A logical model for integration of heterogeneous experimental data in soil research

N.A. Vasilyeva, A.A. Vladimirov, T.A. Vasiliev

FGBNU Federal Research Center "V.V. Dokuchaev Soil Science Institute", Moscow, Russia

Abstract. The undoubted challenge for science is the extraction of knowledge from fast growing heterogeneous datasets. Particularly, details of experimental setups are insufficiently formalized and cannot be easily inserted into databases. Thus, there is a problem of using these details in the process of data integration and meta-analyses. For this purpose, we developed a scheme of formalization for object descriptions with its origination, protocols for field and laboratory measurements (including instruments and experimental conditions). It allows the integration of larger amounts of data accounting for its specifics of acquisition, for example, by applying adjustments, assigning weights to data sources (based on its reliability, method precision and experimental uncertainty) or directly accounting for experimental conditions in models. This formalization is currently used to develop an electronic laboratory journal for soil research, intended for detailed description of a conducted or planned experiment. The study aims to: increase the re-producibility of scientific research results; allow automatic data processing and error detection, and most importantly; effective soil data mining for decision support systems.

Keywords: *heterogeneous data sources, formalization, standards integration, soils, reproducibility.*
DOI: 10.14357/20790279230116

Introduction

An undoubted challenge for soil science is extracting knowledge and relevant information from the ever-growing, diverse and complex soil data sets [1]. For parameterization and validation of predictive models in analytical systems, it is required to identify the most complete set of relevant data in the database. While the relevance of data is determined by the details of their acquisition, i.e., the setting and conditions of the experiment, these are not formalized enough to be taken into account when choosing data. "Data acquisition" in the current study includes field descriptions of soils, field and laboratory measurements, digitization of archive materials (such as, for example, legacy soil data, maps and thin soil sections).

The task of formalizing a scientific experiment and organizing machine-analyzable data flows coming from various sources, is important for the development of scientific activity in the digital world. Journals, research institutes, universities and manufacturers of laboratory equipment make their own sparse attempts [2-3]. However, it is conducted intensively yet only from the perspective of increasing the reproducibility of the results of scientific research, which is only one of the goals. Currently, more or less formalized templates for describing measurement methods are created for shared use and interlaboratory exchange, such as, *Nature Protocols exchange, OpenWetWare Protocol Categories, Protocol Online: Search Protocols, A secure platform for developing and sharing reproducible methods, A peer-reviewed protocol journal Bio-protocol, Optimized Lab Protocols for Testing Soils, JoVe*. However, typically they represent a set of text descriptions of protocol steps, not suitable for automated processing and comparison. We aim for a formalization of the measurement protocol which would be an unambiguous and machine-readable description of the necessary conditions for performing the measurement, the measurement process itself, the results obtained and their mathematical or algorithmic processing. Formalization of a measurement method together with a detailed protocol of the experiment (which includes a certain instrument, its settings, current calibration, etc.) allows the reproducibility of the results, i.e., increase confidence in the data and make data FAIR [4], i.e., it improves and simplifies the exchange and development of research methods. At the stage of data meta-analysis, this makes it possible to

identify and take into account experimental errors. If random errors are detected by repeated measurements, systematic errors (laboratory, operator, instrumental) can only be detected when analyzing large datasets. When systematic errors are found in the conduct of an experiment, one can also see what results they could affect. In some cases, it is possible to recalculate (correct) measurement results (for example, by recalibration). Another important goal is to use exactly the same information about experimental settings in protocol steps descriptions and in data processing scripts to avoid possible errors.

Existing systems for the formal description of experimental protocols – *Electronic Laboratory Journals (ELN) and Laboratory Information Management Systems (LIMS), namely, OSF, Labcollector, Hivebench, SciCloud, Accelrys (BIOVIA), Elabwtf, SciNote, Senaite, Bikalims, Occhiolino (GNU LIMS)* are either paid or shareware (free limited functionality or limited amount of storage, paid technical support), have limited options for embedding calculation functions, export options, and are focused specifically on laboratory analysis of physical samples, having no soil specificity. Abstract field objects, such as "terrain", "surface" and "soil profile" or a "trench", as well as long-term field experiments, for which there may be descriptions and measurements, are not included in the formalization scheme of such standard systems. Even though existing ELNs have convenient constructors for creating formalized protocols of experiments, they are designed only for the convenience of each individual user or group of researchers with their objects. Therefore, when implementing such a product by research institutes, a database collected from a set of ELNs of all employees will not be suitable for further effective joint analysis of the collected data. Moreover, a detailed formalization scheme is needed at least to evaluate existing open-source software suitability as components of a developed information system. Thus, soil research requires a specific implementation of such an information system with at least a soil-specific scripts and models library.

