographic coordinates were not available during the original compilation stage of the project, either because they were not recorded or the projection could not be identified with any degree of certainty.

Table 1: Number of SPADE Measured Profiles Plots per Country

Figure 2: Distribution of SPADE/M Profile Plots

Some national institutions provided the original data in electronic form, but others only as hardcopies following the spreadsheet format (Figure 1). The digitization of the data from the

hardcopies was performed manually by an operator. Measures of quality assessment and control for the digitised data are not reported. While all profile data are made available in digital format, the integration of the data stored in the separate files into a single structure was expected to be achieved through an additional contract (Daroussin, 1999b).

SPADE/M Database Model

A simple structure was adopted for the SPADE/M data, which is largely comparable to the original spreadsheet format. This unsophisticated approach was adopted instead of a data model using full normalization to encourage the use of the data and facilitate users not trained in database management. As file storage format, the dBase dbf format (Version IV) was chosen, as this is compatible with most geographic information systems, database management systems, spreadsheets and statistical software packages.

A schematic overview of the data model used for SPADE/M is given in Figure 3. The file names used in the figure are further explained in Tables 2 to 4.

Figure 3: Schematic Data Model for Soil Profile Analytical Database of Europe of Measured Profiles (SPADE/M)

The main elements of the data model consist of two tables containing the measured or observed values:

- **PLOT** table (PLOT_DAT)
- **HORIZON** table (HOR_DAT)

The PLOT table contains the parameters characterizing the plot or site, where samples were taken. In the spreadsheet pages, these data are generally stored on the same page as the measured results, but with a more ambiguously defined structure and format. The HORIZON table contains the parameters characterizing the various soil layers or horizons identified at a plot location. In the spreadsheet pages, these data were generally stored in the form of a split

table. The PLOT table uses a unique identifier key (on field PLOT_ID) to link to records in the HORIZON table through a one-to-many relationship. The correct format and content of the data in the tables are then validated for each field to achieve a standardized database of profile measurements.

The data tables are linked to tables containing the definition of a parameter, where such an approach is appropriate. The corresponding files are identifiable by their _DEF ending of the file name. The structure of each of the measured or observed values together with the description data units and of field names are stored in the files PLOT_STR and HOR_STR.

SPADE/M Field Properties

The data tables contain several fields, where a plot attribute is expressed in more than one format. For example, soil is specified by name, but also as a coded value according to an external legend. To distinguish between different forms of expressing an attribute or measurement, the naming of fields follows a standard convention by suffix. An overview of field name suffixes and their signification is given in Table 2.

Table 2: SPADE/M Field Naming Convention

Naming SPADE/M field names differently from those specified in the SPADE metadata document became necessary, because the dBase format (dbf) restricts naming data fields and storage types. Using the dbf format, field names are limited to 10 characters and a field name like DEPTH_OTHOBS exceeds this limit. In the format alpha-numeric data are stored in the *character* format. Integer values are generally stored in *float* format, while the *number* format is used for any rational figure. For binary data the *float* format is used in preference to the *logical* format of dBase. The translation of the logical format by other programs is not always consistent (True/False, Yes/No or 1/0 can be used). File names follow the DOS convention of an 8.3 format (8 character file name and 3 character file suffix, separated by a full stop). Although this convention is no longer in universal use, some software still limits file names to 8

names to 8 characters.

An overview of the fields of the PLOT table and descriptive names are presented in Table 3.

Table 3: *Structure of PLOT Table*

Data stored in the fields of the PLOT table correspond to an actual expression of a plot attribute in the analogous spreadsheet cell. The table contains the filed names of the SPADE data where appropriate. The fields SURV_NO and SURV_DATE were added to allow storing the results of more than one survey for a plot position.

The field names and descriptions of the HORIZON table are presented in Table 4.

Table 4: Structure of HORIZON Table

Data stored in the fields of the horizon data table correspond to an actual expression of a horizon attribute in the columns in the spreadsheet tables. All fields of the original SPADE data were used and some fields had to be added to allow for the storage of multiple-survey data (SMPL_NO, ANLS_DATE) or specific situations found in the data, which could not be adequately stored in the original structure (e.g. SILT2_V/ESD or SAR_V/X).

Methodology for Data Transfer and Validation

Data were transferred from the spreadsheet pages to a common structure using a semi-automatic procedure, implemented in form of macros of the spreadsheet package used. Due to the variety of entries found, data were generally transferred to the database tables in alpha-numeric format, even in cases where only numeric entries were foreseen.

The validation process of the data was performed in stages:

- Verification of data position (in the spreadsheet page)
- Authentication of data format
- Substantiation of data value

The checks performed during the validation stages are presented hereafter.

1. Verification of Data Position

On the spreadsheet page data should have been entered into pre-defined cells. However, in practice the information was not recorded consistently in these fixed positions. Some variations from standard positions are arbitrary, such as leaving one or more empty cells beneath the field descriptor, whereas others are inevitable, e.g. when a profile contains more than the predefined 7 horizons.