The disadvantage of existing soil and soil-geographic databases is that, firstly, they do not contain the history of the origin of objects and the sequence of actions on them (both in field, for example, technological maps of crop cultivation, and laboratory, for example, various treatments and fractionation of samples). This leads to data fragmentation and lack of relationships (potentially relevant data are lost). Secondly, they do not contain details of data acquisition methodology. Thus, only data obtained by one widely used method are selected, which cuts off all other data obtained by other or similar methods. At the same time, different experimental conditions are unavoidably mixed. Realizing the existing problem of formalizing laboratory measurement details, in the recent years International Soil Data Center in Wageningen has started to request information about methods steps from each laboratory which provided them legacy data, to evaluate datasets considering accuracy and precision [5]. While it is rather difficult to do it in detail for legacy data, it is possible to supply all the newly generated data with formalized data acquisition procedures.

The solution may be to have formalized data acquisition protocols that allow the maximum use of all available related data, for example assigning data sources different weights (calculated based on the accuracy of the method, the reliability of the data source, experimental errors or processing errors), homogenizing data to comparable values, by introducing corrections for experimental conditions (or experimental conditions could be directly used in mathematical models) or in other ways taking into account the differences in obtaining data. An ensemble statistical approach, using the entire available data set and models, while assimilating data coming from heterogeneous sources over time, is considered to be more informative for predictive modeling and estimating its uncertainties than the use of narrow subsampling [6-7]. The presence of links between objects allows combining initially independent experiments to analyze soil properties variability in space and time. For example, to make generalizations to obtain regional/global dependencies necessary for predictive models.

Formalized research protocols are published by a number of authoritative specialized journals *Nature protocols, Springer Protocols, Cell Protocols*. When formalized protocols are used, automatic processing of results and calculation of errors is possible. It becomes possible to transfer the entire database from one classification/description system to another according to established rules. One can identify intersecting sets in soil descriptions, measured properties and experimental parameters, as well as have control over consumables, the state of the instrument base, workload and etc.

The aim of this work was to develop a conceptual scheme for formal description of heterogeneous data and methods for its acquisition to ensure the possibility of organized collection and storage of all soil research results. It gives data reliability estimation for further analyses and generalizations, while providing reproducibility for research studies.

1. Results and discussions

The developed formalization scheme is shown in the Figure 1. This scheme shows logical elements of the database necessary for the coherent collection

of complete and formalized information about experimental studies and further analysis of this information. The proposed scheme allows us to formally describe standard and non-standard methods as a sequence of simple actions (method elements). This scheme makes it possible to link descriptions and measurements carried out in the field and in the laboratory according to any formally described methods on such objects as "terrain", "soil surface", "transect", "soil profile", "soil horizon", "sample", "thin section" and etc. a single spatial database with a history of filling the object with data. This allows different researchers to supplement objects in the system with soil studies at any time in an arbitrary order and subsequently carry out meta-analysis on the required spacetime scale.

The database contains four main logical blocks: a block of reference information, a block of data, a block of methods and a user-specific/inventory block (Fig. 1).

1.1. Database structure

Information block

The reference information block contains a single expandable list of soil properties and a table describing various groupings of those properties, as well as a list of measurement units and their conversions. For example, grouping can be according to an object under study (area, profile, sample, etc.), field of knowledge (physical, chemical, etc.), description standard (properties of FAO, EGRPR, WoSIS, WISE etc.). In this case, one property can belong to several groups. Such grouping structure of soil property is universal (compatible with other standards) and is supposed to be extensible by adding new groups. Any legacy data can be imported "as is", extending existing templates for entry of new data. The idea is not to create another new standard and not create data homogenization in advance but use homogenization scripts at export of data according to user specifications, using the advantage of formalized data acquisition.