For the identification of the correct position of data in the database fields a procedure based on manual inspection was used. Data were identified by starting from the nominal cell co-ordinate of the top-left corner of a data block as a first approximation. All other data were then identified relative to this reference position on the page. However, in all cases the actual position of data in the spreadsheet was verified manually and adjusted were necessary.

At this stage only the actual reproduction of data from a cell position in the spreadsheet to a corresponding field and record in the data table can be established. The actual content of the data transferred is preserved by using an alpha-numeric format for all data.

2. Authentication of Data Format

Data formats were authenticated by a procedure, which evaluated the conformity of the expected field format with the contents in the imported alpha-numeric entry. For numeric fields, the effect of changing the format of the transferred alpha-numeric value was evaluated. All problem cases were highlighted and examined manually. For alpha-numeric field entries, any leading or trailing spaces were removed, as were more than one space between alpha-numeric characters.

During the data authentication stage, it became evident that in some cases the original data structure had to be adjusted to store entries more consistently. Concerned were those fields where a numeric entry was defined, but exceptions to the normal conditions required highlighting the condition by entering an alpha-numeric code in the numeric field. This situation occurred, for example, when it is specified that a parameter does not meet or exceeds a defined value, such as for the sodium adsorption ratio. In the database, the situation is represented by creating a specific flag-field to describe the situation. However, in cases where the meaning of entries for a parameter was particularly confusing, the flag value was not retained in the database. An example is the parameter “Exchangeable sodium percentage of the CEC (%)”, where an entry of -10 should have signified “Less than 15% (humid areas)”. Nevertheless, also found in the field are entries of “<10”, which could mean either.

3. Substantiation of Data Value

While the previous checks mainly concern correctly identifying the entry intended to be associated with a parameter, the checks for substantiating data values relate to the actual figures provided. For this purpose, the data values are evaluated with respect to permissible or plausible entries. Permissible entries are defined in the specification document for the SPADE data. An example is the method field associated with various parameters indicating the method of measurement. For each method field, the permissible entries are pre-defined and the field should contain no entries other than the ones defined and in exactly the form specified. In some cases the field entries were modified to comply with the specifications, but without changing the actual meaning. For example, the method data were adjusted to always use a capital A.

Checks on the plausibility of entries are more complex and require backing up the checks with thematic information. Data plausibility was evaluated by comparing the data values with a range of likely figures for minimum and maximum values, which may define hard or soft boundaries. A hard boundary is a terminator value for a plausible range, such as 0-1

0-1 or 0-100% for relative values, leading to either rejecting or accepting a value. A soft boundary is one of diminishing probability for finding a value outside a given range. Examples for soft boundaries are ranges for pH-values or for bulk density.

Plausibility checks can be applied to a single parameter, but also to a combination of parameters. A simple check is the completeness of the texture data: the sum of all texture components should be 100. This was found to be not always the case. Whenever the texture sum deviated by more than 1% from the expected value the situation was investigated. One cause found for failing the check was that the sum of all the sand fractions was recorded in the field intended to hold the largest sand fraction while individual values for smaller sand fractions were also entered in the appropriate separate fields.

Some additional modifications to the data became indispensable to maintain consistency of the values reported. For example, where only a single value for sand content was reported this was generally moved to the field with the ESD of less than 2000µm. For the silt fraction an additional field had to be inserted. Otherwise measurements of a second silt value, mainly 20-50µm or 20-60µm, would have been recorded as a sand fraction.

A specific problem in the original data is the representation of missing values. A data item may be missing for several reasons because:

- it was not reported, e.g. the value exists but it is not available or has been lost;
- it was not measured, e.g. because of lack of time or the expense of the analysis;
- it could not be measured, e.g. particle–size grades cannot be measured in a soil comprising 100% organic material;
- it should not be measured, e.g. organic carbon is rarely measured, as a matter of routine, in the deeper subsoil horizons of mineral soils, because the content is usually extremely small.

While the coding of missing data or non-measurable properties is specified in the documentation it was not generally followed. Numerous cases exist where missing data were not coded, but indicated by a zero, '-1' or another flag value outside the permissible range of values. In particular a zero entry can pose a serious problem of ambiguity with respect to the significance of the value, e.g. where it was also used to indicate actual absence of a parameter.

Following the ambiguity of the coding and inconsistency in applying codes at all to mark missing data it was decided to not explicitly code any missing information. All obvious codes for missing data, mainly negative entries for numeric values or a derivate of an 'N/A' entry for alpha-numeric data, are not recorded in the database tables. Subsequently, zero entries are removed in cases where they could only be interpreted as indicating a missing value, e.g. for bulk density or pH. In cases, where the meaning of a zero entry could not be established with certainty the values are retained. Thus, any data stored in SPADE/M could signify a measured value. Referential integrity between the data and the definition tables was established in the working environment before the data were exported. Non-specified codes used in the data tables were added to the definition tables and commented in the corresponding field.

Results

The SPADE/M database provides a more universally serviceable structure for storing the measured profile data than the collection of spreadsheet files in the original version of SPADE. Due to the variability of data entries in the original forms, data could not be simply copied from spreadsheet cells to database records. Furthermore, some adaptations in the database structure were needed to represent the conditions reported for a plot or horizon in a consistent form. The checks on permissible and plausible entries together with the exclusion of entries for missing data resulted in a higher degree of harmonization of values recorded in the database.

The completeness of the information available to the user was assessed for the main fields in the plot and horizon tables. In this context completeness refers to the number of valid entries

over the total number of records for the parameter. The results are presented in Table 5 and Table 6. As the tables indicate, the plot and horizon information are not complete. The degree to which data are available depends very much on the parameter.

Table 5: Completeness of Plot Data Fields

Table 5 shows the completeness of the information in the plot table. For all except one data sheet the country code was indicated. The missing country code could be recovered from the file name. Serviceable geographic co-ordinates could be established for 82.3%of the plots. This restricts the use of the database for validation purposed of spatial layers of soil properties to 408 plots. A soil name or code according to FAO convention is given for all plots. However, the information is provided for some plots following the FAO74 convention (FAO, 1974) and for other plots according to the FAO90 legend (FAO-UNESCO-ISRIC, 1990). Groundwater levels are stated mainly for the mean lowest level (79.4%), but less so for the mean highest level (66.5%) and for less than half the plots (47.8%) the normal level is indicated.

Table 6: Completeness of Horizon Data Fields

The wide variation in the completeness of parameters reported is also apparent in the horizon table presented in Table 6. Depth limits could be defined for all horizons and a horizon name is given in 96.9% cases. Well defined soil properties are texture (91.6%) and pH (82.0%). For more than half the horizons values for a parameter are given, with the notable exception of $CaSO_4$ (4.2%), electrical conductivity (13.7%) and sodium adsorption ratio (10.9%).

For some parameters, e.g. for soil structure in the horizon table, the completeness of data availability cannot be established by merely relating the number of entries to the total number of records. This could only be achieved if a reliable indicator for missing data was available. However, the original data do not contain a consistent approach to separate, for example ‘no structure’ from ‘no measurement’.

Discussion and Conclusions

The time it has taken (almost 20 years) for the SPADE/M database to pass from the proposal of building a soil profile database of measured parameters (in 1986) to the realization of the task (database V. 1.0 in 2005) may not be representative of similar activities of collecting data at a multi-national level. Yet, the scarcity of comparable databases with multi-national coverage suggests the hidden complexity of storing data from different sources in a coherent form. This makes the broader availability of data on measured soil profiles in Europe and support to extend the range of profiles to a larger coverage the more significant.

Collecting soil profile data is a time-consuming task. For the SPADE data, a harmonization approach was added. No specific methods were detailed for sampling and measuring soil parameters. Instead, the methods of measurement or analysis used should be recorded and stored with the data in a common format. This approach allows collecting data *a posteriori*, i.e. from surveys already conducted. Defining stringent rules on data collection would have excluded many data from being included, thus restricting the number of plots in the database.

For the SPADE database, differences in specifications for the estimated and the measured profiles have lead to some confusion as to which parameter was recorded on the Proforma II sheets. Examples are electrical conductivity, where class symbols are specified for the estimated profiles and measured values for the measured data, and organic material, which is organic matter for estimated profiles and organic carbon for measured ones. These parameters were unified in SPADE/M.

The specifications governing data storage were more detailed than the data collection and analysis rules. The information from the plots should be entered in a fixed form on the pages of an electronic spreadsheet. The advantage of this approach is the very low overhead in terms of technical requirements for data capture. Data could be entered on a hardcopy or directly in the cells of the spreadsheet page by soil scientists. None of these methods puts any restrictions on the content or format of the information entered.

The advantage of simplified data entry is outweighed by the resulting low level of standardization of the data entered. This has proven to be a major obstacle to transferring the information stored on individual data sheets to a common structure. The flexibility of entering data has lead to information being stored erratically on the data entry form, to variations in data formats and to non-conforming values. In consequence, the transfer of the data from the spreadsheet pages to the database needed extensive manual intervention. A specific area of uncertainty affecting most parameters is the format used to indicate missing values. In the original data the recording of missing data is inconsistent and the cause of data not being recorded is not specified. This situation occurred despite the clear guidance given for recording missing data in the original procedure (Madsen & Jones, 1995b).

The design of SPADE/M was governed by the aim of providing easy access to harmonized data. The structure is familiar to users of spreadsheets, but as a consequence the model does not prevent redundancies. These issues and referential integrity were addressed as processing steps in the preparation of the database. The design is further based on the assumption that there is either only one dataset per plot or that plot and survey data are of the same quality, i.e. either all observations are constant or all are potentially variable between surveys. When storing more than one dataset per plot the former situation leads to data redundancy, while the later can cause data inconsistencies between surveys. In addition, the data of the PLOT table are not all of the same quality. Some parameters must be considered constant to define the plot, e.g. the plot coordinates. Yet, other parameters determined at a plot could in reality change over time, e.g. land use or groundwater tables.

During the process of harmonizing the data, some elements of the original data were not transferred to the new tables. In principle, all values positively identified as not representing a valid measurement were excluded. Yet, this does not imply that all values stored in SPADE/M represent actual measurements, because values which could either signify a valid entry or be missing data were retained. This situation is an improvement over the original dataset, but still requires conscientiousness in the analysis to avoid generating spurious results.