Data block

The data block contains information about soil samples and other objects and the results of experiments. Data can be entered in any degree of detail, starting from a simple structure as a table "object-property-value" to a detailed description of the experiment. With a detailed description, all data is stored in the form of a history of accumulated events (creating a soil profile or a plot, description, sampling, incubation, measurement and so on). *The database does not impose strict requirements on the chronology of events, for example, measurements can precede the field description of a profile.* Each event is described by an entry in the e-journal (table "Journal entries"). The entry contains information about:

- described objects (relationship with the "Objects" table);
- created and destroyed objects by the method and relations between objects (table "Objects origin"). For example, a sample taken from a soil profile will be created; the one processed according to a sample preparation protocol will be destroyed and several new ones created: a reduced initial sample and, for example, several fractions;
- values of soil properties and their relations to objects through the table "Object-property-value" and "Values" of different types;
- optional raw data (if a processing script is applied by the method).

Each journal entry is linked to a method block (protocol and protocol steps).

Methods block

The block of data acquisition methods contains formalized methods and specific implementation protocols covering all three stages of data acquisition: preparation/preprocessing, measurement and data processing. The formalized methodology is described by a "Protocol" which consist of "Protocol steps" which are grouped and ordered. Each step represents an application of some method with certain optional settings and parameters. Each method has information on its applicability (types of source and result objects, result property), human-readable description, and optional detailed protocol which is recursively formally described as a sequence of steps (and listed in "Protocols"), i.e. each step itself can be represented by another protocol. Every step can have its processing script ("Model_id" in the "Data processing models" table) which acts on raw data. Journal data entries are allowed to be associated with a certain protocol or just with a single step. When a user changes step parameters a new step is created and is written to the journal, while the original step gets the status "modified". Protocols can be created from existing base protocol steps (from database or from user's own journal), while new protocol steps can be created from methods, adding instrumental setting/parameters and a processing model. Parameters depend on the calibration and settings of the instruments and the conditions of the experiment.

Thus, the whole variety of techniques is reduced to a manageable number of basic elements with parameters (such as temperature, duration, rotation speed, reagent, etc.). Each protocol step is associated with information about consumables and equipment that can be used at this step.

The user interface for data entry is created automatically based on the formalized method detail. The concept allows protocols with any degree of detail, assuming continuously increasing formaliza-