The completeness of the information stored in the database varies widely with the parameter recorded. A soil name or code, given for all plots and coordinates, could be recovered for over 80% of the plots. Less information is available on the groundwater levels and the depth of soil. Horizons are best described with respect to texture (88%) and pH value (82%). Other parameters are reported with less data entries.

The main recommendation for compiling future versions of SPADE and similar databases is to extend the spatial coverage to European countries not yet included. Adding data from different areas would broaden the basis of typical soil profile characteristics. Enlarging the SPADE/M database with additional soil profile information of comparable characteristics can be achieved by simply entering the data into the relevant fields under consideration of the data format definitions and conserving data integrity. The translation tables explaining codes for country names, soil names and land cover (according to the Corine nomenclature) already include the range of possible entries in the code fields in the present version.

The scope of the database could be enlarged to include not only typical conditions, but to provide a general structure for storing soil profile data. For example, a survey on soil horizons can be performed repeatedly on the same plot and the same sample can be analysed by more than one laboratory. Such data could support estimating the variation in horizon characteristics for a given soil.

Compiling profiles according to the spatial representation of plot positions should not be a requirement to extend the number of profiles. The guiding principle should be to cover the main European soil types under different conditions, e.g. according to climatic zone, land use, etc., to support the refinement of the SGDBE. The process of extending the database could be very much improved by providing a computer-based utility for entering data with built-in validity checks. The checks should include a definition of mandatory entries (plot fields, soil name), controls on permissible entries (format, codes), limiting values to defined ranges (minimum, maximum) and some assessment of plausibility (texture content sum). This approach would enhance the possibility of verifying any queries with the field scientist and improve the reliability of the information stored in the database.

Acknowledgements

We thank all our colleagues in European Soil Bureau Network for their past and continuing collaboration in providing soil data and expertise for construction of the European Soil Database. We also express ours thanks to Luca Montanarella, DG Joint Research Centre, Ispra and Secretary of the European Soil Bureau Network for his support and encouragement in conducting this study.

Bibliography

Batjes, N.H. (1995): A homogenised soil data file for global environmental research: a subset of FAO, ISRIC and NRCS profiles (Version 1.0). Working Paper and Preprint 95/10, International Soil Reference and Information Centre, Wageningen.

Batjes, N.H. (1997): A world data set of derived soil properties by FAO-UNESCO soil unit for global modelling. Soil Use and Management (13):9-16.

Batjes, N.H., E.M. Bridges and F.O. Nachtergaele (1995): World Inventory of Soil Emission Potentials: development of a global soil database of process-controlling factors. In: Climate Change and Rice (eds S. Peng *et al*.), Springer-Verlag, Heidelberg, pp. 110-115.

Breuning-Madsen, H. and R.J.A. Jones (1995): Soil profile analytical database for the European Union. Danish Journal of Geography (95):49-57.

Breuning-Madsen, H. and R.J.A. Jones (1998): Towards a European Soil Profile Analytical Database. pp. 43-50. In: H.J. Heineke, W. Eckelmann, A.J. Thomasson, R.J.A. Jones, L. Montanarella and B. Buckley (eds.): Land Information Systems: Developments for planning the sustainable use of land resources. European Soil Bureau Research Report No.4. EUR 17729 EN. Office for Official Publications of the European Communities, Luxembourg.

CEC (1985): Soil Map of the European Communities, 1:1,000,000. Office for Official Publications of the European Communities, Luxembourg.

Codd, E.F. (1970): A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, Vol.13(6): 377-387.

Daroussin, J. (1999a): Metadata: Soil Profile Analytical Database of Europe, Version 2.1.0.0, 29/03/1999. Document available on: http://eusoils.jrc.it/ESDB_Archive/ESDBv2/popup/pt_meta.htm. Site last accessed: July, 2005.

Daroussin, J. (1999b): Attribute Coding for the Soil Profile Analytical Database of Europe, Version 2.1.0.0, 29/03/1999. Document available on: http://eusoils.jrc.it/ESDB_Archive/ESDBv2/popup/pt_spec.htm. Site last accessed: July, 2005.

FAO-ISRIC (1989). FAO-ISRIC Soil Database (SDB). World Soil Resources Reports No.64, Food and Agriculture Organisation of the United Nations, Rome, Italy.

FAO–UNESCO (1974): FAO–UNESCO Soil Map of the World. Volume I: Legend. UNESCO, Paris, France

FAO-UNESCO-ISRIC (1990): FAO-UNESCO Soil Map of the World: Revised Legend. World Soil Resources Report 60. FAO, Rome.

Jamagne, M. and D. King (1991): Mapping methods for the 1990s and beyond. pp. 181-196. In: J.M. Hodgson (ed.): Soil and Groundwater Research Report I. EUR 13340 EN. Office for Official Publications of the European Communities, Luxembourg

Jamagne, M., C. Le Bas, M. Berland and W. Eckelmann (1995): Extension of the EU database for the soils of Central and Eastern Europe. pp. 85-99. in: D. King, R.J.A. Jones and A.J. Thomasson (eds.): European Land Information Systems for Agro-environmental Monitoring. EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg.

Jones, R.J.A., B. Buckley, and M.G. Jarvis (1998): European Soil Database: Information access and data distribution procedures: pp. 19-31. In: H.J. Heineke, W. Eckelmann, A.J. Thomasson, R.J.A. Jones, L. Montanarella, & B. Buckley (eds). Land Information Systems: Developments for Planning the Sustainable Use of Land Resources. European Soil Bureau Research Report No 4, EUR 17729 EN, Office for Official Publications of the European Communities, Luxembourg.

King, D. (1995): Foreword. pp. 5-6. In: D. King, R.J.A. Jones & A.J. Thomasson (eds.): European Land Information Systems for Agro-environmental Monitoring. EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg,

King, D., A. Burrill, J. Daroussin, C. Le Bas, R. Tavernier and E. Van Ranst (1995): The EU Soil Geographic Database. pp. 43-60. In: D. King, R.J.A. Jones, and A.J. Thomasson (eds.). European Land Information Systems for Agro-environmental Monitoring. EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg,

Madsen, H.B. (1991): The principles for construction of an EC-soil database system. pp. 173-180. In: J.M. Hodgson (ed.): Soil survey – a basis for soil protection. Soil and Groundwater Research Report I. EUR 13340 EN. Office for Official Publications of the European Communities, Luxembourg.

Madsen, H.B. and R.J.A. Jones (1995a): The establishment of a soil profile analytical database for the European Union. pp. 55-63. In: D. King., R.J.A. Jones & A.J. Thomasson (eds.): European Land Information Systems for Agro-environmental Monitoring. EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg.

Madsen, H.B. and R.J.A. Jones (1995b): Guidelines for completing profile proformas. pp. 277-284. In: D. King, R.J.A. Jones & A.J. Thomasson (eds.): European Land Information Systems for Agro-environmental Monitoring. EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg,

Platou, S.W., A.H. Nørr and H.B. Madsen (1989): Digitisation of the EC Soil Map. pp. 12-24. In: Jones, R.J.A. and B. Biagi (eds.): Agriculture: computerization of land use data. EUR 11151 EN. Office for Official Publications of the European Communities, Luxembourg.

Van Engelen, V.W.P. and T.T. Wen (1995): Global and National Soils and Terrain Digital Databases (SOTER), Procedures Manual (revised edition). United Nations Environmental Programme, Food and Agriculture Organization of the United Nations, International Society of Soil Science and International Soil Reference and Information Centre, Wageningen.

Van Ranst, E., A.J. Thomasson, J. Daroussin, J.M. Hollis, R.J.A. Jones, M. Jamagne, D. King, and L. Vanmechelen (1995): Elaboration of an extended knowledge database to interpret the 1:1,000,000 EU Soil Map for environmental purposes. pp. 71-84. In: D. King, R.J.A. Jones & A.J. Thomasson (eds.): European Land Information Systems for Agro-environmental Monitoring. EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg.

Vossen, P. and J. Meyer-Roux (1995): Crop monitoring and yield forecasting activities of the MARS project. pp. 11-29. In: D. King, R.J.A. Jones & A.J. Thomasson (eds.): European Land Information Systems for Agro-environmental Monitoring. EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg.

Table 1 Number of SPADE Measured Profiles Plots per Country

Country	Plots*
Albania	15
Belgium	34
Denmark	8
Estonia	37
France	33
Greece	19
Hungary	39
Italy	13
Luxembourg	13
Netherlands	20
Portugal	7
Romania	61
Slovak Republic	18
Slovenia	22
Spain	25
Switzerland	40
United Kingdom	86
TOTAL	**496**

* No of plots for which at least some data were reported in the forms

Table 2 SPADE/M Field Naming Convention

Field Name Suffix	Signification
_C	**c**lassified or coded entry
_ESD	**e**quivalent **s**pherical **d**iameter
_ID	key **id**entifier field, used for index
_KPA	**kPa** value for measurement
_M	measurement **m**ethod
_NAME	describing **name**
_V	continuous **v**alue
_X	binary field expressing presence or absence of an attribute