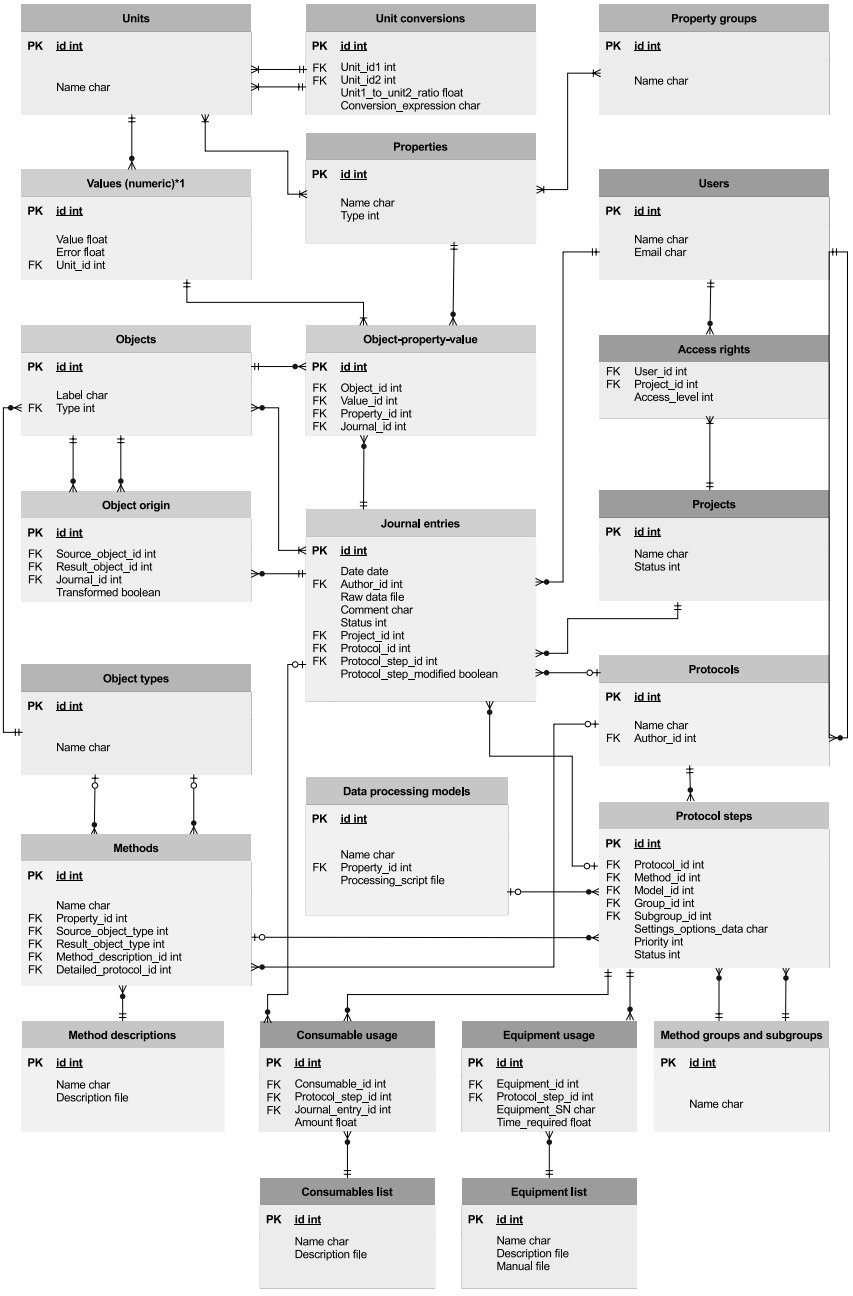

Fig. 1. Logical scheme of formalization for heterogeneous experimental data in Crow's Foot notation[1]

tion down to basic steps. Protocol has a branching structure and data entry can be performed by the user at any level of detail. However, the final estimates of dataset reliability, accuracy and reproducibility in the system are assigned according to the extent to which raw data were supplied. We believe that due to the presence of calculating scripts at each step (when required), overall standardization of data in e-journal (potential bonuses for automation in analyses), protocol reusability, collaborative mode and various "helpers" (statistical quality control, checks and availability of templates) will encourage users to enter raw data into the e-journal.

Along with standardized methods, in many cases of scientific research there is a need to store and use non-standard (author's or temporary), experimental methods and modifications. If for standard methods it

[1] The diagram contains a table "Values (numeric)" for numerical values, at the same time, it is provided that the database contains several similar tables for values of different types. For example, character strings for use in soil descriptions or bibliographic information, geographical information (points, contours) for introducing data from soil maps.

is enough for the user to indicate method name, since their formalization can be entered at any time after, for non-standard methods – formalization is the responsibility of the author.

It is known that in addition to the information described in state standards and other methodological manuals, "many different factors can affect the variability of measurement results performed using the same method, those including: operator; equipment; equipment calibration; environmental parameters; time interval between measurements." [8]. *The proposed logical model allows to save all the details of the experiment in a formalized form.* In recent years, journals such as Science and Nature, as well as The Transparency and Openness Promotion (TOP) Committee, have been actively urging scientists to make their work transparent so that their experiments can be repeated "at least in theories." They are developing increasingly stringent criteria for journal publications regarding the provision of formalized methods and detailed experimental protocols to improve the reproducibility of scientific results, and are even promoting testing for a pre-published experimental design (study pre-registration), which reduces the bias towards publishing results with a certain effect detection relative to experiments with a negative result, in which the intended effect was not detected with the corresponding study protocol [9-12].

Inventory block

The inventory block contains general lists of equipment (table "Equipment list") and consumables (table "Consumables list"), as well as inventory records. Table "Consumable usage" contains information about amount and type of consumables required for each protocol step (when linked to protocol step), as well as amount, actually used in experiment (when linked to journal entry). Information about instrument being used in a given protocol step is stored in the "Equipment usage" table.

1.2. Advantages

This way of presenting data using the developed conceptual scheme differs in that it allows:

- to create and store any type of described object – from plot as a result of field partition to a soil fraction/solution or other objects obtained in complex experimental procedures, retaining the full chain of objects origin and treatments;
- to store a complex data structure: during the experiment, many measurements of the same value for one sample can be made, all of them can be stored in the presented concise scheme. An example of a multivariable dependence, or dependence of several values from several variables, can be the measurement of various gases emissions at changing soil temperature and moisture;
- to save the history of actions on the object, for example, store and update new data in a complex and long-term field experiment: the dynamics of carbon content in soil during changes in vegetation and fertilizers inputs is recorded as a sequence of single actions;
- to estimate reliability and accuracy of the dataset based on raw data provision, methods information and usage of automated data processing;
- to perform various analyses of protocols. For example, comparative analysis of by machine learning approaches to assess the dependence of results on protocol peculiarities, develop protocols, select a protocol for a specific task, taking into account such characteristics as applicability and accuracy.