Table 3 Structure of PLOT Table

Field Name		Content
SPADE/M	**SPADE**	
PLOT_ID	-	Internal sequential ID identifying plot
PLOT_NO	-	Number of plot
SURV_NO	-	*Number of survey on plot*
CNTY_C	-	Eurostat country code
LOC_NAME	-	Location or identifier of plot
LON_COOR_V	LONG	Longitude coordinates of plot position
LAT_COOR_V	LAT	Latitude coordinates of plot position
PRJ_C	-	Coordinate projection code
ALT_V	ALT	Single altitude values, averaged in case of range
SURV_DATE	-	*Date of survey*
SOIL_NAME	SOIL	Soil name as given by author
SOIL_C	-	SOIL_NAME according to FAO coding
GWL_NM_V	-	Normal level of a permanent or perched groundwater table in cm, class value converted to class mean
GWL_NM_C	-	Normal level of a permanent or perched groundwater table, class value
GWL_HI_C	GWL_HI	Mean highest level of a permanent or perched groundwater table
GWL_LO_C	GWL_LO	Mean lowest level of a permanent or perched groundwater table
LU_NAME	LU	Dominant land use at plot as defined by author
LU_CLC_C	-	Land use class value according to CORINE legend
PM_NAME	PM	Dominant parent material
D_ROO_V	DEPTH_ROC	Depth of soil available for rooting
D_ROO_X	-	Depth of soil available for rooting exceeds value
D_ROC_V	DEPTH_ROO	Depth to a rock obstruction to rooting
D_ROC_X	-	Depth to a rock obstruction to rooting exceeds value
D_OTH_V	DEPTH_OTHOBS	Depth to any obstruction to rooting other than rock
D_OTH_X	-	Depth to any obstruction to rooting other than rock exceeds value
ORIG_C	-	Measurement origin
COMMENT	-	Additional comments

Field in *italics*: extension to allow more than one survey per plot

Table 4 Structure of HORIZON Table

Field Name		Content
SPADE/M	**SPADE**	
HOR_ID	-	Internal sequential ID identifying horizon
PLOT_ID	-	Internal ID identifying plot
SMPL_NO		*Number of sample for horizon*
ANLS_DATE	-	*Date of analysis*
LAB_C	-	*Code for analysing laboratory*
HOR_ID	HOR_NUM	Sequential internal ID identifying horizon within plot
HOR_NAME	HOR_NAME	Horizon name as given by author
HOR_BEG_V	DEPTH_HOR_START	Begin of horizon
HOR_END_V	DEPTH_HOR_END	End of horizon
CLAY_V	CLAY	Clay particle content
CLAY_ESD	CLAY_ESD	Clay particle size
SILT1_V	SILT	Silt content of first silt particle size
SILT1_ESD	SILT_ESD	First particle size for silt content
SILT2_V	-	Silt content of second silt particle size
SILT2_ESD	-	Second particle size for silt content
SAND1_V	SAND1	Sand content of first sand particle size
SAND1_ESD	SAND1_ESD	First particle size for sand content
SAND2_V	SAND2	Sand content of second sand particle size
SAND2_ESD	SAND2_ESD	Second particle size for sand content
SAND3_V	SAND3	Sand content of third sand particle size
SAND3_ESD	SAND3_ESD	Third particle size for sand content
GRAV_C	GRAVEL	Class percentage of stones and gravel in the soil
STRU_C	STRUCT	Structure class
OC_V	OC	Soil organic carbon content
OC_M	OC_M	Soil organic carbon measurement method
N_V	N	Soil nitrate
N_M	N_M	Soil nitrate measurement method
CACO3_V	CACO3_TOT	$CaCO_3$ equivalent value (weight %)
CACO3_M	CACO3_TOT_M	$CaCO_3$ measurement method
CASO4_V	CASO4	$CaSO_4$ value (weight %)
CASO4_M	CASO4_M	$CaSO_4$ measurement method
PH_V	PH	pH value
PH_M	PH_M	pH measurement method
EC_V	EC	Electrical conductivity value (dS/m range at 25 °C)
EC_C	EC	Electrical conductivity class (dS/m range at 25 °C)
EC_M	EC_M	Electrical conductivity method
SAR_V	-	Sodium adsorption ratio (%)
SAR_X	-	SAR less than 4 (humid areas)
ESP_V	EXCH_NA_P	Exchangeable Sodium Percentage of the CEC
EXC_CA_V	EXCH_CA	Calcium exchangeable base value
EXC_CA_M	EXCH_CA_M	Calcium exchangeable base measurement method
EXC_MG_V	EXCH_MG	Magnesium exchangeable base value
EXC_MG_M	EXCH_MG_M	Magnesium exchangeable base measurement method
EXC_CAMG_V	-	Combined calcium + magnesium exchangeable base value
EXC_CAMG_M	-	Combined calcium + magnesium exchangeable base measurement method
EXC_K_V	EXCH_K	Potassium exchangeable base value
EXC_K_M	EXCH_K_M	Potassium exchangeable base measurement method
EXC_NA_V	EXCH_NA	Sodium exchangeable base value
EXC_NA_M	EXCH_NA_M	Sodium exchangeable base measurement method
CEC_V	CEC	Cation exchange capacity value
CEC_M	CEC_M	Cation exchange capacity measurement method
BS_V	BS	Base saturation (%) as a proportion of the CEC taken up by exchangeable bases (TEB/CEC)
BS_M	BS_M	Base saturation measurement method
WC1_V	WC_1	First value of soil water retention value (volume % of water)
WC1_KPA	WC_1_M	Suction value for WC1_V
WC2_V	WC_2	Second soil water retention value (volume % of water)
WC2_KPA	WC_2_M	Suction value for WC2_V
WC3_V	WC_3	Third soil water retention value (volume % of water)
WC3_KPA	WC_3_M	Suction value for WC3_V
WC4_V	WC_4	Fourth soil water retention value (volume % of water)

WC4_KPA	WC_4_M	Suction value for WC4_V
WCFC_V	WC_FC	Soil water retention at field capacity (volume % of water)
WCFC_KPA	WC_FC_M	Suction value for soil water retention at field capacity
POR_V	POR	Total porosity value
POR_M	POR_M	Total porosity measurement method
BD_V	BD	Bulk density value
BD_M	BD_M	Bulk density measurement method

Field in *italics*: extension to allow more than one sample per survey

Table 5 Completeness of Plot Data Fields

Field	Entries	Completeness
Country indication	495	99.8%
Name for plot location	120	24.2%
Coordinates in geographic system	408	82.3%
Side of meridian indicated or ascertained	385	77.6%
Altitude information	422	85.1%
Soil name or code	496	100.0%
Ground water level, normal	237	47.8%
Ground water level, mean highest	330	66.5%
Ground water level, mean lowest	394	79.4%
Land use information	480	96.8%
Land use information transferable to CORINE Land Cover	399	80.4%
Parent Material information	488	98.4%
Depth of soil available for rooting	226	45.6%
Depth to a rock obstruction to rooting	152	30.6%
Depth to any obstruction to rooting other than rock	104	21.0%
Origin	389	78.4%

Note: total number of plots: 496

Table 6 Completeness of Horizon Data Fields

Field	Entries		Completeness	
Horizon name as given by author	2292		96.9%	
Horizon limits	2366		100.0%	
Clay content	2102		91.6%	*
Silt content	2103		91.6%	*
Sand content	2107		91.8%	*
Texture sum = 100%	1825		86.8%	**
Organic carbon content	1809	***	76.5%	
Soil nitrate	1274		53.8%	
$CaCO_3$ equivalent value	1312		55.5%	
$CaSO_4$ value	99		4.2%	
pH value	1941		82.0%	
Electrical conductivity value	323		13.7%	
Sodium adsorption ratio	259		10.9%	
Calcium exchangeable base value	1387		58.6%	
Magnesium exchangeable base value	1441		60.9%	
Potassium exchangeable base value	1470		62.1%	
Sodium exchangeable base value	1386		58.6%	
Cation exchange capacity value	1674		70.8%	
Base saturation value	1592		67.3%	
Total porosity value	1255		53.0%	
Bulk density value	1221		51.6%	

Note: total number of horizons: 2366
* Calculated over 2296 mineral horizons
** Calculated over mineral horizons with texture information
*** Includes converted values of organic carbon content

Figure 1 Data Entry Form for Plot and Measure Soil Profile Data (Proforma II)

Proforma II for Soil Analytical Data : Measured

Soil name : Jeg 2/4 1010 Wallasea series

Country : UK

Groundwater level :

Highest : -1

Lowest : -1

Longitude : 0deg 33min 59sec E

Latitude : 51deg 28 min 39sec N

Parent material : Estuarine/Marine Alluvium

Altitude : 2m OD

Landuse : Agriculture (Pasture)

Horizon	Depth (cm)	Texture_1 CLAY	esd	Texture_2 SILT	esd	Texture_3 SAND	esd	Texture_4 SAND	esd	Texture_5 SAND	esd	Stones GRAVEL	Struct COD	OM VAL	COD	N VAL	COD	CaC0 3 VAL	COD	CaS0 4 VAL	COD	pH VAL	COD	EC VAL	COD	
Ah	0-10	54	<2um	45	2um-60	1	60-200	0	200-600	0	600-2	0	5	20		-1		0		0		6.1	A12	3.61	A17	
Bg1	10-26	55	<2um	42	2um-60	2	60-200	1	200-600	0	600-2	0	4	1		-1		0		0		7.6	A12	2.25	A17	
Bg2	26-42	46	<2um	49	2um-60	4	60-200	1	200-600	0	600-2	0	4	1		-1		0		0		8.1	A12	2.31	A17	
Bg3	42-56	48	<2um	50	2um-60	2	60-200	0	200-600	0	600-2	0	2	1		-1		0		0		8.3	A12	2.56	A17	
BCg1	56-75	59	<2um	40	2um-60	1	60-200	0	200-600	0	600-2	0	4	1		-1		0		0		8.5	A12	2.67	A17	
BCg2	75-100	49	<2um	50	2um-60	1	60-200	0	200-600	0	600-2	0	4	1		-1		0		0		8.5	A12	2.89	A17	