In the case when the details of some experiment are not available, the database schema also allows to store information just as "object-property-value" as in most existing databases. Thus, the proposed scheme is compatible with known formats, for example, those used in the Soil Geographical Database of Russia (PGBD RF) [13], the Unified State Register of Soil Resources (EGRPR)[14], WISE Soil Property Database [15], the International Soil Carbon Network (ISCN), an intercontinental aggregator and provider of soil data for the Information System (WoSIS) of the International Soil Data Center (ISRIC) [16] and etc., while creating many new opportunities.

The developed scheme for formalization of heterogeneous soil data: 1) increases the reproducibility of scientific research results, 2) allows automatic data processing, and most importantly, 3) allows effective data mining and, thus, is an important base part in creating analytical systems for modeling scenarios and decision making.

1.3. Options for data entry protocols

Data entry is proposed to be carried out in 3 general stages – minimal, extend-ed and detailed (Table 1).

The minimal description allows to quickly make: an inventory of all the objects, involved instruments, a general overview of the methods, and is also used to enter data from literary sources when they do not contain detailed study protocols.

The extended description information already allows to fill up the Unified State Register of Soil Resources. From the point of view of instrumental base, it is already possible at this level to monitor the state and involvement of instruments in specific research protocols. It becomes possible to evaluate the types and volumes of produced data, workload of devices and employees, time ranges and costs of measurements. It allows to optimize work and carry out quality control (errors and systematic errors of personnel and devices with a possibility of its localization and correction).

Table 1

Types of protocols for data entry from external sources

Description\Type	Minimal	Extended	Detailed
Soil profile or sample description (level “data”)	User name; geolocation; sampling/description date; soil name (if description); measurement or sampling depth.	Minimal description; and Names and depths of soil horizons	Minimal description; and Any properties of the objects (location, soil profile, horizon, sample)
Methods of data acquisition	Method name; property name; source object type; result object type; file with description.	Minimal description; and List of instruments with details (serial number, condition, precision, accuracy, etc.); list of consumables with usage	Extended description; and Formalized protocol steps; Instruments settings; Data processing model with parameters.

A detailed description is produced continuously and is the result of a full-fledged electronic laboratory journal.

2. Case of implementation (work in progress)

The presented scheme of formalization is being implemented for the development of Information system that provides integration and multi-level presentation of legacy and current data (See Fig. 2). The top panel shows user interfaces for data entry, as well as a dashboard with reports and statistics, and the bottom panel shows related components of the database. Interaction between the interfaces and the database occurs through the electronic document management system. User (web) and program (web API) interfaces are divided into interfaces:

a) for data entry (such as field soil description helper, request forms for laboratory analyzes containing