Depth (cm)	Exch Ca VAL	COD	Exch Mg VAL	COD	Exch K VAL	COD	Exch Na VAL	COD	CEC VAL	COD	BS VAL	COD	WC-1 VAL	COD	WC-2 VAL	COD	WC-3 VAL	COD	WC-4 VAL	COD	WC-FC VAL	COD	TOT VAL	POR COD	DB VAL	COD
0-10	22.8	A19	14.0	A19	2.3	A19	4.7	A19	49.5	A22	88	A24	48	Th10	42	Th40	30	Th200	23	1500	51	Th5	75	A26	0.52	A28
10-26	9.5	A19	9.3	A19	2.6	A19	4.2	A19	32	A23	-1		52	Th10	49	Th40	43	Th200	32	1500	52	Th5	51	A26	1.31	A28
26-42	8.5	A19	8.9	A19	2.5	A19	4.6	A19	28.6	A23	-1		-1	Th10	-12	Th40	-1	Th200	-1	1500	-1	Th5	-1		-1	
42-56	5.7	A19	6.9	A19	2.1	A19	4.6	A19	25.7	A23	-1		49	Th10	47	Th40	38	Th200	37	1500	49	Th5	48	A26	1.38	A28
56-75	6.2	A19	7.4	A19	2.3	A19	5.5	A19	20.2	A23	-1		48	Th10	47	Th40	43	Th200	33	1500	49	Th5	51	A26	1.30	A28
75-100	10.0	A19	11.7	A19	2.8	A19	3.8	A19	23.7	A23	-1		-1	Th10	-1	Th40	-1	Th200	-1	1500		Th5	-1		-1	

Root Depth (cm)	150
D_Rock	999
D_Oth_Obs	999

Origin of Data e.g Code 1 or 2	2

Analytical codes, e.g. "A12", are listed in Madsen & Jones (1995b)
Th5, 10, 40, 200, 1500 are volumetric water contents (%) retained at these suctions in kPa

Figure 2 Distribution of SPADE/M Profile Plots

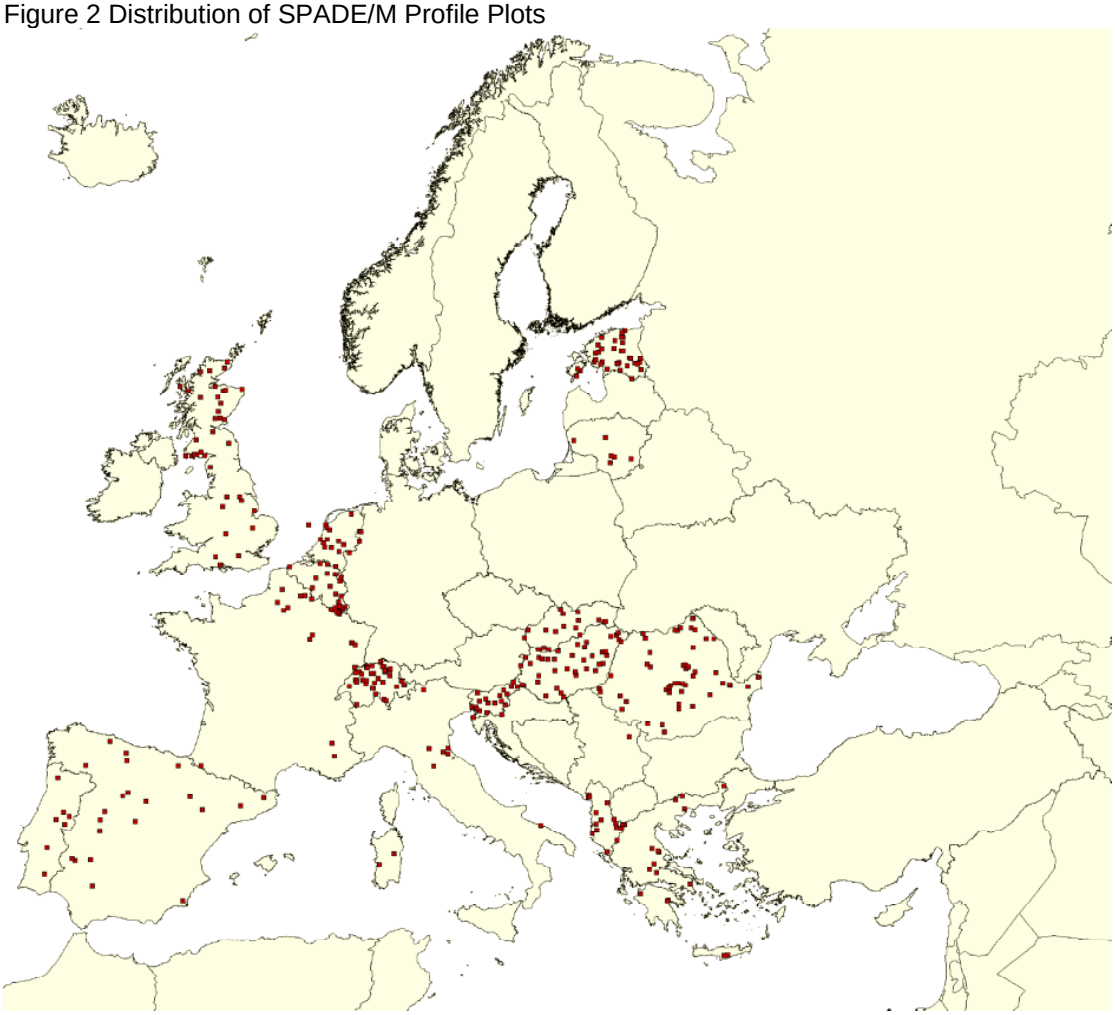

Figure 3 Schematic Data Model for Soil Profile Analytical Database of Europe of Measured Profiles (SPADE/M)