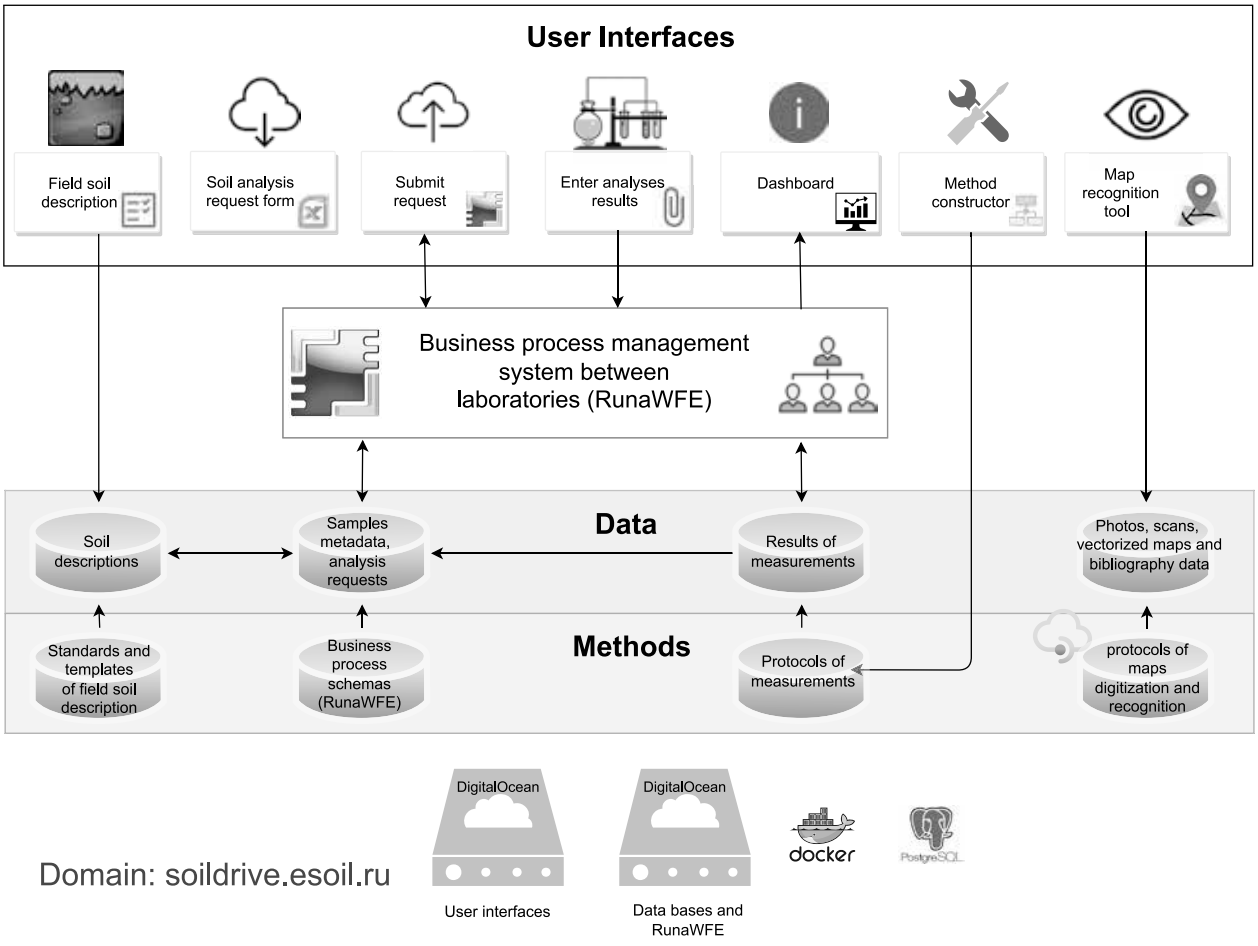

Fig. 2. General scheme of the information system

metadata of soil samples, upload of requests into the system and upload of measurement results by operator, constructor of methods and a subsystem for map recognition),

b) to search and create a data sample, generate statistics and reports, work with models in the information panel (data meta-analysis, generate scenarios, etc.).

The database consists of two levels, the first is "data" level (Fig. 2), which includes field soil descriptions, laboratory requests, metadata of the analyzed samples, the results of measurements, as well as photographs (profiles, thin sections, maps, microscopic, etc.), scans and vectorized maps, maps metadata. The second level is "methods of data acquisition" (Fig. 2), which includes:

– methods/schemas of field descriptions, schemas of electronic document management processes, methods of measurements and data processing models, methods of digitization and image recognition;
– detailed protocols for data acquisition with a description of certain measuring instruments or data processing (characteristics and settings of instruments, calibration curves and model parameters) and software.

Conclusions

The proposed formalization scheme makes it possible to store structured information about soils in various levels of detail. Formalization of data acquisition is the basis for the creation of an electronic laboratory journal containing the un-ambiguous formulation of a conducted or a planned experiment. The scheme provides the ability to search for experimental time series or compile pseudo-time experiments to provide simulation models with a relevant set of data for initialization and parameterization. This makes it possible to further tackle such an important scientific problem as estimating the effects of parametric and structural uncertainties in projections of ecosystem models. This work is part of the scientific rationale for the creation of a multi-level analytical system "Soil and land resources of Russia for agricultural production".

References

1. *Wadoux, A.M.J.-C., M. Roman-Dobarco, and A.B. McBratney.* 2021. Perspectives on data-driven soil research. European Journal of Soil Science 72:1675–1689. doi: 10.1111/ejss.13071.
2. *Giraldo, O, A. Garcia, and O. Corcho.* 2018. A guideline for reporting experimental protocols in life sciences. PeerJ 6:P.e4795. doi: 10.7717/peerj.4795.
3. *Halbritter, A.H, H.J. De Boeck, A.E. Eycott et al.* 2020. The handbook for standardized field and laboratory measurements in terrestrial climate change experiments and observational studies (ClimEx). Methods Ecol Evol. 11:22– 37. doi: 10.1111/2041-210X.13331.
4. *Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3(1):160018. doi: 10.1038/sdata.2016.18.
5. *Ribeiro, E., N.H. Batjes, and A.J.M. van Oostrum, eds.* 2020. World Soil Information Service (WoSIS) – Towards the standardization and harmonization of world soil profile data. Procedures manual 2020, Report 2020/01, Wageningen: ISRIC – World Soil Information 145 p. doi: http://doi.org/10.17027/isric-wdc-2020-01.
6. *Niu, S., Y. Luo, M.C. Dietze, T.F. Keenan, Z. Shi, J. Li and III F.S., Chapin.* 2014. The role of data assimilation in predictive ecology. Ecosphere 5(5):1-16. doi: 10.1890/ES13-00273.1.
7. *Martre, P., D. Wallach, S. Asseng, F. Ewert et al.* 2015. Multimodel ensembles of wheat growth: many models are better than one. Glob Change Biol. 21:911-925. doi: 10.1111/gcb.12768.
8. GOST R ISO 5725-1-2002. 2009. Tochnost' (pravil'nost' i pretsizionnost') metodov i rezul'tatov izmereniy. Chast' 1. Osnovnyye polozheniya i opredeleniya [Accuracy (correctness and precision) of measurement methods and results. Part 1. Basic provisions and definitions]. Moscow: StandartinformPubls. 24 p.
9. *Buck S.* 2015. Solving reproducibility. Science 348(6242):1403. doi: 10.1126/science.aac8041.
10. *Alberts, B., R.J. Cicerone, S.E. Fienberg, A. Kamb, M. McNutt, R.M. Nerem et al.* 2015. Self-correction in science at work. Science 348(6242):1420-1422. doi: 10.1126/science.aab3847.
11. *Belyaev, I.* 2015. Kharuko Obokata ne obzhalovala zaklyuchenie o fal'sifikatsii eyu rabot po sozdaniyu STAP-kletok, TASS. Available at: https://nauka.tass.ru/nauka/1685497 (accessed November 18 2022).
12. *Nosek, B.A., G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler et al.* 2015. Promoting an open research culture. Science 348(6242):1422-1425. doi: 10.1126/science.aab2374.
13. *Golozubov, O.M., V.A. Rozhkov, I.O. Alyabina, A.V. Ivanov, V.M. Kolesnikova, S.A. Shoba*. 2015. Technologies and Standards in the Information Systems of the Soil-Geographic Database of Russia. Eurasian Soil Science 48(1):1-10. doi: 10.1134/S1064229315010068.
14. *Alyabina, I.O., V.A. Androkhanov, V.V. Vershinin, S.N. Volkov, N.F. Ganzhara, G.V. Dobrovol'skii,*

A.V. Ivanov, A.L. Ivanov, E.A. Ivanova, L.I. Il'in, M.L. Karpachevskii, A.N. Kashtanov, V.I. Kiryushin et al. 2014. Edinyi gosudarstvennyi reestr pochvennykh resursov Rossii. Versiya 1.0. Available at: http://egrpr.soil.msu.ru/ (accessed November 18 2022).

15. *Batjes, N.H.* 2009. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use and Management 25:124-127. doi: 10.1111/j.1475-2743.2009.00202.x.
16. *Harden, J.W., G. Hugelius, A. Ahlström et al.* 2018. Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter. Glob Change Biol. 24:e705–e718. doi: 10.1111/gcb.13896.

Vasilyeva N.A. PhD, Federal Research Center "V.V. Dokuchaev Soil Science Institute", Pyzhevsky lane 7/2, Moscow, 119017, Russian, e-mail: nadezda.a.vasilyeva@mail.ru

Vladimirov A.A. PhD, Federal Research Center "V.V. Dokuchaev Soil Science Institute", Pyzhevsky lane 7/2, Moscow, 119017, Russian, e-mail: artem.a.vladimirov@gmail.com

Vasiliev T.A. Federal Research Center "V.V. Dokuchaev Soil Science Institute", Pyzhevsky lane 7/2, Moscow, 119017, Russian, e-mail: TarasVasiliev44@gmail.com